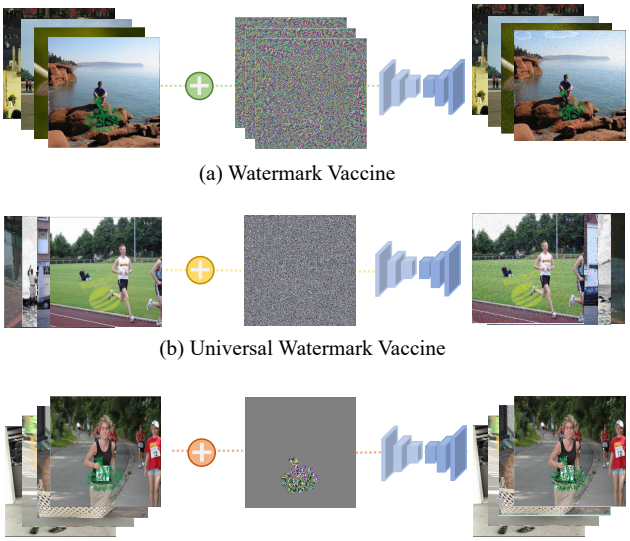
Universal Watermark Vaccine: Universal Adversarial Perturbations for Watermark Protection

Jianbo Chen[1] Xinwei Liu[2,3]* Siyuan Liang[2,3] Xiaojun Jia[2,3] Yuan Xun[2,3]
[1]College of Computer Science and Electronic Engineering, Hunan University, Hunan, China
[2]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[3]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
jianbo@hnu.edu.cn, {liuxinwei, liangsiyuan, xunyuan}@iie.ac.cn, jiaxiaojunqaq@gmail.com

Abstract

As computing ability continues to rapidly develop, neural networks have found widespread use in various fields. However, in the realm of visible watermarking for image copyright protection, neural networks have made image protection through watermarking less effective. Some research has even shown that watermarks can be removed without damaging to the original image, posing a significant threat to digital copyright protection. In response, the community has introduced adversarial perturbations for watermark protection, but these are sample-specific and time-consuming in real-world scenarios. To address this issue, we propose a new universal adversarial perturbation for watermark removal networks that offers two options. The first option involves adding perturbations to the entire host image, bringing the output of the watermark removal network closer to the original image and providing protection. The second option involves adding perturbations only to the watermark position, reducing the impact of the perturbation on the image and enhancing stealthiness. Our experiments demonstrate that our method effectively resists watermark removal networks and has good generalizability across different images.



(a) Watermark Vaccine

(b) Universal Watermark Vaccine

(c) Universal Watermark Vaccine with Mask

Figure 1. We demonstrate the effectiveness of the Watermark Vaccine technique [22] in (a). Our proposed UWV and MUWV, illustrated in (b) and (c), respectively, enable the use of a single universal perturbation instead of multiple distinct watermarks.

1. Introduction

The extensive proliferation of personal computers, the internet, and multimedia technology has enabled the sharing of digital data across the globe. However, the accessibility and usability of image processing tools have made it effortless to duplicate or modify digital data, raising concerns about illegal replication and tampering [17]. Digital watermarking is a popular technique for protecting the ownership and authenticity of digital images [2]. However, visible watermarks can be easily removed by some advanced watermark-removal techniques [6, 16, 26]. Marcelo Bertalmio *et al*. [1] proposed a new algorithm for digital inpainting of still images, which automatically fills in damaged or unwanted regions with surrounding information. Liang *et al*. [19] proposed a two-stage network effectively removes watermarks from images by addressing incomplete detection and degraded texture quality issues. With the rapid development of deep learning, blind visible watermark-removal deep neural networks have been proposed, allowing for the reconstruction of watermarked images without any prior information about the watermarks [4, 5, 14]. These techniques enable individuals to use or distribute digital assets without permission or attribution, pos-

*Corresponding author: liuxinwei@iie.ac.cn

ing a significant threat to digital copyright protection.

Due to the rapid development of advanced watermark-removal technologies, traditional watermarking methods have become increasingly vulnerable in protecting the copyrights of image owners. Recently, Liu *et al*. [22] proposes a defense mechanism by using adversarial machine learning to prevent visible watermark removal. They optimized an imperceptible adversarial perturbation on the host images to proactively attack watermark-removal networks, dubbed Watermark Vaccine and demonstrated its effectiveness in preventing watermark removal. Their method, however, has limitations as it requires generating a specific perturbation for each image, which differs from the practical scenarios encountered in reality. Inspired by recent advancements in Universal Adversarial Perturbation [21, 24, 30], which have shown the existence of a universal perturbation vector capable of causing misclassification of natural images by deep neural networks regardless of the specific image, we extend this paradigm to the domain of watermarking. Building upon the groundwork laid by Liu [22], we propose a novel defense mechanism called Universal Watermark Vaccine (UWV). It enables faster and more convenient generation of perturbations and improves universality compared to previous works, making it more applicable to real-life scenarios.

