Evaluation of Security of ML-based Watermarking: Copy and Removal Attacks

Vitaliy Kinakh
Department of Computer Science
University of Geneva
Geneva, Switzerland
vitaliy.kinakh@unige.ch

Brian Pulfer
Department of Computer Science
University of Geneva
Geneva, Switzerland
brian.pulfer@unige.ch

Yury Belousov
Department of Computer Science
University of Geneva
Geneva, Switzerland
yury.belousov@unige.ch

Pierre Fernandez
Meta, FAIR
University of Rennes, Inria, CNRS, IRISA
pierre.fernandez@inria.fr

Teddy Furon
University of Rennes, Inria, CNRS, IRISA
Rennes, France
teddy.furon@inria.fr

Slava Voloshynovskiy
Department of Computer Science
University of Geneva
Geneva, Switzerland
svolos@unige.ch

***Abstract*—The vast amounts of digital content captured from the real world or AI-generated media necessitate methods for copyright protection, traceability, or data provenance verification. Digital watermarking serves as a crucial approach to address these challenges. Its evolution spans three generations: handcrafted, autoencoder-based, and foundation model based methods. While the robustness of these systems is well-documented, the security against adversarial attacks remains underexplored. This paper evaluates the security of foundation models' latent space digital watermarking systems that utilize adversarial embedding techniques. A series of experiments investigate the security dimensions under copy and removal attacks, providing empirical insights into these systems' vulnerabilities. All experimental codes and results are available in the repository.**

***Index Terms*—digital watermarking, watermarking attack, self-supervised learning, latent space.**

I. Introduction

The emergence of a vast amount of content is reshaping our digital landscape. This content is either captured directly from the real world, i.e., physically produced, or created via digital algorithms, i.e., synthetically generated. This spans various media, including images, videos, audio, and text.

In this new landscape, verifying the integrity, authenticity, and provenance poses significant challenges to maintaining trust, preventing misinformation, preserving the integrity of legal evidence, and upholding ethical standards. Notably, the EU AI Act recognizes the risks linked with the recent machine learning (ML) models and the content they generate [1].

Digital watermarking is a crucial technical means in copyright protection and traceability. This technology aims to meet four primary requirements: imperceptibility, payload, robustness and security. While its robustness is well-documented, the security aspects, particularly of recent schemes based on ML, remain underexplored.

Foundation Models (FMs) and, notably, Vision Foundation Models (VFMs) are central to this evolving digital ecosystem [2], [3]. They represent a significant advancement in ML capabilities. These large pre-trained neural networks, refined on extensive and diverse datasets, are versatile tools. Many downstream applications use VFMs for analyzing content, like image classification, semantic segmentation, object detection, content retrieval, and tracking.

Based on this idea, a similar trend in watermarking [4], [5] aims to leverage the robustness and performance of these models. They usually utilize adversarial embedding techniques to hide information in VFMs' latent spaces. It makes the resulting watermarking robust and very versatile: able to operate on images with different resolutions, with a variable payload and a manually defined trade-off between robustness and quality. This paper evaluates and highlights the brittle security of these methods. Addressing this gap enhances the understanding and development of secure digital watermarking in our increasingly digital world.

The main contributions are as follows: a) We introduce two classes of attacks against latent space watermarking, specifically focusing on copy and removal attacks; b) We investigate the performance of these attacks on a state-of-the-art technique within this class of watermarking, evaluating both zero-bit and multi-bit watermarking schemes; c) We demonstrate the impact of target selection strategies in the effectiveness of removal attacks; d) We provide a comprehensive analysis of the vulnerability of DINOv1 [6], highlighting the necessity for future research on a broader range of foundation models.

II. State of the Art of Watermarking

Digital watermarking embeds information within digital media, balancing (1) *imperceptibility* - the distortion induced by the watermark is not perceptible for a human observer, (2) *payload* - the amount of data embedded in the content, (3) *robustness* - the ability to retrieve the hidden message under a given set of distortions and (4) *security* - the ability

This research was partially funded by the SNF Sinergia project (CRSII5-193716): Robust deep density models for high-energy particle physics and solar flare analysis (RODEM).

to withstand attacks exploiting the system's vulnerability. Techniques vary from *zero-bit watermarking*, where a mark is embedded into a content using a secret key and the detection assesses the presence of this mark within the content, to *multi-bit watermarking*, which encodes a message in content, and the decoder retrieves the embedded message bit by bit.

Digital watermarking has evolved across three generations differentiated by their embedding domains:

1) $\mathcal{DW}_1$: Techniques in this category embed watermarks in the *spatial* or *transform* domains, including DFT [7], [8], DCT [9], [10], Fourier-Melline [11], and DWT [12] domains, with both *zero-bit* [13] and *multi-bit* watermarking [14], [15]. These methods aim for invisibility and basic robustness, employing additive or quantization-based embedding techniques [16], [17].
2) $\mathcal{DW}_2$: This group jointly trains ML-based encoder and decoder for adaptive embedding [18]–[20], focusing on content-driven robustness enhancements. These methods involve training under differentiable distortions, including adversarial settings [21], [22], and require adaptation to new types of datasets and distortions.
3) $\mathcal{DW}_3$: The most recent advancement explores watermarking by using iterative adversarial-like embeddings in the latent spaces of pre-trained models, either trained on a supervised task [4] or with VFMs [5]. In this paper, we consider DINOv1 model [6]. DINOv1 is a self-supervised learning computer vision model, that uses student-teacher framework, the student predicts teacher's output for different image augmentations. DINOv1 captures semantic information and performs well on tasks like image classification and object detection.

