Watermark Vaccine: Adversarial Attacks to Prevent Watermark Removal

Xinwei Liu[1,2], Jian Liu [3], Yang Bai[4] , Jindong Gu[5], Tao Chen[3], Xiaojun Jia[1,2] *, Xiaochun Cao[1,6]

[1] SKLOIS, Institute of Information Engineering, CAS, Beijing, China
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[3] Ant Group, Beijing, China
[4] Tencent Security Zhuque Lab, Beijing, China
[5] University of Munich, Munich, Germany
[6] School of Cyber Science and Technology, Shenzhen Campus, Sun Yat-sen University, Shenzhen 518107, China
{liuxinwei, jiaxiaojun}@iie.ac.cn {rex.lj,boshan.ct}@antgroup.com
mavisbai@tencent.com jindong.gu@outlook.com
caoxiaochun@mail.sysu.edu.cn

Abstract. As a common security tool, visible watermarking has been widely applied to protect copyrights of digital images. However, recent works have shown that visible watermarks can be removed by DNNs without damaging their host images. Such watermark-removal techniques pose a great threat to the ownership of images. Inspired by the vulnerability of DNNs on adversarial perturbations, we propose a novel defence mechanism by adversarial machine learning for good. From the perspective of the adversary, blind watermark-removal networks can be posed as our target models; then we actually optimize an imperceptible adversarial perturbation on the host images to proactively attack against watermark-removal networks, dubbed ***Watermark Vaccine***. Specifically, two types of vaccines are proposed. Disrupting Watermark Vaccine (DWV) induces to ruin the host image along with watermark after passing through watermark-removal networks. In contrast, Inerasable Watermark Vaccine (IWV) works in another fashion of trying to keep the watermark not removed and still noticeable. Extensive experiments demonstrate the effectiveness of our DWV/IWV in preventing watermark removal, especially on various watermark removal networks. The Code is released in https://github.com/thinwayliu/Watermark-Vaccine.

Keywords: Visible Watermark Removal, Watermark Protection, Adversarial Attack

1 Introduction

With the rapid development of digital media and the increasing dependence of deep neural networks (DNNs) on enormous training data, copyright protection

* Corresponding Author

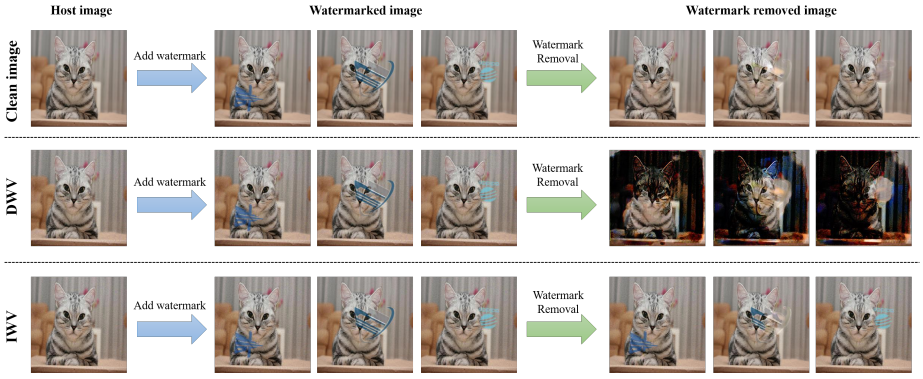


Fig. 1. The protective effects of our watermark vaccines on different watermark patterns or parameters. The current blind watermark-removal technique, such as WDNet[32], can effectively remove the watermarks (**top**). When the host images are equipped with Disrupting Watermark Vaccine (DWV), the watermark-removed images will be ruined (**middle**). However, when the host images are equipped with Inerasable Watermark Vaccine (IWV), the results can not be purified successfully as the host images (**bottom**).

attracts great attention especially for image data [45]. Visible watermarking thus becomes an essential technique [3]. It prevents illicit users from obtaining some critical information and using copyrighted high-quality images. As a result, it can reduce illegal theft and play a role in publicity and warning. Mintzer *et al.* [35] posed two characteristics for visible watermark, that is, it can be recognized by human eyes but will not significantly obscure the details.

However, visible watermark is in face of security issues as it can be effectively removed by some watermark-removal techniques [10,2,42,46,53]. Among these techniques, some require the location area of the watermark. Huang *et al.* [21] propose to remove the watermark using image inpainting. Park *et al.* [40] propose to formulate it into a feature matching problem. With the rapid development of deep learning, the community has proposed some blind visible watermark-removal DNNs, which can reconstruct watermarked images end-to-end without any information about watermarks [4,8,11,20,29,30,32]. These works usually adopt two-stage strategies. In the first stage, the networks aim to predict the watermark region. After that, in the second stage, they work on recovering the background of such a watermark region. Without strong assumptions, the blind watermark-removal network has become a mainstream method. In Figure 1 (a), we take WDNet [32] as an example to show its performance in both identifying and removing the watermark.

Due to these advanced watermark-removal technologies, traditional watermarking methods can no longer effectively protect the copyrights of picture owners. Recently, Khachaturov *et al.* [25] proposed to fool the inpainting-based removal networks to protect the watermark. However, this type of networks

