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] ⋆, and 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. This document supplements paper Watermark Vaccine: Adversarial Attacks to Prevent Watermark by providing theoretically analysis, dataset details, additional experiments and more results, and more visualizations.

1 Theoretically Analysis

In this section, we will theoretically analyze why our proposed vaccines work effectively and give the lower bound or upper bound of the watermark protection for DWV and IWV.

Definition 1. *Let x be a host image, w be a watermark, and δ be a watermark vaccine. For simplicity, we make plus denote the watermarking and vaccination operations. Thus, the watermarked image with the perturbation can be written as $(x + w + \delta)$. Let f denotes the watermark-removal network, then we can let $f(x + w, \delta)$ denote the watermark removed image of $(x + w + \delta)$. Next, define $\|f(x+w, \delta) - (x+w+\delta)\|$ as the distance between the watermark removed image with vaccine and the ground truth, which is used to evaluate the effectiveness of the watermark vaccine.*

Assumption 1 *We assume that the f satisfies a local Lipschitz condition on a set Ω such that*

$$\forall x, \exists \delta \quad \|f(x, \delta) - f(x + w, \delta)\| \leq L \|w\|, \tag{1}$$

where for the any x, there exists a δ, and after adding a w, there exists $(x + w, \delta) \in \Omega$. In particular, we assume that the watermark is small enough, so the watermarking variation $\|w\|$ of an image is also small.

⋆ Corresponding Author

Proposition 1. *Assume that f satisfies Assumption 1, let the DWV δ_1 disturbs the removed image of the host image x, which is equal to maximize the distance between the watermark removed image $f(x, \delta_1)$ and the ground truth $(x + \delta_1)$, such that*

$$\|f(x, \delta_1) - (x + \delta_1)\| \geq M, \tag{2}$$

where M is the lower bound for optimization. And for any random noise δ_0, there always exists $\|f(x, \delta_0) - (x + \delta_0)\| \ll M$. Thus, there exists the lower bound for the distance between the watermark removed image and watermarked images:

$$\|f(x + w, \delta_1) - (x + w + \delta_1)\| \geq M - (L + 1)\,\|w\| \gg \|f(x, \delta_0) - (x + \delta_0)\|. \tag{3}$$

Proposition 2. *Assume that f satisfies Assumption 1, let the IWV δ_2 disturbs the removed image of the host image x, which is equal to minimize the distance between the removed image $f(x, \delta_2)$ and $x + \delta_2$, such that*

$$\|f(x, \delta_2) - (x + \delta_2)\| \leq N, \tag{4}$$

where N is the upper bound for optimization. And for any random noise δ_0, there always exists $\|f(x, \delta_0) - (x + \delta_0)\| \gg N$. Thus, there exists the upper bound for the distance between the removed image with watermarks and watermarked images:

$$\|f(x + w, \delta_2) - (x + w + \delta_2)\| \leq N + (L + 1)\,\|w\| \ll \|f(x, \delta_0) - (x + \delta_0)\|. \tag{5}$$

Proof. For the Proposition 1, by assumption there exists Lipschitz constant L such that $\|f(x, \delta_1) - f(x + w, \delta_1)\| \leq L\,\|w\|$. Then, so long as DWV δ_1 satisfies $\|f(x, \delta_1) - (x + \delta_1)\| \geq M$, we can inductively get:

$$\begin{aligned}
\|f(x + w, \delta_1) - (x + w + \delta_1)\| &= \|f(x + w, \delta_1) - f(x, \delta_1) \\
&\qquad + f(x, \delta_1) - (x + \delta_1) - w\| \\
&\geq |\|(x + w, \delta_1) - f(x, \delta_1)\| - \\
&\qquad \|f(x, \delta_1) - (x + \delta_1) - w\|| \\
&\geq |M - \|f(x, \delta_1) - (x + \delta_1) - w\|| \\
&= M - \|f(x, \delta_1) - (x + \delta_1) - w\| \quad (*) \\
&\geq M - \|f(x, \delta_1) - (x + \delta_1)\| - \|w\| \\
&\geq M - L\|w\| - \|w\| \\
&= M - (L + 1)\|w\| \quad .
\end{aligned}$$

In the $(*)$ step, because the lower bound M is expected to be larger when generating the DWV, and $\|w\|$ is assumed to be very small in the Assumption, the term in the absolute value is always a positive number. In addition, due to $\|f(x, \delta_0) - (x + \delta_0)\| \ll M$, we can get:

$$\|f(x + w, \delta_1) - (x + w + \delta_1)\| \geq M - (L + 1)\|w\| \gg \|f(x, \delta_0) - (x + \delta_0)\|.$$

