# Rethinking the Vulnerability of DNN Watermarking: Are Watermarks Robust against Naturalness-aware Perturbations?

Run Wang[*,†]
Wuhan University
Wuhan, Hubei, China
wangrun@whu.edu.cn

Haoxuan Li[†]
Wuhan University
Wuhan, Hubei, China
haoxuanli@whu.edu.cn

Lingzhou Mu[†]
Wuhan University
Wuhan, Hubei, China
mlzmlz@whu.edu.cn

Jixing Ren[†]
Wuhan University
Wuhan, Hubei, China
2017301500283@whu.edu.cn

Shangwei Guo
Chongqing University
Chongqing, Chongqing, China
swguo@cqu.edu.cn

Li Liu[‡]
Fudan University
Shanghai, Shanghai, China
liuli@fudan.edu.cn

Liming Fang[§]
Nanjing University of Aeronautics and Astronautics
Nanjing, Jiangsu, China
fangliming@nuaa.edu.cn

Jing Chen[†]
Wuhan University
Wuhan, Hubei, China
chenjing@whu.edu.cn

Lina Wang[†,¶]
Wuhan University
Wuhan, Hubei, China
lnwang@whu.edu.cn

## ABSTRACT

Training Deep Neural Networks (DNN) is a time-consuming process and requires a large amount of training data, which motivates studies working on protecting the intellectual property (IP) of DNN models by employing various watermarking techniques. Unfortunately, in recent years, adversaries have been exploiting the vulnerabilities of the employed watermarking techniques to remove the embedded watermarks. In this paper, we investigate and introduce a novel watermark removal attack, called *AdvNP*, against all the existing **four** different types of DNN watermarking schemes via input preprocessing by injecting Adversarial Naturalness-aware Perturbations. In contrast to the prior studies, our proposed method is the first work that generalizes all the existing four watermarking schemes well without involving any model modification, which preserves the fidelity of the target model. We conduct the experiments against **four** state-of-the-art (SOTA) watermarking schemes on **two** real tasks (*e.g.*, image classification on ImageNet, face recognition on CelebA) across multiple DNN models. Overall, our proposed AdvNP significantly invalidates the watermarks against the four watermarking schemes on two real-world datasets, *i.e.*, **60.9%** on the average attack success rate and up to **97%** in the worse case. Moreover, our AdvNP could well survive the image denoising techniques and outperforms the baseline in both the fidelity preserving and watermark removal. Furthermore, we introduce two defense methods to enhance the robustness of DNN watermarking against our AdvNP. Our experimental results pose real threats to the existing watermarking schemes and call for more practical and robust watermarking techniques to protect the copyright of pretrained DNN models. The source code and models are available at https://github.com/GitKJ123/AdvNP.

## CCS CONCEPTS

• **Security and privacy** → **Human and societal aspects of security and privacy**; • **Information systems** → **Multimedia information systems**; • **Computing methodologies** → Artificial intelligence.

## KEYWORDS

DNN watermarking, naturalness-aware perturbations, relighting

**ACM Reference Format:**
Run Wang, Haoxuan Li, Lingzhou Mu, Jixing Ren, Shangwei Guo, Li Liu, Liming Fang, Jing Chen, and Lina Wang. 2022. Rethinking the Vulnerability of DNN Watermarking: Are Watermarks Robust against Naturalness-aware Perturbations?. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3503161.3548390

## 1 INTRODUCTION

In recent years, DNNs have achieved tremendous success in many cutting-edge fields [33], such as image classification [22], speech recognition [37], and natural language processing [15], *etc.*. However, training DNN models is time-consuming and computationally

[*] Run Wang is the corresponding author (wangrun@whu.edu.cn)
[†] The Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University
[‡] Fudan Development Institute
[§] Shenzhen Research Institute
[¶] Zhengzhou Xinda Institute of Advanced Technology

*MM '22, October 10–14, 2022, Lisboa, Portugal.*

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548390

**Table 1: Comparison with the existing three popular DNN watermarking attacks in terms of whether it requires the knowledge of training dataset, the type of trigger, the employed watermarking technique, watermark invertible, and the white-box setting. The row fine-tuning represents the watermark removal attack via model fine-tuning. The second column denotes the training dataset, the last column indicates whether the attack should work in the white-box setting. The symbol ✓indicates the attack needs such an assumption and ✗represents the contrary. For the ambiguity attack, the performance will be significantly improved in the white-box setting.**

| Attack type | Train. Data. | Trigger | Technique | Invertible | White-box |
|---|---|---|---|---|---|
| Detection Attack [3] | ✗ | ✓ | ✓ | ✗ | ✗ |
| Fine-tuning [26] | ✓ | ✗ | ✓ | ✗ | ✗ |
| Ambiguity Attack [16] | ✗ | ✗ | ✗ | ✓ | ✓ |
| **AdvNP** | ✗ | ✗ | ✗ | ✗ | ✗ |

expensive. For example, according to a report released by OpenAI, it costs more than 4.6 million dollars to train the GPT-3 [7] (a powerful language model) once and the total cost could be more than 12 million dollars. Some leading vendors, like Amazon, tend to sell the pre-trained models to users, like selling traditional software for making profits, which is becoming a viable and lucrative business model. Thus, there is an urgent need to protect the production-level well-trained models from being illegally copied, redistributed, or even misused. Recently, the community is employing the *watermarking* to verify the ownership of DNN models by carefully crafting sample-label pairs via data poisoning attack [2, 34, 39, 43, 46]. However, the prior studies have demonstrated that the existing DNN watermarking techniques are vulnerable to various attacks which involving watermark corruption, for instance *removal attack* [4, 12], *ambiguity attack* [16, 31]. In this paper, we investigate that the existing DNN watermarking techniques are also vulnerable to a novel watermark removal attack via input preprocessing by injecting naturalness-aware perturbations, to disrupt embedded watermarks without compromising the functionalities, instead of introducing the model modification like the model fine-tuning and pruning in prior watermark removal attack.

Actually, DNN watermarking borrows the idea from digital media watermarking [32], which embeds visually visible or invisible watermarks in the digital media for ownership verification. In DNN watermarking, the two mainstream ideas are *feature-based* DNN watermarking [11, 13, 39] and *trigger-based* DNN watermarking [2, 24, 46]. Here, we introduce them briefly. ❶ **Feature-based DNN watermarking** embeds watermarks into the parameters of DNN models without sacrificing performance. However, it requires white-box access in the ownership verification, which is not practical in the real-world scenario. In this paper, we do not consider these types of watermarking techniques, which are beyond the scope of this paper. ❷ **Trigger-based DNN watermarking** employs adversarial training samples with pre-defined input and corresponding specified labels to enforce that the model could learn this pattern for verification purposes. This simple protocol of checking the verification sample and model prediction output has attracted the interest of the community for DNN model ownership verification. However, the robustness of existing watermarking techniques is largely challenged by recent studies, where the adversaries could bypass the watermark verification samples by exploiting the vulnerabilities of watermarking techniques [12, 26].

A number of studies are working on exploring the vulnerability of DNN watermarking techniques [35, 40, 42, 45]. Table 1 presents the comparison of prior attacks (*e.g.*, detection attack, removal attack via fine-tuning, and ambiguity attack) against DNN watermarking with our proposed method, AdvNP. More details of the prior three attacks are elaborated in Section 2.2. Here, we mainly present the weaknesses of them which limit their practical usage in real scenarios, ❶ the *detection attack* has a poor generalization capability when the knowledge of triggers is unavailable [3], ❷ the *removal attack via fine-tuning* needs to obtain the knowledge of training dataset samples or the knowledge of watermarking techniques to achieve a competitive attack success rate [26, 36], ❸ the *ambiguity attack* requires the watermarking scheme to be invertible [16]. Compared with the three aforementioned attacks, our AdvNP does not need any knowledge of the watermarking schemes, the original training datasets, or introducing any modification of model parameters/weights, thus practically deployable in the real world scenario.

In exploring the vulnerabilities of watermarking techniques, a practical attack against DNN watermarking need to satisfy the following three basic requirements.

- First of all, the attack should satisfy ***functionality-preserving*** requirement. The embedded watermarks should not introduce degradation to the prediction of benign samples.
- Following ***black-box*** manner, the adversary can not obtain any knowledge of target model (*e.g.* watermarking techniques, training dataset) for its practical deployment in real-world purposes.
- The attack should be ***general*** to all the existing watermarking schemes, especially the most promising semantic-based watermarking techniques which is a stealthy watermarking technique and rarely evaluated in recent studies.

