Supplementary Material of Attention-Guided Watermark Vaccine Against Watermark Removal

Fig. 1. The IQAs of adversarial outputs generated by disrupting and keeping watermark vaccines.

I. Differentiable Noise

Due to the inherently discrete nature of Salt&Pepper and JPEG, these noises are non-differentiable, resulting in failure gradient back-propagation. We discuss the corresponding solutions as follows.

A. Salt&Pepper

The Salt&Pepper noise randomly replaces image pixels with either the minimum value (0, representing a black dot) or the maximum value (1, representing a white dot). This process is non-differentiable. To address this issue, we use the Bernoulli distribution [1] to create two binary masks: salt mask and pepper mask. We then apply a smooth function such as the Sigmoid function, to transform these binary values of masks into continuous probability distributions. In this way, we can approximate the Salt&Pepper noise, enabling the gradient back-propagation for ABCG-VE.

B. JPEG

Previous works adopt differentiable JPEG noise as a data augmentation process to train robust models, such as Hidden [2] and MBRS [3]. However, these works only modify the DCT coefficients to simulate JPEG noise, the simulated JPEG noise is still far from real JPEG noise. Reich et al. [4] model JPEG encoding-decoding in a differentiable manner, including quantization, QT scale floor, QT floor, QT clipping, and output clipping. This approach can accurately resemble standard JPEG over the whole compression range. We adopt this work as JPEG noise in Section ABCG-VE. This code can be installed as a Python package: pip install git+https://github.com/necla-ml/Diff-JPEG.

II. Vaccine Evaluation Model

The vaccine evaluation model can be viewed as a discriminator to determine whether the input image is the original image (without watermark), watermark image, or perturbed image. We use Resnet50 as the backbone model. However, the difference between the original image and the watermarked image is much smaller, leading to the performance of Resnet50 being far from satisfactory. Therefore, we add the Spatial Attention Module (SAM) [5] into Resnet50 to make the model learn more discriminative feature representation. The SAM respectively captures the global features F_g and prominent features F_p by average-pooling and max-pooling operations along the channel axis. Those are then concatenated and convolved by the convolution operation with the filter size of 3*3 f^{3*3}. The process of SAM can be formulated as:

$$SAM(F) = \sigma\left(f^{3*3}\left([AvgPool\left(F_g\right); MaxPool\left(F_p\right)]\right)\right) \quad (1)$$

where σ represents the Softmax function. The parameters of the vaccine evaluation model are optimized by minimizing the Cross-Entropy loss, which is represented by:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log\left(p_{i,c}\right) \quad (2)$$

where N represents the number of samples, C represents the number of labels, $y_{i,c}$ determines the sample y_i with the label c, and $p_{i,c}$ represents the predicted probability.

For a given adversarial output, it can be fed into the trained model to compute the probability distribution, thereby predicting labels. The predicted labels of adversarial outputs can be used to determine whether the watermark vaccines successfully protect the watermarked images.

Training-We sample three sets of 1000 images (i.e., original images, watermarked images, and adversarial outputs of disrupting watermark vaccine) to train the vaccine evaluation model. To ensure a fair comparison, these adversarial outputs are derived from different watermark vaccines. We visually inspect the adversarial outputs to identify those that have been protected by the disrupting watermark vaccine, and manually select the training data. This model is trained on the pre-trained Resnet50 by 100 epochs.

III. Ablation Studies

Firstly, we investigate the influence of different λ on the performance of the proposed PFAG-VA. We adopt an adaptive weight rather than a fixed one. As shown in Fig.2, our approach achieves the highest PSRs. This is because different

TABLE I

THE PERFORMANCES OF DIFFERENT COMPONENTS OF THE PROPOSED APPROACH IN TERMS OF IMPERCEPTIBILITY. THE PFAG-VA ACHIEVES THE HIGHEST IMPERCEPTIBILITY. THE AGWV FURTHER OPTIMIZES THE WATERMARK VACCINE TO CORRECT THE ATTENTION BIAS, LEADING TO A SLIGHT LOSS OF IMPERCEPTIBILITY.