Similarly, for Proposition 2, by assumption, there exists Lipschitz constant L such that $\|f(x,\delta_2) - f(x+w,\delta_2)\| \leq L\,\|w\|$. Then, if the IWV δ_2 satisfies the condition $\|f(x,\delta_2) - (x+\delta_2)\| \leq N$, we can also inductively get:

$$\begin{aligned}
\|f(x+w,\delta_2) - (x+w+\delta_2)\| &= \|f(x+w,\delta_2) - f(x,\delta_2) \\
&\quad + f(x,\delta_2) - (x+w+\delta_2)\| \\
&\leq \|f(x+w,\delta_2) - f(x,\delta_2)\| \\
&\quad + \|f(x,\delta_2) - (x+w+\delta_2)\| \\
&\leq L\|w\| + \|f(x,\delta_2) - (x+\delta_2)\| + \|w\| \\
&\leq L\|w\| + N + \|w\| \\
&= N + (L+1)\|w\|.
\end{aligned}$$

due to the $\|f(x,\delta_0) - (x+\delta_0)\| \gg N$ and $\|w\|$ is small enough, then we can get:

$$\|f(x+w,\delta_2) - (x+w+\delta_2)\| \leq N + (L+1)\|w\| \ll \|f(x,\delta_0) - (x+\delta_0)\|$$

Although the above Propositions are only valid under certain assumptions, they explain why our vaccines can successfully prevent the watermark from being removed. According to the Equation (3), if the lower bound M becomes larger or watermarking variation $\|w\|$ becomes smaller, the lower bound for disrupting distance will be larger. Thus, the effect of DWV is better. Similarly, according to the Equation (5) the upper bound N or the watermarking variation $\|w\|$ is smaller, the upper bound for inerasable distance will be smaller, and the effect of IWV will be better. These conclusions are consistent with our intuitions.

2 Dataset Description

To the best of our knowledge, there are several visible watermark datasets synthesized in recent work: LVW [3], LOGO-L [1], LOGO-H [1], and CLWD [4]. However, LWV mainly contains gray-scale watermarks, which is not applicable to real scenes. In addition, LOGO-L and LOGO-H both only contain the processed watermarked images (each watermarked image is fixed in size and position) and do not have the raw watermark patterns. For our experiments, it may not be appropriate to use them. Therefore, we only use the CLWD dataset in our paper to evaluate our watermark vaccines.

CLWD. Colored Large-scale Watermark Dataset (CLWD) [4] contains 60K watermarked images made up of 160 colored watermarks for training, and 10K watermarked images made up of 40 colored watermarks for testing. The host images are all sourced from the PASCAL VOC2012 training and testing sets, while the watermarks are taken from the open-sourced logo images distributed online. A watermarked image in CLWD is synthesized by one PASCAL image and attached to one processed watermark onto it. The size, location, rotation angle, and transparency for each watermark are selected previously. As described in our main paper, we only train the pre-trained model with the watermarked images in CLWD but use the raw watermark patterns to create the watermarked images to evaluate the universality of watermark vaccines in the attack stage.

3 More Visualization Results of Effectiveness

In Section 4.2 of our manuscript, we have shown the effect of the watermark vaccine by a few examples. In order to further verify the effectiveness of the watermark vaccine, in this section, we show more visualization results for the Disrupting watermark vaccine (DWV) and Inerasable watermark vaccine (IWV) under different blind watermark removal networks. Just like in the main paper, we choose the WDNet [4], BVMR [3], and SplitNet [1] as our models. Qualitative visualizations are shown in Figure 3, 4 and 5 respectively. In order to show the diversity, we select different watermark patterns w and parameters θ for each row.

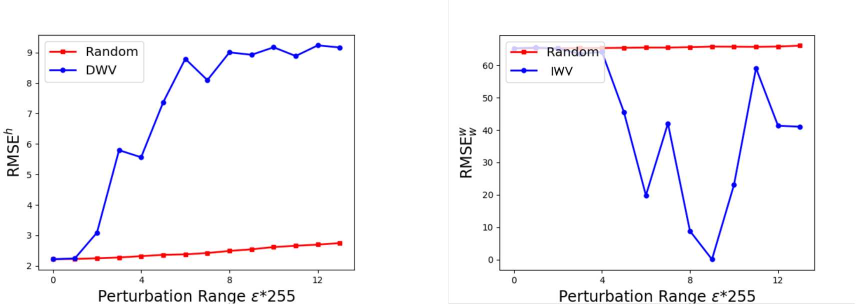
4 Sensitivity to Hyperparameters

In this section, we discuss the sensitivity of watermark vaccines to three hyperparameters: balance parameter β, step size α, and iteration T. We select the β from 1.0 to 3.0, the T from 20 to 60, and the α from 1/255 to 8/255, respectively. When one of the hyperparameters changes, the other parameters will be fixed at the setting value in the main paper. Empirically, we choose the RMSE^h for DWV as the evaluation metrics and the RMSE_w^w for IWV as the evaluation metrics. In addition, we also compare the DWV and IWV with clean input in different watermark-removal networks for a better presentation, and the results are shown in Figure 6, 7 and 8.