Security of digital watermarking: Extensive robustness and security assessments have been conducted on the $\mathcal{DW}_1$ group. These studies pinpoint the difficulty to fight against the *copy attack* [23], the *remodulation attack* [24], and the *sensitivity attack* [25]–[28]. Conversely, the exploration of the security of $\mathcal{DW}_2$ and $\mathcal{DW}_3$ watermarking in the face of adversarial attacks is still in its infancy. This early inquiry phase highlights a significant gap in our understanding of their security, indicating a critical field for research endeavours.

Notations: We denote by $\mathcal{X} = \mathbb{R}^{H\times W\times C}$ the space of images of size $H \times W \times C$. A trained VFM is denoted as $f_\phi : \mathcal{X} \to \mathcal{Z}$ mapping the image space to the latent space $\mathcal{Z} = \mathbb{R}^d$. Notations $\mathbf{x}_0$, $\mathbf{x}_w$, and $\mathbf{x}_a$ stand for the original, watermarked and attacked images in $\mathcal{X}$, $\mathbf{z}_0$, $\mathbf{z}_w$ and $\mathbf{z}_a$ correspond to their latent space representations in $\mathcal{Z}$. We have $\mathbf{x}_w = w(\mathbf{x}_0, m, k)$ where m is the message to be hidden and k the secret key, and $\mathbf{x}_a = t(\mathbf{x}_w)$ where t is an image transformation pertaining to a set of attacks $\mathcal{T}$.

The distortion is measured by $\mathcal{L}_\mathcal{X} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$. In the case of mean square error (MSE), $\mathcal{L}_\mathcal{X}(\mathbf{x}_0, \mathbf{x}_w) = ||\mathbf{x}_0 - \mathbf{x}_w||_2^2/H/W/C \leq D_w$, where D_w defines the embedding distortion budget between the original and watermarked images. If the size and geometry of the image after the attack are preserved, one can also define the attack distortion $\mathcal{L}_\mathcal{X}(\mathbf{x}_w, \mathbf{x}_a)$. The MSE is usually given in log scale by the peak signal-to-noise ratio $\text{PSNR}_w = 10\log_{10}\left(255^2/\mathcal{L}_\mathcal{X}(\mathbf{x}_0, \mathbf{x}_w)\right)$ for measuring quality of watermarked imaged and $\text{PSNR}_a = 10\log_{10}\left(255^2/\mathcal{L}_\mathcal{X}(\mathbf{x}_w, \mathbf{x}_a)\right)$ for attacked images.

III. VFM-based Adversarial Embedding Watermarking

This section summarizes the watermarking method [5] by first accounting for the detection/decoding stage.

A. Detection and Decoding

We consider two scenarios: zero-bit (detection only) and multi-bit watermarking (decoding the hidden message).

Zero-Bit. Given a secret carrier $\mathbf{w} \in \mathcal{Z}$ s.t. $\|\mathbf{w}\| = 1$, generated from the secret key k, that represents a 0-bit watermarking, the detection region is the dual hypercone:

$$\mathcal{D}_k := \{\mathbf{z} \in \mathbb{R}^d : |\mathbf{z}^T\mathbf{w}| > \|\mathbf{z}\|\cos(\gamma)\}. \tag{1}$$

The angle γ is defined by the targeted false acceptance rate P^t_{fa}, that is theoretically given for a non-watermarked $\mathbf{x}$ as:

$$P^t_{\text{fa}} := \mathbb{P}\left[f_\phi(\mathbf{x}) \in \mathcal{D}_K | K \sim \mathcal{U})\right] = 1 - I_{\cos^2(\gamma)}\left(\frac{1}{2}, \frac{d-1}{2}\right), \tag{2}$$

where $I_\tau(\alpha, \beta)$ is the regularized Beta incomplete function. The following function gauges how $\mathbf{z}$ is close to $\mathcal{D}_k$:

$$\mathcal{L}^I_\mathcal{Z}(\mathbf{z}, \mathbf{w}) = \|\mathbf{z}\|^2\cos^2(\theta) - (\mathbf{z}^T\mathbf{w})^2. \tag{3}$$

Its sign indicates whether $\mathbf{z}$ lies inside $\mathcal{D}_k$, its amplitude indicates how far $\mathbf{z}$ is from $\mathcal{D}_k$ or deep inside $\mathcal{D}_k$.