In this paper, we propose two novel methodologies for protecting digital images against watermark attacks. The first approach, Universal Watermark Vaccine (UWV) with the full image, uses neural network training to generate adversarial perturbations across all host images, providing universal protection against watermark removal attacks. The second approach, UWV with mask (MUWV), builds on the first by incorporating watermark mask information during training to reduce the extent of interference with the host image's information. Experimental evaluations show the efficacy of both UWV and MUWV in mitigating watermark attacks while minimizing loss of image quality. These methodologies have promising practical applications in digital image protection. Our key contributions are summarized as follows:

- We introduce the concept of universal classification networks into generative networks to address the shortcomings of previous methods in terms of generality.
- We present two methods for generating vaccines to protest visible watermarks. The first involves adding perturbations to the entire image, while the second adds perturbations only to the watermark mask, leaving the remaining regions free and perturbed.
- Our experiment results demonstrate that the proposed UWV and MUWV exhibit strong generality and adaptability, making them highly effective in preventing watermark removal across a variety of watermark removal networks.

2. Related Work

2.1. Watermark Vaccine

The Watermark Vaccine [22] is a proactive defense mechanism that targets blind watermark-removal networks used by adversaries [15] to remove watermarks from digital images, akin to a traditional vaccine that prevents infection before it occurs. It achieves this by adding an imperceptible adversarial perturbation to the host images before their release, which can take the form of two different types of vaccines: the Disrupting Watermark Vaccine (DWV) that ruins the host image and watermark after passing through watermark-removal networks [23], and the Inerasable Watermark Vaccine (IWV) that prevents watermark removal while still being noticeable. Experimental results demonstrate that the Watermark Vaccine is effective at preventing watermark removal, particularly on various watermark-removal networks, and it offers a promising solution for protecting digital images and maintaining their ownership and copyright protection. These results highlight the potential of adversarial machine learning techniques for good and their potential applications in cybersecurity.

2.2. Universal Adversarial Perturbation

Universal adversarial perturbations are a type of adversarial perturbation *et al*. [20, 27, 31] that can be applied to multiple different images without the need to generate a new perturbation for each image. They were first proposed by Moosavi-dezfooli *et al*. [24] in their paper for image classification tasks and have also been applied to semantic segmentation tasks. The perturbations are usually larger than those generated for individual images. Universal perturbations can be used to protect digital copyrights by making it difficult for watermark removal networks to accomplish their original task. Hendrik *et al*. [12] also proposed an attack against semantic image segmentation. They use uniform adversarial perturbations that can produce a desired target segmentation as an output. Xie *et al*. [29] proposed a systematic algorithm for generating adversarial samples for object detection and segmentation tasks in their paper.

2.3. Visible Watermark Removal

Visible watermark removal techniques have been developed to evaluate and enhance the resilience of visible watermarks. Previous methods [13, 25] required user interaction to remove watermarks, necessitating the identification of the watermark location and subsequent recovery of the area. Others [8, 9, 32] made strong assumptions about the watermark, limiting their applicability in real-world scenarios. Deep learning has emerged as a powerful tool in