In Figure 6, we can see that the RMSE_w^w has an oscillating change as the β changes. However, whatever the β is, the RMSE_w^w for them are always much lower than those for clean input, which means the IWV is always effective regardless of β. In Figure 7, the iteration T shows a similar oscillating change to β for DWV and IWV. Thus, the watermark vaccines are also not sensitive to the iteration T. However, in Figure 8, the effects of the watermark vaccine diminish as the step sizeα increases, and the worst result is obtained when $\alpha = 8$. At this time, the Projected Gradient Descent method [5] in our algorithm has degenerated into the Fast Gradient Sign method [2]. Therefore, we conclude that our watermark vaccines are not sensitive to balance parameter β and iteration T but are less effective if the step size α is too large. In this way, our watermark vaccines are easy to apply in reality without excessive tuning of hyperparameters for every host image.

5 Budgets

In this section, we explore the impact of the perturbation budget ϵ on the performance of the watermark vaccine. The same as our manuscript, we choose



(a) DWV budget (b) IWV budget

Fig. 1. The impact of watermark vaccine budgets on metrics on WDNet [4]. The x-axis shows the perturbation budgets ϵ *255 and the vertical axis shows the value of RMSE^h for DWV or RMSE^w_w for IWV.

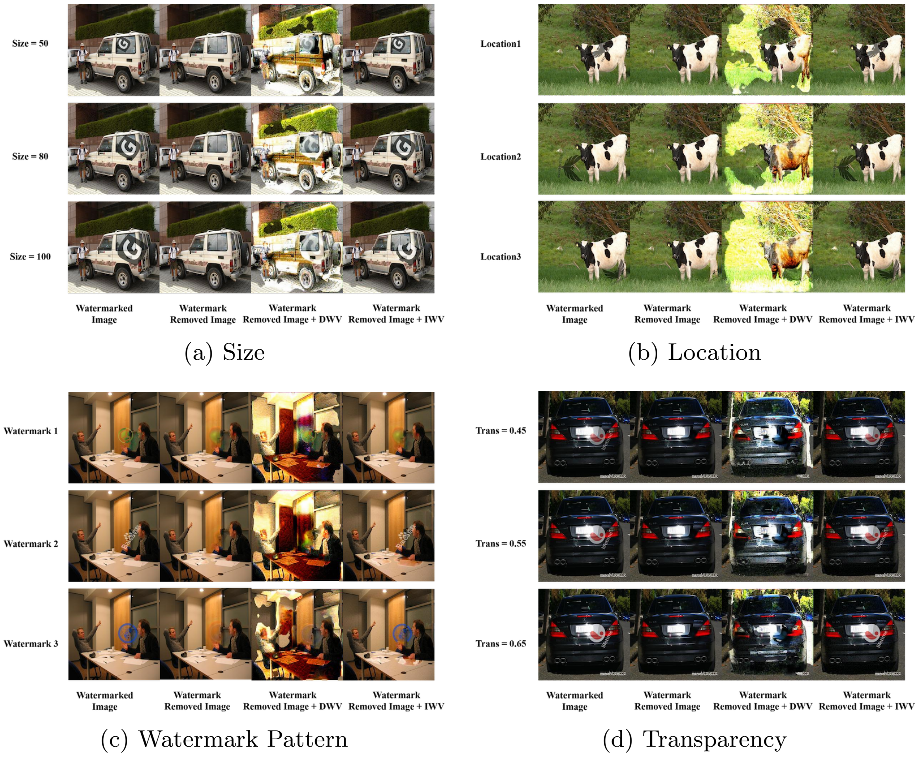
the WDNet as an example and choose the RMSE^h for DWV as the evaluation metrics and the RMSE^w_w for IWV as the evaluation metrics. For quantitative analysis, we calculate the metrics averaging 1,000 images with different budgets and plot the Figure 1. It can be seen that the larger perturbation budgets ϵ cause RMSE^h to change dramatically for the DWV, and then it flattens out. Therefore, the DWV can generate an adversarial effect with only a small budget. However, for the IWV, there is a noticeable drop until the budget ϵ is larger than 4/255 and then oscillate.

We also give a qualitative study on the budget's impact ϵ. In Figure 9 and 10, the top row shows the watermarked image with DWV/IWV as the budget increases. The middle row shows the corresponding watermark removed images, and the bottom row shows their masks. Usually, we choose the perturbation less than 8/255 because such perturbations can be imperceptible. In Figure 9, we can find the image has been destroyed when $\epsilon = 2/255$, and most areas of the watermark removed image are distorted when $\epsilon = 6/255$. In Figure 10, the watermark removed image can not be purified by IWV until $\epsilon = 4/255$, and there is a best protection when $\epsilon = 8/255$.

Combing the quantitative and qualitative analysis, we can learn that the DWV can ruin the watermark removed images even if the perturbation budget is small, but the protection of IWV is not the best even if the budget is large enough because its effect trends to be oscillating. Anyway, our watermark vaccine can still be effective on a small budget that is imperceptible to humans.