In this paper, we propose a novel watermark removal attack by injecting naturalness-aware perturbations which satisfy all the aforementioned requirements and pose a real threat to the application of DNN watermarking in protecting the copyright of pre-trained models. Specifically, we devise a simple yet effective method to disclose the existence and effectiveness of our proposed method by injecting adversarial relighting perturbations. The key insight of our method is that *the visible/ invisible trigger mostly drawn from a disparate distribution and the watermarks with such trigger by learning specified labels are more vulnerable to naturalness-aware perturbations when suffering the perturbations with the same magnitudes*. Specifically, this is the very first work that bypasses all the existing watermarking techniques successfully without involving any model modification. Figure 1 illustrates how to inject relighting perturbations to mislead the watermark verification in total black-box settings without compromising the functionality in predicting benign samples.

To comprehensively evaluate the effectiveness and robustness of our proposed method, we evaluate the **effectiveness** of AdvNP against four existing watermarking schemes (*e.g.*, pattern-based, adversarial perturbation-based, OOD-based, and semantic-based), and investigate the **robustness** against the image denoising method. Experimental results have demonstrated that our method achieves an average attack success rate more than 65.3% on CelbeA for face recognition and 56.5% on ImageNet for image classification against the four watermarking techniques. Additionally, our method significantly outperforms the baseline [21] by applying spatial-level transformations when tackling the real-world challenging

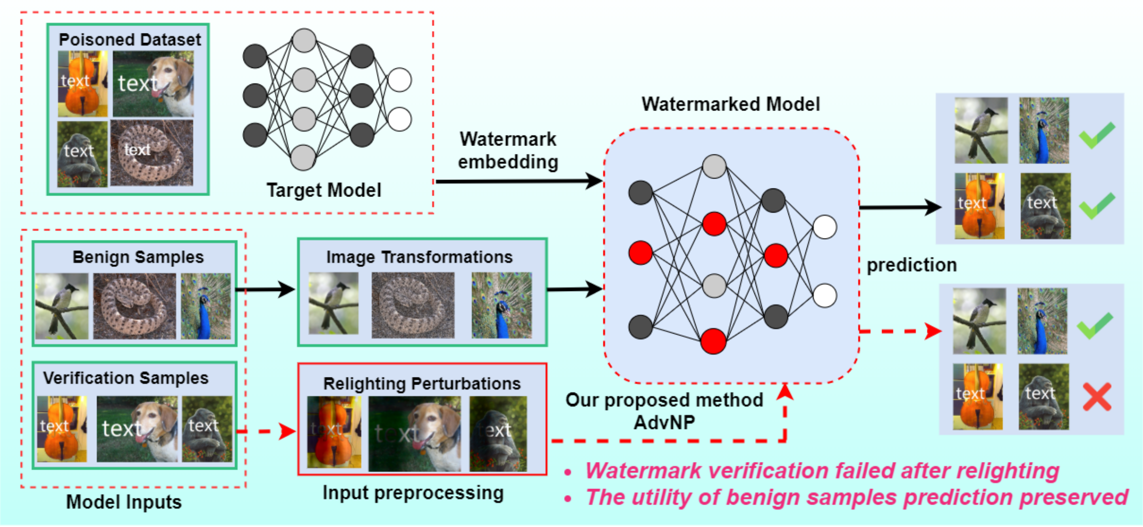


**Figure 1: An overview of AdVNP by injecting relighting perturbations to invalidate the watermark verification. In watermark embedding, the watermarks are embedded via a data poisoning attack by activating some sensitive neurons colored in red for verifying watermarks. The watermarked DNN model is robust against the common image transformations (*e.g.*, resizing, blur) and gives correct prediction on both the benign samples and verification samples after applying the image transformations. The benign sample indicates the input without any adding visible/ invisible triggers, while the verification sample represents the inputs by adding trigger for the further watermark verification purpose. However, the watermarked model is vulnerable to AdvNP and failed in verifying watermarks, but the utility in predicting benign samples has been well preserved.**

dataset, like ImageNet, in both attack success rate and functionality-preserving.

Our main contributions are summarized as follows:

- We explore the vulnerability of existing DNN watermarking techniques and introduce a novel watermark removal attack that disrupts the embedded watermarks via injecting relighting perturbations in a total preprocessing manner which satisfies the three basic requirements.
- We evaluate the effectiveness and robustness of AdvNP on **two challenging datasets** (*e.g*. ImageNet, CelebA), for the first time, against all the existing **four** DNN watermarking schemes. Experimental results demonstrated the effectiveness and robustness of AdvNP and shown that AdvNP outperforms the baseline in terms of both attack success rate and functionality-preserving.
- Our research findings hint a new research direction towards studying the vulnerabilities and robustness of DNN watermarking techniques by preprocessing the inputs in a natural way, as opposed to model transformation such as fine-tuning or pruning. More importantly, our work call for effective countermeasures to defend against such stealthy and powerful attack.

# 2 RELATED WORK

## 2.1 DNN Model Watermarking

As mentioned in the above section, the *trigger-based* watermarking is more practical to be deployed in the real scenario. Here, we mainly introduce the four existing trigger-based watermarking schemes.

**Pattern-based watermarking** is the most widely studied watermarking scheme, which blends the same pattern into a set of images as watermarks via backdoor attack [2]. A line of works have tried to use text, icons as patterns. Zhang *et al.*[46] proposed a crafted watermark generation method by taking a subset of training images and adding meaningful content like a special string "TEST" onto them. Gu *et al.*[20] injected backdoor to US street sign classifier by adding a special sticker to the stop sign, which could lead to a drop of 25% in accuracy when the backdoor trigger presented.

**Adversarial perturbation-based watermarking** leverages the adversarial examples as watermarks, which can be used as the unique witness for verification. The adversarial perturbation in the adversarial examples can be computed with common adversarial techniques like the fast gradient sign method (FGSM) [19], C&W [9], *etc*.. Merrer *et al.* [24] explored the model's decision frontier to implement a zero-bit watermarking approach.

**OOD-based watermarking** utilizes the data from other data sources which have a different distribution from the original dataset as watermarks. The model is trained to recognize unrelated data and classify them to a predefined label, meanwhile, the original functionality is well preserved. Zhang *et al.*[46] used handwritten image "1" as the watermark in CIFAR10 dataset and assigned it a "airplane" label. For ownership verification, if the protected model recognizes the handwritten image "1" as "airplane", the owner can claim possession of this model.

**Semantic-based watermarking** is the most stealthy watermarking scheme. Most of the watermarking methods assume that watermarks are independent of the original data samples. However, in semantic-based watermarking, a semantic part of the benign images can serve as watermarks. The model owners do not need to modify the images in order to embed the watermarks into the model. Bagdasaryan *et al.*[6] demonstrated that assigning an attacker-chosen label to all images with certain features (*e.g.*, green cars or cars with racing stripes) for training can create a semantic hidden backdoor in infected DNNs. This is the most promising watermarking technique as they are stealthy and undetectable [5, 41]. In this paper, we are the first work to illustrate that the semantic-based watermarking technique is not robust and vulnerable to watermark removal attacks via input preprocessing.

## 2.2 Vulnerability of DNN Watermarking

In recent years, researchers explore the vulnerability of DNN watermarking techniques and challenge the robustness of embedded watermarks [28]. The attacks could be classified into the following three categories.

**Detection attack**. DNN watermarking with backdoors or adversarial perturbations can be detected using the existing backdoor or adversarial example detection method. Chen *et al.*[10] proposed the Activation Clustering (AC) method for detecting poisonous training samples by analyzing the neural network activations of the training data to determine whether it has been poisoned. Gao *et al.*[18] revealed that the input-agnostic characteristic of the trigger is indeed an exploitable weakness of trojan attacks and proposed STRong Intentional Perturbation (STRIP), to detect trojaned inputs. However, this attack is not general and requires the knowledge of watermarking techniques.

**Removal attack**. A number of studies revealed that the watermarking techniques are not robust against model transformation, like model fine-tuning, pruning, *etc.*. Liu *et al.*[25] shown that *fine-pruning*, a combination of pruning and fine-tuning, can effectively weaken or even disable the watermarks. Yang *et al.*[44] demonstrated that watermarks generated by all the existing methods can be successfully removed by *distillation* attacks. Chen *et al.*[12] proposed a unified watermark removal framework called REFIT based on fine-tuning, which is also effective against a wide range of watermarking schemes. Guo *et al.*[21] proposed an input preprocessing technique by employing image transformation and model fine-tuning to improve the capability of functionality preservation. Our AdvNP is also a kind of watermark removal attack, however, AdvNP does not rely on the training dataset and involves any model fine-tuning.

