BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 01/01 pp1–16
DOI: 10.26599/BDMA.2025.9020070

Image Copyright Dual-Protection Based on Extractable and Imperceptible Adversarial Watermark

Yuming Liu, Shan Ai*, Zhili Zhou*, Wei Pang, Changyu Dong,
Huilin Ge, Daizhi Liao, and Yongfeng Huang

Abstract: Generally, there are two popular ways to protect image copyright, i.e., proactive protection (preventing illegal use via adversarial perturbation) and passive protection (verifying ownership by digital watermarking). However, since the perturbation and watermark embedded into an image will interfere with each other, directly embedding them into the image cannot achieve the proactive protection and passive protection, simultaneously. To address this issue, we propose an image copyright dual-protection approach, which embeds an extractable and imperceptible adversarial watermark (EIAW) in the image frequency-domain. Specifically, the adversarial watermark is automatically embedded and optimized in the manner of allowing for effectively attacking the deep neural networks (DNNs) and accurately extracting the embedded watermark, simultaneously. Moreover, instead of using the pixel-domain constraints, i.e., L_p norms, we introduce a frequency-domain constraint to optimize the watermark embedding locations. Experiments on ImageNet and CIFAR-10 demonstrate that the proposed EIAW achieves high attack effectiveness (up to 100%) and extraction accuracy (up to 93%), while maintaining good watermark imperceptibility.

Key words: adversarial attack; digital watermark; adversarial watermark; copyright protection; dual-protection

1 Introduction

With the widespread use of portable devices and online platforms, social media platforms, such as Facebook, Twitter, and Instagram, have become the main channels for sharing images. However, sharing extensive images comes with two main risks. First, user-uploaded images may be automatically recognized and analyzed by deep neural networks (DNNs) to obtain sensitive semantic information, such as age and gender[1]; Second, images distributed on networks could be illegally copied for unauthorized use. Therefore, in the era of artificial intelligence, it is urgently required to protect the image copyright both before and after the illegal use of images. The image copyright dual-protection including the proactive

• Yuming Liu and Daizhi Liao are with School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou 510000, China. E-mail: lym@e.gzhu.edu.cn; ldz@e.gzhu.edu.cn.
• Shan Ai, Zhili Zhou, and Changyu Dong are with Institute of Artificial Intelligence, Guangzhou University, Guangzhou 510000, China. E-mail: aishan@gzhu.edu.cn; zhou_zhili@163.com; changyu.dong@gzhu.edu.cn.
• Wei Pang is with School of Mathematical and Computing Sciences, Heriot-Watt University, Edinburgh EH14 4AS, UK. E-mail: w.pang@hw.ac.uk.
• Huilin Ge is with College of Automation, Jiangsu University of Science and Technology, Zhenjiang 212013, China. E-mail: ghl1989@just.edu.cn.
• Yongfeng Huang is with Zhongguancun Laboratory, Beijing 100094, China, and also with Department of Electronic Engineering, Tsinghua University, Beijing 100084, China. E-mail: yfhuang@tsinghua.edu.cn.
* To whom correspondence should be addressed.
Manuscript received: 2025-02-21; revised: 2025-05-20; accepted: 2025-06-09

protection (preventing the illegal use) and the passive protection (verifying image ownership after the illegal use) has become more and more important.

Generally, there are two popular ways to protect image copyright, i.e., adversarial attack and digital watermarking. The adversarial attack[2−5] usually adds an adversarial perturbation into a copyrighted image to mislead the model inference process, thereby preventing the illegal use of the image. However, such adversarial perturbations are added into images in an irreversible way and thus cannot be extracted accurately for verifying the image ownership. The digital watermarking[6−9] can embed a watermark into an image to verify ownership after unauthorized use. However, digital watermarking cannot prevent illegal use or copyright infringement in advance. To achieve copyright dual-protection, the most intuitive manner is to embed both a digital watermark and an adversarial perturbation into an image, sequentially. However, the embedded digital watermark and adversarial perturbation may affect each other, which makes the watermark difficult to be extracted and the adversarial perturbation ineffective. In addition, some adversarial watermarking approaches[10, 11] have been proposed, which embed a watermark as the adversarial perturbation into images. However, they often fail to enable accurate watermark extraction. Therefore, existing methods struggle to achieve effective dual-protection of image copyright.

To solve the above issues, in this paper, we propose a dual-protection approach for image copyright based on an extractable and imperceptible adversarial watermark (EIAW). Specifically, we design a strategy for embedding and optimizing the adversarial watermark using modulo operations in the frequency domain. It first generates and embeds a base adversarial watermark in the frequency embedding space, and then automatically optimizes it to attack DNNs within an embedding subspace without affecting the extractability of the watermark. As a result, this approach can effectively attack DNNs while allowing the embedded watermark to be accurately extracted. Additionally, Unlike traditional methods that rely on L_p norms for optimizing the adversarial watermark in the pixel domain[3−5], which has been proven to poorly align with human perception[12]. we introduce a frequency domain constraint to optimize watermark embedding locations, which ensures the final watermarked images maintain high quality. We illustrate the differences among the adversarial attack, digital watermark, and the proposed EIAW in Fig. 1.

In summary, Our main contributions are summarized as follows:

- **The EIAW is proposed for image copyright dual-protection:** Unlike existing adversarial attack or digital watermark methods, EIAW allows for effectively attacking DNNs and accurately extracting the watermark, enabling image copyright dual-protection.
- **The frequency domain constraint is designed to optimize the imperceptibility of the adversarial watermark:** Instead of using L_p norms to limit the perturbation in the pixel domain, we introduce a frequency domain constraint to optimize the watermark embedding locations, achieving more imperceptible adversarial watermarked images.
- **The superiority of the proposed approach is proven by extensive experiments:** Extensive experiments show that the proposed EIAW method not only achieves comparable attack effectiveness and efficiency to state-of-the-art methods, i.e., PGD, but also maintains high watermark imperceptibility, accurate watermark extraction, and robustness against various noise attacks.

The structure of this paper is organized as follows. Section 2 discusses the related work. Section 3 describes the proposed EIAW method in detail. Section 4 presents and analyzes the experimental results. Section 5 concludes the paper.

2 Related Work

In this section, we review relevant works in three areas: adversarial attack, digital watermarking, and adversarial watermark attacks.

2.1 Adversarial attack

Adversarial attack deceives DNNs by adding small perturbations to images. These perturbations can alter the model's predictions, preventing DNNs from accurately analyzing the image and thus protecting the image's semantic information. Early research by Szegedy et al.[2] introduced adversarial perturbations using back-propagation and gradient-based algorithms. Since then, various methods have been proposed to generate adversarial examples[3−5, 13, 14]. A well-known gradient-based method is the fast gradient sign method (FGSM)[5], which crafts perturbations by taking a step

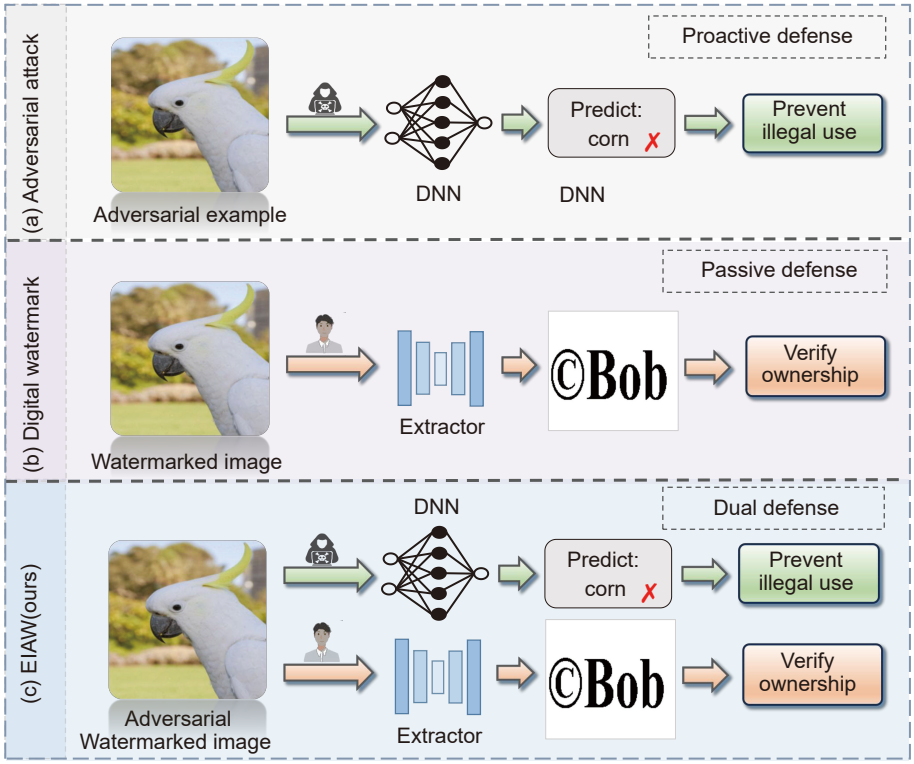


Fig. 1 Comparison of different approaches: (a) proactive protection via adversarial attack is able to prevent illegal use; (b) passive protection via digital watermark is able to verify ownership; and (c) our dual-protection approach based on EIAW can prevent illegal use and verify ownership, simultaneously.

in the direction of the gradient. Similarly, the projected gradient descent (PGD) attack[3], often considered state-of-the-art, updates perturbations iteratively starting from random initialization.

