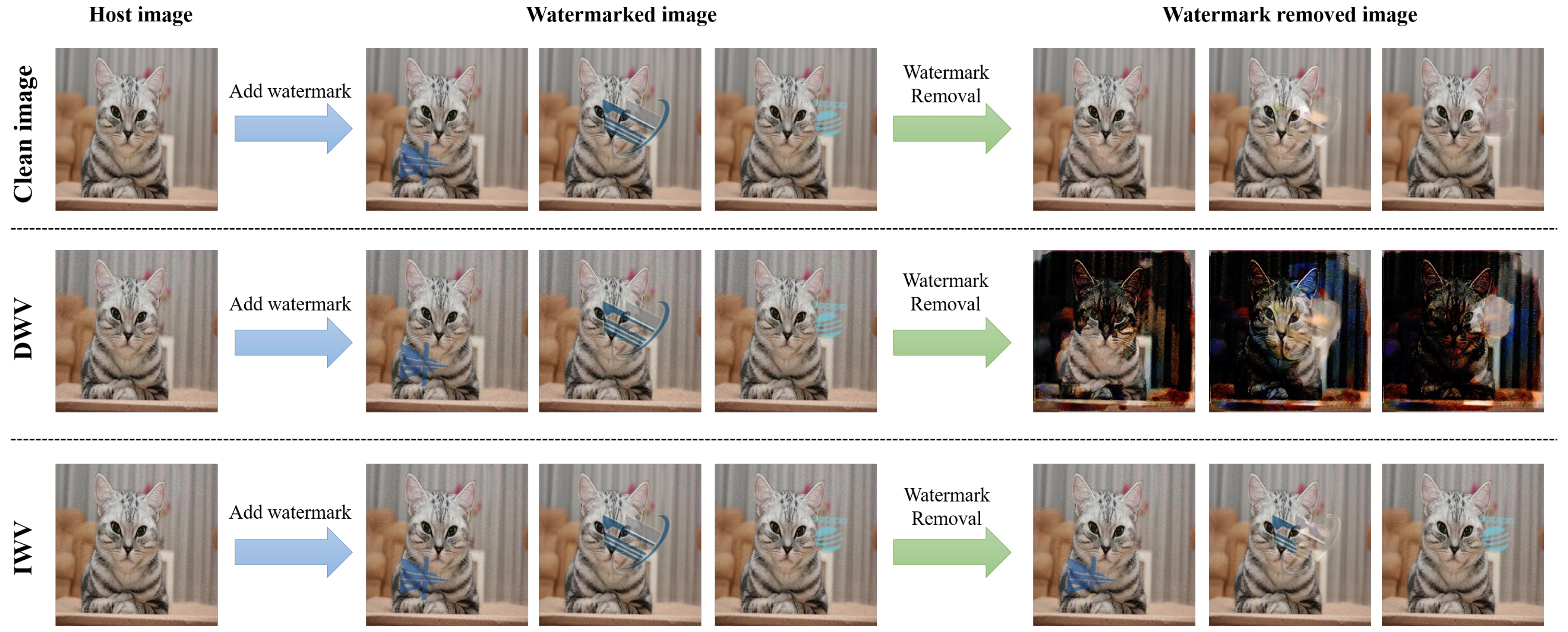
ECCV EUROPEAN CONFERENCE ON COMPUTER VISION TEL AVIV 2022
October 23-27, 2022, Tel Aviv

Watermark Vaccine: Adversarial Attacks to Prevent Watermark Removal

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

1. SKLOIS, Institute of Information Engineering, CAS 2. School of Cybers Security, University of Chinese Academy of Sciences 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

Motivation & Contribution



The protective effects of our watermark vaccines on different watermark patterns or parameters.

Motivation: Visible watermarking thus becomes an essential technique. It prevents illicit users from obtaining some critical information and using copyrighted high-quality images. However, visible watermark is in face of security issues as it can be effectively removed by some watermark-removal techniques. Inspired by recent studies on adversary, which show that imperceptible adversarial perturbations can cause some incorrectly outputs for DNNs, 'adversarial for good' is thus a new protection method.

Contribution:

- 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.