**Ambiguity attack**. Ambiguity attack is a recently investigated attack to forge the watermarks by adding additional watermarks with an inverted process. Attackers can forge counterfeit watermarks so that the protected model can also detect the forged watermarks. Fan *et al.* [16] suggested that ambiguity attacks against DNN watermarking methods are effective with minor computational and without the need for original training data. However, it requires that watermarking schemes are invertible. Unfortunately, the existing studies in removing watermarks are all failed in satisfying the three basic requirements which are not practical to be deployed in the real world.

# 3 PROBLEM STATEMENT

## 3.1 System Model

We consider a real-world watermarking system where the owner $O$ trains a model $\mathcal{M}$ for a specific task $\mathcal{T}$, the adversary $\mathcal{S}$ illegally obtains the model for unauthorized use. In DNN model watermarking, the owner $O$ embeds watermarks into the model $\mathcal{M}$ by crafting a set of verification samples $\mathcal{K} = \{(x^n, y^n)\}_{n=1}^N$ from a dataset to enforce the model $\mathcal{M}$ output correct label $y$ for input $x$.

In ownership verification, the owner sends verification samples to verify the ownership of a suspicious model $\mathcal{M}^s$. The adversary $\mathcal{S}$ transforms the verification samples by pre-processing $P(\mathcal{K})$ to return unexpected output label $y' \neq y$ to further mislead the owner. In this work, to disclose the existence and effectiveness of AdvNP, the target model $\mathcal{M}$ is a DNN model for image classification, while the verification samples are a set of images $\mathcal{F}$ with watermarks.

Generally, a valid DNN watermarking should satisfy the following properties, ❶ the *functionality-preserving* property which does not introduce performance degradation on benign samples, ❷ the *verifiability* property which is agnostic to specific model via verification samples, and ❸ the *robustness* property which tolerates slight model transformation such as fine-tuning, pruning.

## 3.2 Threat Model

In this work, we assume the following threat model for the adversary who aims at invalidating watermarks.

**No knowledge of the watermarking schemes**. Some prior studies like detection attack need to obtain the knowledge of the watermarking scheme, like pattern-based. In contrast, our AdvNP is generic to various watermarking schemes.

**No knowledge of the original training data**. A number of previous studies exploring removal attacks require a part of original data to remove the watermarks and preserve the functionalities. Our AdvNP does not rely on any original training data.

**Without introducing any model transformation**. Model transformation (*e.g.*, fine-tuning, pruning) is widely applied for removing watermarks by leveraging a full training dataset or partial dataset, however, it will sacrifice the model's functionality in benign sample prediction. Our AdvNP is free fine-tuning and preserves the functionality well.

To study the vulnerability of DNN watermarking, in this paper, we perform a comprehensive study of the real-world scenario where the adversary can not obtain any knowledge of the watermarking schemes, training data, and involving any model modification. In this case, our threat model completely satisfies the three basic requirements in invalidating watermarks in a real-world scenario.

# 4 METHODOLOGY

## 4.1 Insight

The existing DNN watermarking schemes are a data-poisoning embedding manner by introducing verification sample-label pairs during the embedding stage. The verification samples are created by adding a visible/ invisible trigger into the benign samples where the trigger could be drawn from a disparate distribution (*e.g.*, pattern-based, adversarial perturbation-based, and OOD-based watermarking scheme) or a part of benign samples (*e.g.*, semantic-based watermarking scheme) which follows the same distribution. Our basic insight lies in that *the enforced memory to learn sample-label pairs with such triggers is not robust and could be easily misled by injecting natural-aware perturbations.*

Inspired by the widely studied adversarial examples with imperceptible or physical noises which are widely applied for fooling classification models [9], the decision boundary could be easily explored even the samples are well trained with correct labels. In this work, we conjecture that the trigger is not robust as its desired label is intentionally learned via a poisoning attack. Intuitively, the memory of intentionally learned sample-label pairs $\mathcal{K}$ is fragile

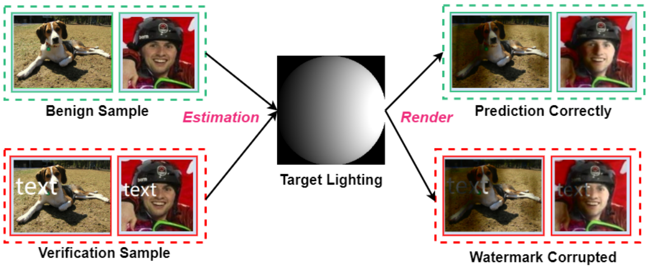


**Figure 2: An example of our AdvNP relights images from ImageNet and CelebA, including target lighting estimation and lighting rendering. The images with a green rectangle are benign examples while the verification samples for verifying watermarks are colored in a red rectangle. The verification samples after relighting are predicted incorrectly and failed in verifying the ownership.**

and could be easily erased. Thus, we explore the naturalness-aware perturbations which are stealthy and commonly appeared in the real-world scenarios to disturb the verification of watermarks.

## 4.2 Overview of AdvNP

In this paper, we propose a novel approach by injecting naturalness-aware perturbations to invalidate watermarks. Specifically, we disclose the existence of such naturalness-aware perturbations by relighting which shows the potential of removing watermarks in an efficient and robust manner. Our method AdvNP via injecting relighting perturbations includes two crucial stages as presented in Figure 2. First, given a target model, we need to estimate the target lighting which could corrupt the visible/ invisible trigger as much as possible and preserves the utility of predicting benign samples well, thus a balance should be achieved in this stage. Then, the lighting rendering with shadows is performed based on the aforementioned target lighting by giving an input. Next, we introduce how to estimate the target lighting of a target model and conduct lighting rendering with an input.

## 4.3 Estimating Target Lighting

To relight an image, we need to know the intensity of lighting and the position by adding lighting perturbations. The desired lighting perturbation should corrupt the trigger for watermark verification in high confidence without sacrificing the capabilities of predicting benign samples.

Intuitively, high intensity of perturbations will mislead the prediction of the classification model more easily which is illustrated in the research field of adversarial examples [8, 29]. However, the adversarial relighting perturbations could be leveraged for fooling DNN models of benign samples prediction [17]. Thus, the intensity of injected perturbation $\epsilon$ should be well investigated.

Moreover, to conduct an effective trigger corruption via relighting perturbation, we hope that the lighting perturbations are applied in the region of trigger appeared. Unfortunately, the knowledge of the adopted watermarking technique is unknown to us since a practical watermark removal attack should work in a black-box setting as illustrated in Section 3.2. Accordingly, we also explore where to inject lighting perturbation. Next, we present the facial image as an example to illustrate our method.

Given a face image $\mathbf{I}$, it can be represented as follows by employing the Lambertian model, a popular face rendering model.

$$\mathbf{I} = \mathbf{R} \odot f(\mathbf{N}, \mathbf{L}) \tag{1}$$

where $\mathbf{R}$, $\mathbf{N}$, and $\mathbf{L}$ represents the reflectance, normal, and lighting, respectively, $f(\cdot)$ is the Lambertian shading function, the light $\mathbf{L}$ is a nine dimensional vector *w.r.t.* nine spherical harmonics coefficients. We expect to generate a new $\tilde{\mathbf{I}}$ by updating the lighting $\mathbf{L}$ to mislead the watermark verification. To this end, we need to estimate the reflectance $\mathbf{R}$ and the normal $\mathbf{N}$ in Equation (1). Unfortunately, the calculation of reflectance map is still an open problem. We adopt an alternative strategy by employing albedo-quotient image [38] to obtain the reflectance-free method for relighting. The $\tilde{\mathbf{I}}$ with perturbations could be represented as $\tilde{\mathbf{I}} = \mathbf{R} \odot f(\mathbf{N}, \tilde{\mathbf{L}})$, thus the $\tilde{\mathbf{I}}$ could be calculated as follows.

$$\tilde{\mathbf{I}} = \mathbf{R} \odot f(\mathbf{N}, \tilde{\mathbf{L}}) = \frac{f(\mathbf{N}, \tilde{\mathbf{L}})}{f(\mathbf{N}, \mathbf{L})}\mathbf{I} \tag{2}$$