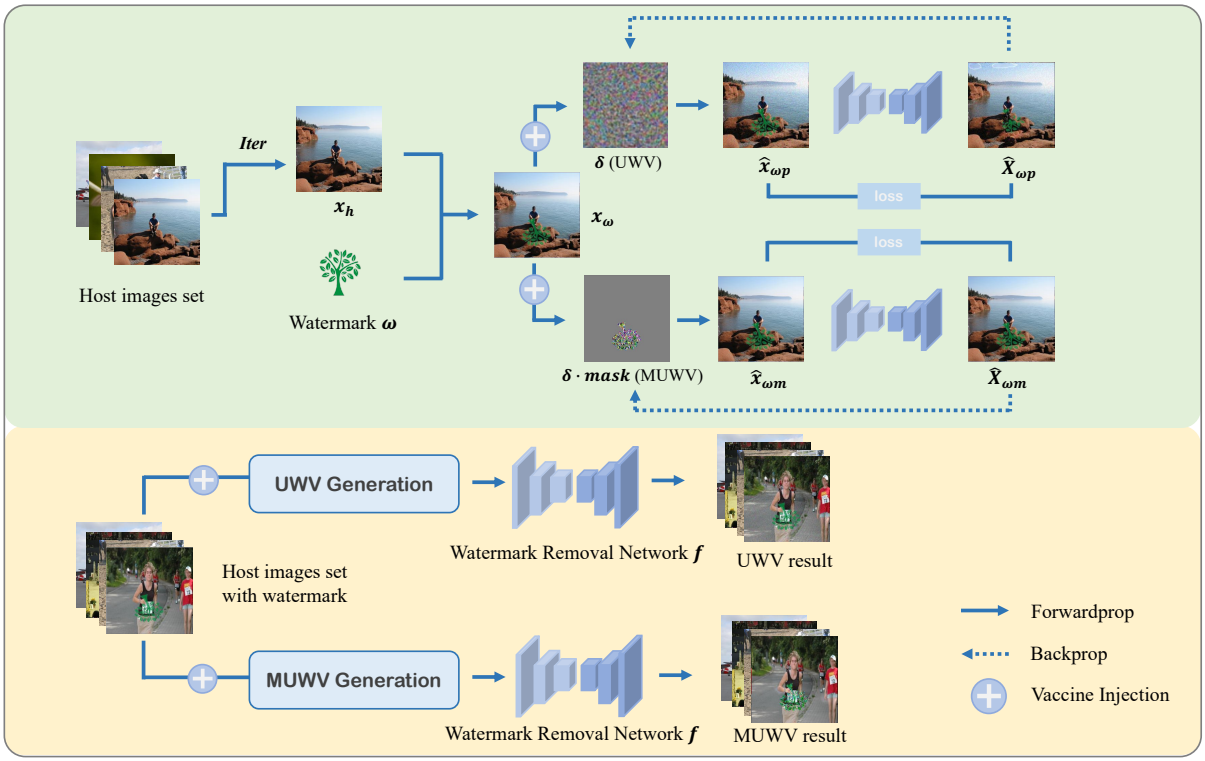


Figure 2. The overview figure depicts the generation (on the first row) and application (on the second row) of our proposed UWV and MUWV. We propose to minimize the loss to generate UWV and MUWV, as shown in the first row. Next, we apply UWV/MUWV to the host images to generate 'protected host images'. When UWV/MUWV is added to those protected host images, they become difficult to remove by a watermark removal network.

computer vision [10,11,28], with researchers exploring two popular strategies for formulating an end-to-end solution to the watermark removal problem. One approach is to treat the task as an image-to-image translation problem [3, 18], while the other involves a two-stage process: first, a mask is used to locate the watermark, and then a network is trained to remove the watermark by restoring the background in the affected area [5, 7, 19, 23]. Experimental results suggest that the latter approach may be more effective in removing watermarks. Thus, the focus of this paper is on preventing the use of these types of networks.

3. Methodology

3.1. Problem Formulation

Based on the problem formalization presented in the "Watermark Vaccine" paper that we followed, the Universal Watermark Vaccine (UWV) can be formulated as a constrained optimization problem. Given a host image x_h and a fixed watermark pattern ω, the operation of adding a watermark can be represented by a function g with parameters θ that specify the position (p, q), size (u, v), and transparency α of the watermark on the host image x_h. The resulting watermarked image x_ω is defined by the equation:

$$x_\omega = g\left(x_h, \omega, \theta\right) \tag{1}$$

Then we assume that there is a UWV δ, which is limited by the infinite parametric L_∞ in ε. After injecting it into the watermarked image, we can get the vaccinated image by,

$$\begin{aligned} \hat{x}_w &= x_w + \delta \\ \|\delta\|_\infty &\leq \varepsilon \end{aligned} \tag{2}$$

We denote the network for watermark removal as f. The model can generate output with the watermark removed image and the corresponding mask, which is defined as:

$$\hat{X}_\omega, \hat{M}_\omega = f\left(\hat{x}_\omega\right) \tag{3}$$

Here, we use $Q(\cdot)$ to represent the measurement of the watermark removal effect, and our objective is to optimize the UWV to suppress the watermark removal effect of the network on all watermarked images by minimizing $Q(\cdot)$. Therefore, the UWV needs to satisfy the following equation in the testing set:

$$\min_{\delta} \sum_{x_h \in X_{test}} Q\left[f\left(g\left(x_h, \omega, \delta\right)\right)\right] \tag{4}$$