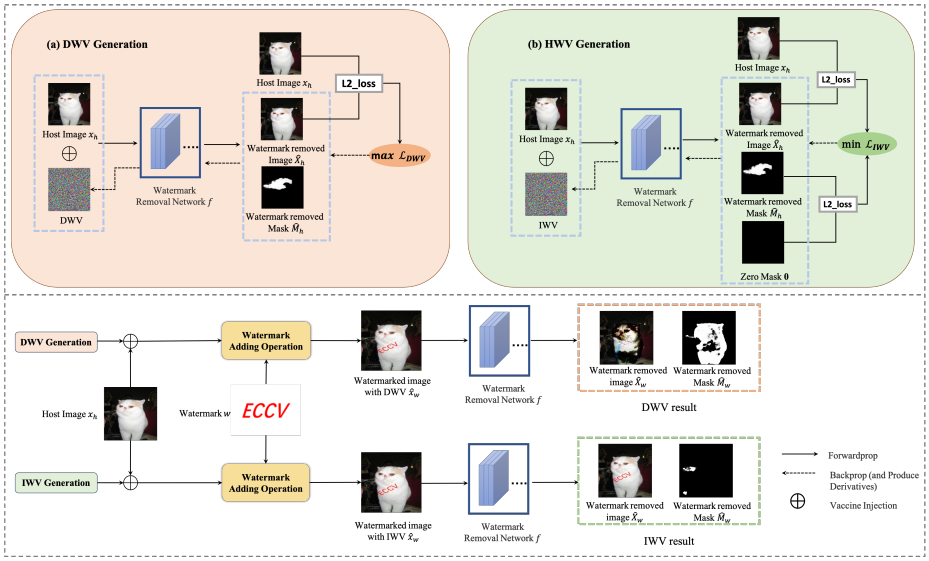


Fig. 2. The overview figure of the generation (on the first row) and application (on the second row) of our proposed watermark vaccine. We propose to maximize $\mathcal{L}_{\mathcal{DWV}}$ to generate DWV and minimize $\mathcal{L}_{\mathcal{IWV}}$ to generate IWV, as shown in (a) and (b), respectively. Then we first apply DWV/IWV on the host images to generate 'protected host images'. When watermarks are added on those protected host images, they are hard to be normally removed, tending to show either ruined or watermark-preserving results.

are demanding and not widely used. So we focus on preventing the blind watermark removal networks in this paper. Inspired by recent studies on adversary [1,5,49,17,56,22], which show that imperceptible adversarial perturbations can cause some incorrectly outputs for DNNs, 'adversarial for good' is thus a new protection method. To note that, generating a simple adversarial perturbation on the watermarked image could not work directly. Because in real-world scenarios, a watermark is always automatically generated at the last step by the system or website, and the protected image is required to finish uploading before it. More importantly, the watermark is not permanent for one host image. The watermark could be changed with some specific circumstances (such as Enterprise renaming, logo changes, year updates, etc.). So it can be costly to regenerate an adversarial perturbation for the same host image yet with a different watermark. Thus, a universal perturbation can be useful and efficient for watermark protection.

In this paper, we propose a watermark-agnostic perturbation against blind watermark-removal network, dubbed ***Watermark Vaccine***, which is injected on host images before adding watermark just like vaccination in reality. Our method is equivalent to a 'double insurance' for copyright protection: the visible watermark serves as a warning and annotating function, telling people not to infringe. The watermark vaccine ensures that the visible watermark won't be re-

moved by blind watermark-removal networks, which can effectively reduce illegal dissemination and other infringements. Then the vaccinated image can protect any watermark from being removed. Specifically, we propose two types of watermark vaccines according to their attack effects: Disrupting Watermark Vaccine (DWV) and Inerasable Watermark Vaccine (IWV). DWV aims to disrupt the watermark-removed image while IWV attempts to still keep the watermark through blind watermark-removal networks. The framework of our proposed watermark vaccine's generation and is shown in Fig. 2. In Fig. 1 (b) and (c), we can see that after injected with DWV or IWV, the watermark-removed images will either be ruined or the watermarks are not completely removed. In addition, the masks are disrupted for DWV or are induced elsewhere for IWV. Both results demonstrate the effectiveness of our proposed watermark vaccine in successfully preventing watermark removal.

Our key contributions are summarized as follows:

- We are the first to propose the watermark-agnostic perturbations for blind watermark-removal networks, dubbed *Watermark Vaccine*, to prevent the watermark removal from host images.
- We present two types of effective and powerful watermark vaccines (DWV and IWV), which aim to either disrupt the watermark-removed images or keep the watermarks uncleared respectively.
- We evaluate the effectiveness and universality of two vaccines. The results demonstrate that they generalize well on different watermark patterns, sizes, locations as well as transparencies. In addition, our watermark vaccine can also resist some common image processing operations.

2 Related Work

Visible Watermark Removal. Visible watermark-removal techniques are proposed to evaluate and improve the robustness of visible watermarks at the beginning. In the earlier works [21,40,41], the user's interaction is always required to remove watermark. Namely, they require the location of the watermark and recover that area. However, it can be not practical when processing massive images without location information. In [12] and [15], they assume that the same watermark is added to all host images. Nevertheless, this assumption is also too strong to apply in real scenarios.

With the development of deep learning, neural networks show a great power in computer vision tasks [18,27,57,16]. Several works try to apply neural networks to formulate an end-to-end problem, and there are two popular ways applied to solve it. One way is to directly formulate the watermark removal as an image-to-image translation task [4,29]; the other way is adopting a two-stage strategy to formulate the problem: the first step is to locate the by a mask, and the second step is to recover the background in the watermark area and train a network to solve both at the same time [11,20,32,30]. The latter method was found in experiments can be more effective in watermark removal, so we mainly focus on preventing the second type of networks in this paper.

Adversarial Attacks on Generative Models. Szegedy *et al.* [49] first found and proposed the adversarial examples. In [17], Goodfellow *et al.* proposed the Fast Gradient Sign Method (FGSM), which is a one-step gradient attack. After that, some stronger generation methods are proposed like I-FGSM [28], M-FGSM [13], and Projected Gradient Descent (PGD) [34]. In the black-box setting, adversarial examples can transfer across the different models [39,38,14,31,52] or can be generated by approximating gradients [6,51]. However, most of attack and defense works are about the classification problem[23,24].

Recent works have focused more on attacks on generative models. In [26] and [50], the authors firstly explore adversarial attacks against Variational Autoencoders (VAE) and VAE-GANS. Ruiz *et al.* [43,44] apply transferable adversarial attacks to disrupt facial manipulation systems. From then on some other works [55,7,47,54] adopt adversarial machine learning in translation-based deepfake models. These works show that adversarial attacks can defend against some malicious generative networks to protect users' privacy. Khachaturov *et al.* [25] find that adversarial perturbations can fool an inpainting system into generating a patch similar to random pictures. Although this work can be applied to protect watermarks, the inpainting-based watermarking method needs obtaining a mask in image, and is not practical for big watermarks. Therefore, attacking the inpainting-based watermarking method [25] cannot protect the watermark in natural scenes.