Methods:

- General Problem Formulation:

1.Inject Vaccine: $\hat{x}_h = x_h + \delta$, $\|\delta\|_\infty \le \varepsilon.$

2.Add Watermark: $\hat{x}_w = g\left(\hat{x}_h, \omega, \theta\right),$

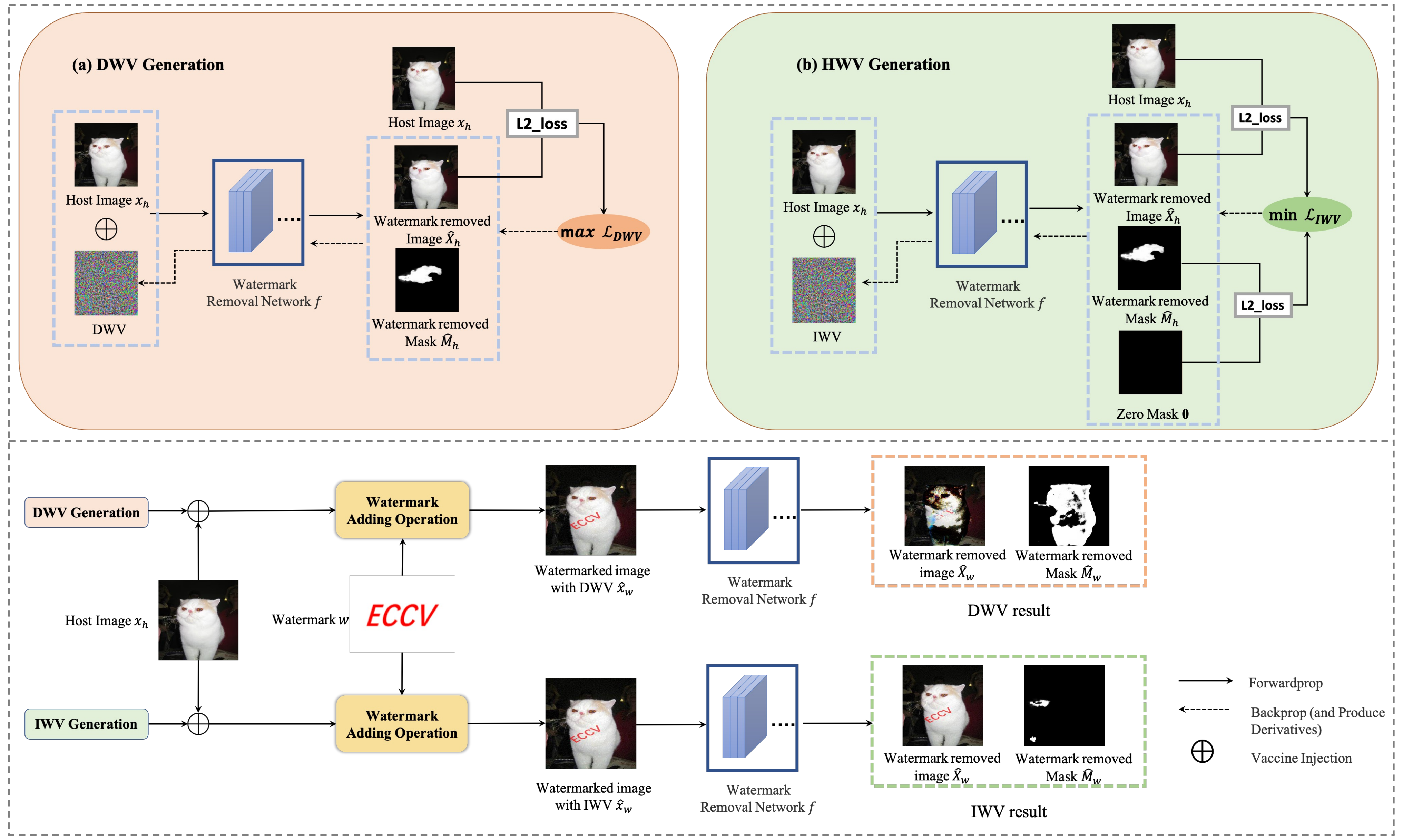
3.Watermark Removal: $X_w, M_w = f\left(x_w\right),$ $\hat{X}_w, \hat{M}_w = f\left(\hat{x}_w\right).$

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

Two challenges:

◆ The effect of watermark removal Q(·) can be customized in a variety of different ways, but it is required to be differentiable during optimization.