Multi-Bit. The hidden message is $m = (m_1, \ldots, m_\ell) \in \{-1, 1\}^\ell$. The random generator seeded with the secret key k produces an orthogonal family of carriers $\{\mathbf{w}_1, \ldots, \mathbf{w}_\ell\} \subset \mathcal{Z}$. The decoder retrieves $\hat{m}$ as the sign of the projections:

$$\hat{m} = \left(\text{sign}\left(f_\phi(\mathbf{x})^\top\mathbf{w}_1\right), \ldots, \text{sign}\left(f_\phi(\mathbf{x})^\top\mathbf{w}_\ell\right)\right).$$

The following function gauges how $\mathbf{z}$ lies deep inside the decoding region within a margin $\mu \geq 0$ on the projections.

$$\mathcal{L}^{II}_\mathcal{Z}(\mathbf{z}, m) = \frac{1}{\ell}\sum_{i=1}^{\ell}\max\left(0, \mu - \left(\mathbf{z}^\top\mathbf{w}_i\right)\cdot m_i\right). \tag{4}$$

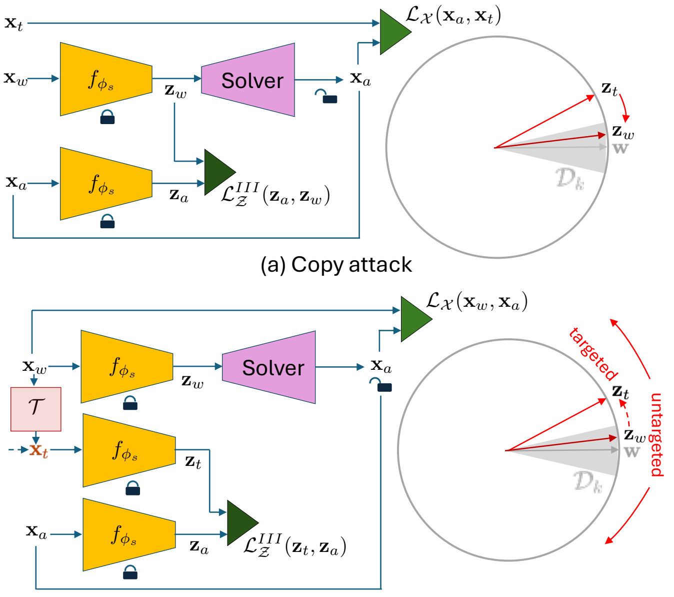
B. Watermark embedding

The embedding takes an original image $\mathbf{x}_0 \in \mathcal{X}$ and outputs a visually similar image $\mathbf{x}_w \in \mathcal{X}$. The previous section defines a loss function $\mathcal{L}_\mathcal{Z}$ in the latent space, be it (3) or (4). The embedding aims at minimizing this loss under the constraint of distortion defined in the image domain. Augmentations are introduced to make the watermark signal more robust. These are image modifications belonging to a set $\mathcal{T}$ of typical attacks with a range of parameters, such as rotation, crops and blur. The application of attack $t \in \mathcal{T}$ to image $\mathbf{x}$ writes as $t(\mathbf{x}) \in \mathcal{X}$.

The losses $\mathcal{L}_\mathcal{Z}$ and $\mathcal{L}_\mathcal{X}$ are combined as follows:

$$\mathcal{L}_\mathcal{W}(\mathbf{x}, \mathbf{x}_0, t) := \lambda\mathcal{L}_\mathcal{Z}(f_\phi(t(\mathbf{x}))) + \mathcal{L}_\mathcal{X}(\mathbf{x}, \mathbf{x}_0), \tag{5}$$

where λ controls the trade-off between two terms: $\mathcal{L}_\mathcal{Z}$ aims to push the feature of any transformation of $\mathbf{x}_w$ deep inside the detection/decoding region, while $\mathcal{L}_\mathcal{X}$ favors low distortion.


$\mathbf{x}_t$
$\mathbf{x}_w$
f_{ϕ_s}
Solver
$\mathbf{z}_w$
$\mathbf{x}_a$
$\mathcal{L}_{\mathcal{X}}(\mathbf{x}_a, \mathbf{x}_t)$
$\mathbf{z}_t$
$\mathbf{z}_w$
$\mathbf{w}$
$\mathcal{D}_k$
$\mathbf{z}_a$
$\mathcal{L}_{\mathcal{Z}}^{III}(\mathbf{z}_a, \mathbf{z}_w)$
$\mathcal{L}_{\mathcal{X}}(\mathbf{x}_w, \mathbf{x}_a)$
targeted
untargeted
$\mathcal{T}$
$\mathcal{L}_{\mathcal{Z}}^{III}(\mathbf{z}_t, \mathbf{z}_a)$

(a) Copy attack

(b) Removal attack

Fig. 1. Generalized diagram explaining the proposed (a) copy and (b) untargeted and targeted removal attacks (on the example of zero-bit watermarking in the latent space). The secret carrier $\mathbf{w}$ and the decision region $\mathcal{D}_k$ (show in gray) are unknown for the attacker.

The embedding is typical from the adversarial ML literature minimizing an Expectation over Transformation (EoT) [29]:

$$\mathbf{x}_w := \arg\min_{\mathbf{x} \in C(\mathbf{x}_0)} \mathbb{E}_{T \sim \mathcal{U}(\mathcal{T})}[\mathcal{L}_{\mathcal{W}}(\mathbf{x}, \mathbf{x}_0, T)], \tag{6}$$

where $C(\mathbf{x}_0) \subset \mathcal{X}$ is the set of admissible images w.r.t. the original one. It is defined by two steps of normalization applied to the pixel-wise difference $\boldsymbol{\delta}_0 = \mathbf{x} - \mathbf{x}_0$: (1) we apply a SSIM [30] heatmap attenuation, which scales $\boldsymbol{\delta}_0$ pixel-wise to hide the information in perceptually less visible areas of the image; (2) we set a target PSNR and rescale $\boldsymbol{\delta}_0$ accordingly.