The above methods are usually carried out in the pixel domain, recent research has also explored adversarial attacks in the frequency domain. For example, Wang et al.[15] demonstrated the sensitivity of DNNs to high-frequency components, leading to the development of frequency domain adversarial attacks. Luo et al.[16] proposed a method that enhances imperceptibility by minimizing differences between the low-frequency components of clean and adversarial images. S2I-FGSM[17] enhances transferability through spectrum transformations in the frequency domain. AdvDrop[18] takes a unique approach by crafting adversarial examples through information removal rather than addition, demonstrating that minimal frequency component manipulation can significantly impact model performance. Guo et al.[19] constrained perturbations to specific frequency bands to reduce the number of black-box queries required for attacks. While our work focuses on optimizing perturbations in specific frequency bands for both attack effectiveness and watermark imperceptibility.

Despite these advancements, L_p norms are commonly used to constrain perturbations for generating more imperceptible adversarial examples. However, recent studies show that L_p norms do not align well with human perception[12]. Therefore, other perceptual distances, such as structural similarity index measure[20] (SSIM) and learned perceptual image patch similarity[21] (LPIPS), have been used to improve imperceptibility. Inspired by this, our work introduces a frequency domain constraint to generate perturbations that are less perceptible to human vision, achieving highly imperceptible adversarial watermarks.

2.2 Digital watermarking

Digital watermarking embeds meaningful information, such as copyright ownership, content description into digital content. It provides an effective solution for copyright protection, and is widely used to track and prevent copyright infringement[6, 22–24].

In general, digital watermarks can be classified into visible and invisible watermarks based on their visual effects. Visible watermarks[7] are more perceptible to human eyes, making them more susceptible to targeted

attacks and modifications. In contrast, invisible watermarks[8, 9, 25] use data redundancy techniques that typically bypass the human visual system. Recent advances in invisible watermarking have focused on enhancing robustness against complex distortions. For instance, Li et al.[26] proposed a grayscale deviation simulation method to resist screen-shooting distortions, while Fu et al.[27] developed a wavelet-based recovery network to maintain watermark integrity under various image processing operations. These works demonstrate the importance of domain-specific transformations for robust watermarking, which aligns with our frequency domain approach. Furthermore, Liao et al.[28] showed that robust feature extraction in compressed media can effectively preserve critical information, further supporting our choice of frequency domain embedding for adversarial watermarking. Additionally, depending on whether the original image is required for extraction, watermarking methods are classified as blind or non-blind. Non-blind methods merge the original and watermark images, requiring the original image during extraction. Blind watermarking embeds watermarks directly using specific rules, allowing extraction without the original image. In this work, we employ an invisible, blind-extractable watermark in the frequency domain to craft adversarial watermarks.

2.3 Adversarial watermark attacks

Recent studies have explored using watermarks as adversarial perturbations to attack DNNs. Jia et al.[10] proposed the basin hopping evolution (BHE) algorithm to embed visible watermarks into images, causing inference errors in DNNs. Jiang et al.[29] improved this with a fast differential evolution method. Zuo et al.[30] enhanced the multi-swarm particle swarm optimization (MPSO) algorithm to conduct extensive attack experiments. However, all above methods add watermarks in the pixel domain and use visible watermarks to attack. It is clear that visible watermarks degrade the quality of the protected images and are easy to notice. To generate more imperceptible images, Zhang et al.[11] introduced a frequency domain adversarial watermarking framework, which embeds adversarial watermarks in the frequency coefficients. Although this watermark is invisible, they still need the original image for extraction, limiting its practical use.

Some studies have used adversarial watermarks to protect image copyright. For example, Zhu et al.[31] proposed adversarial examples with embedded watermarks to stop generative models from copying unauthorized images. Wang et al.[32] designed an end-to-end adversarial watermark fusion model (AWFM) that combines watermark embedding and adversarial perturbations into a single task to generate invisible adversarial watermarked images. However, training these models is challenging because they require optimizing multiple objectives simultaneously.

In this paper, we propose a dual-protection adversarial watermarking framework. It embeds extractable watermarks into the frequency coefficients under the constraints, ensuring imperceptibility while enabling watermarks can be extracted accurately.

3 Methodology

3.1 Overview

Given an original watermark image $W_{\text{in}} \in \mathbf{R}^{N\times N}$, a cover image $I \in \mathbf{R}^{N\times N}$, its ground-truth label y, and a DNN classifier f: $I \to \mathbf{R}^k$ that maps the image to an output vector representing a probability distribution over the discrete label set $\{1, 2, \ldots, k\}$, where k is the number of classes. We use the cross-entropy as the loss function, denoted by L. The formula for the cross-entropy loss is

$$L = -\sum_{i=1}^{k} y_i \log(f(I)_i)$$

where y_i is the ground-truth label for class i, and $f(I)_i$ is the predicted probability for class i.

Let I_{adv} denote the adversarial example, I_{w} denote the watermarked image after embedding the base watermark, W_{out} denote the extracted watermark, F represents the frequency domain coefficient matrix of the cover image, F_{w} denotes the frequency-domain matrix after watermark embedding, and F_{aw} denotes the adversarial frequency domain matrix generated during the attack process. We use the discrete cosine transform (DCT) to convert the cover image from the pixel domain to the frequency domain, and employ the inverse DCT (IDCT) to revert it back.

The goal of adversarial attacks is to craft I_{adv} such that $f(I_{\text{adv}}) \neq y$, while the aim of blind watermarking is to embed a watermark W_{in} into the cover image to obtain the watermarked image I_{w}, from which the watermark information W_{out} can be extracted accurately. Different from the above two methods, our goal is to embed and optimize a watermark into the image I while allowing for effectively attacking DNNs

and accurately extracting the watermark, simultaneously. In this paper, we propose a novel frequency domain adversarial watermark algorithm EIAW, which can generate extractable and imperceptible adversarial watermark. The pipeline of the EIAW approach is shown in Fig. 2.

Next, we will first introduce the frequency domain constraint, followed by a detailed presentation of our proposed EIAW, approach by three parts: watermark embedding, watermark optimization, and watermark extraction.

3.2 Frequency domain constraint

In digital watermarking, watermarks are often embedded in the frequency domain. This approach inspires us to craft watermark perturbations directly in the frequency domain. To evaluate the effectiveness of perturbations in different frequency regions, we calculate the gradients of the loss function L with respect to both the frequency domain image F and the pixel domain image I,

$$\nabla_F L = \frac{\partial L(\theta, \mathrm{IDCT}(F), y)}{\partial I} \cdot \frac{\partial I}{\partial F} \tag{1}$$

$$\nabla_I L = \frac{\partial L(\theta, I, y)}{\partial I} \tag{2}$$

where θ represents the model parameters, $\nabla_F L$ and $\nabla_I L$ are the gradients of the loss function L with respect to the input F (in the frequency domain) and I (in the pixel domain), respectively, and y is the true label of the cover image.

The magnitude of these gradients indicates how changes at each location affect the loss. In other words, regions with larger gradients have a greater influence on the model's inference.