3 Methodology

3.1 Preliminary

The watermark vaccine generation can be formalized as an optimization problem with constraints. We assume a host image as x_h, and the invisible watermark vaccine δ is injected onto it before adding watermark, which is restricted in L_∞ norm bound ϵ. Thus, we get the vaccinated image $\hat{x}_h$ by

$$\begin{aligned} &\hat{x}_h = x_h + \delta \\ &\|\delta\|_\infty \leq \varepsilon. \end{aligned} \tag{1}$$

Then we select a watermark sample w from the watermark set W, where $w \in W$. The adding watermark operation can be assumed as g. And g requires some parameters θ to identify the location (p, q), the size (u, v) and the transparency α of the watermark w injected on the host image x_h. So a watermarked image $\hat{x}_w$ with vaccine can be defined as follows,

$$\hat{x}_w = g\left(\hat{x}_h, \omega, \theta\right), \tag{2}$$

where $\theta = (p, q, u, v, \alpha)$. In practice, the parameters p, q, n, m and α are always random. Similarly, the watermarked image without vaccine x_w can also be obtained in the same way. Then we assume the blind watermark-removal network

as f, and we can get the watermark removed images and masks of x_w and $\hat{x}_w$ respectively through the network, which is defined as

$$\begin{aligned} X_w, M_w &= f\left(x_w\right), \\ \hat{X}_w, \hat{M}_w &= f\left(\hat{x}_w\right). \end{aligned} \tag{3}$$

Here, we denote the measurement of watermark removal effect as $Q(\cdot)$, and the goal of the vaccine is to degrade the removal effect of watermark removed images by minimizing $Q(f(g(x_h+\delta, w, \theta)))$. In addition, we desire our watermark vaccine to be universal for different watermark patterns, positions, sizes, transparencies, thus the expected $Q(f(g(x_h+\delta, w, \theta)))$ over different watermark w and adding parameters θ is required to make vaccine watermark-agnostic. Our watermark vaccine generation can be formulated as follows,

$$\min_{\delta} \quad \mathbb{E}_{w \sim W}\, \mathbb{E}_{\theta \sim \Theta}\left[Q(f(g(x_h+\delta, w, \theta)))\right]. \tag{4}$$

Unfortunately, there are two challenges in solving the above optimization problem. First, the effect of watermark removal $Q(\cdot)$ can be customized in a variety of different ways, but it is required to be differentiable during optimization. In addition, the two expectation over W and Θ is hard to optimize by considering the loss of all combinations simultaneously. Although we can refer to 'universal adversarial perturbation' in [36,19,48,37], it is still time-consuming and difficult to obtain the optimal vaccine. To address these issues, we further propose two types of watermark vaccine in the following.

3.2 Disrupting Watermark Vaccine (DWV)

One way to protect the watermark is to disrupt the watermark-removed images, which means that as long as the watermarked image with the watermark vaccine passes through the watermark-removal networks, the output image will be ruined and could never be used. Thus, we call this vaccine as **Disrupting Watermark Vaccine (DWV)**. Next, in order to avoid the two expectations in Equation 4, we decide to generate the watermark-agnostic vaccine on the host images instead of watermarked images. Therefore, we inject vaccine δ on the clean host image x_h and get the vaccinated image $\hat{x}_h$. After passing through the network f, we get the watermark removed image and watermark removed mask, which is denoted as $\hat{X}_h$ and $\hat{M}_h$, because they are generated on clean host images without a watermark. Such operation can be formulated as,

$$\hat{X}_h, \hat{M}_h = f(\hat{x}_h), \tag{5}$$

Then, we define the $\mathcal{L}_{\mathcal{DWV}}$ to measure the distance between the watermark removed image $\hat{X}_h$ and the clean host image x_h,

$$\mathcal{L}_{\mathcal{DWV}}\left(x_h, \delta\right) = \left\| \hat{X}_h - x_h \right\|^2, \tag{6}$$

where the image distance adopt the mean-square error to measure.

The objective here is to maximize $\mathcal{L}_{\mathcal{DWV}}$ such that the watermark removed image is significantly different from the host image. Thus, the watermark removed image of the host image is severely ruined by watermark-removal networks. Naturally, whatever watermark is added onto the image, the benign areas (the areas without the watermark) will be destroyed as well. As a result, the watermark removed images of watermarked images sufficiently deteriorate such that it has to be discarded or such that the modification is perceptually evident The problem can be formally expressed as follows,

$$\begin{array}{ll} \max_{\delta} & \mathcal{L}_{\mathcal{DWV}}\left(x_h, \delta\right) \\ \text{s.t.} & \hat{x}_h = x_h + \delta \\ & \|\delta\|_{\infty} \leqslant \varepsilon, \end{array} \tag{7}$$

To be consistent with the previous requirements, we restrict the perturbations δ in L_∞ norm bound ϵ. We use projected gradient descent (PGD) [34] to solve the optimization problem in Equation 8, and we can get the optimal δ according to the following iterative formula,

$$\delta^{t+1} = \text{Proj}\left(\delta^t + \alpha \,\text{sign}\left(\nabla_{\delta^t} \mathcal{L}_{\mathcal{DWV}}\left(x_h, \delta^t\right)\right)\right), \tag{8}$$

where $\nabla_{\delta^t} \mathcal{L}_{\mathcal{DWV}}\left(x_h, \delta^t\right)$ is the gradient of the disrupting loss w.r.t δ^t. α is the step size, Proj() denotes project the δ^t within the norm bound $(-\epsilon, \epsilon)$ and project the $x + \delta^t$ within the valid space $(0, 1)$. In Fig. 2 (a), we can see the framework of DWV generation, and at the bottom shows the inference of DWV.

3.3 Inerasable Watermark Vaccine (IWV)

Contrasted with DWV to protect the watermark by ruining the watermark removed images, another solution is to prevent the watermark from being identified and removed. To this end, as an alternative to DWV, we propose another vaccine in this section, **Inerasable Watermark Vaccine (IWV)**. It aims to make the watermarks hard to be detected and removed. As a result, the watermark patterns can not be erased completely on the watermark removed images. Inspired by the Equation (6), we design the $\mathcal{L}_{\mathcal{IWV}}$ as follow:

$$\mathcal{L}_{\mathcal{IWV}}\left(x_h, \delta\right) = \frac{1}{2}\left(\beta \left\|\hat{X}_h - x_h\right\|^2 + \|\hat{M}_h - \mathbf{0}\|^2\right), \tag{9}$$

where $\hat{X}_h$ and $\hat{M}_h$ is the output of the blind watermark-removal network f, x_h is the host image, $\mathbf{0}$ is a zero matrix, which is the same size as the predicted mask. There are two distance terms in the loss $\mathcal{L}_{\mathcal{IWV}}$: image term and mask term. The image term is equal to the Equation (6), and the mask term measures the distance between the predicted mask $\hat{M}_h$ and a zero matrix $\mathbf{0}$. The β is the hyperparameter to balance two loss terms.