IV. Attacks against ML-based Digital Watermarking

This paper assumes the attacker knows neither the secret key k nor the message m. However, the main brick of the system is the foundation model f_ϕ which is open-sourced and therefore a white-box for the attacker.

A. Watermark Copy Attack

The objective of a *copy attack* is to maximize the probability of falsely accepting a non-watermarked image as a watermarked one. Given a watermarked image $\mathbf{x}_w$ and a target image $\mathbf{x}_t$, the attack seeks to transfer the watermark from $\mathbf{x}_w$ to $\mathbf{x}_t$ without knowledge of the message m or the key k.

In contrast to the traditional copy attack [23], Fig. 1a proposes a generalization across various embedding domains that does not necessitate the additivity of the embedding.

Given the watermarked image $\mathbf{x}_w$ and the target image $\mathbf{x}_t$, our copy attack generates an attacked image $\mathbf{x}_a$ that is perceptually close to $\mathbf{x}_t$ according to the loss function $\mathcal{L}_{\mathcal{X}}(\mathbf{x}_t, \mathbf{x}_a)$. Concurrently, the latent representation $\mathbf{z}_a$ of the attacked image is driven towards the latent representation $\mathbf{z}_w$ of the watermarked image as per a loss function $\mathcal{L}_{\mathcal{Z}}^{III}(\mathbf{z}_a, \mathbf{z}_w)$. The total loss for the generalized copy attack is formulated as:

$$\mathcal{L}_{\mathcal{A}}^{\mathrm{C}}(\mathbf{x}_a, \mathbf{x}_w, \mathbf{x}_t) = \mathcal{L}_{\mathcal{X}}(\mathbf{x}_a, \mathbf{x}_t) + \lambda \mathcal{L}_{\mathcal{Z}}^{III}(\mathbf{z}_a, \mathbf{z}_w), \tag{7}$$

where λ is a weighting factor that balances the contributions of the perceptual and latent similarity terms. The latent space loss is defined as $\mathcal{L}_{\mathcal{Z}}^{III}(\mathbf{z}_a, \mathbf{z}_w) = -\frac{\mathbf{z}_a^T \mathbf{z}_w}{\sqrt{\|\mathbf{z}_a\|_2^2 \|\mathbf{z}_w\|_2^2}}$ for both zero-bit and multi-bit watermarking. Minimization is achieved via gradient descent over N iterations. Similar to the watermark embedding (6), the attack also involves two normalization steps applied to the difference $\boldsymbol{\delta}_{at} = \mathbf{x}_a - \mathbf{x}_t$, i.e. the SSIM masking and the rescaling to impose a certain PSNR$_a$. The final image is rounded to quantized pixels. The algorithm of the proposed copy attack is presented below.

Algorithm 1 Copy Attack

```
Input: $\mathbf{x}_w$: watermarked image, $\mathbf{x}_t$: target image; $f_\phi$: feature extractor (FM)
$\mathbf{z}_w \leftarrow f_\phi(\mathbf{x}_w)$, $\mathbf{x}_a \leftarrow \mathbf{x}_t$                // initialize
for $t = 0, \ldots, N-1$ do
    $\mathbf{x}_a \xleftarrow{\text{constraints}} \mathbf{x}_a$               // impose constraints via $\boldsymbol{\delta}_{at}$
    $\mathbf{z}_a \leftarrow f_\phi(\mathbf{x}_a)$               // compute latent representation
    $\mathbf{x}_a \leftarrow \mathbf{x}_a + \eta \times \mathrm{Adam}\left(\mathcal{L}_{\mathcal{A}}^{\mathrm{C}}(\mathbf{x}_a, \mathbf{x}_w, \mathbf{x}_t)\right)$
                                                   // update the image
end for
$\mathbf{x}_a \xleftarrow{\text{constraints}} \mathbf{x}_a$       // impose constraints via $\boldsymbol{\delta}_{at}$, rounding
Return: Attacked image $\mathbf{x}_a$
```

Extension to multiple watermarked images. When multiple images $\{\mathbf{x}_{wn}\}_{n=1}^{L}$ watermarked with the same key and the same message (in the case of multi-bit watermarking) are available to the attacker, one can compensate the lack of knowledge of the acceptance region $\mathcal{D}$ by solving the following optimization problem: for $\mathbf{z}_{wn} = f_\phi(\mathbf{x}_{wn})$, $\forall n \in [L]$,

$$\mathcal{L}_{\mathcal{A}}^{\mathrm{C}}(\mathbf{x}_a, \mathbf{x}_w, \mathbf{x}_t) = \mathcal{L}_{\mathcal{X}}(\mathbf{x}_a, \mathbf{x}_t) + \frac{\lambda}{L} \sum_{n=1}^{L} \mathcal{L}_{\mathcal{Z}}^{III}(\mathbf{z}_a, \mathbf{z}_{wn}). \tag{8}$$

In our experiments, we observe the very high success rates of the targeted attacks in the setup where $L = 1$. Thus, we do not experiment with these attacks in Sec. V.