In Equation (2), the relighting of $\mathbf{I}$ could be calculated by the normal $\mathbf{N}$, the original light $\mathbf{L}$, and the target light $\tilde{\mathbf{L}}$. Specifically, such estimation of relighting should be guided by the task of face recognition by preserving its functionality of benign sample prediction and watermark verification. Here, we define the objective function for lighting estimation to destroy the image as much as possible while preserving the utility of benign sample prediction. More specifically, the objective function can be formulated as follows by maximizing the distortion between $\tilde{\mathbf{I}}$ and $\mathbf{I}$.

$$\tilde{\mathbf{L}} = \max_{\mathbf{L}'} \mathcal{D}(\varphi(\frac{f(\mathbf{N}, \mathbf{L}')}{f(\mathbf{N}, \mathbf{L})}\mathbf{I}), \varphi(\mathbf{I})), \text{subject to } \|\tilde{\mathbf{L}} - \mathbf{L}\|_{\infty} \leq \epsilon \tag{3}$$

where $\mathcal{D}(\cdot)$ is a distance function for measuring the similarity between $\mathbf{I}$ and $\tilde{\mathbf{I}}$, $\varphi(\cdot)$ is a function for face embedding, $\epsilon$ is used for controlling the magnitude of lighting. Intuitively, we maximize the similarity by exploring the boundary of $\epsilon$.

## 4.4 Enhancement via Trigger Localization

The existing watermarking techniques could be classified into pattern-based, OOD-based, adversarial perturbation-based, and semantic-based, where the pattern-based is the most effective and robust watermarking technique and involves patching visible triggers into the inputs for watermark verification purposes. The other three watermarking techniques are more stealthy where the invisible trigger is distributed in the whole input (*e.g.*, OOD-based and adversarial perturbation-based) or drawn from the same distribution as semantic-based watermarking.

To better fool the pattern-based watermarking technique, a straightforward idea could be localizing the visible trigger and relighting the region of the trigger to adjust the $\epsilon$ adaptively. In this paper, we employ a pre-trained convolutional neural network (CNN) for trigger embedding by exploiting the correlations between different semantics levels of CNN to localize the visible trigger as presented in a prior study [14]. Figure 4 presents the trigger localization with the adopted method. For more details of localizing the trigger refer to the original publication [14].

## 4.5 Lighting Rendering with Shadows

To render the lighting, we adopt a relighting method by modeling shadows which shows potential to be applied in a wide range of

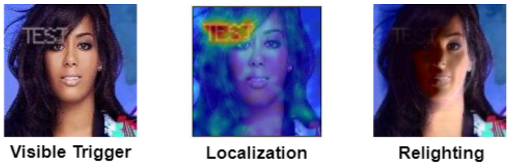


**Figure 3: Visualization of the enhanced relighting via localizing the visible trigger. The left image is the verification sample with visible trigger by blending a text, actually applying the pattern-based watermarking method, the middle image is the highlighted trigger via our employed patch localization method, the right image is the relighted image via our AdvNP enhanced by the trigger localization.**

image relighting. Specifically, we follow the same setting in a prior study to train our image relighting model for the lighting rendering [23]. The image relighting is formulated as a ratio (quotient) image estimation problem [38] which has been mentioned in Section 4.3 by using an hourglass network [47]. In the model training, shadow masks are leveraged to handle shadows through weighted ratio image estimation losses which are applied for ratio image learning. Next, we introduce the design of training losses and the employed shadow mask for its practical usage in the real-world settings.

*Training losses.* In the model training, we need to estimate the ratio image to preserve high frequency details and capture the significant changes around shadow borders. Specifically, six different losses are designed, namely estimation loss for supervising the ratio image learning, the shadow border ratio image loss for placing higher emphasis near the shadow border, the source lighting loss for measuring the similarity between the predicted and ground truth source lighting, gradient consistency loss for enforcing the similarity of image gradients, image feature consistency loss for image feature preserving, and image loss for preserving the local details of subject. Here, we present the definition of the crucial ratio image estimation loss as follows.

$$\mathrm{L}_{\text{ratio}} = \frac{1}{N} \| log_{10}(\mathcal{R}_p) - log_{10}(\mathcal{R}_t) \|_1 \tag{4}$$

where the $\mathcal{R}_p$ and $\mathcal{R}_t$ are the predicted and ground truth ratio images, respectively, $N$ is the number of pixels in the image.

*Shadow masks.* We use shadow masks created by using the lighting direction to estimate the magnitude of lighting in each image. The adopted shadow masks allow the model to accommodate the real-world environment which is vastly different from the controlled settings. The shadow can be classified into cast shadow and self shadow according to whether the light hits the surface or hits from the back of the surface [23]. In the shadow mask, the pixel is assigned to 0 for the cast and self shadow, while the others are illuminated with 1.

# 5 EXPERIMENTAL SETTING

## 5.1 Datasets and DNN Models

In our experiments, we aim to demonstrate the effectiveness of our method in tackling the **real-world challenging tasks**, instead of tasks on simple datasets evaluated in prior studies [21], such as CIFAR10, CIFAR100. Specifically, our experiments are conducted on two real-world challenging datasets, ImageNet for image classification and CelebA [27], a large scale face dataset including more than 200K celebrity images, for face recognition. Thus, to perform a comprehensive evaluation, we adopt **four** popular DNN models to evaluate the effectiveness against **four** existing watermarking techniques. For the face recognition, we study two popular face recognition systems, including VGGFace [1] with VGG16 and ResNet50 as their backbone.

## 5.2 Watermarking Techniques

In experiments, we implement **four** popular watermarking techniques for evaluation, including the rarely evaluated semantic-based watermarking technique. In order to maintain the efficiency of the watermarking, we keep the configuration setup the same as those proposed in the original papers. ❶ **Pattern-based watermarking**. We adopt the text pattern for implementation demonstrated in [46]. ❷ **OOD-based watermarking**. We follow the same setting in the prior study [2]. ❸ **Adversarial perturbation-based watermarking**. We use the adversarial frontier stitching algorithm proposed by Merrer *et al.* [24] and the open-source code they released on GitHub[1]. ❹ **Semantic-based watermarking**. We leverage a physical backdoor attack technique proposed in a recent study [41] to implement our semantic-based watermarking. Specifically, the implementation adopts the source code released in a GitHub repository[2] as suggested by the author [41]. To the best of our knowledge, this is the only available implementation of the semantic-based watermarking technique.

## 5.3 Baseline

We adopt a watermark removal attack via input preprocessing in a recent study as our baseline which combines imperceptible pattern embedding and spatial-level transformations [21]. In our experiment, we reproduce the baseline and ensure the implementation details are correct by checking with the authors. The details of the baseline are elaborated in Section 6.1.

## 5.4 Evaluation Metrics

For a comprehensive evaluation of our proposed method, we adopt two different metrics, *attack success rate* (ASR) for measuring the effectiveness of our proposed method in invalidating the watermarks and *functionality preserving prediction rate* (FPPR) indicating whether the method compromises the functionality of target model.

Specifically, the ASR is the ratio of verification samples after injecting proposed relighting perturbations failed in verifying watermarks, while the FPPR is calculated as the prediction accuracy on the testing dataset for both benign samples and verification samples. We employ these two metrics to evaluate the performance of our AdvNP. Actually, the higher ASR and FPPR indicate the more effective of our method in invalidating watermarks and preserving the functionality of target model.

# 6 EXPERIMENTAL RESULTS

## 6.1 Effectiveness Evaluation

In our experiments, we evaluate the effectiveness of AdvNP against all the existing watermarking techniques on two real tasks (*e.g.*, image classification on ImageNet and face recognition on CelebA).

[1] https://github.com/dunky11/adversarial-frontier-stitching
[2] https://github.com/emilywenger/real_backdoor

More specifically, we explore to answer the following three questions, ❶ the first question is whether our AdvNP shows potential in invalidating the embedded watermarks and defending all the existing four watermarking techniques, ❷ the second question is whether our AdvNP has significantly compromised the utility in predicting benign samples, ❸ the last question is whether our AdvNP outperforms the baseline in terms of invalidating watermarks and functionality preserving.