6 More Results of Universality

In Section 4.3 of the main paper, we have shown a good universality of our watermark vaccines on WDNet [4]. In this section, we show the visualization of universality results in Figure 2 and present the universality of watermark vaccine on BVMR [3] and SplitNet [1]. As said in the main paper, we test 1,000 host images with ten random sizes/ locations/ patterns/ transparencies of watermarks



(a) Size

(b) Location

(c) Watermark Pattern

(d) Transparency

Fig. 2. The visualization of watermark vaccines on different settings on WDNet [4]. Each row shows the watermark removed images under different sizes/ locations/ patterns/ transparencies of watermarks.

and then calculate the mean and variance of these results from different settings. The final results are shown in Table 1 for BVMR and Table 2 for SplitNet.

In Figure 2, the DWV can always ruin the watermark removed image regardless of size/ location/ pattern/ transparency of the watermarks on it, while the IWV can keep the watermark on them under different settings. It is worth noting that in Figure 2 (a), when the size of the watermark is larger, the effect of IWV has a slight reduction (e.g. the watermark pattern is not fully preserved with size = 100). This is also consistent with our analysis in the main paper. In Table 1 and 2, the large means difference compared to the clean input and the small standard deviation prove that our watermark vaccine can be universal for different watermark settings. In addition, we can see that the standard deviation for the size of watermarks is larger than others, which is probably because the watermarking variation $\|w\|$ is quite important for the watermark vaccine.

Table 1. Mean and standard deviation over evaluation metrics of DWV and IWV on BVMR [3] for random patterns/ locations/ sizes/ transparencies of watermarks. Both two vaccines are compared with the clean input.

	Watermark		Location		Size		Transparency	
Metrics	Clean	DWV	Clean	DWV	Clean	DWV	Clean	DWV
PSNR^h	40.82±0.09	29.42±0.02	42.59±0.20	29.34±0.02	38.64±0.58	29.52±0.06	40.70±0.04	29.34±0.01
SSIM^h	0.9947±0.0002	0.6642±0.0026	0.9960±0.0004	0.6478±0.0023	0.9861±0.0024	0.7064±0.0085	0.9933±0.0002	0.6600±0.0007
RMSE^h	2.37±0.02	8.72±0.02	1.98±0.04	8.78±0.01	3.26±0.19	8.59±0.05	2.42±0.01	8.78±0.01
RMSE^h_w	25.04±1.00	27.56±0.92	25.31±1.48	27.95±0.81	27.53±1.19	31.02±1.37	26.74±0.54	30.45±0.41

(a) DWV

	Watermark		Location		Size		Transparency	
Metrics	Clean	IWV	Clean	IWV	Clean	IWV	Clean	IWV
PSNR^w	40.81±0.13	42.56±0.29	42.95±0.22	44.53±0.22	39.29±0.61	40.55±0.72	41.36±0.10	43.64±0.08
SSIM^w	0.9874±0.0007	0.9914±0.0006	0.9925±0.0007	0.9954±0.0006	0.9723±0.0040	0.9817±0.0036	0.9888±0.0004	0.9945±0.0002
RMSE^w	2.40±0.03	2.05±0.05	1.95±0.05	1.63±0.04	3.12±0.19	2.82±0.19	2.31±0.02	1.84±0.01
RMSE^w_w	42.54±1.47	36.42±1.66	39.52±2.59	34.39±2.16	38.59±2.41	34.73±2.21	38.26±0.97	29.04±0.84

(b) IWV.

7 More Results of Resistance to Image Processing Operations

We have shown that our watermark vaccines can resist JPEG compression operation and Gaussian Blur operation in the main paper. In this section, we will give more results for SplitNet [1] and BVMR [3], and we will consider four other operations: brightness adjustment, contrast adjustment, saturation adjustment and hue adjustment. Empirically, we choose the RMSE^h for DWV as the evaluation metrics and the RMSE^w_w for IWV for a better illustration.

Figure 11 and 12 show the effect of JPEG Compression and Gaussian Blur on DWV/IWV for BVMR and SplitNet. We can find that our watermark vaccines can resist a higher 70% compression rate or lower 1 radius of a gaussian blur for BVMR, while they only resist a higher 80% compression rate or lower 0.75 radius of a gaussian blur for SplitNet. Although an excessive compression ratio or a big radius of blur will reduce the performance of the watermark vaccines, as the main paper says, these excessive operations will also degrade the image quality at the same time.

In Figure 13, 14, 15 and 16, we can see that our watermark vaccines can also resist other image operations. Among these operations, we find contrast adjustment and saturation adjustment have little effect on watermark vaccine for three networks, while if the hue adjustment is too strong, the effect of watermark vaccine is significantly reduced like JPEG compression and Gaussian Blur. Therefore, we conclude that our watermark vaccines can resist the moderate change by most common image operations.

Table 2. Mean and standard deviation over evaluation metrics of DWV and IWV on SplitNet [1] for random patterns/ locations/ sizes/ transparencies of watermarks. Both two vaccines are compared with the clean input.