B. Watermark Removal Attack

The watermark removal damages the watermarked image to maximize the probability of miss detection (zero-bit watermarking), or the bit error rate (BER) (multi-bit watermarking).

Our proposal is to jeopardize the latent space representation with the hope of diminishing the presence of the watermark. Specifically, given a watermarked image $\mathbf{x}_w$, the attack generates an attacked image $\mathbf{x}_a$ perceptually similar to $\mathbf{x}_w$ while ensuring that its latent representation $\mathbf{z}_a$ is far from $\mathbf{z}_w$. This strategy does not require an additive approximation of the embedding. Neither the watermark detector/decoder output nor the secret key k is required.

Technically, the watermark removal can be achieved by a) *untargeted attack* (removal-untargeted, R-U) or b) *targeted attack* (R-T). In the untargeted case, the loss function is defined $\mathcal{L}_{\mathcal{Z}}^{IV}(\mathbf{z}_a, \mathbf{z}_w) = \frac{(\mathbf{z}_a^T \mathbf{z}_w)^2}{\sqrt{\|\mathbf{z}_a\|_2^2 \|\mathbf{z}_w\|_2^2}}$ for both zero-bit and multi-bit watermarking.

$$\mathcal{L}_{\mathcal{A}}^{\mathrm{R-U}}(\mathbf{x}_w, \mathbf{x}_a) = \mathcal{L}_{\mathcal{X}}(\mathbf{x}_w, \mathbf{x}_a) - \lambda \mathcal{L}_{\mathcal{Z}}^{IV}(\mathbf{z}_w, \mathbf{z}_a). \quad (9)$$

The targeted removal attack generates an attacked image $\mathbf{x}_a$ that is perceptually close to the watermarked image $\mathbf{x}_w$ while its latent representation $\mathbf{z}_a$ gets away from $\mathbf{w}$ and instead aligns with the latent representation of a target image $\mathbf{z}_t$:

$$\mathcal{L}_{\mathcal{A}}^{\mathrm{R-T}}(\mathbf{x}_w, \mathbf{x}_t, \mathbf{x}_a) = \mathcal{L}_{\mathcal{X}}(\mathbf{x}_w, \mathbf{x}_a) + \lambda \mathcal{L}_{\mathcal{Z}}^{III}(\mathbf{z}_t, \mathbf{z}_a), \quad (10)$$

Minimization of the total loss is achieved via stochastic gradient descent over N iterations. The final image is obtained with the SSIM masking and scaling of the perturbation $\boldsymbol{\delta}_{aw} = \mathbf{x}_a - \mathbf{x}_w$ to achieve a given PSNR$_a$, and rounding.

Algorithm 2 Watermark Removal Attack

```
Input: $\mathbf{x}_w$: watermarked image, $\mathbf{x}_t$: target image; $f_\phi$: feature extractor (FM), $attack\_type$: type of attack (targeted or untargeted)
Compute: $\mathbf{z}_t = f_\phi(\mathbf{x}_t)$
Initialize: $\mathbf{x}_a \leftarrow \mathbf{x}_w$
for $t = 0, \dots, N-1$ do
    $\mathbf{x}_a \xleftarrow{\text{constraints}} \mathbf{x}_a$    // impose constraints via $\boldsymbol{\delta}_{aw}$
    $\mathbf{z}_a \leftarrow f_\phi(\mathbf{x}_a)$    // compute latent representation
    if $attack\_type$ == “untargeted” then
        $\mathbf{x}_a \leftarrow \mathbf{x}_a + \eta \times \mathrm{Adam}(\mathcal{L}_{\mathcal{A}}^{\mathrm{R-U}}(\mathbf{x}_w, \mathbf{x}_a))$
            // update the image according to untargeted attack
    else if $attack\_type$ == “targeted” then
        $\mathbf{x}_a \leftarrow \mathbf{x}_a + \eta \times \mathrm{Adam}(\mathcal{L}_{\mathcal{A}}^{\mathrm{R-T}}(\mathbf{x}_w, \mathbf{x}_t, \mathbf{x}_a))$
            // update the image according to targeted attack
    end if
end for
$\mathbf{x}_a \xleftarrow{\text{constraints}} \mathbf{x}_a$    // impose constraints via $\boldsymbol{\delta}_{aw}$, rounding
Return: Attacked image $\mathbf{x}_a$
```

The target selection during the removal attack plays an important role for the success of the attack. Three strategies are being considered. 1) Choosing any random non-watermarked image $\mathbf{x}_t$. 2) Setting target to be a heavily degraded version of $\mathbf{x}_w$ for which the watermark is no longer detected. Then, the optimization (10) restores a better image quality. 3) Selecting random watermarking carrier as the new target.

V. Experimental Results

The implementation of the studied zero-bit and multi-bit watermarking is based on the paper [5]. The ResNet-50 trained with DINOv1 [6] is used as the vision backbone. All experiments are performed on the DIV2K dataset [31] with typical image size 2000 × 1500. Unless specified otherwise, the experiments are repeated using 10 different keys for watermark embedding and detection on a subset of 800 images from DIV2K. In all experiments, the PSNR$_w$ of the original watermarked image is fixed at 42 dB, and the target PSNR$_a$ varies from 30 to 45 dB. For most of the attacks, the actually achieved PSNR$_a$ is higher than the above target value.