To make it more intuitive, we visualize the gradient heatmaps in Fig. 3. From the heatmaps, we observe that the gradient distribution in the pixel domain lacks distinct patterns. In contrast, the frequency domain gradient distribution shows a clear trend. Specifically, mid-to-low frequency regions (top-left) exhibit larger gradients, while high-frequency regions (bottom-right) have smaller gradients. This means that perturbations in mid-to-low frequency regions achieve better attack performance. Even small perturbations in these regions can significantly affect the model's predictions. Although these perturbations may cause some degradation in image quality, the improved attack effectiveness outweighs the visual loss. Overall, disturbances in low and medium frequencies are more cost-effective.

Based on this observation, we design a frequency domain constraint to limit watermark perturbations to mid-to-low frequency regions. Specifically, we assign the value "1" to locations within the mid-to-low frequency areas, allowing perturbations to be added. All other regions are assigned the value "0", creating a binary mask M,

$$M = \begin{cases} 1, & \text{if } m \leqslant u,\ v \leqslant n; \\ 0, & \text{otherwise,} \end{cases}$$

where $m, n \in [0, N]$, and N represents the dimension of the frequency-domain coefficient matrix F, equivalently, the dimension of the pixel domain image I. And u and v denote the horizontal and vertical coordinates in the frequency domain. The size of the mask is denoted as $\alpha = (n-m)/N \in [0, 1]$, while the starting position of the mask is defined as $\beta = m/N$. For images with multiple color channels, the same masking strategy is applied independently to each channel.

During the watermark embedding and optimization stages, we only modify the DCT coefficients in regions where the mask value is 1. This ensures that the watermark perturbations are confined to the mid-to-low

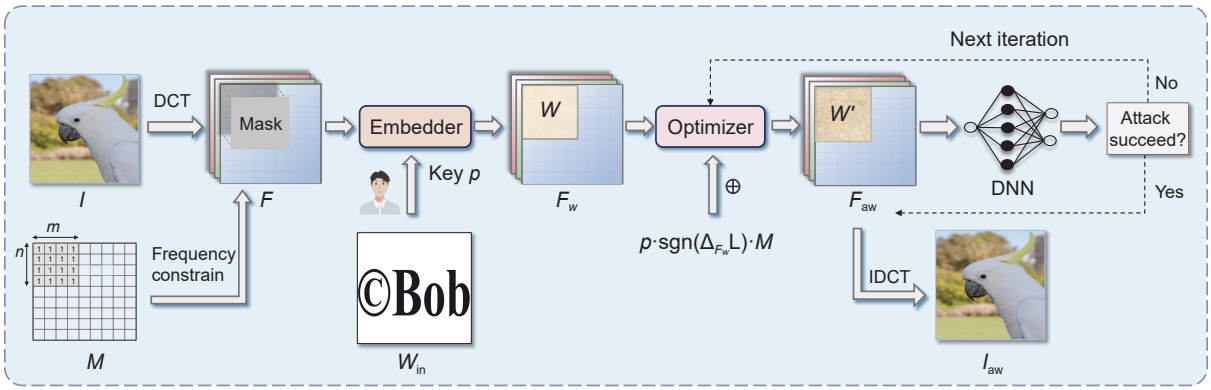


Fig. 2 Pipeline of the proposed EIAW approach.

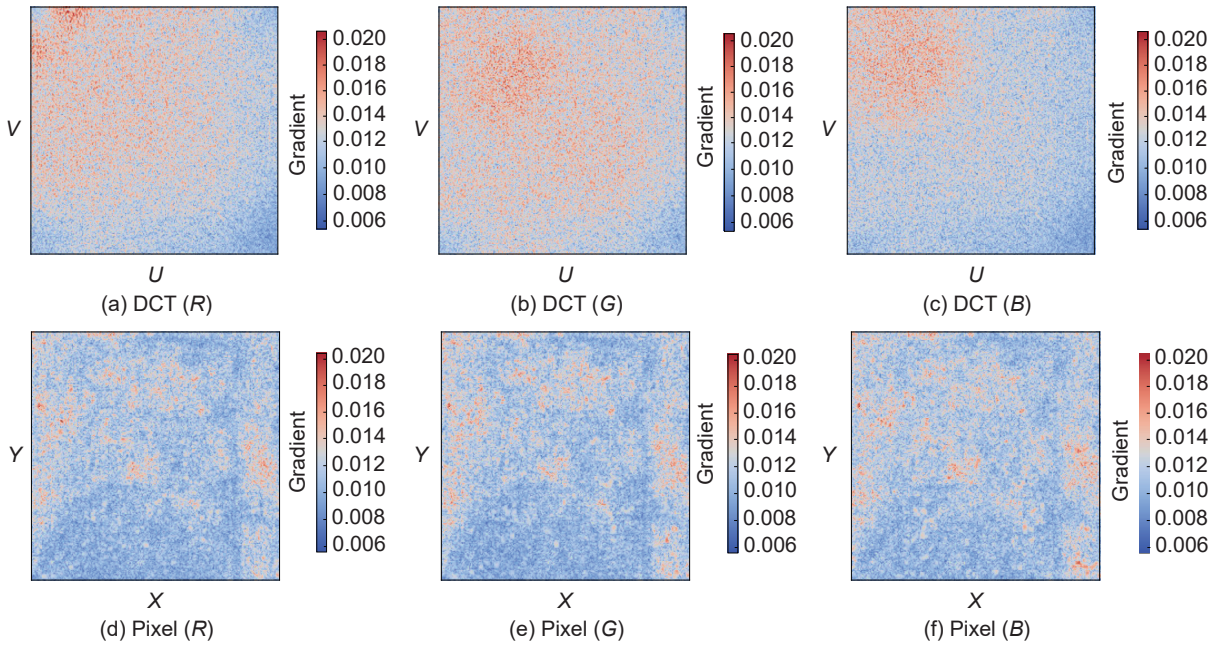


Fig. 3 Gradient heatmaps of the RGB channel in the frequency domain (a-c) and pixel domain (d-f), respectively. In the pixel domain, the axes X and Y denote the spatial coordinates of the image. In the frequency domain, the axes U and V denote the horizontal and vertical frequency coordinates obtained after applying the DCT.

frequency areas.

3.3 Watermark embedding

To enable accurate watermark extraction without requiring the cover image, we introduce the watermark embedding and extracting strategy based on modulo operation in the frequency domain. First, the image is transformed from the pixel domain to the frequency domain, where watermark bits are embedded by adjusting frequency coefficients based on the remainder of the modulo operation. More details are given as follows.

Within the masked region, where the watermark bit w_{in}=1, the DCT coefficients are modified according to the following strategy:

$$\begin{aligned}&\text{if } F(u,\ v)\ \bmod\ p \geqslant \frac{p}{2},\\&\text{then } F(u,\ v) \longleftarrow F(u,\ v)+\frac{p}{2},\\&\text{else } F(u,\ v) \longleftarrow F(u,\ v)\end{aligned} \tag{3}$$

Conversely, when the watermark bit w_{in}=0, we perform the following strategy:

$$\begin{aligned}&\text{if } F(u,\ v)\ \bmod\ p < \frac{p}{2},\\&\text{then } F(u,\ v) \longleftarrow F(u,\ v)+\frac{p}{2},\\&\text{else } F(u,\ v) \longleftarrow F(u,\ v)\end{aligned} \tag{4}$$

where $F=\text{DCT}(I)$ represents the frequency-domain coefficient matrix obtained by applying the DCT to image I. The entry at position $(u,\ v)$ is written as $F(u,\ v)$. The parameter p is the embedding key, which must be an even integer.

By modifying the DCT coefficients according to the above embedding strategy, we successfully embed the watermark and obtain the base watermarked image I_{w}. The embedding process is described in Algorithm 1.

Algorithm 1 EIAW embedding process

Input: Cover image I, watermark w_{in}, modulus p, and mask M

Output: Watermarked image I_{w}

$F \leftarrow \text{DCT}(I)$;
for $(u,\ v)$ in the mask region, where M=1 **do**
  if $w_{\text{in}}(u,\ v)=1$ **then**
    if $F(u,\ v)\ \bmod p \geqslant p/2$ **then**
      $F(u,\ v) \leftarrow F(u,\ v)+p/2$;
    end if
  else
    if $F(u,\ v)\ \bmod p < p/2$ **then**
      $F(u,\ v) \leftarrow F(u,\ v)+p/2$;
    end if
  end if
end for
$I_w \leftarrow \text{IDCT}(F)$;
return I_w

Notably, since explicit for-loops are computationally inefficient, we leverage PyTorch's array slicing operations to achieve implicit parallel computation, accelerating the watermark embedding.