	Watermark		Location		Size		Transparency	
Metrics	Clean	DWV	Clean	DWV	Clean	DWV	Clean	DWV
PSNR^h	43.01±0.16	33.68±0.08	44.67±0.15	33.78±0.09	40.50±0.40	33.67±0.13	42.47±0.02	33.64±0.15
SSIM^h	0.9963±0.0003	0.8846±0.0026	0.9976±0.0002	0.8858±0.0019	0.9917±0.0011	0.8952±0.0037	0.9960±0.0001	0.8887±0.0025
RMSE^h	1.84±0.03	5.43±0.04	1.53±0.03	5.40±0.04	2.66±0.13	5.44±0.08	1.96±0.01	5.47±0.07
RMSE^h_w	21.72±1.60	44.26±2.83	20.73±0.77	71.51±1.69	21.88±0.73	62.46±2.04	20.44±0.77	69.44±2.73

(a) DWV

	Watermark		Location		Size		Transparency	
Metrics	Clean	IWV	Clean	IWV	Clean	IWV	Clean	IWV
PSNR^w	41.72±0.19	44.05±0.36	43.47±0.13	45.78±0.22	39.34±0.48	41.50±0.63	41.70±0.08	45.11±0.22
SSIM^w	0.9875±0.0005	0.9927±0.0005	0.9928±0.0003	0.9958±0.0004	0.9687±0.0036	0.9766±0.0035	0.9887±0.0004	0.9951±0.0001
RMSE^w	2.14±0.04	1.77±0.07	1.76±0.03	1.44±0.04	3.09±0.17	2.71±0.19	2.15±0.01	1.62±0.04
RMSE^w_w	50.03±1.74	5.16±0.79	47.84±1.53	28.63±1.51	48.01±1.48	33.83±2.14	46.16±1.41	24.64±0.85

(b) IWV.

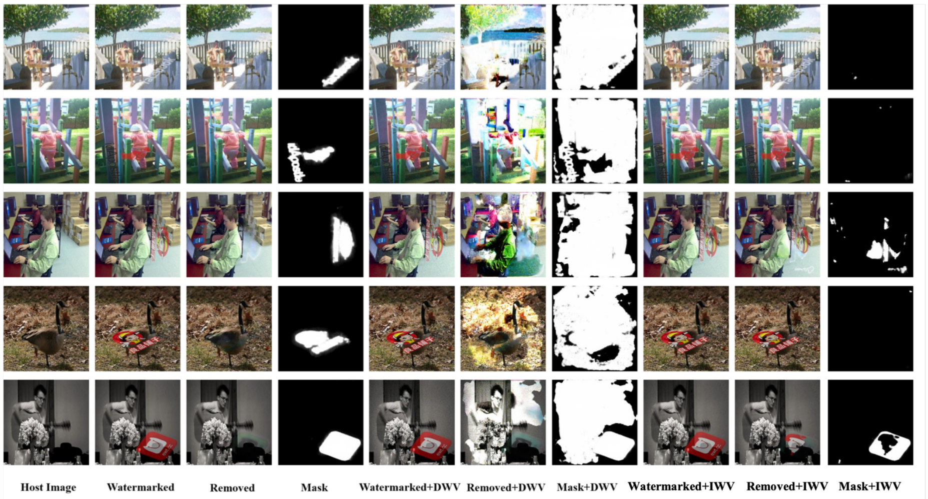


Fig. 3. Qulitative comparison of the protective effects of DWV and IWV under the WDNet [4]. In each row, we show the watermark removal results on the images without vaccine, the images with DWV, and the images with IWV under the same model.

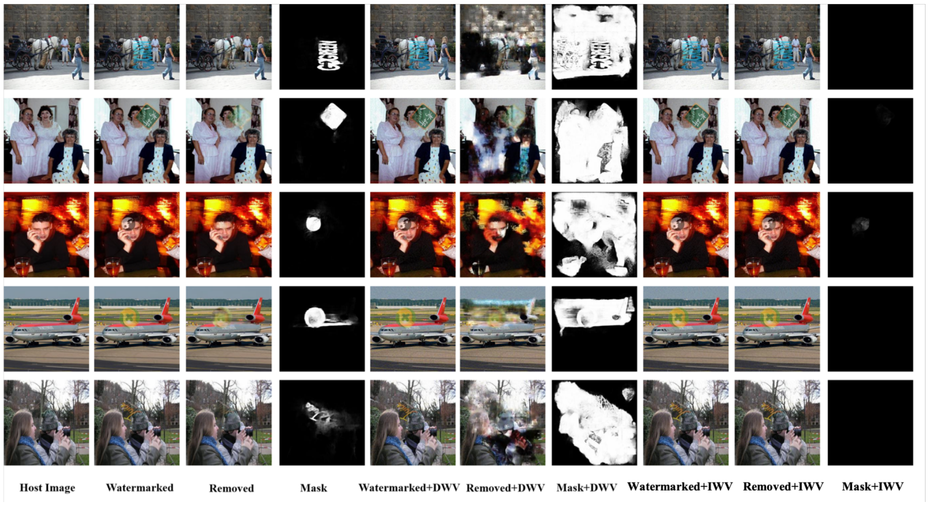


Fig. 4. Qulitative comparison of the protective effects of DWV and IWV under the BVMR [3]. In each row, we show the watermark removal results on the images without vaccine, the images with DWV, and the images with IWV under the same model.