Algorithm 1: The Generation process of (Mask) Universal Watermark Vaccine

Input: The set of the host images X;

Quality function $Q(\cdot)$; Iteration T;

Perturbation bound ε; Watermark Mask $\mathcal{M}$

Output: UWV/MUWV δ;

initialize $\delta \leftarrow 0$;

if δ == *'MUWV'* **then**

  $m = \mathcal{M}, Q = Q_m$

else

  $m = \mathcal{I}, Q = Q_p$

end

for $i = 1$ *To* T **do**

  if $Q\left(f\left(g\left(x_i + \delta \cdot m\right)\right)\right) > \xi$ **then**

    Compute the minimimal perturbation:

    $\Delta\delta_i \leftarrow \underset{r}{\operatorname{argmin}} \|r\|_2$

    s.t. $\mathrm{Q}(f\left(g\left(x_i + \delta \cdot m\right)\right)) \leq \psi$

    Update the perturbation:

    $\delta \leftarrow \operatorname{project}\left(\delta + \Delta\delta_i\right)$

  end

end

3.2. UWV with full images

The first approach is creating a UWV by identifying a universal perturbation that can be added to any watermarked image. Our goal is to find a perturbation δ that satisfies the condition $|\delta|_\infty \leq \varepsilon$. The algorithm starts by initializing $\delta = randInit$ and iteratively updates the minimum perturbation $\Delta\delta_i$ for each image x_i. The objective is to minimize $Q(\cdot)$ on validation data, which is defined as follows:

$$\min_{\delta} \sum_{x_h \in X_{val}} Q\left[f\left(g\left(x_h, \omega, \delta\right)\right)\right] \tag{5}$$

To measure the effectiveness of the watermark removal network, we use the Protection Loss as the effect function Q. The Protection Loss is defined as:

$$Q_P = \left\| \hat{X}_{\omega p} - \hat{x}_{\omega p} \right\|^2 \tag{6}$$

The iteration stops when a termination condition is satisfied. We define the loop termination condition as:

$$Q_p(f(g(x_i + \delta))) \leq \xi \tag{7}$$

where ξ is a small positive value (e.g., 0.03) that indicates δ satisfies this condition for image x_i.

Our optimization objective is to compute the minimal perturbation $\Delta\delta_i$ that satisfies our optimization target:

$$\Delta\delta_i \leftarrow \underset{r}{\operatorname{argmin}} |r|_2 \text{ s.t. } Q_p\left(f\left(g\left(x_i + \delta\right)\right)\right) \leq \psi \tag{8}$$

Here, ψ is not a specific value but a value infinitely close to 0, representing our expectation that Q is as small as possible.

We aim to minimize Q so that $\left\| \hat{X}_{\omega p} - \hat{x}_{\omega p} \right\|^2$ approaches 0, indicating that the output image is almost identical to the original image, and the watermark removal fails. Algorithm 1 presents the pseudocode for our approach.

3.3. UWV with mask only

The second method, called MUWV, only applies perturbation to the watermarked part of host images. This method doesn't affect the rest of the image, resulting in a cleaner image. By focusing the perturbation on the mask, the watermark protection party can reduce its impact on the image and improve its sharpness, making it more difficult for the watermark to be detected and removed. To implement this method, we improve the first approach by multiplying the applied perturbation with the mask vector.

$$\min_{\delta} \sum_{x_h \in X_{val}} Q_M\left[f\left(g\left(x_h, \omega, \delta \cdot m\right)\right)\right] \tag{9}$$

We use a novel effect function Q of the watermark removal network to improve the effectiveness, the Protection Loss is defined as follows:

$$Q_M = \left\| \hat{X}_{\omega p} \cdot m - \hat{x}_{\omega p} \cdot m \right\|^2 \tag{10}$$

The optimization objective and loop termination condition for MUWV is similar to UWV, which are defined as follows:

$$\Delta\delta_i \leftarrow \underset{r}{\operatorname{argmin}} |r|_2 \text{ s.t. } Q_M\left(f\left(g\left(x_i + \delta \cdot m\right)\right)\right) \leq \psi \tag{11}$$

where ψ is a value infinitely close to 0, representing our expectation that Q is as small as possible.