3.4 Watermark optimization

To obtain the adversarial watermark without affecting the watermark extraction, we optimize the watermark by changing the modification magnitudes of frequency coefficients to attack DNNs within the embedding subspace. In such manner, the remainders of DCT coefficients when divided by p can be kept unchanged, and thus the extractability will not be affected. Specifically, to optimize the watermark in the frequency domain embedding subspace, we propagate the gradient of the loss using the following chain rule in Eq. (1). Then, we update the DCT coefficients by stepping in the direction of the gradient. To avoid affecting the extractability of the watermark, the remainders of the DCT coefficients should remain unchanged. Therefore, the step size must be an integer multiple of p. The specific formula is as follows:

$$F_{\mathrm{aw}}^{t+1} = F_{\mathrm{aw}}^{t} + p \times \mathrm{sgn}(\nabla_F L) \times M \tag{5}$$

where $\mathrm{sgn}(\cdot)$ is the sign function, M represents the frequency-domain mask. F_{aw}^{t} and F_{aw}^{t+1} denote the adversarial frequency-domain coefficient matrices at the t-th and $(t+1)$-th iteration, respectively.

Assuming the attack succeeds after $t+1$ iterations, $F_{\mathrm{aw}} = F_{\mathrm{aw}}^{t+1}$ denotes the frequency domain adversarial watermarked image. Then, we convert it back to the pixel domain to obtain the final image I_{aw},

$$I_{\mathrm{aw}} = \mathrm{Clip}(\mathrm{IDCT}(F_{\mathrm{aw}}), 0, 1) \tag{6}$$

where $\mathrm{Clip}(\cdot)$ is the operation that restricts the resulting values to the range [0, 1]. Algorithm 2 describes the pseudo-code of optimization algorithm. By optimizing the watermark to attack DNNs within a frequency domain embedding subspace without affecting the extractability of the watermark, we achieve the dual goals of preventing illegal use and verifying ownership, simultaneously. To provide a clear visualization of the constrained optimization process, we illustrate the optimization process in Fig. 4.

3.5 Watermark extraction

Watermark extraction is the inverse process of embedding. Specifically, given the adversarial watermarked image I'_{aw}, we first convert it to frequency domain, then, in the region where $M=1$, the watermark is extracted using the following strategy:

$$w_{\mathrm{out}} = \begin{cases} 1, & \text{if } \sum_{0}^{c-1}\left((F'_{\mathrm{aw}} \mod p) < \frac{p}{2}\right) \geqslant \lceil \frac{c}{2} \rceil, \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

Algorithm 2 EIAW optimization process

Input: Classifier f with loss function L, cover image I, label y, watermark W_{in}, key p

Output: Adversarial watermarked image I_{aw}

$F = \mathrm{DCT}(I)$;
$F_{\mathrm{w}} = \mathrm{Embed}(F, W_{\mathrm{in}}, p, M)$;
Initialize iter $\leftarrow 0$, attack_successful $\leftarrow$ false;
while iter $\leqslant$ Max_iter **do**
  iter $\leftarrow$ iter $+ 1$;
  $F_{\mathrm{w}} \leftarrow F_{\mathrm{w}} + p \times \mathrm{sgn}(\nabla_{F_{\mathrm{w}}} L) \times \mathrm{M}$;
  $I_{\mathrm{aw}} \leftarrow \mathrm{IDCT}(F_{\mathrm{w}})$;
  if $\arg\max_{\hat{y}} f(I_{\mathrm{aw}}) \neq y$ **then**
    attack_successful $\leftarrow$ true;
    break
  end if
end while
if attack_successful **then**
  return I_{aw};
else
  return false;
end if

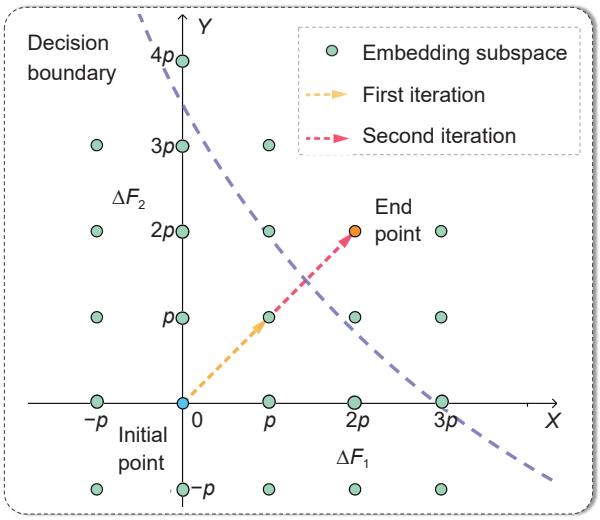


Fig. 4 Visualization of the watermark optimization process. The X and Y axes in the 2D space represent the modification magnitudes of DCT coefficients, i.e., ΔF_1 and ΔF_2, respectively. The 2D space represents the embedding space, and the set of green points in the 2D space represents the embedding subspace in which the adversarial watermark can be optimized without affecting its extractability.

where $F'_{aw} = \mathrm{DCT}(I'_{aw})$ denotes the DCT matrix, and c denotes the number of pixel channels, with $c = 3$ for RGB images. Finally, W_{out} is extracted to recover the original watermark W_{in} without the cover image.

Notably, in addition to embedding 2D binary watermark images, our method also supports direct embedding of arbitrary-length bit-streams. To achieve this, we employs Zigzag scanning[33] to convert DCT coefficients into a 1D sequence. The bit-stream is then directly embedded into a selected mid-low frequency band whose length matches that of the bit-stream, enabling flexible and scalable embedding. The embedding, optimization, and extraction procedures for bit-streams follow the same strategy as those used for image-based watermarks, ensuring consistency and robustness.

4 Experiment

In this section, we evaluate and compare the proposed approach to demonstrate its performances in terms of attack effectiveness, efficiency, watermark extraction, and imperceptibility. Typical experimental results of EIAW approach are shown in Fig. 5. The original images can be correctly classified by ResNet101. After embedding the adversarial watermarks generated by EIAW, they can mislead the pre-trained ResNet101 while maintain good visual quality.

4.1 Experiment settings

4.1.1 Dataset and models

We evaluate the proposed EIAW approach on 1500 images from the ImageNet-1K[34] and CIFAR-10 datasets. We use ResNet101[35], AlexNet[36], VGG19[37], SqueezeNet1_0[38], MobileNet_V2[39], and Inception_V3[40] as the target models.

4.1.2 Baselines

To compare the effectiveness and efficiency of the watermark attack, we use several adversarial watermark attack methods: Adv-watermark [10], AFW[11], and MISPSO[30]. We also compare with typical gradient-based attack methods under the L_p norms, including PGD[3] and FGSM[5], C&W[4]. Additionally, we compare with the two-phase method that adds perturbations and watermarks sequentially.

4.1.3 Evaluation metrics

We use the following metrics to evaluate the attack capability, extraction accuracy, and imperceptibility of different methods.

(1) Attack capability: Attack effectiveness and efficiency are evaluated using the attack success rate (ASR) and attack time (AT), respectively. ASR measures the percentage of images for which the pre-trained model changes its prediction after the images are altered. AT measures the average time required to attack a single image. In addition, we also report the

Fig. 5 Experimental results of the EIAW method. The first row shows the original images, the second row displays the corresponding true labels, the third row presents the protected images generated by EIAW, along with their predicted labels, PSNR, and SSIM metrics, and the last row shows the extracted watermark images.

increase in cross-entropy loss (LOSS), where a larger LOSS indicates a greater degradation of model prediction accuracy, thus reflecting a more successful attack.

(2) Imperceptibility: To sufficiently evaluate the watermark's imperceptibility, we use the peak signal-to-noise ratio (PSNR), SSIM[20], and LPIPS[21]. These metrics compare the quality of the watermarked image with the original image to assess how perceptible the watermark.