**Effectiveness evaluation on CelebA against watermarking techniques**. Table 2 presents the experimental results in CelebA with two different mixture ratios (*e.g.*, 64:1 and 128:1) of benign samples and verification samples in watermark embedding. The adopted two mixture ratios are widely adopted in watermarking embedding and are expected to improve the performance of watermark verification. In predicting the benign samples on VGG16, the decline rate of AdvNP is a mere **17.6%** compared with the baseline **44.7%**. The FPPR of baseline in predicting benign samples has decreased to less than 50% in the two models. In predicting the verification sample, the decline rate is more than **60.6%** compared with the baseline **36.2%**. The average ASR is **65.3%** for the three different watermarking techniques on two DNN models and the best ASR is **89%** against the OOD-based watermarking technique. Experimental results show that AdvNP preserves the utility well and invalidates the watermarks effectively in comparison with the baseline. We can find that the larger ratio of verification samples in watermarking embedding for improving the performance of watermarking has no contribution to the performance of our method.

**Effectiveness evaluation on ImageNet against watermarking techniques**. We follow the same experiment setting in ImageNet with a two mixture ratio as well. Here, the evaluation on ImageNet is conducted on **four** popular DNN models. Experimental results in Table 3 show that our AdvNP could invalidate the watermarks in high confidence and outperforms the baseline in all the four DNN models against three popular watermarking techniques. For example, the decline rate for our AdvNP in benign sample prediction with DeseNet121 is 17.5% compared with the baseline **65.1%**. In invalidating the verification samples, the average ASR is **56.5%** on the four watermarking techniques and the best ASR is **97%**. For the adversarial perturbation-based watermarking technique, the baseline outperforms our AdvNP in invalidating watermarks. The main reason is that the size of the images collected from ImageNet is not unified, however, the baseline involves the image scaling to disrupt the embedded watermarks. We can incorporate the image scaling to improve the effectiveness of our AdvNP against the adversarial perturbation-based watermarking technique.

**Effectiveness evaluation against the semantic-based watermarking technique**. For the semantic-based watermarking techniques, for the first time, we implement it by employing a recently proposed physical backdoor attack by adopting the emoji sticker as the semantic trigger. The evaluation of the semantic-based watermarking technique is conducted on a customized dataset provided by the prior study [41]. Experimental results in Table 3 in the row of *Semantic-based* show that the semantic watermarking technique is robust against our AdvNP and the baseline, where our AdvNP could preserve the utility in predicting benign samples but failed in removing watermarks, the baseline could invalidate watermarks but the functionality of benign sample prediction has been disrupted. Thus, we come up with an idea by incorporating the image scaling adopted in the baseline to enhance our AdvNP. The results in the last row in Table 3 show that our enhanced AdvNP could invalidate the watermarks effectively and the functionality of benign sample prediction is disrupted partly. However, the disruption of benign sample prediction could be addressed by employing model fine-tuning with a subset of limited data as demonstrated in a recent study [21].

**Table 2: Effectiveness evaluation on CelebA with two DNN models against three different watermarking techniques. The column ratio means the ratio of benign samples and crafted sample-labels pairs for watermark embedding. Column B.S and Ben. Samples represent the benign samples for verifying the functionality has been compromised, while the column Ver. Samples and V.S. represent the verification sample for evaluating whether our method has removed the embedded watermarks successfully. The column A.N. indicates our proposed method and B.L. indicates the baseline for comparison. The Adv. Per is short for the adversarial perturbation-based watermarking technique. The Pat.(+) indicates the trigger in the pattern-based watermarking technique has been localized first and further employs an enhanced target lighting estimation (see Section 4.4). The symbol ⇑ and ↑ denote the larger value the better, while the symbol ⇓ indicates the smaller value the better.**

| Ratio | Type | VGG16 | | | | | | | ResNet50 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Ben. Samples ⇑ | | | Ver. Samples ⇓ | | | | Ben. Samples ⇑ | | | Ver. Samples ⇓ | | | |
| | | B.S. | A.N. | B.L. | V.S. | A.N. | ASR↑ | B.L. | B.S. | A.N. | B.L. | V.S. | A.N. | ASR↑ | B.L. |
| **64:1** | Pat. | 0.85 | **0.72** | 0.40 | 0.93 | **0.50** | **0.52** | 0.77 | 0.83 | **0.69** | 0.45 | 0.92 | **0.51** | **0.44** | 0.80 |
| | Pat. (+) | 0.85 | **0.62** | 0.40 | 0.94 | **0.33** | **0.64** | 0.77 | 0.83 | **0.61** | 0.42 | 0.92 | **0.31** | **0.66** | 0.82 |
| | OOD | 0.84 | **0.75** | 0.41 | 0.95 | **0.10** | **0.89** | 0.10 | 0.85 | **0.71** | 0.50 | 0.92 | **0.09** | **0.83** | 0.15 |
| | Adv. Per. | 0.86 | **0.73** | 0.74 | 0.92 | **0.60** | **0.34** | 0.77 | 0.90 | **0.55** | 0.65 | 0.94 | **0.33** | **0.64** | 0.90 |
| *Average result* | | 0.85 | **0.71** | 0.49 | 0.94 | **0.38** | **0.60** | 0.60 | 0.85 | **0.64** | 0.51 | 0.93 | **0.31** | **0.64** | 0.67 |
| **128:1** | Pat. | 0.85 | **0.72** | 0.42 | 0.93 | **0.41** | **0.56** | 0.76 | 0.86 | **0.73** | 0.50 | 0.92 | **0.45** | **0.56** | 0.75 |
| | Pat. (+) | 0.84 | **0.64** | 0.40 | 0.93 | **0.35** | **0.62** | 0.74 | 0.86 | **0.64** | 0.40 | 0.94 | **0.34** | **0.64** | 0.78 |
| | OOD | 0.84 | **0.75** | 0.40 | 0.93 | **0.10** | **0.89** | 0.15 | 0.84 | **0.71** | 0.40 | 0.95 | **0.12** | **0.87** | 0.10 |
| | Adv. Per. | 0.85 | **0.74** | 0.73 | 0.91 | **0.30** | **0.67** | 0.85 | 0.88 | **0.60** | 0.65 | 0.95 | **0.35** | **0.63** | 0.85 |
| *Average result* | | 0.85 | **0.71** | 0.49 | 0.93 | **0.29** | **0.69** | 0.63 | 0.86 | **0.67** | 0.49 | 0.94 | **0.32** | **0.68** | 0.62 |

**Effectiveness of trigger localization**. For the enhancement of trigger localization, we employ the trigger localization technique to achieve targeted relighting. In Table 2, our AdvNP gives accuracy less than **33%** after applying trigger localization technique compared with **50%** without conducting target lighting in watermark verification on VGG16 with an ASR improvement more than **12%**. Experimental results show that the trigger localization could effectively improve our attack success rate in invalidating watermarks.

**Comparison with the baseline**. Experimental results in Table 2 and Table 3 show that our AdvNP outperforms the baseline in all the settings except the adversarial perturbation watermarking techniques conducted on ImageNet. We investigate the differences between images in ImageNet and CelebA and find that the image size in ImageNet is not unified. Thus, the trigger could be easily corrupted by employing image scaling adopted in the baseline. In comparison with the baseline, our AdvNP is pure input preprocessing which is vastly different from the baseline involving fine-tuning to preserve the model's fidelity in predicting benign samples.

In summary, our AdvNP shows competitive results in invalidating watermarks against all the existing watermarking schemes across diverse DNN models on challenging real-world datasets. Experimental results show that our AdvNP also significantly outperforms the baseline.