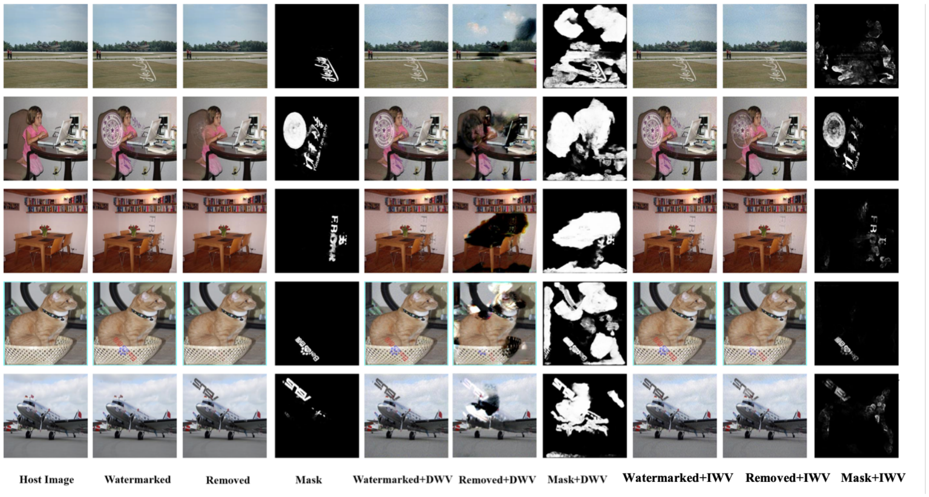
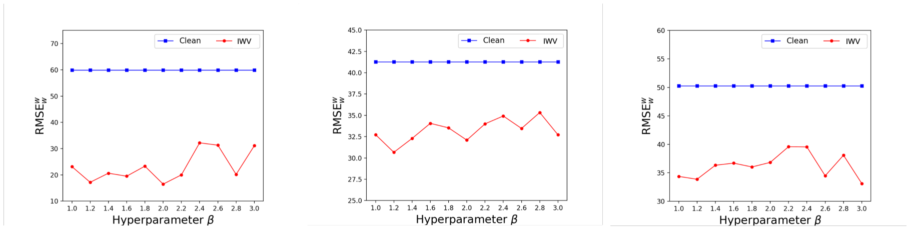
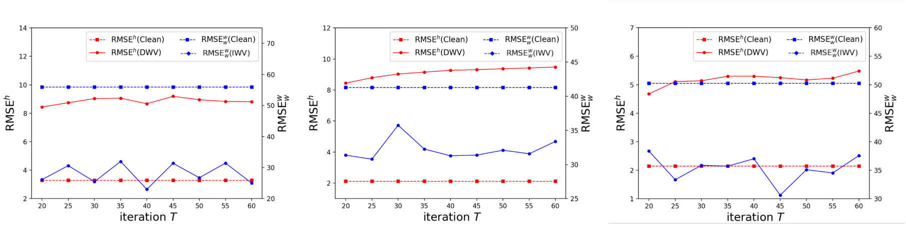


Fig. 5. Qulitative comparison of the protective effects of DWV and IWV under the SplitNet [1]. In each row, we show the watermark removal results on the images without vaccine, the images with DWV, and the images with IWV under the same model.



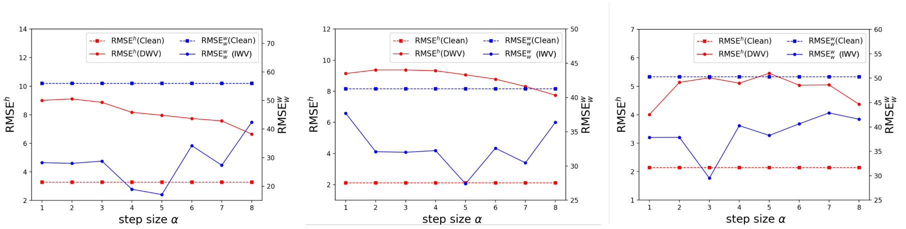
(a) WDNet (b) BVMR (c) SplitNet

Fig. 6. The effect of IWV on three watermark-removal network with different balance parameters β. We choose the RMSE^w_w as an evaluation metric for IWV. The blue lines show the results of clean input, and the red lines show the results of IWV.



(a) WDNet (b) BVMR (c) SplitNet

Fig. 7. The effect of DWV/IWV on three watermark-removal network with different iteration T. We choose the RMSE^h as an evaluation metric for DWV, and the RMSE^w_w for IWV. The solid lines show the results with watermark vaccines, while the dashed lines show the results of clean input. The red lines show the results of RMSE^h, and the blue lines show the results of RMSE^w_w.