(3) Extraction accuracy: The extraction accuracy rate (EAR) measures the watermark's extractability. It is calculated as the ratio of the number of correctly extracted bits to the total number of bits. A higher EAR value, closer to 1, indicates more accurate extraction.

4.1.4 Evaluation environment

To make a fair comparsion, all experiments are conducted on an NVIDIA GeForce RTX 3090 GPU using the PyTorch framework in Python.

4.2 Parameter setting

In the proposed EIAW method, there are two key parameters: p, which is the key to conducted modular operation, and M, which determines where the watermark perturbation is applied. We evaluate the impact of these parameters on the method's performance in terms of attack capability and imperceptibility. The impacts of parameter p is shown in Table 1. As it illustrates, increasing p slightly improves attack efficiency by reducing the attack time. However, this improvement comes at huge cost of image quality, as evidenced by the decrease in PSNR and SSIM and the increase in LPIPS as p grows. Therefore, we choose the smallest parameter $p = 2$ to minimize quality loss.

The impact of the mask's location is shown in Fig. 6.

Table 1 Impact of key p on attack time and watermark imperceptibility. "↑" means larger is better, while "↓" means smaller is better.

p	Attack time (s)	Imperceptibility		
		PSNR ↑	SSIM[20] ↑	LPIPS[21] ↓
2	0.0911	51.819	0.9963	0.0011
4	0.0711	45.604	0.9888	0.0043
6	0.0695	42.377	0.9782	0.0099
8	0.0652	40.027	0.9650	0.0178
10	0.0634	38.183	0.9501	0.0273

Masks of the same size are applied to different frequency regions: low-frequency (M1), mid-low-frequency (M2), mid-high-frequency (M3), and high-frequency (M4). The plots show how the loss and image quality metrics change over iterations. From Fig. 6, it is clear that low-frequency perturbations are more effective for attacks, as they cause the loss to increase rapidly. However, they also significantly degrade image quality. On the other hand, high-frequency perturbations have a smaller impact on image quality but are less effective for attacks. Therefore, we constrain the perturbations to the mid-low-frequency region, which strikes a balance between attack effectiveness and image quality.

Finally, if not specified, the watermark attack parameters in EIAW are set as follows: $p = 2$, $\alpha = 0$, $\beta = 0.5$, with a maximum of 20 iterations. Additionally, a monitoring mechanism is used for all iterative attack methods: if the attack is successful, the iteration terminates early.

4.3 Comparison of attacking performance

By embedding an additional watermark into the original image, we disrupt key local regions that are essential for image classification, effectively

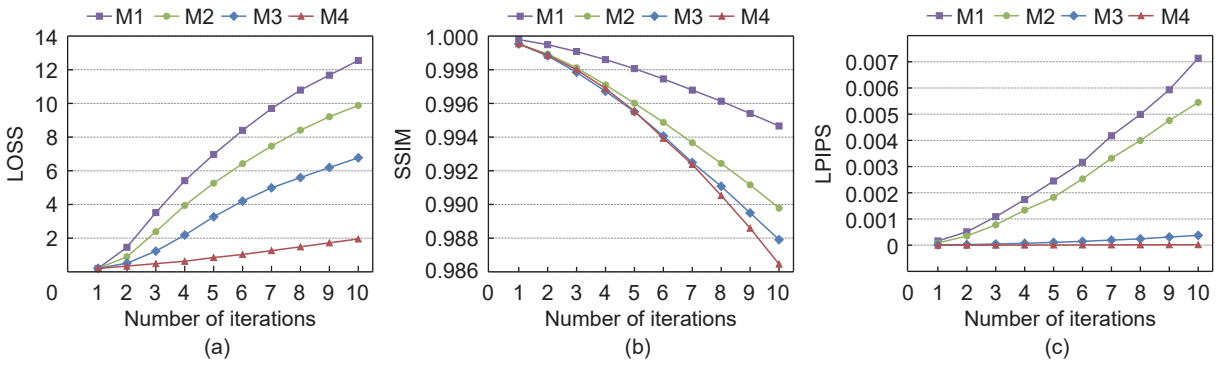


Fig. 6 Impact of the frequency domain mask's location on image quality. The three subfigures show the mask applied to different regions (M1, M2, M3, and M4) and illustrate the changes in LOSS, SSIM, and LPIPS across iterations.

misleading a well-trained neural network. To illustrate the impact of the adversarial watermark, we visualize the attention maps using gradient weighted class activation mapping (Grad-CAM)[41]. As shown in Fig. 7, the clean images, correctly classified by ResNet101, along with their attention maps and labels in green (Fig. 7a). The adversarial watermarked images, along with their attention maps and labels in red (Fig. 7b), by embedding an additional watermark into the original image, we disrupt key local regions that are essential for classification, thereby misleading the well-trained Resnet101[35] model.

To further quantitatively evaluate the attack performance of the proposed EIAW method, we compare six attack methods across five widely used pretrained classification models (ResNet101, AlexNet, VGG19, Inception_V3, and SqueezeNet1_0) on ImageNet-1K[34] and CIFAR-10 datasets. Table 2 shows the attack success rates (ASR) of various attack methods on different neural network models. Traditional attack methods, such as PGD[3] and C&W[4], achieve high ASRs across all models. The Adv-watermark method[10], based on visible watermarking, achieves an average ASR of 71.7%, while the MISPSO[30] method, using the MPSO algorithm, reaches 75.9%. The AFW method, which embeds the watermark in the frequency domain, achieves a high ASR of 97.4%. In comparison, our proposed EIAW method achieves a comparable attack effectiveness to PGD[3], and outperforms other adversarial watermark methods. We further evaluate the black-box transferability of EIAW through comprehensive cross-model validation. As shown in Table 3, when adversarial examples generated from one model architecture are transferred to attack other unseen models, these results demonstrate that our frequency domain perturbations maintain reasonable effectiveness across different architectures, making them practical for real-world scenarios where the target model may be unknown.

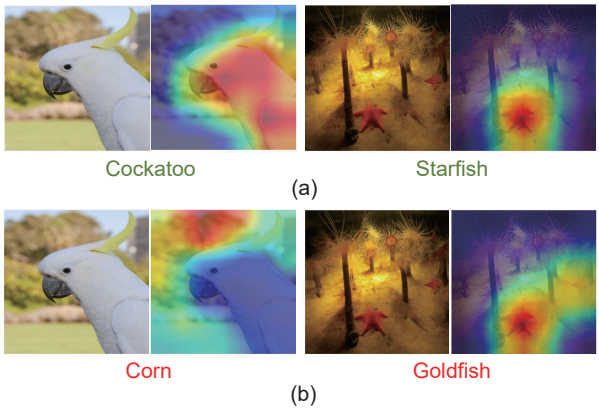


Fig. 7 Effects of EIAW on the attention maps of neural networks (based on ResNet101 predictions). (a) The original images, along with their attention maps and predicted categories (cockatoo and starfish); (b) The adversarial watermarked images, along with their attention maps, and the predicted categories (corn and goldfish).

Next, we evaluate the efficiency of the proposed method. Table 4 compares the attack efficiency of the EIAW method with the state-of-the-art PGD[3] attack and the adversarial watermark method AFW. As shown in Table 4, EIAW achieves an average ASR of 99.5% with a runtime of 0.0467 s. This is slightly slower than PGD, but much faster than AFW. This shows that, despite the conversions between the frequency domain and pixel domain, the additional computational cost is minimal. However, this minor time overhead provides significant benefits, such as improved image quality and the ability to extract the watermark for dual protection, as demonstrated in the following sections.

4.4 Comparison of imperceptibility

To achieve a more secure and practical copyright

Table 2 ASR comparison with various methods on Imagenet against different models. Bold values indicate the best result in each column.

(%)

Attack	Resnet101[35]	Inception_V3[40]	Squeezenet1_0[38]	VGG19[37]	Average
PGD[3]	**100.0**	98.3	**100.0**	99.4	99.4
FGSM[5]	90.2	83.9	99.4	96.0	92.4
C&W[4]	99.8	**98.5**	**100.0**	98.8	**99.5**
Adv-watermark[10]	78.0	73.0	69.1	64.8	71.7
MISPSO[30]	68.9	76.5	85.4	73.1	75.9
AFW[11]	98.2	92.1	**100.0**	99.6	97.4
EIAW (Ours)	**100.0**	98.2	**100.0**	**99.6**	**99.5**

Table 3 Transferability of ASR for adversarial watermarked images across surrogate and attack models. Bold values indicate the best result in each column.