A. Investigation on the Copy Attack

The first experiment investigates the robustness against the copy attack. The goal is to copy the watermark on unwatermarked images from a single watermarked image. The PSNR$_w$ of the original watermarked image is fixed at 42 dB.

For zero-bit watermarking, the attack success rate measures the proportion of crafted images that are wrongly flagged by the watermark detection (1), for different targeted probabilities of false acceptance $P_{\text{fa}}^t \in \{10^{-5}, 10^{-6}, 10^{-7}\}$. The optimization of Alg. 1 achieves the attack success rate equals one for the entire range of studied PSNR$_a$ and targeted false acceptance P_{fa}^t. This confirms the strength of the copy attack.

The second experiment involves multi-bit watermarking. The watermark payload varies $\ell \in \{10, 30, 50, 100\}$ bits. Fig. 2 shows that, at low values of PSNR$_a$ (strong attack distortions), the multi-bit watermarks are perfectly copied. At higher values of PSNR$_a$ (weak attack), the BER naturally increases but not significantly. The increase of message length causes higher value of BER obtained at high PSNR$_a$ = 47.5 dB, but for lower PSNR$_a$ the impact of watermark payload length is insignificant. This demonstrates strong clonability.

B. Investigation on the Removal Attack

This section studies both untargeted and targeted removal attacks against zero-bit and multi-bit watermarking. In contrast to the copy attack, the attack success rate now measures the probability of miss P_{m} for zero-bit watermarking, *i.e.*, the proportion of watermarked images that are no longer detected after the attack, and the BER for multi-bit watermarking.

The untargeted removal attack (9) does not require any target. Fig. 3 reports the observed P_{m} for the zero-bit watermarking detection at different targeted probabilities of false acceptance. On the other hand, Fig. 4 shows the influence on BER for multi-bit watermarking. The untargeted removal

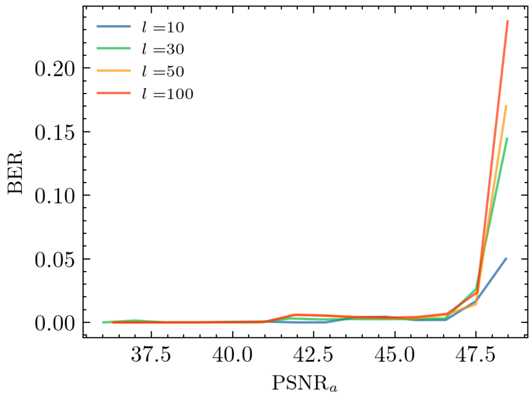


Fig. 2. Bit Error Rate (BER) for multi-bit watermarking under the copy attack with varying PSNR$_a$ and watermark payloads ℓ. The attack can successfully copy the binary message (BER $< 1\%$) of the watermarked image into any non-watermarked image, even at very low distortion budgets (PSNR$_a$ = 45 dB).

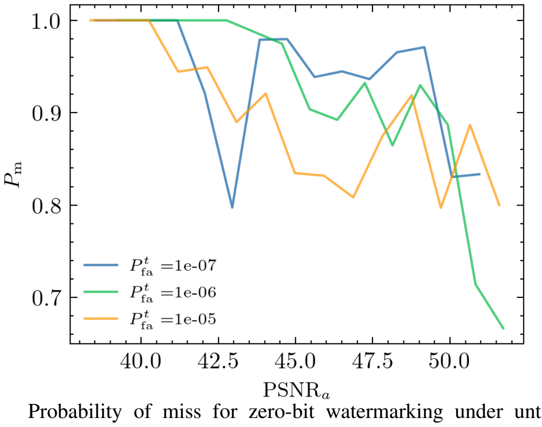


Fig. 3. Probability of miss for zero-bit watermarking under untargeted removal attack against PSNR$_a$ of the attacked image, for varying probability of false acceptance. The untargeted attack achieves P_m close to 1 at lower values of PSNR$_a$ around 40 dB, while P_m decreases with the increase of PSNR$_a$ towards 50 dB.

attack significantly impacts the performance of both watermarking schemes.

In contrast to the untargeted removal attack, the targeted removal attack needs to select the target $\mathbf{x}_t$ and accordingly $\mathbf{z}_t = f_\phi(\mathbf{x}_t)$. The target image selection strategies include random selection of $\mathbf{x}_t$ denoted as "other image", selecting the denoised watermark image as $\mathbf{x}_t = d_\text{Wiener}(\mathbf{x}_w)$, and selecting directly $\mathbf{z}_t$ randomly in the latent space.

Fig. 5 shows the P_m under targeted removal attack for zero-bit watermarking with the required target probability of false acceptance: 10^{-5}, 10^{-6} and 10^{-7}. The selection of the denoised image based on Wiener filter with size 25×25 as a target image provides the best results in maximization of probability of miss for all values of probability of false acceptance. Comparing the results from Fig. 5 and Fig. 3, one can conclude that both untargeted and targeted removal attacks achieve P_m close to 1, for PSNR$_a \leq 41$ dB, that demonstrates high efficiency of both strategies.

As for multi-bit watermarking, the BER evaluates the success of the attack. The watermark payload is fixed at $\ell \in \{10, 30, 50, 100\}$ bits. The results in Fig. 6 demonstrate how the BER depends on the PSNR$_a$ of the attacked image. The removal efficiency decreases with the increase of PSNR$_a$.