(a) WDNet (b) BVMR (c) SplitNet

Fig. 8. The effect of DWV/IWV on three watermark-removal network with different step size α, where $\alpha = 1$ means the step size is $1/255$. We choose the RMSE^h as an evaluation metric for DWV, and the RMSE^w_w for IWV. The solid lines show the results with watermark vaccines, while the dashed lines show the results of clean input. The red lines show the results of RMSE^h, and the blue lines show the results of RMSE^w_w.

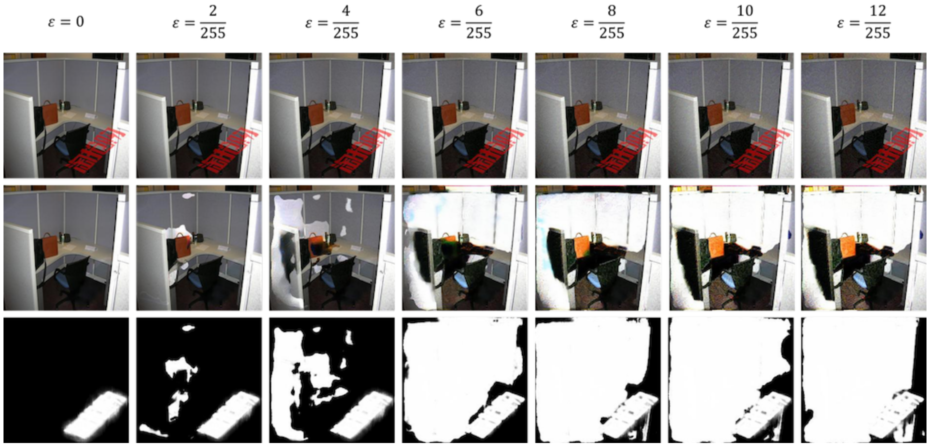


Fig. 9. Watermark removal results with increasing budget ϵ of DWV. The top row shows that the watermarked images with DWV we produced, and the second and the bottom row show the removed images and masks we predict.

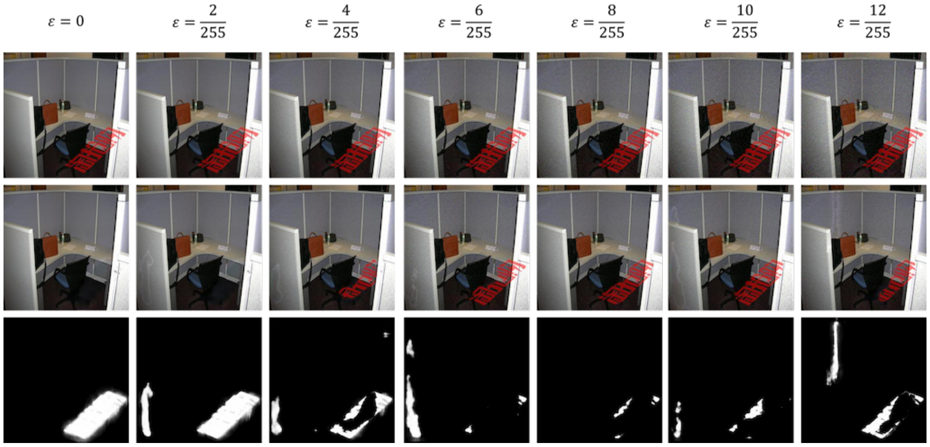


Fig. 10. Watermark removal results with increasing budget ϵ of IWV. The top row shows that the watermarked images with IWV we produced, and the second and the bottom row show the removed images and masks we predict.

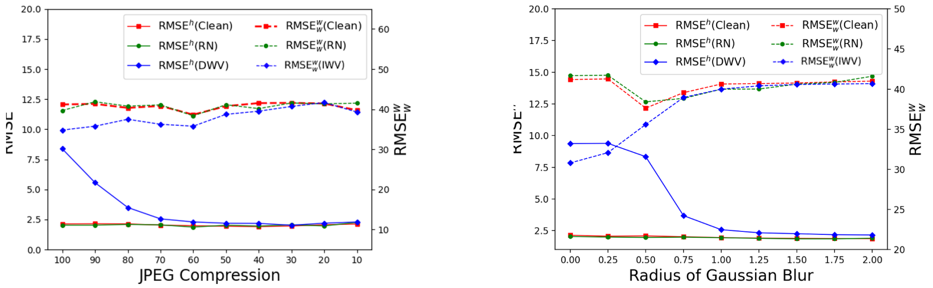


Fig. 11. Effect of two image-based transformation operations (JPEG Compression, Blur) on watermark vaccine for BVMR [3]. The solid lines show the change of RMSE^h, while the dashed lines show the RMSE$_w^w$ change.