Ideally, the predicted image $\hat{X}_h$ should be almost the same as the input image x_h, and the predicted mask $\hat{M}_h$ should be almost black, which means there is no watermark can be detected on $\hat{x}_h$ and the $\mathcal{L}_{\mathcal{IWV}}$ should be very close to 0.

Algorithm 1: Watermark Vaccine Generation

Input: host image x_h, blind watermark-removal network f, iteration T, step size α, perturbation bound ϵ

Output: Host image with watermark vaccine $\hat{x}_h$

$\delta \leftarrow 0, \hat{x}_h \leftarrow x_h + \delta$;
for *i = 1 to T* **do**
  if *vaccine is 'DWV'* **then**
    using Equation (6) to calculate the $\mathcal{L}_{\mathcal{DWV}}$;
    $\delta \leftarrow \delta + \alpha \operatorname{sign}(\nabla_\delta \mathcal{L}_{\mathcal{DWV}}(x_h, \delta))$;
  else
    using Equation (9) to calculate the $\mathcal{L}_{\mathcal{IWV}}$;
    $\delta \leftarrow \delta - \alpha \operatorname{sign}(\nabla_\delta \mathcal{L}_{\mathcal{IWV}}(x_h, \delta))$;
  end
  $\hat{x}_h \leftarrow x_h + \operatorname{clip}(\delta, -\epsilon, \epsilon)$;
end
$\hat{x}_h \leftarrow \operatorname{clip}(\hat{x}_h, 0, 1)$;

However, in reality, these are not 0 and show a large loss actually, as shown in Fig. 2(b). Based on this situation, we decide to minimize the $\mathcal{L}_{\mathcal{IWV}}$ to generate the vaccine that can make the output of the watermark-removal network close to the ideal one. This seems to be a well-intentioned fix for performance, but it is actually superfluous and adversarial. In the test stage, whatever watermark is added to the host image, the IWV will suppress the removal network to recognize it, and the outputs still preserve the watermarks and they tend to be the same as watermarked image inputs. Thus, the IWV generation can be formulated as,

$$\begin{array}{ll} \min_\delta & \mathcal{L}_{\mathcal{IWV}}(x_h, \delta) \\ \text{s.t.} & \hat{x}_h = x_h + \delta \\ & \|\delta\|_\infty \leqslant \varepsilon. \end{array} \tag{10}$$

We also restrict the perturbations δ in L_∞ norm bound ϵ, and solve it by projected gradient descent (PGD) [34] again as follows,

$$\delta^{t+1} = \operatorname{Proj}\left(\delta^t - \alpha \operatorname{sign}\left(\nabla_{\delta^t} \mathcal{L}_{\mathcal{IWV}}(x_h, \delta)\right)\right). \tag{11}$$

In Fig. 2 (b), we can see the framework of IWV generation and the inference of IWV. The pseudocode of our algorithm to generate watermark vaccine including DWV and IWV is shown in Algorithm 1. We use projected gradient descent (PGD) [34] to solve the problem (7) or (10), and then we generate the DWV/IWV. During the inference stage, we inject the DWV and IWV onto host image and add the watermark on it. After that, we evaluate their adversarial results through the watermark-removal networks f and expect to get the damaged or watermark-preserved image. In addition, we theoretically analyze why our vaccines work effectively and give the lower or upper bound of the watermark protection for DWV and IWV in Sec.1 of the Supplementary Material.

4 Experiments

4.1 Experimental Setups

Datasets. We use the CLWD (Colored Large-scale Watermark Dataset) [32] in our experiments, which contains three parts: watermark-free images, watermarks and watermarked images. We first pretrain the watermark-removal networks using watermarked images in the train set of CLWD. Then in the attack stage, we use the watermark-free images as host images to generate watermark vaccines, and then add the watermarks with generated watermark vaccines. The details about the dataset can be checked in Sec.2 of the Supplementary Material.
Models Architectures. We choose three advanced blind watermark-removal networks: BVMR [20], SplitNet [11], and WDNet [32]. We train them on the watermarked images of CLWD and save the best checkpoint parameters.
Evaluation Metrics. Following the previous work [32,11,30], Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM), Root-Mean-Square distance (RMSE), and weighted Root-Mean-Square distance (RMSE_w) are adopted as our evaluation metrics. The difference between RMSE and RMSE_w is that RMSE_w only focuses on the watermarked area. For DWV, we specify these metrics as PSNR^h, SSIM^h, RMSE^h, RMSE_w^h compared with host images for a better illustration. The lower $\text{PSNR}^h/\text{SSIM}^h$ or the higher $\text{RMSE}^h/\text{RMSE}_w^h$ mean the worse results of watermark-removal networks thus the better protection performance of proposed DWV. For IWV, we specify the metrics as PSNR^w, SSIM^w, RMSE^w, RMSE_w^w which are compared with watermarked images. Different from DWV, the higher $\text{PSNR}^w/\text{SSIM}^w$ or the lower $\text{RMSE}^w/\text{RMSE}_w^w$ mean that the more excellent performance on keeping the watermarks on thus the better protection performance of proposed IWV.
Attack Parameters. During attack, we empirically set the L_∞ norm bound ϵ as $8/255$, which is imperceptible by human eyes. We set the step size α as $2/255$, and iteration T as 50. We set the hyperparameter β in Eq. 9 for IWV to 2 initially. We also further discuss the sensitivity of these hyperparameters in the Sec.5 of Supplemental Material.

4.2 Effectiveness of Watermark Vaccine

We evaluate the effectiveness of DWV and IWV on 10,000 random host-watermark image combinations. We test three models with the same watermark parameters θ on the same dataset. We compare the clean input and the input with random noise at the same time, which is also restricted in L_∞ norm bound $(-\epsilon, \epsilon)$.