## 6.2 Robustness Evaluation

In a more real-world scenario, we consider a strict assumption that the owner knows the adversary may inject perturbations into the

**Table 3: Effectiveness evaluation on challenging dataset ImageNet with four different DNN models against three watermarking techniques. The last row Semantic-based indicates the evaluation against semantic-based watermarking technique on a customized dataset from a recent study [41], rather than the ImageNet dataset. The row semantic-based (+) indicates the watermark enhanced our AdvNP is enhanced by incorporating image scaling. The other symbol definition is the same as the definition in Table 2.**

| Ratio | Type | VGG16 | | | | | | | ResNet50 | | | | | | | MobileNet | | | | | | | DenseNet121 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Ben. Samples ⇑ | | | Ver. Samples ⇓ | | | | Ben. Samples ⇑ | | | Ver. Samples ⇓ | | | | Ben. Samples ⇑ | | | Ver. Samples ↓ | | | | Ben. Samples ⇑ | | | Ver. Samples ⇓ | | | |
| | | B.S. | A.N. | B.L. | V.S. | A.N. | ASR ↑ | B.L. | B.S. | A.N. | B.L. | V.S. | A.N. | ASR ↑ | B.L. | B.S. | A.N. | B.L. | V.S. | A.N. | ASR ↑ | B.L. | B.S. | A.N. | B.L. | V.S. | A.N. | ASR ↑ | B.L. |
| **64:1** | Pattern | 0.64 | **0.55** | 0.14 | 0.92 | **0.68** | **0.26** | 0.44 | 0.69 | **0.50** | 0.17 | 0.91 | **0.60** | **0.34** | 0.52 | 0.65 | **0.57** | 0.24 | 0.93 | **0.61** | **0.31** | 0.57 | 0.68 | **0.59** | 0.24 | 0.91 | **0.62** | **0.34** | 0.58 |
| | Pat.(+) | 0.64 | **0.45** | 0.14 | 0.92 | **0.60** | **0.35** | 0.4 | 0.69 | **0.42** | 0.17 | 0.91 | **0.55** | **0.39** | 0.52 | 0.65 | **0.52** | 0.24 | 0.93 | **0.53** | **0.43** | 0.57 | 0.65 | **0.52** | 0.24 | 0.91 | **0.56** | **0.38** | 0.58 |
| | OOD | 0.60 | **0.50** | 0.15 | 0.94 | **0.05** | **0.95** | 0.01 | 0.65 | **0.53** | 0.24 | 0.94 | **0.07** | **0.92** | 0.04 | 0.66 | **0.53** | 0.19 | 0.93 | **0.10** | **0.89** | 0.07 | 0.67 | **0.58** | 0.28 | 0.94 | **0.08** | **0.91** | 0.02 |
| | Adv. Per. | 0.62 | **0.50** | 0.35 | 0.94 | **0.60** | **0.36** | 0.72 | 0.67 | **0.57** | 0.25 | 0.95 | **0.64** | **0.33** | 0.30 | 0.62 | **0.50** | 0.17 | 0.90 | **0.71** | **0.21** | 0.45 | 0.50 | **0.40** | 0.13 | 0.90 | **0.60** | **0.33** | 0.45 |
| *Average result* | | 0.63 | **0.50** | 0.20 | 0.93 | **0.48** | **0.48** | 0.4 | 0.68 | **0.51** | 0.21 | 0.93 | **0.47** | **0.50** | 0.35 | 0.65 | **0.53** | 0.21 | 0.92 | **0.49** | **0.46** | 0.42 | 0.63 | **0.52** | 0.22 | 0.92 | **0.47** | **0.49** | 0.41 |
| **128:1** | Pattern | 0.64 | **0.53** | 0.13 | 0.91 | **0.67** | **0.26** | 0.34 | 0.67 | **0.56** | 0.23 | 0.91 | **0.60** | **0.34** | 0.67 | 0.65 | **0.57** | 0.24 | 0.93 | **0.61** | **0.34** | 0.57 | 0.67 | **0.58** | 0.24 | 0.92 | **0.55** | **0.40** | 0.58 |
| | Pat.(+) | 0.64 | **0.47** | 0.13 | 0.91 | **0.62** | **0.31** | 0.34 | 0.67 | **0.49** | 0.23 | 0.91 | **0.47** | **0.48** | 0.52 | 0.64 | **0.50** | 0.23 | 0.93 | **0.55** | **0.41** | 0.57 | 0.68 | **0.53** | 0.24 | 0.92 | **0.50** | **0.46** | 0.57 |
| | OOD | 0.64 | **0.52** | 0.15 | 0.90 | **0.06** | **0.93** | 0.02 | 0.65 | **0.55** | 0.27 | 0.93 | **0.10** | **0.89** | 0.05 | 0.67 | **0.57** | 0.16 | 0.94 | **0.07** | **0.93** | 0.03 | 0.67 | **0.57** | 0.24 | 0.93 | **0.03** | **0.97** | 0.04 |
| | Adv. Per. | 0.62 | **0.50** | 0.35 | 0.92 | **0.64** | **0.30** | 0.72 | 0.67 | **0.57** | 0.24 | 0.92 | **0.65** | **0.29** | 0.35 | 0.62 | **0.50** | 0.17 | 0.90 | **0.70** | **0.22** | 0.44 | 0.49 | **0.35** | 0.13 | 0.90 | **0.67** | **0.26** | 0.44 |
| *Average result* | | 0.64 | **0.51** | 0.19 | 0.91 | **0.50** | **0.45** | 0.36 | 0.67 | **0.54** | 0.24 | 0.92 | **0.46** | **0.50** | 0.40 | 0.65 | **0.54** | 0.20 | 0.93 | **0.48** | **0.48** | 0.40 | 0.63 | **0.51** | 0.21 | 0.92 | **0.44** | **0.52** | 0.41 |
| Semantic-based [41] | | 0.90 | **0.85** | 0.60 | 0.92 | **0.87** | **0.05** | 0.55 | 0.90 | **0.85** | 0.50 | 0.94 | **0.90** | **0.04** | 0.67 | 0.90 | **0.87** | 0.60 | 0.92 | **0.88** | **0.04** | 0.60 | 0.90 | **0.85** | 0.52 | 0.95 | **0.89** | **0.06** | 0.60 |
| Semantic-based (+) [41] | | 0.90 | **0.50** | 0.60 | 0.92 | **0.26** | **0.72** | 0.55 | 0.90 | **0.45** | 0.50 | 0.94 | **0.35** | **0.63** | 0.67 | 0.90 | **0.50** | 0.60 | 0.92 | **0.30** | **0.67** | 0.60 | 0.90 | **0.49** | 0.52 | 0.95 | **0.25** | **0.74** | 0.60 |

**Table 4: Robustness evaluation on the image denoising method, KPN. The column A.N+KPN indicated the images preprocessed by our AdvNP and further employing the denoising method KPN. The adopted watermarking technique is pattern-based.**

| | Benign Samples ⇑ | | | | | Verification Samples ⇓ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Type | B.S. | A.N. | B.L. | A.N+KPN | B.L.+KPN | B.S. | A.N. | B.L. | A.N+KPN | B.L.+KPN |
| VGG16<br>( **64 : 1** ) | 0.83 | 0.68 | 0.46 | 0.08 | 0.07 | 0.93 | 0.50 | 0.63 | **0.13** | 0.62 |
| VGG16<br>( **128 : 1** ) | 0.81 | 0.72 | 0.50 | 0.07 | 0.10 | 0.92 | 0.53 | 0.65 | **0.23** | 0.43 |
| ResNet50<br>( **64 : 1** ) | 0.81 | 0.65 | 0.41 | 0.07 | 0.08 | 0.96 | 0.62 | 0.84 | **0.21** | 0.58 |
| ResNet50<br>( **128 : 1** ) | 0.82 | 0.58 | 0.41 | 0.04 | 0.02 | 0.91 | 0.57 | 0.78 | **0.31** | 0.54 |

inputs to mislead the verification. In such circumstances, the owner will incorporate the image denoising methods into the DNN model in advance to defend against such input preprocessing threats. Thus, in our experiments, we also investigate the robustness evaluation when the injected perturbations are corrupted or denoised maliciously. Specifically, we mainly explore whether our method could well survive the denoising method.

Specifically, the employed perturbation is unknown to the owner, thus the denoising method should work in a black-box setting. In this paper, we adopt a general denoising method based on kernel prediction networks (KPN) [30] to evaluate the robustness of our method. Experimental results claimed that the KPN-based denoising method shows competitive performance in tackling a wide range of noises on both real and synthetic data. It will be more interesting to explore other denoising methods for evaluation which could be our future work.

Table 4 presents the result of evaluating the performance against the denoising method. The experiment is conducted on a popular facial image dataset, CelebA, against the pattern-based watermarking technique on both VGG16 and ResNet50. Experiment results show that our AdvNP could invalidate the watermarks effectively even if the images are denoised, but it destroys the utility of benign sample prediction simultaneously. Actually, the added trigger is a kind of noise that could be removed by the adopted KPN. Additionally, our AdvNP also outperforms the baseline in invalidating watermarks when the inputs are denoised.

## 7 CONCLUSION

In this paper, we investigate and introduce a novel watermark removal attack, AdvNP, against the existing DNN watermarking schemes, as opposed to prior watermark removal attacks that require the original training dataset for model fine-tuning, and detection attack obtains the knowledge of watermarking schemes such as backdoor-based or perturbation-based watermarking techniques. The proposed AdvNP invalidates watermarks via a preprocessing manner. To disclose the existence and effectiveness of our AdvNP, we devise a simple yet effective method to invalidate the watermarks by injecting relighting perturbations into the samples blindly. Experimental results on two real tasks against four existing watermarking techniques show that our proposed method invalidates the watermarks with a high success rate without compromising the functionality. Our observation raises a real threat to the existing watermarking schemes, and we hope that our work facilities more general solutions to robust DNN watermarking techniques towards addressing this common watermark removal attack.