(%)

Surrogate model	Attack model			
	ResNet101	VGG19	Inception_V3	SqueezeNet1_0
ResNet101	**100.0**	68.3	62.5	72.8
VGG19	65.7	**99.6**	58.9	70.2
Inception_V3	63.2	60.4	**99.6**	68.7
SqueezeNet1_0	70.5	66.8	64.3	**100.0**

Table 4 Comparison of ASR and AT across different models and attack methods on Imagenet and CIFAR-10 datasets.

Dataset	Model	PGD[3]		AFW[11]		EIAW (ours)	
		ASR (%)	AT (s)	ASR (%)	AT (s)	ASR (%)	AT (s)
ImageNet[34]	ResNet101[35]	100.0	0.0563	98.2	0.4302	100.0	0.0910
	AlexNet[36]	100.0	0.0086	99.5	0.2997	100.0	0.0171
	VGG19[37]	99.5	0.0157	99.6	0.3264	99.6	0.0236
	Inception_V3[40]	97.9	0.0633	92.1	0.4446	98.2	0.0806
	Squeezenet1_0[38]	100.0	0.0131	100.0	0.3230	100.0	0.0211
	Average	100.0	0.0314	97.9	0.3647	99.5	0.0467
CIFAR-10[41]	ResNet50[35]	100.0	0.0401	84.3	0.1264	99.8	0.1233
	AlexNet[36]	96.7	0.0384	86.8	0.1334	100.0	0.0245
	VGG19[37]	99.3	0.0514	89.4	0.1611	100.0	0.0293
	Mobilenet_V2[39]	100.0	0.0338	87.4	0.2309	98.7	0.1290
	Average	99.8	0.0414	87.0	0.1620	99.2	0.0765

protection, the protected image should be visually indistinguishable from the original image. In this section, we compare the imperceptibility of watermarks or perturbations generated by different methods. To sufficiently evaluate the imperceptibility, we use three standard metrics: PSNR, SSIM[20], and LPIPS[21]. Higher PSNR and SSIM values indicate better image quality, while lower LPIPS values reflect less perceptual distortion, meaning the watermark or perturbation is less noticeable to the human eye.

As shown in Table 5, for the ResNet101 model, EIAW achieves the highest PSNR of 52.68 and SSIM of 0.9970, outperforming other methods. Its LPIPS score of 0.0006 indicates that the perceptual difference between the original and protected images is minimal. Similar results are observed for other models, where EIAW consistently achieves higher PSNR and SSIM values and lower LPIPS scores compared to other methods. In comparison, attack methods such as PGD[3] and FGSM[5], which constrain perturbation in the pixel domain using L_p norms, cause significant quality degradation. Other adversarial watermark methods, such as Adv-watermark, embed visible watermarks, making them more noticeable and leading to greater quality loss. Although AFW operates in the frequency domain, it applies perturbations to the entire coefficient, resulting in noticeable distortions as well. In contrast, our method uses a frequency domain constraint to limit perturbations to specific regions, achieving better image quality.

Table 5 Comparison of imperceptibility against different models under various attack methods on Imagenet. Bold values indicate the best result in each column. "↑" means larger is better, while "↓" means smaller is better.

Model	Attack method	Imperceptibility		
		PSNR ↑	SSIM[20] ↑	LPIPS[21] ↓
ResNet101	PGD[3]	45.40	0.9896	0.0015
	FGSM[5]	36.21	0.9250	0.0309
	Adv-watermark[10]	25.91	0.9604	0.0739
	AFW[11]	36.07	0.9500	0.0120
	EIAW (ours)	**52.68**	**0.9970**	**0.0006**
VGG19	PGD[3]	45.45	0.9886	0.0013
	FGSM[5]	36.21	0.9206	0.0294
	Adv-watermark[10]	25.11	0.9604	0.0711
	AFW[11]	36.49	0.9545	0.0100
	EIAW (ours)	**51.29**	**0.9968**	**0.0008**

Table 6 Average EAR of EIAW on different models. Bold value indicates the best result.

(%)

Resnet101[35]	alexnet[36]	Inception_V3[40]	Vgg19[37]	Squeezenet1_0[38]
92.10	**93.19**	92.68	91.98	92.24

4.5 Watermark extraction performance

In this section, we evaluate the watermark extraction performance of the proposed method. From Table 7, we can see that EIAW achieves a high watermark EAR of over 92% across different models. This level of performance is sufficient for practical applications, such as copyright verification. We do not compare our method with Adv-watermark and AFW here, as both require the original image for watermark extraction, while our approach can extract the watermark based solely on the secret key p, which is more suitable for real-world applications.

The watermark should remain extractable even when the image is distorted by various attacks. Therefore, to further demonstrate the robustness of the watermark extraction, we conduct both qualitative and quantitative evidence of the watermark's resilience to common image processing operations. As shown in Fig. 8 and Table 9, we apply several common image distortions, such as JPEG compression, and cropping, and evaluate the watermark extraction performance under these conditions. The results show that, although the extracted watermark images exhibit some distortion after noise is added, the watermark remains recognizable. This demonstrates that our method can preserve copyright information even under distortions.

Table 7 Bit-stream embedding performance.

Bit length	PSNR	EAR (%)	ASR (%)
512-bit	53.3	91.3	94.9
1024-bit	53.8	92.1	98.3
2048-bit	53.8	92.4	99.7
4096-bit	53.1	93.6	100.0

4.6 Bit-stream embedding evaluation

To further validate the flexibility and effectiveness of our method, we conduct an additional experiment using bit-stream embedding instead of binary watermark images. Specifically, we adopt a Zigzag-based encoding[33] strategy to directly embed arbitrary-length bit-streams into the mid-low frequency band of the DCT domain, the length of this band is the same as the length of the bit-stream. The experimental results are presented in Table 7. Across different bit-stream lengths (512 bits to 4096 bits), the proposed method consistently achieves high PSNR values, indicating excellent imperceptibility. Notably, the watermark extraction remains robust, with EAR improving as the bit length increases, reaching 93.6% in the 4096-bit case. Similarly, the ASR increases with the embedding payload, achieving 100% when embedding 4096 bits.

From a theoretical perspective, our method modifies DCT coefficients through constrained optimization, which is agnostic to the semantic content of the embedded data. Whether the watermark is a 2D binary image or a 1D bit-stream, they are mathematically isomorphic in the frequency domain. Moreover, the energy distribution in the selected frequency band remains consistent across different input formats, which ensures that the embedding process is format-independent. These demonstrates that our method not

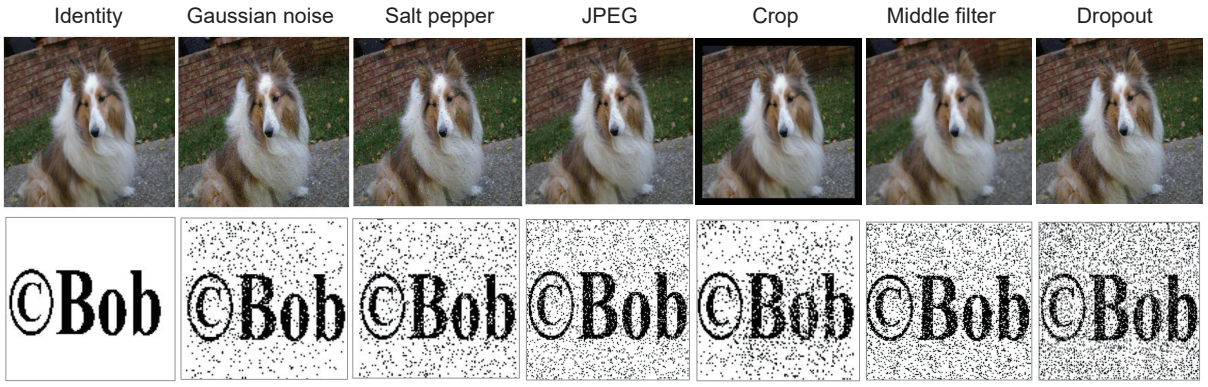


Fig. 8 Robustness of the watermark under different noises. The first row shows the protected images with different editing operations, and the second row shows the extracted watermark images.

only supports 2D binary watermark images, but also generalizes well to arbitrary bit-streams, further extending its applicability to real-world digital copyright protection.