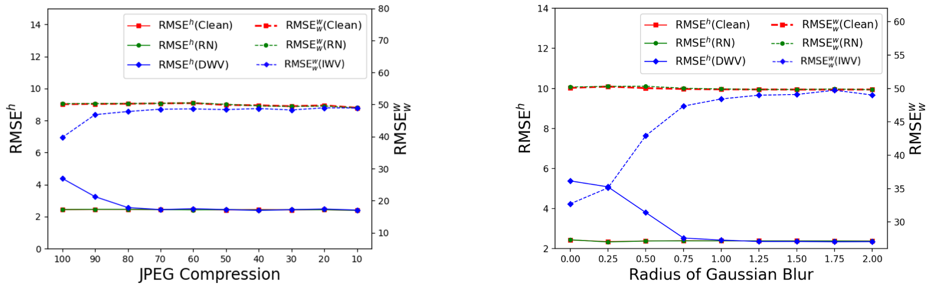
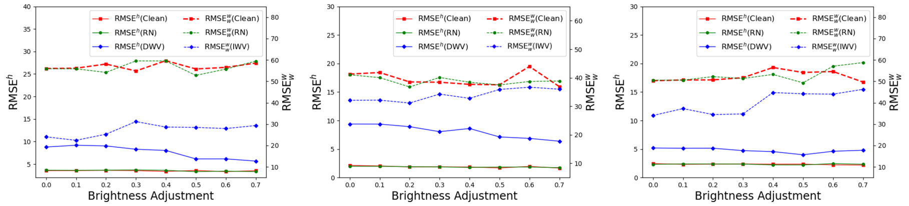
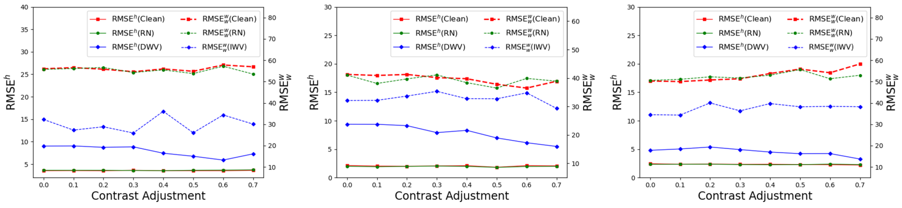


Fig. 12. Effect of two image-based transformation operations (JPEG Compression, Blur) on watermark vaccine for SplitNet [1]. The solid lines show the change of RMSE^h, while the dashed lines show the RMSE$_w^w$ change.



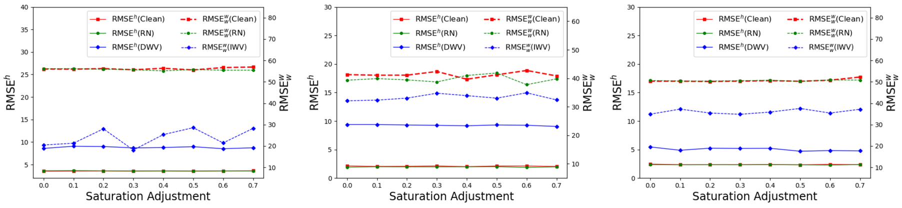
(a) WDNet (b) BVMR (c) SplitNet

Fig. 13. Effect of Brightness Adjustment on watermark vaccine for three models. The solid lines show the change of RMSE^h, while the dashed lines show the RMSE$_w^w$ change.



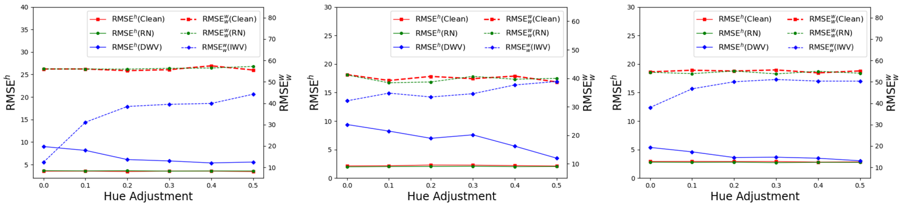
(a) WDNet (b) BVMR (c) SplitNet

Fig. 14. Effect of Contrast Adjustment on watermark vaccine for three models. The solid lines show the change of RMSE^h, while the dashed lines show the RMSE^w_w change.



(a) WDNet (b) BVMR (c) SplitNet

Fig. 15. Effect of Saturation Adjustment on watermark vaccine for three models. The solid lines show the change of RMSE^h, while the dashed lines show the RMSE^w_w change.



(a) WDNet (b) BVMR (c) SplitNet

Fig. 16. Effect of Hue Adjustment on watermark vaccine for three models. The solid lines show the change of RMSE^h, while the dashed lines show the RMSE^w_w change.

References

1. Cun, X., Pun, C.M.: Split then refine: Stacked attention-guided resunets for blind single image visible watermark removal. In: AAAI (2021)
2. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: ICLR (2015)
3. Hertz, A., Fogel, S., Hanocka, R., Giryes, R., Cohen-Or, D.: Blind visual motif removal from a single image. In: CVPR (2019)
4. Liu, Y., Zhu, Z., Bai, X.: Wdnet: Watermark-decomposition network for visible watermark removal. In: WACV (2021)
5. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: ICLR Poster (2018)