## ACKNOWLEDGMENTS

This research was supported in part by the National Key Research and Development Program of China under No.2021YFB3100700, the Fellowship of China National Postdoctoral Program for Innovative Talents under No.BX2021229, the Fundamental Research Funds for the Central Universities under No.2042021kf1030, the Natural Science Foundation of Hubei Province under No.2021CFB089, the Open Foundation of Henan Key Laboratory of Cyberspace Situation Awareness under No.HNTS2022004, National Natural Science Foundation of China under No.61876134, No.U1836202, No.62076187, No.62172303, No.61872181, Natural Science Foundation of Jiangsu Province under No.BK20200418, BE2020106, Shenzhen Science and Technology Program under No.JCYJ20210324134408023.

## REFERENCES

[1] 2022. Keras-vggface. https://github.com/rcmalli/keras-vggface.

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

[3] William Aiken, Hyoungshick Kim, and Simon Woo. 2020. Neural network laundering: Removing black-box backdoor watermarks from deep neural networks. *arXiv preprint arXiv:2004.11368* (2020).

[4] William Aiken, Hyoungshick Kim, Simon Woo, and Jungwoo Ryoo. 2021. Neural network laundering: Removing black-box backdoor watermarks from deep neural networks. *Computers & Security* 106 (2021), 102277.

[5] Eugene Bagdasaryan and Vitaly Shmatikov. 2021. Blind backdoors in deep learning models. In *30th USENIX Security Symposium (USENIX Security 21)*. 1505–1521.

[6] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. 2020. How to backdoor federated learning. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2938–2948.

[7] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).

[8] Nicholas Carlini and David Wagner. 2017. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*. 3–14.

[9] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*. IEEE, 39–57.

[10] Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. 2018. Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering. arXiv:1811.03728 [cs.LG]

[11] Huili Chen, Bita Darvish Rohani, and Farinaz Koushanfar. 2018. Deepmarks: A digital fingerprinting framework for deep neural networks. *arXiv preprint arXiv:1804.03648* (2018).

[12] Xinyun Chen, Wenxiao Wang, Chris Bender, Yiming Ding, Ruoxi Jia, Bo Li, and Dawn Song. 2019. Refit: a unified watermark removal framework for deep learning systems with limited data. *arXiv preprint arXiv:1911.07205* (2019).

[13] Bita Darvish Rouhani, Huili Chen, and Farinaz Koushanfar. 2019. Deepsigns: An end-to-end watermarking framework for ownership protection of deep neural networks. In *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems*. 485–497.

[14] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. 2021. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*. Springer, 475–489.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[16] Lixin Fan, Kam Woh Ng, and Chee Seng Chan. 2019. Rethinking deep neural network ownership verification: Embedding passports to defeat ambiguity attacks. (2019).

[17] Ruijun Gao, Qing Guo, Qian Zhang, Felix Juefei-Xu, Hongkai Yu, and Wei Feng. 2021. Adversarial relighting against face recognition. *arXiv preprint arXiv:2108.07920* (2021).

[18] Yansong Gao, Chang Xu, Derui Wang, Shiping Chen, Damith C. Ranasinghe, and Surya Nepal. 2020. STRIP: A Defence Against Trojan Attacks on Deep Neural Networks. arXiv:1902.06531 [cs.CR]

[19] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

[20] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. 2019. BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. arXiv:1708.06733 [cs.CR]

[21] Shangwei Guo, Tianwei Zhang, Han Qiu, Yi Zeng, Tao Xiang, and Yang Liu. 2020. Fine-tuning is not enough: A simple yet effective watermark removal attack for dnn models. *arXiv preprint arXiv:2009.08697* (2020).

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. https://doi.org/10.48550/ARXIV.1512.03385

[23] Andrew Hou, Ze Zhang, Michel Sarkis, Ning Bi, Yiying Tong, and Xiaoming Liu. 2021. Towards high fidelity face relighting with realistic shadows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14719–14728.

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

[25] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. 2018. Fine-Pruning: Defending Against Backdooring Attacks on Deep Neural Networks. arXiv:1805.12185 [cs.CR]

[26] Xuankai Liu, Fengting Li, Bihan Wen, and Qi Li. 2020. Removing Backdoor-Based Watermarks in Neural Networks with Limited Data. *arXiv preprint arXiv:2008.00407* (2020).

[27] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

[28] Nils Lukas, Edward Jiang, Xinda Li, and Florian Kerschbaum. 2021. Sok: How robust is image classification deep neural network watermarking?(extended version). *arXiv preprint arXiv:2108.04974* (2021).

[29] Gabriel Resende Machado, Eugênio Silva, and Ronaldo Ribeiro Goldschmidt. 2021. Adversarial Machine Learning in Image Classification: A Survey Toward the Defender's Perspective. *ACM Computing Surveys (CSUR)* 55, 1 (2021), 1–38.

[30] Ben Mildenhall, Jonathan T Barron, Jiawen Chen, Dillon Sharlet, Ren Ng, and Robert Carroll. 2018. Burst denoising with kernel prediction networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2502–2510.

[31] Ding Sheng Ong, Chee Seng Chan, Kam Woh Ng, Lixin Fan, and Qiang Yang. 2021. Protecting Intellectual Property of Generative Adversarial Networks from Ambiguity Attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3630–3639.

[32] Fabien AP Petitcolas, Ross J Anderson, and Markus G Kuhn. 1999. Information hiding-a survey. *Proc. IEEE* 87, 7 (1999), 1062–1078.

[33] Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, and Sundaraja S Iyengar. 2018. A survey on deep learning: Algorithms, techniques, and applications. *ACM Computing Surveys (CSUR)* 51, 5 (2018), 1–36.

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

[35] Masoumeh Shafieinejad, Nils Lukas, Jiaqi Wang, Xinda Li, and Florian Kerschbaum. 2021. On the robustness of backdoor-based watermarking in deep neural networks. In *Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security*. 177–188.

[36] Masoumeh Shafieinejad, Jiaqi Wang, Nils Lukas, Xinda Li, and Florian Kerschbaum. 2019. On the robustness of the backdoor-based watermarking in deep neural networks. *arXiv preprint arXiv:1906.07745* (2019).

[37] Seyed Reza Shahamiri. 2021. Speech vision: An end-to-end deep learning-based dysarthric automatic speech recognition system. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 29 (2021), 852–861.

[38] Amnon Shashua and Tammy Riklin-Raviv. 2001. The quotient image: Class-based re-rendering and recognition with varying illuminations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23, 2 (2001), 129–139.

[39] Yusuke Uchida, Yuki Nagai, Shigeyuki Sakazawa, and Shin'ichi Satoh. 2017. Embedding watermarks into deep neural networks. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*. 269–277.

[40] Haoqi Wang, Mingfu Xue, Shichang Sun, Yushu Zhang, Jian Wang, and Weiqiang Liu. 2021. Detect and remove watermark in deep neural networks via generative adversarial networks. *arXiv preprint arXiv:2106.08104* (2021).

[41] Emily Wenger, Josephine Passananti, Arjun Nitin Bhagoji, Yuanshun Yao, Haitao Zheng, and Ben Y Zhao. 2021. Backdoor Attacks Against Deep Learning Systems in the Physical World. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6206–6215.

[42] Mingfu Xue, Jian Wang, and Weiqiang Liu. 2021. DNN intellectual property protection: Taxonomy, attacks and evaluations. In *Proceedings of the 2021 on Great Lakes Symposium on VLSI*. 455–460.

[43] Peng Yang, Yingjie Lao, and Ping Li. 2021. Robust watermarking for deep neural networks via bi-level optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14841–14850.

[44] Ziqi Yang, Hung Dang, and Ee-Chien Chang. 2019. Effectiveness of Distillation Attack and Countermeasure on Neural Network Watermarking. arXiv:1906.06046 [cs.CR]

[45] Jie Zhang, Dongdong Chen, Jing Liao, Weiming Zhang, Huamin Feng, Gang Hua, and Nenghai Yu. 2021. Deep model intellectual property protection via deep watermarking. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[46] Jialong Zhang, Zhongshu Gu, Jiyong Jang, Hui Wu, Marc Ph Stoecklin, Heqing Huang, and Ian Molloy. 2018. Protecting intellectual property of deep neural networks with watermarking. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*. 159–172.