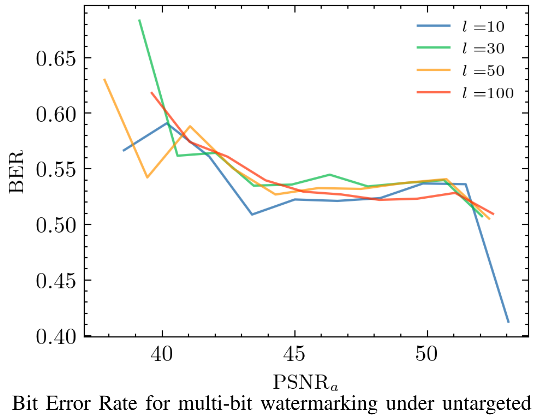


Fig. 4. Bit Error Rate for multi-bit watermarking under untargeted removal attack against PSNR$_a$ at varying payload of ℓ bits. The attack increases the BER significantly, inverting the majority of the hidden bits.

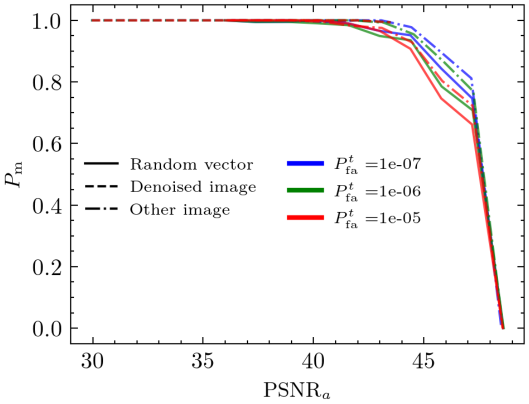


Fig. 5. Probability of miss for zero-bit watermarking under targeted removal attack with different target image selection strategies. All kinds of targeted attacks achieve better success rates than the untargeted ones.

The choice of target in the targeted removal attack dictates the different attack efficiency in terms of effective PSNR$_a$ and achievable BER for different watermark message lengths. The "other image" target selection requires largest PSNR$_a$, i.e., highest possible distortions, to maximally damage the watermarked message for the range of 30-37 dB. The random vector subset space target allows achieves similar values of BER starting at 37 dB but with considerably higher variability of BER values for different message lengths. Finally, the "denoised image" selection as a target for the considered removal attack achieves similar results starting from 39 dB under the same impact of message length on BER variability. The overall increase of PSNR$_a$ leads to the decrease of BER due to the reduction of allowable distortion budget.

One can observe that under the untargeted attacks, the results are somewhat unstable under different PSNR$_a$. We argue that this is due to the nature of untargeted attacks. Unlike targeted attacks, which push the image latent representation

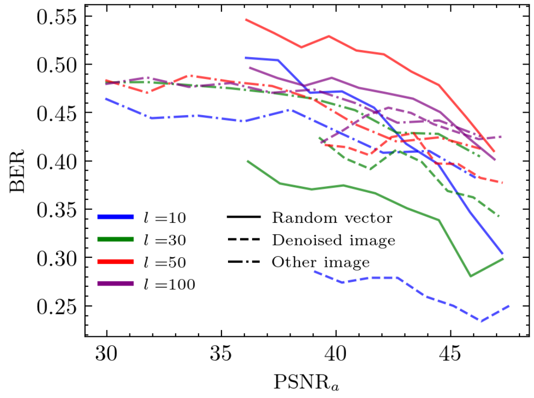


Fig. 6. Bit Error Rate for multi-bit watermarking under targeted removal attack with different target image selection strategies. The best results correspond to BER=0.5 (random chance).

to be as close as possible to the selected target latent representation, the untargeted attacks push the attacked image latent representation far from the watermarked image (cosine similarity between representations is 0). Thus, it can result in an infinite number of optimal solutions.

VI. Conclusion

This paper investigates the efficacy of copy and removal attacks against a watermarking technique based on the foundation model's latent space. The results demonstrate that the effectiveness of these attacks increases with the level of adversarial distortions applied. Among the two types of attacks, removal attacks have proven to be more efficient against both watermarking schemes. Copy attacks are relatively easier to perform on zero-bit watermarking. This is attributed to the more complex nature of multi-bit watermarking latent space spanning.

It is important to note that all experimental results were obtained using the DINOv1 model. This demonstrates its high vulnerability attacks, and its use for watermarking is not recommended. Consequently, a future research direction involves investigating a broader class of foundation and autoencoder models in the context of digital watermarking, as well as comparison with classical schemes like Broken Arrows [32]. This would help determine whether such vulnerabilities are specific to certain types or consistent across different models. The latter case implies that watermarking is a specific downstream task that cannot be solved with a public foundation model.

References

[1] European Commission, "EU AI act," https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai, 2024, accessed: 2024-03-14.

[2] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "Dinov2: Learning robust visual features without supervision," 2023.

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[4] V. Vukotić, V. Chappelier, and T. Furon, "Are classification deep neural networks good for blind image watermarking?" *Entropy*, vol. 22, no. 2, p. 198, 2020.