Model	WDNet				SplitNet			
	PSNR ↑	SSIM ↑	RMSE ↓	PSR(%) ↑	PSNR ↑	SSIM ↑	RMSE ↓	PSR(%) ↑
Disrupting watermark vaccine								
PGD	32.82	0.83	5.12	96.8	35.09	0.94	4.55	97.1
PFAG-VA	**41.85**	**0.98**	**2.26**	94.5	**42.12**	**0.99**	**2.18**	95.5
ABCG-VE	32.06	0.81	5.62	97.1	34.28	0.93	4.92	97.8
AGWV	40.76	0.97	2.58	95.8	41.52	0.98	2.51	96.4
Keeping watermark vaccine								
PGD	34.91	0.92	3.93	77.8	35.31	0.94	4.45	87
PFAG-VA	**42.78**	**0.99**	**2.19**	76.3	**43.45**	**0.99**	**1.89**	85.1
ABCG-VE	34.62	0.92	4.15	78.6	34.83	0.94	4.76	86.2
AGWV	41.78	0.98	2.07	78.5	42.68	0.98	1.92	85.5

TABLE II

THE PROTECTIVE PERFORMANCES OF DIFFERENT COMPONENTS OF THE PROPOSED APPROACH ON WDNET IN TERMS OF ROBUSTNESS. THE ABCG-VE ACHIEVES THE HIGHEST ROBUSTNESS. THE AGWV RESTRICTS THE WATERMARK VACCINE TO BE ADDED IN THE PROPER REGIONS, LEADING TO A SLIGHT LOSS OF PSR.

Noise	Gaussian (G)		JEPG (J)		Salt&Pepper (S)		Uniform (U)		G, J	G, J, S	G, J, S, U
	0.5s	0.8s	80	50	2%	5%	10	20	0.5, 80	0.5, 80, 2%	0.5, 80, 2%, 10
Disrupting watermark vaccine (PSR %)											
PGD	67.4	35.7	69.6	21.1	65.6	35.8	65.7	36.5	30.5	10.4	0
PFAG-VA	60.5	28.6	52.3	19.5	52.5	26.4	62	30.1	18.6	0	0
ABCG-VE	**96.5**	**90.5**	**97**	**92.8**	**95.2**	**89.9**	**84.8**	**70.3**	**92.2**	**70.4**	**34.9**
AGWV	94.4	87.6	95.5	78.4	94.3	86.3	80.1	65	90.2	68.4	29.8
Keeping watermark vaccine (PSR %)											
PGD	60.3	—	64.7	—	50.4	—	53.7	—	36.1	—	—
PFAG-VA	52.4	—	53.1	—	45.9	—	38.6	—	30.5	—	—
ABCG-VE	**78.1**	—	**78.5**	—	**78.2**	—	**70.9**	—	**69.8**	—	—
AGWV	75.4	—	76	—	76.3	—	68.3	—	68.1	—	—

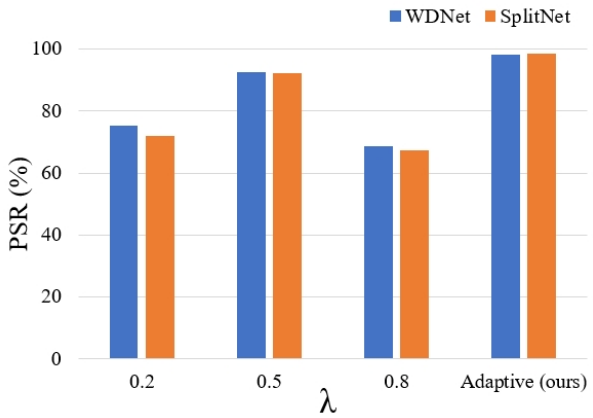

Fig. 2. The protective performance of PFAG-VA using different λ.

attention mechanisms have different importance when processing different images. The proposed approach dynamically adjusts the attention mechanisms during the optimization process of the watermark vaccine, thereby improving its effectiveness.

Then, we conduct extensive ablation studies to investigate the influence of different components of the proposed approach. To validate the effectiveness of PFAG-VA, we first utilize the PFA to restrict the watermark vaccine to be added within the proper region. As shown in Table I, the watermark vaccines optimized by PFAG-VA achieve the highest imperceptibility, while the ABCG-VE performs poorly in imperceptibility. This is because the ABCG-VE further optimizes the noise-tolerant watermark vaccine based on PGD, increasing the adversarial perturbation strength and resulting in a massive loss of image quality. Compared to PGD and ABCG-VE, the AGWV is still superior in imperceptibility. This convincingly validates that perturbing only the proper region can significantly improve the vaccinated image quality.