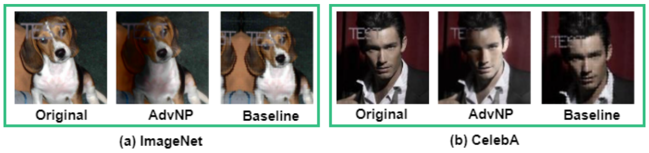
[47] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W Jacobs. 2019. Deep single-image portrait relighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7194–7202.

**Table 5: Effectiveness of applying data augmentation to defend AdvNP. The column Data augmentation indicates the watermark embedding enhanced after applying the data augmentation method. The column increase rate represents the intensity of improvement after applying the enhancement method.**

| | Benign Samples | | Verification Samples | | | |
|---|---|---|---|---|---|---|
| Type | B.S. | A.N. | V.S. | A.N. | Data augmentation | Increase rate |
| Pattern-based | 0.85 | 0.72 | 0.93 | 0.50 | **0.55** | **10%** |
| OOD-based | 0.84 | 0.75 | 0.95 | 0.10 | **0.86** | **760%** |

**Table 6: Effectiveness of employing multiple triggers to defend AdvNP. The column Multiple triggers denotes the watermark embedding enhanced by adopting multiple triggers.**

| | Benign Samples | | Verification Samples | | | |
|---|---|---|---|---|---|---|
| Ratio | B.S. | A.N. | V.S. | A.N. | Multiple triggers | Increase rate |
| 64:1 | 0.85 | 0.72 | 0.93 | 0.50 | **0.62** | **24.0%** |
| 128:1 | 0.84 | 0.73 | 0.93 | 0.41 | **0.56** | **36.6%** |



(a) ImageNet

(b) CelebA

**Figure 4: Visualization of images by applying AdvNP and baseline. The original indicates the images with a pattern-based trigger, the AdvNP represents the images preprocessed by relighting, and the Baseline denotes the image preprocessed by adopting a series of image transformations, like scaling.**

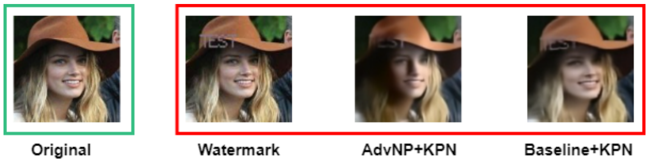


**Figure 5: Visualization of the images preprocessed by our AdvNP and the baseline with denoising method KPN.**

# TECHNICAL APPENDIX

## A OVERVIEW

In this supplementary material, we present the experimental results on watermark enhancement for defending our proposed AdvNP, visualization in comparison with the baseline, discussion, and societal impact. Specifically, we explore two potential defense methods by employing perturbation augmentation and leveraging multiple patches for enhancing our watermarks. Experimental results show that the two defense methods show the potential to enhance the watermark against AdvNP.

## B EXPERIMENTAL RESULTS

### B.1 Watermark Enhancement for Defending AdvNP

As discussed in the aforementioned sections, our method by employing the relighting perturbations could remove the embedded watermarks in an effective and robust manner. In this paper, we seek potential solutions to defend against such attacks and enhance the robustness of watermarking techniques. Inspired by the data augmentation which is widely applied for enhancing the generalization capabilities of the DNN model. Here, we propose the *perturbations augmentation* to defend the threat. Moreover, for the pattern-based watermarking which is the most popular and effective DNN watermarking scheme, a straightforward idea could be to blend multiple triggers into the sample to resist the corruption via naturalness-aware perturbations. Thus, we propose the *multiple triggers* to defend our proposed AdvNP against the pattern-based watermarking scheme.

**Proposed method via perturbation augmentation and multiple triggers**. For the perturbation augmentation defense, we generate large numbers of sample-label pairs $\mathcal{K}$ by injecting relighting perturbations with diverse magnitudes in the watermarking embedding stage. For the multiple triggers defense, the samples for watermark verification are blended with at least two same triggers. The experiments are conducted on the popular VGG16 with a facial dataset, CelebA.

**Experimental results**. Experimental results in Table 5 illustrate that the perturbation augmentation could enhance the robustness of the model against AdvNP, especially the OOD-based watermarking technique with a watermark verification success rate of more than 86% in compared with 10% without enhancement. The increase rates for the pattern-based and OOD-based are **10%** and **760%**, respectively. However, this requires the owner knows the type of adopted naturalness-ware perturbations which is not feasible most of the time. For the multiple trigger enhancement, we employ two mixtures of benign samples and verification samples and experimental results reveal that it can enhance the model against the AdvNP with an increase rate of more than **24%** and **36.6%**, respectively, however, the performance on predicting verification samples is only 60% and works for the pattern-based watermarking techniques merely. In summary, experimental results in Table 5 and Table 6 demonstrated that these two enhancement methods could improve the robustness in defending our AdvNP in some particular experimental settings. However, our proposed AdvNp still poses real threat to the community as the two aforementioned defense methods are not ideal due to its performance and generalized capabilities in tackling unseen watermarking techniques.

## C VISUALIZATION

Figure 4 visualizes the performance by employing our AdvNP and the baseline on ImageNet and CelebA. We can easily observe that our AdvNP corrupts the trigger in high intensity compared with the baseline. Figure 5 visualizes the image by employing the denoising method KPN.

## D DISCUSSION

Our proposed method achieved competitive results in terms of both attack success rate and functionality-preserving. However, our method also exhibits some limitations. First, our AdvNP injects adversarial relighting perturbations to corrupt the embedded triggers. However, the owner could design a specific denoising method for the lighting noises if the owner knows our adopted relighting perturbations. To address this issue, we can combine other naturalness-aware perturbations (*e.g.*, flare, exposure) by employing random selection since deploying a general denoising method is difficult in the real-world scenario. Our work is the first attempt the demonstrate the existence of watermark removal attack by employing relighting perturbations but not limited to this. It might be interesting to explore more naturalness-aware perturbations which could be adopted for disrupting watermarks. Secondly, the semantic trigger drawn from the same distribution of the benign input could evade our AdvNP as the utility of benign sample prediction could be disrupted simultaneously. However, we can apply the object segmentation techniques to infer the possible semantic trigger to achieve targeted relighting.

## E SOCIETAL IMPACT

In this work, we make an early attempt to investigate the vulnerability of DNN watermarking to sample preprocessing, which is a common phenomenon in the real-world and pose a real threat to existing watermarking schemes. We present very first watermark removal attack via preprocessing samples via injecting relighting perturbations without obtaining any knowledge of target model, training datasets, and type of watermarking schemes. Through comprehensive experiments, we have demonstrated that very successful attack can be well disguised in naturalness-aware perturbations, unveiling the vulnerabilities of existing trigger-based DNN watermarking techniques in data-poisoning embedding manner.

Considering that powerful DNN models are critical assets that attracts the illegally copied, distributed, and misused. However, the verification samples spread in the real-world scenarios where various degradation will introduce. The samples could be easily processed in a natural way to conduct a successful attack on DNN watermarking schemes. This work is the first attempt to identify and showcase that such an attack based on injecting relighting perturbations is not only feasible, but also leads to a high attack success rate without compromising the functionality of benign sample prediction. In a larger sense, this work can provide new thinking into how to better design the DNN watermarking pipeline in order to mitigate potential risk caused by the vulnerabilities discussed herein, especially for DNN models deployed in safety-critical applications, such as face recognition, self-driving, *etc.*.

Bad actors can potentially make use of this newly proposed watermark removal attack mode as a wheel to pose security risks on existing DNN watermarking schemes that are not yet well prepared for this new type of attack. We, as researchers, believe that our proposed watermark removal attack can accelerate the research and development of more robust DNN watermarking techniques and effective measures for defending against this real threat. Therefore, our work can serve as an asset and a stepping stone for the future-generation trustworthy design of DNN watermarking techniques.

In addition to the social impact discussed above, the proposed method can also influence various research directions. For example, our proposed AdvNP:

- hints new adversarial training or data augmentation techniques for training robust DNN models in protecting its copyright.
- hints developing new watermarking techniques to be resilience against this novel watermark removal attack.
- hints new directions in exploring the vulnerability of DNN watermarking schemes in suffering other preprocessing methods on a wide range tasks.