[5] P. Fernandez, A. Sablayrolles, T. Furon, H. Jégou, and M. Douze, "Watermarking images in self-supervised latent spaces," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 3054–3058.

[6] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9650–9660.

[7] M. Urvoy, D. Goudia, and F. Autrusseau, "Perceptual dft watermarking with improved detection and robustness to geometrical distortions," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 7, pp. 1108–1119, 2014.

[8] S. Voloshynovskiy, Z. Grytskiv, Y. Rytsar, M. Shovgenuk, and M. Kozlovskiy, "The means of visual data encryption," Patent, 1997.

[9] A. G. Bors and I. Pitas, "Image watermarking using dct domain constraints," in *Proceedings of 3rd IEEE International Conference on Image Processing*, vol. 3. IEEE, 1996, pp. 231–234.

[10] S. Pereira, S. Voloshynovskiy, and T. Pun, "Effective channel coding for dct watermarks," in *Proceedings 2000 International Conference on Image Processing (Cat. No. 00CH37101)*, vol. 3. IEEE, 2000, pp. 671–673.

[11] S. Pereira, J. J. O. Ruanaidh, F. Deguillaume, G. Csurka, and T. Pun, "Template based recovery of fourier-based watermarks using log-polar and log-log maps," in *Proceedings IEEE international conference on multimedia computing and systems*, vol. 1. IEEE, 1999, pp. 870–874.

[12] X.-G. Xia, C. G. Boncelet, and G. R. Arce, "Wavelet transform based watermark for digital images," *Optics Express*, vol. 3, no. 12, pp. 497–511, 1998.

[13] T. Furon, "A constructive and unifying framework for zero-bit watermarking," *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 2, pp. 149–163, 2007.

[14] J. R. Hernández, F. Pérez-González, J. M. Rodriguez, and G. Nieto, "Performance analysis of a 2-d-multipulse amplitude modulation scheme for data hiding and watermarking of still images," *IEEE Journal on Selected areas in Communications*, vol. 16, no. 4, pp. 510–524, 1998.

[15] S. Voloshynovskiy, F. Deguillaume, and T. Pun, "Multibit digital watermarking robust against local nonlinear geometrical distortions," in *IEEE Int. Conf. On Image Processing ICIP2001*, Thessaloniki, Greece, October 2001, pp. 999–1002.

[16] B. Chen and G. W. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," *IEEE Transactions on Information theory*, vol. 47, no. 4, pp. 1423–1443, 2001.

[17] J. J. Eggers and B. Girod, "Quantization effects on digital watermarks," *Signal Processing*, vol. 81, no. 2, pp. 239–263, 2001.

[18] H. Kandi, D. Mishra, and S. R. S. Gorthi, "Exploring the learning capabilities of convolutional neural networks for robust image watermarking," *Computers & Security*, vol. 65, pp. 247–268, 2017.

[19] J.-E. Lee, Y.-H. Seo, and D.-W. Kim, "Convolutional neural network-based digital image watermarking adaptive to the resolution of image and watermark," *Applied Sciences*, vol. 10, no. 19, p. 6854, 2020.

[20] J. Zhu, R. Kaplan, J. Johnson, and L. Fei-Fei, "Hidden: Hiding data with deep networks," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 657–672.

[21] X. Luo, R. Zhan, H. Chang, F. Yang, and P. Milanfar, "Distortion agnostic deep watermarking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 548–13 557.

[22] B. Wen and S. Aydore, "Romark: A robust watermarking system using adversarial training," *arXiv preprint arXiv:1910.01221*, 2019.

[23] M. Kutter, S. Voloshynovskiy, and A. Herrigel, "Watermark copy attack," in *IS&T/SPIE's 12th Annual Symposium, Electronic Imaging 2000: Security and Watermarking of Multimedia Content II*, vol. 3971, San Jose, California USA, 23–28 jan 2000.

[24] S. Voloshynovskiy, S. Pereira, T. Pun, J. J. Eggers, and J. K. Su, "Attacks on digital watermarks: classification, estimation based attacks, and benchmarks," *IEEE communications Magazine*, vol. 39, no. 8, pp. 118–126, 2001.

[25] J.-P. M. Linnartz and M. v. Dijk, "Analysis of the sensitivity attack against electronic watermarks in images," in *International Workshop on Information Hiding*. Springer, 1998, pp. 258–272.

[26] J. W. Earl, "Tangential sensitivity analysis of watermarks using prior information," in *Security, Steganography, and Watermarking of Multimedia Contents IX*, vol. 6505. SPIE, 2007, pp. 449–460.

[27] P. Comesana, L. Pérez-Freire, and F. Pérez-González, "Blind newton sensitivity attack," *IEE Proceedings-Information Security*, vol. 153, no. 3, pp. 115–125, 2006.

[28] M. El Choubassi and P. Moulin, "Sensitivity analysis attacks against randomized detectors," in *2007 IEEE International Conference on Image Processing*, vol. 2. IEEE, 2007, pp. II–129.

[29] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *International conference on machine learning*. PMLR, 2018, pp. 284–293.

[30] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[31] E. Agustsson and R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 126–135.

[32] T. Furon and P. Bas, "Broken arrows," *EURASIP Journal on Information Security*, vol. 2008, pp. 1–13, 2008.