Next, different image noises are applied to perturb the vaccinated images to evaluate the performance of ABCG-VE. As shown in Table II, the performances of PGD and PFAG-VA are quite poor, meaning that watermark vaccines with image noise cannot defend against watermark removal. Then, the watermark vaccine is optimized by ABCG-VE and achieves the highest robustness against image noise. This is because ABCG-VE can correct the attention bias between the vaccinated images with and without image noise, maintaining the effectiveness of the noisy watermark vaccine.

Overall, the ablation studies prove the effectiveness of the proposed PFAG-VA and ABCG-VE, while the AGWV finds a good trade-off between imperceptibility and robustness.

REFERENCES

[1] Albert W Marshall and Ingram Olkin, "A family of bivariate distributions generated by the bivariate bernoulli distribution," *Journal of the American Statistical Association*, vol. 80, no. 390, pp. 332–338, 1985.

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

[3] Zhaoyang Jia, Han Fang, and Weiming Zhang, "Mbrs: Enhancing robustness of dnn-based watermarking by mini-batch of real and simulated jpeg compression," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 41–49.

[4] Christoph Reich, Biplob Debnath, Deep Patel, and Srimat Chakradhar, "Differentiable JPEG: The Devil is in the Details," in *WACV*, 2024.

[5] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

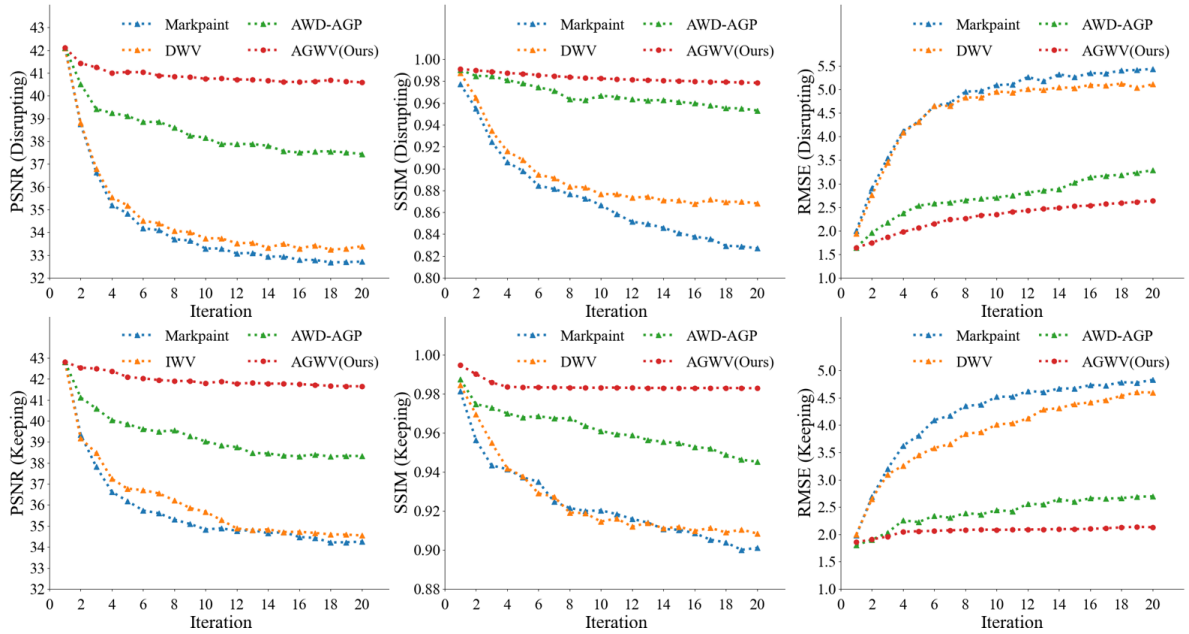

Fig. 3. The IQA values of vaccinated images generated by attacking WDNet when using different iterations. Higher PSNR and SSIM values indicate better imperceptible, while lower RMSE is preferable.

Fig. 4. Examples of watermark vaccines against watermark removal. For watermark vaccines, the higher PSNR value indicates better imperceptibility. Top row: vaccinated images. Second row: protective results. Remaining rows: protective results under different noises. The green border indicates successful protection, while the red border indicates failed protection.