In Fig. 3, we show the qualitative visualization of different networks and their corresponding results. It shows that no matter which network it is, the watermark removed images can be ruined if the host images are injected with DWV, although the watermarks can be successfully removed. For IWV, there are still noticeable all or part of watermarks on the watermark removed images, and other parts of the images are not damaged. On the contrary, the inputs with random noise present no protective effect on the watermark removed results for

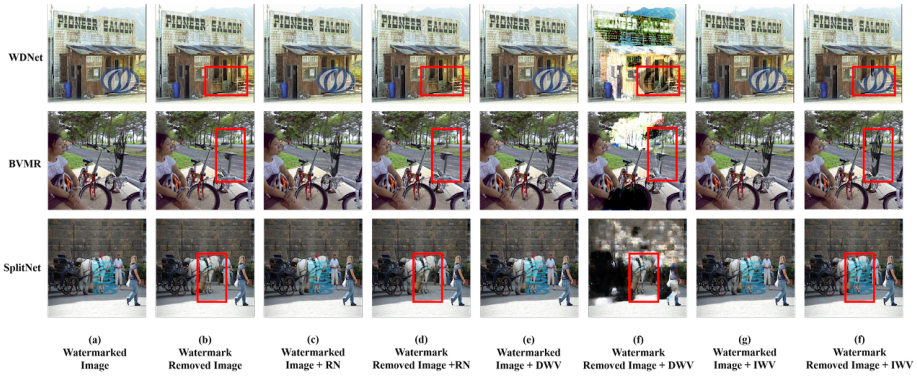


Fig. 3. Qualitative comparison of DWV and IWV. In each row, we show the watermark removal effects on the images without any vaccine, the images with random noises, the images with DWV, and the images with IWV under the same network.

any watermark-removal network. We show more visualization results in Sec.3 of the Supplementary Material.

Tab. 1 demonstrates the quantitative results of watermark vaccines on different watermark-removal networks. In Tab. 1(a), we can find that random noise could not disrupt the watermark removed image, while the DWV can significantly degrade the quality of watermarked removed images with the lower $\text{PSNR}^h/\text{SSIM}^h$ and the higher $\text{RMSE}^h/\text{RMSE}_w^h$ than others. On the other hand, by observing Tab. 1(b), the watermark removed images with IWV have a better similarity with watermarked input with a little higher PSNR/SSIM. It is noticeable that the RMSE_w^w for IWV is much lower than others, which is to evaluate whether the watermark part is well preserved. Moreover, the above phenomena in different watermark-removal networks tend to be the same, and the quantitative results are consistent with the qualitative visualization.

Although DWV can ruin the watermark removed images, the watermark patterns can also be removed. On the contrary, the watermarks with IWV could be still noticeable on the watermark removed images by human eyes. Therefore, which type of vaccine to choose depends on the need for protection.

4.3 Universality of Watermark Vaccine

As mentioned in Sec. 3.2 and 3.3, the watermark vaccine we proposed can adapt to different watermarks and parameters and has a good universality. To illustrate this, we investigate the different watermark patterns, sizes, locations and transparency of watermarks for 1,000 host images. The WDNet [32] is selected as the model for an example. Other models can be found in the Supplementary Material. We test every host image with ten random-selected watermark patterns and fix other parameters. Similarly, we test every host image with ten random locations with a fixed watermark and other fixed parameters for the location.

Table 1. Impact of the two vaccines on WDNet, BVMR, and SplitNet on the same dataset and with the same parameters θ. The perturbations and random noises are restricted in L_∞ norm bound 8/255. **Clean** denotes the watermarked image with no vaccines, and **RN** denotes the watermarked images with random noise. For DWV, the lower PSNR^h/SSIM^h or the higher RMSE^h/RMSE^h_w the better. For IWV, the higher PSNR^w/SSIM^w or the lower RMSE^w/RMSE^w_w the better. The best-protection results are denoted in boldface.

	WDNet[32]				BVMR[20]				SplitNet[11]			
Metrics	PSNR^h	SSIM^h	RMSE^h	RMSE^h_w	PSNR^h	SSIM^h	RMSE^h	RMSE^h_w	PSNR^h	SSIM^h	RMSE^h	RMSE^h_w
Clean	38.62	0.9946	3.09	16.25	41.96	0.9955	2.09	23.86	42.32	0.9939	2.12	21.86
RN	38.19	0.9938	3.23	17.06	42.48	0.9957	1.98	24.13	42.73	0.9943	2.07	21.33
DWV(Ours)	**29.68**	**0.6360**	**8.47**	**41.36**	**29.43**	**0.6462**	**8.68**	**26.85**	**34.12**	**0.8951**	**5.18**	**67.68**

(a) The effect of DWV on different watermark-removal networks.

	WDNet[32]				BVMR[20]				SplitNet[11]			
Metrics	PSNR^w	SSIM^w	RMSE^w	RMSE^w_w	PSNR^w	SSIM^w	RMSE^w	RMSE^w_w	PSNR^w	SSIM^w	RMSE^w	RMSE^w_w
Clean	37.76	0.9788	3.42	52.77	41.88	0.9893	2.13	42.68	40.91	0.9788	2.53	49.67
RN	37.53	0.9755	3.50	52.95	42.59	0.9917	2.00	42.73	41.59	0.9795	2.41	49.29
IWV(Ours)	**45.16**	**0.9831**	**2.24**	**28.00**	**43.31**	**0.9926**	**1.86**	**37.42**	**42.79**	**0.9834**	**2.23**	**35.00**

(b) The effect of IWV on different watermark-removal networks.

Table 2. Mean and standard deviation over evaluation metrics of DWV and IWV for random watermark patterns and location parameters.

	Watermark		Location	
Metrics	Clean	DWV	Clean	DWV
PSNR^h	39.12±0.02	29.40±0.03	40.82±0.04	28.95±0.01
SSIM^h	0.9957±0.0000	0.6021±0.0028	0.9974±0.0001	0.5288±0.0020
RMSE^h	2.85±0.01	8.74±0.02	2.37±0.01	9.15±0.01
RMSE^h_w	17.15±0.18	42.88±0.77	16.67±0.11	52.02±0.68

(a) DWV

	Watermark		Location	
Metrics	Clean	IWV	Clean	IWV
PSNR^w	38.42±0.02	47.35±0.21	40.37±0.03	52.30±0.30
SSIM^w	0.9874±0.002	0.9938±0.0003	0.9956±0.0001	0.9981±0.0001
RMSE^w	3.08±0.01	1.63±0.03	2.48±0.01	1.08±0.03
RMSE^w_w	54.25±0.54	20.67±0.39	48.28±0.38	10.39±0.55

(b) IWV.