◆ Two expectation over w and θ is hard to optimize by considering the loss of all combinations simultaneously.


(a) DWV Generation
Host Image x_h
L2_loss
Host Image x_h
Watermark removed Image $\hat{X}_h$
max $\mathcal{L}_{DWV}$
DWV
Watermark Removal Network f
Watermark removed Mask $\hat{M}_h$
(b) HWV Generation
Host Image x_h
L2_loss
Host Image x_h
Watermark removed Image $\hat{X}_h$
min $\mathcal{L}_{IWV}$
IWV
Watermark Removal Network f
Watermark removed Mask $\hat{M}_h$
L2_loss
Zero Mask 0
DWV Generation
Watermark Adding Operation
Watermarked image with DWV $\hat{x}_w$
Watermark Removal Network f
Watermark removed image $\hat{X}_w$
Watermark removed Mask $\hat{M}_w$
DWV result
Host Image x_h
Watermark w
ECCV
IWV Generation
Watermark Adding Operation
Watermarked image with IWV $\hat{x}_w$
Watermark Removal Network f
Watermark removed image $\hat{X}_w$
Watermark removed Mask $\hat{M}_w$
IWV result
Forwardprop
Backprop (and Produce Derivatives)
⊕ Vaccine Injection

The overview figure of the generation and application of our proposed watermark vaccine.

- Disrupting Watermark Vaccine (DWV)

1. Inject Vaccine: $\hat{x}_h = x_h + \delta$, $\|\delta\|_\infty \le \varepsilon.$

2. Watermark Removal: $\hat{X}_h, \hat{M}_h = f(\hat{x}_h),$

3. Goal: $\mathcal{L}_{\mathcal{DWV}}(x_h, \delta) = \left\|\hat{X}_h - x_h\right\|^2,$

$$\begin{aligned} \max_\delta \quad & \mathcal{L}_{\mathcal{DWV}}(x_h, \delta) \\ \text{s.t.} \quad & \hat{x}_h = x_h + \delta \\ & \|\delta\|_\infty \leqslant \varepsilon, \end{aligned}$$

- Inerasable Watermark Vaccine (IWV)