4.7 Comparison with two-phase method

To achieve copyright dual-protection, the most intuitive manner is to embed the watermark and perturbation sequentially, which we denote as the two-phase method. To demonstrate the superiority of our method over the two-phase approach, we conducted experiments in terms of the ASR, EAR, and imperceptibility.

The results are shown in Table 8, as we can see, in terms of EAR, the proposed EIAW method achieves an extraction accuracy of 92.34%, outperforming the two-phase method. This is mainly because, when the perturbation and watermark are added sequentially, they interfere with each other, making extraction more difficult. And in terms of imperceptibility, because the two-phase method requires two modifications, it leads to greater quality degradation. In contrast, EIAW embeds a single adversarial watermark and improves image quality using a frequency domain constraint, thereby achieving better performance.

4.8 Ablation study

In this section, we conduct an ablation study to prove the effect of the frequency domain constraint on image quality. We compare three strategies: (1) frequency domain constrained attack (F-c), which applies perturbations to specific frequency components using the mask, (2) frequency domain full coverage attack (F-f), which applies perturbations to all frequency components, and (3) pixel domain full coverage attack (P-f), which applies perturbations to the entire image in the pixel domain, and using L_p norms to constrain the perturbations.

As shown in Fig. 9, the frequency domain constrained attack (F-c) achieves the best image quality among the three strategies. In contrast, the frequency domain full coverage attack (F-f), which applies perturbations to all frequencies, causes more image quality degradation, even worse than the pixel domain full coverage attack (P-f). It further validates the effectiveness of our frequency domain constraint in improving image quality.

5 Conclusion

This paper introduces a novel approach for dual-protection of image copyright, based on an extractable and imperceptible adversarial watermark, EIAW. It automatically embeds and optimizes an adversarial watermark to prevent illegal use and verify ownership simultaneously. Additionally, we propose a frequency domain constraint to optimize the locations for watermark embedding. Experimental results show that the proposed EIAW approach achieves attack effectiveness comparable to state-of-the-art methods while enabling accurate watermark extraction and maintaining high image quality. As a result, the EIAW approach is well-suited for a wide range of real-world applications, providing comprehensive protection for image copyright.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Nos. 62372125 and 62476113), the Guangdong Natural Science Funds for Distinguished Young Scholar (No. 2023B1515020041), the Ministry of Science and Technology Xiongan New Area Science and Technology Innovation Special Sub-course (No. 2022XAGG0126), the Liaoning Collaboration Innovation Center For CSLE, the National College Student Innovation Training Program of Guangzhou University (No. 202411078002), and the Provincial College Student Innovation Training Program of Guangzhou University (No. 202511078077).

Table 8 Comparison with the two-phase method in terms of ASR, EAR and imperceptibility. Bold values indicates the best performance among different methods. "↑" means larger is better, while "↓" means smaller is better.

Method	ASR (%)	EAR (%)	Imperceptibility		
			PSNR ↑	SSIM ↑	LPIPS ↓
PGD[3]	**100.00**	50.00	45.40	0.988	0.0015
PGD + Watermarking	76.42	91.57	44.41	0.985	0.0016
Watermarking + PGD	**100.00**	65.23	44.39	0.986	0.0017
EIAW (ours)	**100.00**	**92.34**	**52.68**	**0.997**	**0.0006**

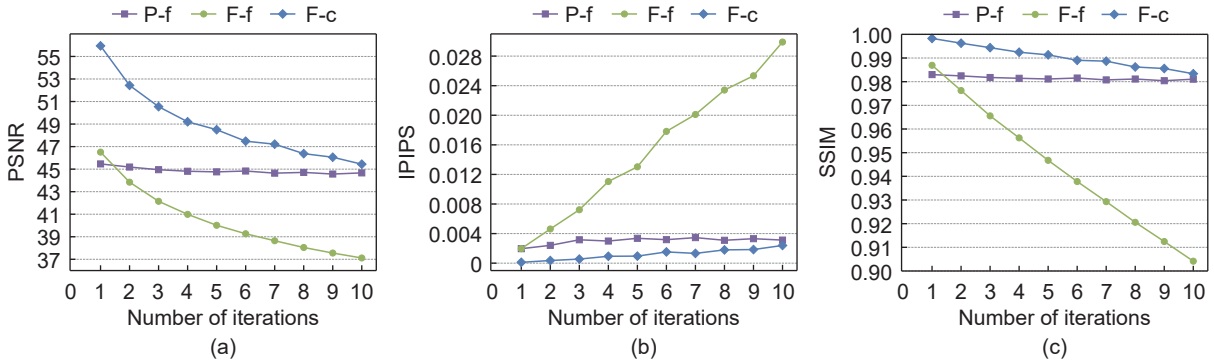


Fig. 9 Comparison of image quality over iterations for three strategies.

Table 9 Robustness of watermark EAR and ASR on ImageNet.

Group	Type	EAR (%)	ASR (%)
Identity	/	92.34	100.00
	JPEG	78.52	82.44
Attacks	Crop	82.43	92.46
	Middle Filter	77.26	83.45

References

[1] S. Mavaddati, Voice-based age, gender, and language recognition based on ResNet deep model and transfer learning in spectro-temporal domain, *Neurocomputing*, vol. 580, p. 127429, 2024.

[2] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, Intriguing properties of neural networks, arXiv preprint arXiv: 1312.6199, 2013.

[3] I. J. Goodfellow, J. Shlens, and C. Szegedy, Explaining and harnessing adversarial examples, arXiv preprint arXiv: 1412.6572, 2014.

[4] N. Carlini and D. Wagner, Towards evaluating the robustness of neural networks, in *Proc. 2017 IEEE Symp. Security and Privacy*, San Jose, CA, USA, 2017, pp. 39–57.

[5] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, Towards deep learning models resistant to adversarial attacks, arXiv preprint arXiv: 1706.06083, 2017.

[6] D. Awasthi, A. Tiwari, P. Khare, and V. K. Srivastava, A comprehensive review on optimization-based image watermarking techniques for copyright protection, *Expert Syst. Appl.*, vol. 242, p. 122830, 2024.

[7] A. V. Nadimpalli and A. Rattani, ProActive DeepFake detection using GAN-based visible watermarking, *ACM Trans. Multimedia Comput., Commun. Appl.*, vol. 20, no. 11, p. 344, 2024.

[8] H. K. Singh and A. K. Singh, Digital image watermarking using deep learning, *Multimedia Tools Appl.*, vol. 83, no. 1, pp. 2979–2994, 2024.

[9] Z. Jia, H. Fang, and W. Zhang, MBRS: Enhancing robustness of DNN-based watermarking by mini-batch of real and simulated JPEG compression, in *Proc. 29th ACM Int. Conf. Multimedia*, Virtual Event, 2021, pp. 41–49.

[10] X. Jia, X. Wei, X. Cao, and X. Han, Adv-watermark: A novel watermark perturbation for adversarial examples, in *Proc. 28th ACM Int. Conf. Multimedia*, Seattle, WA, USA, 2020, pp. 1579–1587.

[11] H. Zhang, G. Cao, X. Zhang, J. Xiang, and C. Wu, Making adversarial attack imperceptible in frequency domain: A watermark-based framework, in *Proc. 2023 IEEE Int. Conf. Multimedia and Expo*, Brisbane, Australia, 2023, pp. 43–48.

[12] M. Sharif, L. Bauer, and M. K. Reiter, On the suitability of Lp-norms for creating and preventing adversarial examples, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops*, Salt Lake City, UT, USA, 2018, pp. 1686–16868.

[13] J. Chen, X. Liu, S. Liang, X. Jia, and Y. Xun, Universal watermark vaccine: Universal adversarial perturbations for watermark protection, in *Proc. 2023 IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops*, Vancouver, Canada, 2023, pp. 2322–2329.

[14] M. Macas, C. Wu, and W. Fuertes, Adversarial examples: A survey of attacks and defenses in deep learning-enabled cybersecurity systems, *Expert Syst. Appl.*, vol. 238, p. 122223, 2024.