To show their universality, we calculate the mean and variance of these results from different settings. Concerning about the watermark size and transparency, we select six sizes: 60×60, 70×70, 80×80, 90×90, 100×100 and six transparency parameters: $\alpha = 0.45, 0.50, 0.55, 0.60, 0.65$. We fix the $\alpha = 0.55$ if the size varies, and fix size $= 80 \times 80$, if the transparency varies. Finally, we calculate their evaluation metrics respectively. The above quantitative results are shown in Tab. 2 and 3.

In Tab. 2, compared to the means of the clean input, the means of DWV and IWV show that our watermark vaccines are still effective, and the minor variances of vaccines prove that our vaccines can be universal among different watermark patterns or locations. Tab. 3 shows that the metrics in each row are not much different, although they are under different sizes and transparencies of the watermark. Hence, the above results indicate that DWV and IWV have good universality, regardless of the watermark pattern, position, size and transparency. The visualization of universality can be seen in the Sec.6 of Supplementary Material.

Table 3. The evaluation metrics for the Clean/DWV/IWV under different size and transparency of watermarks. Each row shows the results under different watermark sizes or different transparencies. The best-attacking results are denoted in boldface.

Metrics	$PSNR^h$		$SSIM^h$		$RMSE^h$		$RMSE^h_w$		$PSNR^w$		$SSIM^w$		$RMSE^w$		$RMSE^w_w$	
Input	Clean	DWV	Clean	DWV	Clean	DWV	Clean	DWV	Clean	IWV	Clean	IWV	Clean	IWV	Clean	IWV
Size=60	39.91	**29.16**	0.9967	**0.5610**	2.62	**8.95**	18.02	**48.87**	39.36	**50.28**	0.9927	**0.9968**	2.76	**1.31**	51.55	**13.70**
Size=70	39.53	29.25	0.9962	0.5784	2.72	8.86	17.73	45.59	38.87	50.04	0.9901	0.9958	2.92	1.35	53.31	16.14
Size=80	39.14	29.25	0.9957	0.5963	2.84	8.78	17.03	41.13	38.39	47.52	0.9868	0.9936	3.09	1.63	55.16	20.71
Size=90	38.67	29.54	0.9950	0.6261	3.00	8.58	17.02	38.11	37.82	45.69	0.9833	0.9911	3.29	1.90	56.35	26.03
Size=100	38.25	29.67	0.9945	0.6455	3.15	8.47	16.81	37.32	37.32	43.06	0.9792	0.9864	3.49	2.34	57.29	32.87
α=0.45	39.24	**29.31**	0.9961	**0.5984**	2.81	**8.80**	15.51	**49.44**	38.35	**48.28**	0.9896	**0.9948**	3.10	**1.52**	45.92	**18.00**
α=0.50	39.19	29.43	0.9959	0.6129	2.83	8.70	16.28	44.32	38.36	46.27	0.9883	0.9940	3.10	1.73	50.69	20.67
α=0.55	39.14	29.25	0.9957	0.5963	2.84	8.78	17.03	41.13	38.39	47.52	0.9868	0.9936	3.09	1.63	55.16	20.71
α=0.60	39.08	29.37	0.9955	0.6004	2.86	8.75	17.79	39.26	38.34	47.70	0.9855	0.9934	3.10	1.59	59.44	22.14
α=0.65	38.99	29.32	0.9952	0.6048	2.89	8.79	18.53	39.31	38.27	47.19	0.9841	0.9934	3.13	1.54	62.54	21.40

Table 4. Vaccines Transferability. The columns correspond to the target model, while the rows correspond to the source model. For brevity, we show the $RMSE^h$ for DWV and the $RMSE^w_w$ for IWV.

Target Model		WDNet	BVMR	SplitNet
Clean		3.14	2.78	2.25
RN		3.29	2.70	2.22
Source Model	WDNet	**8.30**	2.80	2.34
	BVMR	3.35	**8.61**	2.47
	SplitNet	3.23	2.79	**5.17**

(a) $RMSE^h$ of DWV

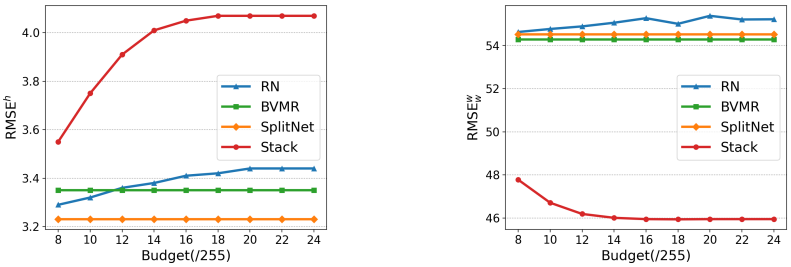
Target Model		WDNet	BVMR	SplitNet
Clean		54.47	43.84	51.10
RN		54.63	43.85	51.15
Source Model	WDNet	**30.14**	43.80	50.80
	BVMR	54.38	**40.13**	50.93
	SplitNet	54.52	43.50	**37.87**

(b) $RMSE^w_w$ of IWV

Interestingly, according to Tab. 3, we find that when the size of the watermark becomes larger, the performance of the DWV and IWV has dropped. Moreover, if the transparency parameter α of the watermark becomes larger, the effect of the protection will be worsen, especially for IWV. This phenomenon is consistent with our analysis in Sec.1 of Supplementary Material, that the watermarking variation $\|w\|$ is one of the factors that determine the effectiveness of watermark protection. The better performance of the watermark vaccine depends on a smaller variation of $\|w\|$. Therefore, it can be a challenge for the copyright owners to choose a suitable size and transparency for the watermark, where a larger and low-transparency watermark is convenient for copyright identification. In comparison, a smaller and high-transparency watermark is more beneficial to protect the watermark vaccine.