1. Inject Vaccine: $\hat{x}_h = x_h + \delta$, $\|\delta\|_\infty \le \varepsilon.$

2. Watermark Removal: $\hat{X}_h, \hat{M}_h = f(\hat{x}_h),$

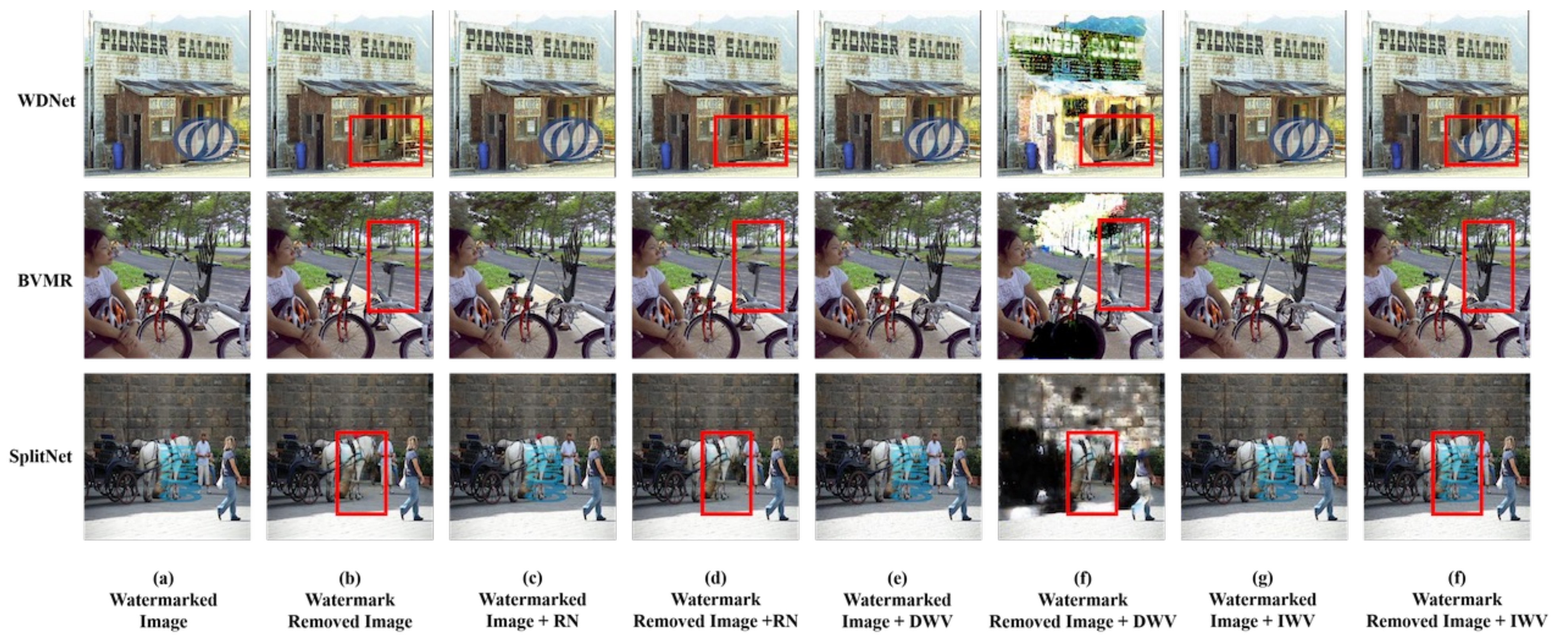
3. Goal: $\mathcal{L}_{\mathcal{IWV}}(x_h, \delta) = \frac{1}{2}\left(\beta\left\|\hat{X}_h - x_h\right\|^2 + \|\hat{M}_h - \mathbf{0}\|^2\right),$

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

```
Algorithm 1: Watermark Vaccine Generation
Input: host image $x_h$, blind watermark-removal network $f$, iteration T, step
       size $\alpha$, perturbation bound $\epsilon$
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 $\mathcal{L}_{\mathcal{DWV}}$;
        $\delta \leftarrow \delta + \alpha\ \mathrm{sign}(\nabla_\delta \mathcal{L}_{\mathcal{DWV}}(x_h, \delta))$;
    else
        using Equation (9) to calculate the $\mathcal{L}_{\mathcal{IWV}}$;
        $\delta \leftarrow \delta - \alpha\ \mathrm{sign}(\nabla_\delta \mathcal{L}_{\mathcal{IWV}}(x_h, \delta))$;
    end
    $\hat{x}_h \leftarrow x_h + \mathrm{clip}(\delta, -\epsilon, \epsilon)$;
end
$\hat{x}_h \leftarrow \mathrm{clip}(\hat{x}_h, 0, 1)$;
```

Experiments & Results

Effectiveness


WDNet
BVMR
SplitNet
(a) Watermarked Image
(b) Watermark Removed Image
(c) Watermarked Image + RN
(d) Watermark Removed Image +RN
(e) Watermarked Image + DWV
(f) Watermark Removed Image + DWV
(g) Watermarked Image + IWV
(f) Watermark Removed Image + IWV

	WDNet[32]				BVMR[20]				SplitNet[11]			
Metrics	PSNR^h	SSIM^h	RMSE^h	RMSE$_w^h$	PSNR^h	SSIM^h	RMSE^h	RMSE$_w^h$	PSNR^h	SSIM^h	RMSE^h	RMSE$_w^h$
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**

	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**

Universality

	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$_w^h$	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.

Metrics	PSNR^h		SSIM^h		RMSE^h		RMSE$_w^h$		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

Conclusion

- Our watermark vaccine is obtained by optimizing **adversarial perturbations** to attack the blind watermark removal network.
- 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.
- The code is released at *https://github.com/thinwayliu/Watermark-Vaccine.*