[15] H. Wang, X. Wu, Z. Huang, and E. P. Xing, High-frequency component helps explain the generalization of convolutional neural networks, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 8681–8691.

[16] C. Luo, Q. Lin, W. Xie, B. Wu, J. Xie, and L. Shen, Frequency-driven imperceptible adversarial attack on semantic similarity, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 2022, pp. 15294–15303.

[17] Y. Long, Q. Zhang, B. Zeng, L. Gao, X. Liu, J. Zhang, and J. Song, Frequency domain model augmentation for adversarial attack, in *Proc. 17th European Conf. Computer Vision*, Tel Aviv, Israel, 2022, pp. 549–566.

[18] R. Duan, Y. Chen, D. Niu, Y. Yang, A. K. Qin, and Y. He, AdvDrop: Adversarial attack to DNNs by dropping information, in *Proc. IEEE/CVF Int. Conf. Computer Vision*, Montreal, Canada, 2021, pp. 7486–7495.

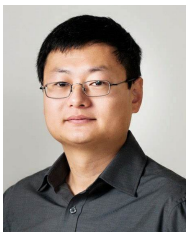
[19] C. Guo, J. S. Frank, and K. Q. Weinberger, Low frequency adversarial perturbation, in *Proc. 35th Conf. Uncertainty in Artificial Intelligence*, Tel Aviv, Israel, 2019, pp.

1127−1137.
[20] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, Image quality assessment: From error visibility to structural similarity, *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.
[21] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 586−595.
[22] S. Sharma, J. J. Zou, G. Fang, P. Shukla, and W. Cai, A review of image watermarking for identity protection and verification, *Multimedia Tools Appl.*, vol. 83, no. 11, pp. 31829–31891, 2024.
[23] C. Qiu, G. Nan, R. Liang, W. Deng, Y. Zhang, Y. Gao, D. Wang, M. Qu, Z. Duan, Q. Sun, et al., Plugging and breathing on the air: A practical defense system for deep learning-based wireless semantic communications, *IEEE Trans. Mobile Comput.*, vol. 24, no. 9, pp. 8683–8699, 2025.
[24] Y. Rong, G. Nan, M. Zhang, S. Chen, S. Wang, X. Zhang, N. Ma, S. Gong, Z. Yang, Q. Cui, et al., Semantic entropy can simultaneously benefit transmission efficiency and channel security of wireless semantic communications, *IEEE Trans. Inf. Forensics Secur.*, vol. 20, pp. 2067–2082, 2025.
[25] Y. Zhang, D. Ye, C. Xie, L. Tang, X. Liao, Z. Liu, C. Chen, and J. Deng, Dual defense: Adversarial, traceable, and invisible robust watermarking against face swapping, *IEEE Trans. Inf. Forensics Secur.*, vol. 19, pp. 4628–4641, 2024.
[26] Y. Li, X. Liao, and X. Wu, Screen-shooting resistant watermarking with grayscale deviation simulation, *IEEE Trans. Multimedia*, vol. 26, pp. 10908–10923, 2024.
[27] L. Fu, X. Liao, J. Guo, L. Dong, and Z. Qin, WaveRecovery: Screen-shooting watermarking based on wavelet and recovery, *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 4, pp. 3603–3618, 2025.
[28] X. Liao, Y. Wang, T. Wang, J. Hu, and X. Wu, FAMM: Facial muscle motions for detecting compressed deepfake videos over social networks, *IEEE Trans. Circuits Syst. Video Technol*., vol. 33, no. 12, pp. 7236−7251, 2023.
[29] H. Jiang, J. Yang, G. Hua, L. Li, Y. Wang, S. Tu, and S. Xia, FAWA: Fast adversarial watermark attack, *IEEE Trans. Comput*., vol. 73, no. 2, pp. 301−313, 2024.
[30] X. Zuo, X. Wang, W. Zhang, and Y. Wang, MISPSO-Attack: An efficient adversarial watermarking attack based on multiple initial solution particle swarm optimization, *Appl. Soft Comput*., vol. 147, p. 110777, 2023.
[31] P. Zhu, T. Takahashi, and H. Kataoka, Watermark-embedded adversarial examples for copyright protection against diffusion models, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2024, pp. 24420−24430.
[32] J. Wang, H. Wang, J. Zhang, H. Wu, X. Luo, and B. Ma, Invisible adversarial watermarking: A novel security mechanism for enhancing copyright protection, *ACM Trans. Multimedia Comput., Commun. Appl.*, vol. 21, no. 2, p. 43, 2024.
[33] A. A. Mohammed, D. A. Salih, A. M. Saeed, and M. Q. Kheder, An imperceptible semi-blind image watermarking scheme in DWT-SVD domain using a zigzag embedding technique, *Multimedia Tools Appl.*, vol. 79, no. 43, pp. 32095–32118, 2020.
[34] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., ImageNet large scale visual recognition challenge, *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.
[35] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 770−778.
[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ImageNet classification with deep convolutional neural networks, in *Proc. 26th Int. Conf. Neural Information Processing Systems*, Lake Tahoe, NV, USA, 2012, pp. 1097–1105.
[37] K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv: 1409.1556, 2014.
[38] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size, arXiv preprint arXiv: 1602.07360, 2016.
[39] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, MobileNetV2: Inverted residuals and linear bottlenecks, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 4510−4520.
[40] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, Rethinking the Inception architecture for computer vision, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 2818−2826.
[41] A. Krizhevsky, *Learning Multiple Layers of Features from Tiny Images*. Toronto: University of Toronto, 2009.
[42] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in *Proc. IEEE Int. Conf. Computer Vision*, Venice, Italy, 2017, pp. 618−626.

Yuming Liu is currently an undergraduate student at School of Computer Science and Cyber Engineering, Guangzhou University, China. Her main research interests include artificial intelligence security, multimedia security, information hiding, computer vision, and deep learning.

Shan Ai received the BEng and MEng degrees from Harbin Engineering University, China in 2010 and 2013, respectively. He is currently an associate professor at School of Artificial Intelligence, Guangzhou University, China. His main research interest is artificial intelligence security.

Zhili Zhou received the MEng and PhD degrees in computer application from Hunan University, China in 2010 and 2014, respectively. He is currently a professor at School of Artificial Intelligence, Guangzhou University, China. Also, he was a postdoctoral fellow at Department of Electrical and Computer Engineering, University of Windsor, Canada. 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*, *Journal of Real-Time Image Processing*, *International Journal on Semantic Web and Information Systems*, and *CMC-Computers Materials & Continua*. He has been selected as "World's Top 2% Scientists" from 2020 to 2023 by Stanford University and Elsevier. He received ACM SIGWEB Rising Star Award and got Guangdong Natural Science Funds for Distinguished Young Scholar. His main research interests include multimedia security, artificial intelligence security, and information hiding.



Changyu Dong received the PhD degree from Imperial College London, UK in 2009. He is currently a professor at School of Artificial Intelligence, Guangzhou University, China. He has authored over 70 publications in international journals and conferences. His recent work focuses mostly on designing practical secure computation protocols. The application domains include trustworthy machine learning, secure cloud computing, and privacy-preserving data mining. His research interests include applied cryptography, AI security, data privacy, and machine learning.

Wei Pang received the PhD degree in computer science from University of Aberdeen, UK in 2009. He is currently a professor in computer science and bicentennial research at School of Mathematical and Computing Sciences, Heriot-Watt University, Edinburgh, UK. He has authored more than 160 papers, including more than 80 journal papers. His research interests include bioinspired computing, data mining, machine learning, and explainable and accountable AI.

Huilin Ge received the MEng degree in control theory and control engineering and the PhD degree in naval architecture and ocean engineering from Jiangsu University of Science and Technology, China in 2013 and 2023, respectively. He is a distinguished researcher at Jiangsu University of Science and Technology, China. With a solid academic foundation in control theory and control engineering, his career is marked by significant contributions to scientific research and project management, having successfully directed two national-level, two provincial-level, and three municipal-level research projects. In recognition of his academic excellence and leadership, he holds the esteemed title of "Vice President of Science and Technology". His research interests encompass deep learning, underwater information perception, and artificial intelligence, with a particular emphasis on the integration of low-level features to derive abstract high-level representations. His research interests include robotics, speech recognition, image recognition, and natural language processing.