4.4 Transferability of Watermark Vaccine

First, we explore the transferability of our vaccines across different watermark removal networks, and Tab. 4 shows the results. We find that the vaccines show limited transferalibity across different watermark removal methods, which is the common problem of the adversarial examples on generative models. The reason may be related to the different procedures and network structures of removal



(a) RMSE^h of DWV (b) RMSE^w_w of IWV

Fig. 4. Testing stacked vaccines on WDNet [32]. For a comparison, we first add a random perturbation baseline and also test the vaccines generated by BVMR and SplitNet respectively. The stacked ones perform clearly the best.

networks. e.g., BVMR [20] is a one-stage method of predicting watermark removed images, while SplitNet [11] and WDNet [32] contain detection, removal and refinement steps. In addition, SplitNet [11] adopts stacked attention-guided ResUNets, but other models do not.

In real world, humans can be protected against different kinds of viruses by inoculating different vaccines. Inspired by this, we study the stacked vaccine assembled by three networks. We test the stacked vaccine on different watermark removal networks (see the partial test results on WDNet in Fig. 4). Compared to the random perturbation baseline and the vaccines generated by other source models, the stacked vaccines perform better under various perturbation budgets. In future work, we will explore how to improve the transferability across different watermark removal networks/frameworks.

4.5 Resistance to Image Processing Operations

In this section, we explore whether our vaccine can resist the common image processing operation. We select two common transformations: JPEG compression [33] and Gaussian blur [9], and take the WDNet as an example for the model. We average the results of 1,000 watermarked images and plot them as Fig. 5. For brevity, we only show the RMSE^h of DWV and the RMSE^w_w of IWV, and other metrics can be found in Sec.7 of the Supplementary Material.

In Fig. 5 (a), as the degree of JPEG compression ratio increases, it shows that the RMSE^h for DWV is declined and higher than random noise at first. Then it gradually rises and approaches the variation of random noise finally. It is possibly because the performance degradation of the watermark vaccine is stronger than the image deterioration at the early stage, and when the degradation is strong enough, the result of DWV is similar to random noise. Regarding the IWV, we can find that the RMSE^w_w of IWV has a sharp rise when the compression ratio increases, then it flattens out. It is worth noting that our watermark vaccines still have effects if the compression ratio is less than 80. The phenomenon in Gaussian blur is quite the same as that in JPEG compression in Fig. 5(b), and

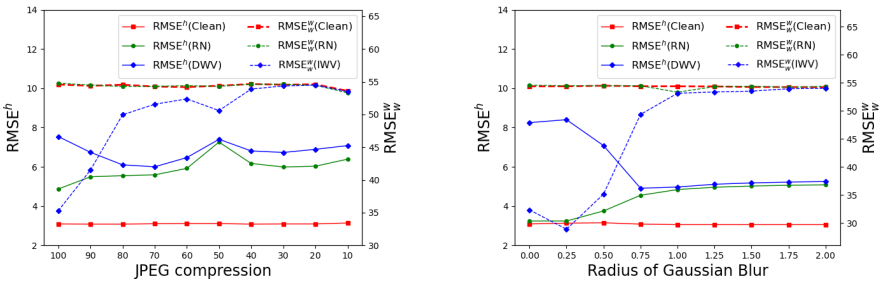


Fig. 5. Effect of two image-based transformation operations (JPEG Compression, Blur) on watermark vaccine. The solid lines show the change of RMSE^h, while the dashed lines show the RMSE_w^w change.

our watermark vaccines can resist the blur operation if the radius of Gaussian blur is less than 0.75. Besides the two image processing operations described above, we also consider some other operations that may affect our watermark vaccines, which will be present in the supplementary material. In conclusion, although some image-based transformation operations could reduce the effect of the watermark vaccine if their degradation is too substantial, they could also result in a lower quality of the image. Therefore, to some degree, our watermark vaccine can effectively resist some image processing operations.

5 Conclusions

Watermarking is an important and effective tool to protect copyright yet in face of the watermark removal threat. In this paper, we develop an idea of a watermark vaccine to protect watermarks. Our watermark vaccine is obtained by optimizing adversarial perturbations to attack the blind watermark removal network. Specifically, we propose two types of vaccine, dubbed disrupting watermark vaccine (DWV) and inerasable watermark vaccine (IWV). When malicious removal is presented, DWV will bring catastrophic damage to the host image, while IWV will keep the watermarks still clearly noticeable to human eyes. Both theoretical analysis and empirical experiments show that our vaccines is universal to different watermark patterns, sizes, locations, and transparencies, and they can also resist typical image transformation operations to a certain extent. This work makes the first exploration to protect watermarks from malicious removal. There is still space to improve our approach, e.g. by improving the transferability of watermark vaccines across target models. We leave further explorations in future work.

Acknowledgement
Supported by the National Key R&D Program of China under (Grant 2019YFB 1406500), Sponsored by Ant Group Security and Risk Management Fund.

References

1. Akhtar, N., Mian, A.: Threat of adversarial attacks on deep learning in computer vision: A survey. IEEE Access (2018)
2. Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: SIGGRAPH (2000)
3. Braudaway, G.W.: Protecting publicly-available images with an invisible image watermark. In: ICIP (1997)
4. Cao, Z., Niu, S., Zhang, J., Wang, X.: Generative adversarial networks model for visible watermark removal. IET Image Processing (2019)
5. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: IEEE Symposium on Security and Privacy (2017)
6. Chen, P.Y., Zhang, H., Sharma, Y., Yi, J., Hsieh, C.J.: Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In: Proceedings of the 10th ACM workshop on artificial intelligence and security (2017)
7. Chen, Z., Xie, L., Pang, S., He, Y., Zhang, B.: Magdr: Mask-guided detection and reconstruction for defending deepfakes. In: CVPR (2021)
8. Cheng, D., Li, X., Li, W.H., Lu, C., Li, F., Zhao, H., Zheng, W.S.: Large-scale visible watermark detection and removal with deep convolutional networks. In: PRCV (2018)
9. Cohen, J., Rosenfeld, E., Kolter, Z.: Certified adversarial robustness via randomized smoothing. In: ICML (2019)
10. Cox, I., Miller, M., Bloom, J., Fridrich, J., Kalker, T.: Digital watermarking and steganography. Morgan kaufmann (2007)
11. Cun, X., Pun, C.M.: Split then refine: Stacked attention-guided resunets for blind single image visible watermark removal. In: AAAI (2021)
12. Dekel, T., Rubinstein, M., Liu, C., Freeman, W.T.: On the effectiveness of visible watermarks. In: CVPR (2017)
13. Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., Li, J.: Boosting adversarial attacks with momentum. In: CVPR (2018)
14. Dong, Y., Pang, T., Su, H., Zhu, J.: Evading defenses to transferable adversarial examples by translation-invariant attacks. In: CVPR (2019)
15. Gandelsman, Y., Shocher, A., Irani, M.: " double-dip": Unsupervised image decomposition via coupled deep-image-priors. In: CVPR (2019)
16. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NeurIPS (2014)
17. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: ICLR (2015)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
19. Hendrik Metzen, J., Chaithanya Kumar, M., Brox, T., Fischer, V.: Universal adversarial perturbations against semantic image segmentation. In: ICCV (2017)
20. Hertz, A., Fogel, S., Hanocka, R., Giryes, R., Cohen-Or, D.: Blind visual motif removal from a single image. In: CVPR (2019)
21. Huang, C.H., Wu, J.L.: Attacking visible watermarking schemes. TMM (2004)
22. Jia, X., Wei, X., Cao, X., Han, X.: Adv-watermark: A novel watermark perturbation for adversarial examples. In: ACMMM (2020)
23. Jia, X., Zhang, Y., Wu, B., Ma, K., Wang, J., Cao, X.: Las-at: Adversarial training with learnable attack strategy. In: CVPR (2022)

24. Jia, X., Zhang, Y., Wu, B., Wang, J., Cao, X.: Boosting fast adversarial training with learnable adversarial initialization. TIP (2022)
25. Khachaturov, D., Shumailov, I., Zhao, Y., Papernot, N., Anderson, R.: Markpainting: Adversarial machine learning meets inpainting. ICML (2021)
26. Kos, J., Fischer, I., Song, D.: Adversarial examples for generative models. In: IEEE Symposium on Security and Privacy Workshops (2018)
27. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. NeurIPS (2012)
28. Kurakin, A., Goodfellow, I., Bengio, S., et al.: Adversarial examples in the physical world. In: ICLR Workshop (2017)
29. Li, X., Lu, C., Cheng, D., Li, W.H., Cao, M., Liu, B., Ma, J., Zheng, W.S.: Towards photo-realistic visible watermark removal with conditional generative adversarial networks. In: ICIG (2019)
30. Liang, J., Niu, L., Guo, F., Long, T., Zhang, L.: Visible watermark removal via self-calibrated localization and background refinement. In: ACM MM (2021)
31. Lin, J., Song, C., He, K., Wang, L., Hopcroft, J.E.: Nesterov accelerated gradient and scale invariance for adversarial attacks. ICLR (2020)
32. Liu, Y., Zhu, Z., Bai, X.: Wdnet: Watermark-decomposition network for visible watermark removal. In: WACV (2021)
33. Liu, Z., Liu, Q., Liu, T., Xu, N., Lin, X., Wang, Y., Wen, W.: Feature distillation: Dnn-oriented jpeg compression against adversarial examples. In: CVPR (2019)
34. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: ICLR Poster (2018)
35. Mintzer, F., Braudaway, G.W., Yeung, M.M.: Effective and ineffective digital watermarks. In: ICIP (1997)
36. Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. In: CVPR (2017)
37. Mopuri, K.R., Uppala, P.K., Babu, R.V.: Ask, acquire, and attack: Data-free uap generation using class impressions. In: ECCV (2018)
38. Papernot, N., McDaniel, P., Goodfellow, I.: Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. arXiv preprint arXiv:1605.07277 (2016)
39. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B., Swami, A.: Practical black-box attacks against machine learning. In: AsiaCCS (2017)
40. Park, J., Tai, Y.W., Kweon, I.S.: Identigram/watermark removal using cross-channel correlation. In: CVPR (2012)
41. Pei, S.C., Zeng, Y.C.: A novel image recovery algorithm for visible watermarked images. IEEE Transactions on information forensics and security (2006)
42. Qin, C., He, Z., Yao, H., Cao, F., Gao, L.: Visible watermark removal scheme based on reversible data hiding and image inpainting. Signal Processing: Image Communication (2018)
43. Ruiz, N., Bargal, S.A., Sclaroff, S.: Disrupting deepfakes: Adversarial attacks against conditional image translation networks and facial manipulation systems. In: ECCV Workshops (2020)
44. Ruiz, N., Bargal, S.A., Sclaroff, S.: Protecting against image translation deepfakes by leaking universal perturbations from black-box neural networks. arXiv preprint arXiv:2006.06493 (2020)
45. Samuel, S., Penzhorn, W.: Digital watermarking for copyright protection. In: IEEE Commun. Mag. (2004)

46. Santoyo-Garcia, H., Fragoso-Navarro, E., Reyes-Reyes, R., Sanchez-Perez, G., Nakano-Miyatake, M., Perez-Meana, H.: An automatic visible watermark detection method using total variation. In: IWBF (2017)
47. Segalis, E., Galili, E.: Ogan: Disrupting deepfakes with an adversarial attack that survives training. arXiv e-prints (2020)
48. Shafahi, A., Najibi, M., Xu, Z., Dickerson, J., Davis, L.S., Goldstein, T.: Universal adversarial training. In: AAAI (2020)
49. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. In: ICLR (2014)
50. Tabacof, P., Tavares, J., Valle, E.: Adversarial images for variational autoencoders. arXiv preprint arXiv:1612.00155 (2016)
51. Uesato, J., O'donoghue, B., Kohli, P., Oord, A.: Adversarial risk and the dangers of evaluating against weak attacks. In: ICML (2018)
52. Xie, C., Zhang, Z., Zhou, Y., Bai, S., Wang, J., Ren, Z., Yuille, A.L.: Improving transferability of adversarial examples with input diversity. In: CVPR (2019)
53. Xu, C., Lu, Y., Zhou, Y.: An automatic visible watermark removal technique using image inpainting algorithms. In: ICSAI (2017)
54. Yang, C., Ding, L., Chen, Y., Li, H.: Defending against gan-based deepfake attacks via transformation-aware adversarial faces. In: 2021 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2021)
55. Yeh, C.Y., Chen, H.W., Tsai, S.L., Wang, S.D.: Disrupting image-translation-based deepfake algorithms with adversarial attacks. In: WACV Workshops (2020)
56. Yuan, X., He, P., Zhu, Q., Li, X.: Adversarial examples: Attacks and defenses for deep learning. IEEE transactions on neural networks and learning systems (2019)
57. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR (2017)