4. Experiments

4.1. Experimental Setups

Datasets. To remain consistent with the paper "Watermark Vaccine" [22], we conducted our experiment based on the same settings described in the paper. We used CLWD [23], a Color Large Scale Watermarking Dataset, which consists of unwatermarked images, watermarks, and watermarked images. We pre-trained the watermark removal network using watermarked images from the

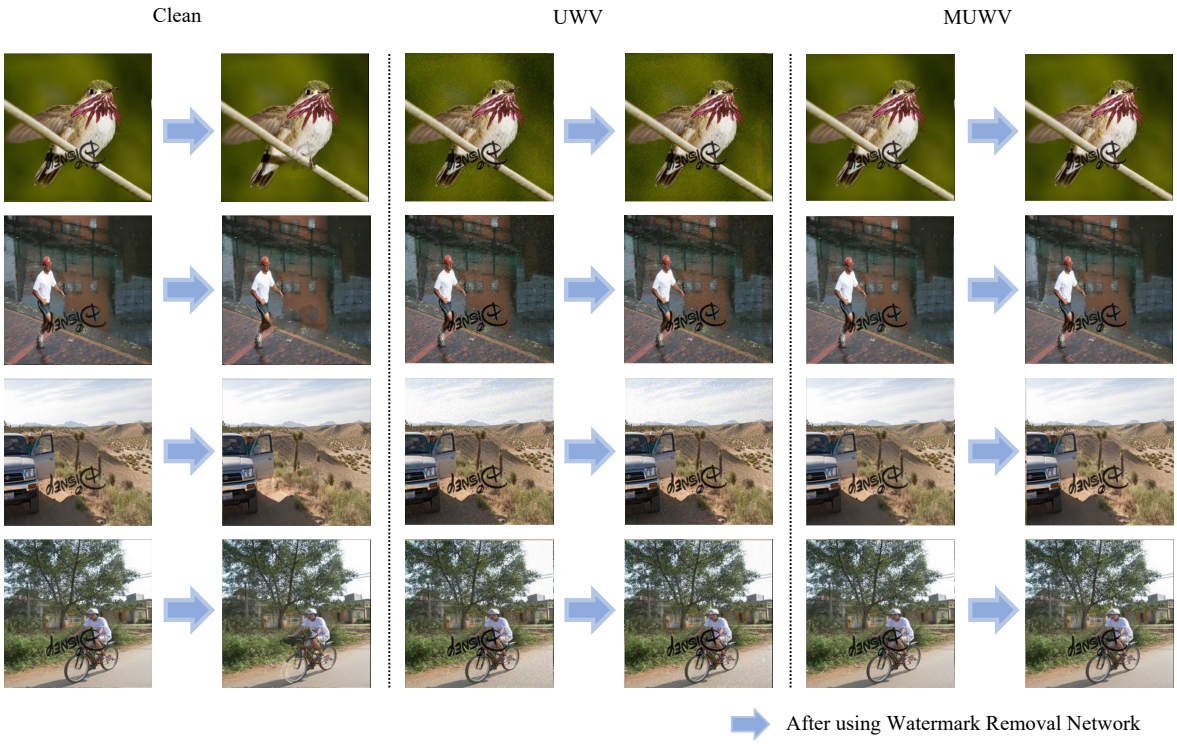


Figure 3. Universality of UWV/MUWV compared to Clean images. In each row, we show the watermark removal effects on the images without any vaccine(Clean), the images with UWV, the images with MUWV under the same watermark removal network.

CLWD training set, and in the attack phase, we added the generated perturbations as a watermark to the host image after generating UWV/MUWV using the unwatermarked image as the host image.

Models Architectures. We used an advanced network, WDNet [23] for blind watermark removal on watermarked images of CLWD. The optimal checkpoint parameters are saved after the training is completed.

Evaluation Metrics. Following the paper "Watermark Vaccine", We plan to use PSNR, SSIM, RMSE and RMSE_w as evaluation metrics, based on previous works and studies [19, 22, 23]. RMSE_w only focuses on the watermark region, while RMSE evaluates the entire image. For protection loss, we compare PSNR, SSIM, RMSE, and RMSE_w with the host image, where higher PSNR/SSIM or lower RMSE/RMSE_w indicates better performance in preserving the watermark. By analyzing the results, we can draw a conclusion on better protection performance.

4.2. Effectiveness of Watermark Vaccine

In Table 1, we compare four perturbation methods: Clean, Random Noise, UWV, and MUWV, and we also present the results in Figure 3. The main objective of our perturbation methods is to improve the quality and robustness of input images against various distortions. The clean input serves as a baseline for comparison, and we can observe that our perturbation methods, UWV and MUWV, improve the image quality metrics significantly. In particular, we can see that the PSNR and SSIM values increase for both UWV and MUWV compared to the clean input. This indicates that our perturbations lead to an increase in image quality by reducing noise and preserving the structural information of the image.

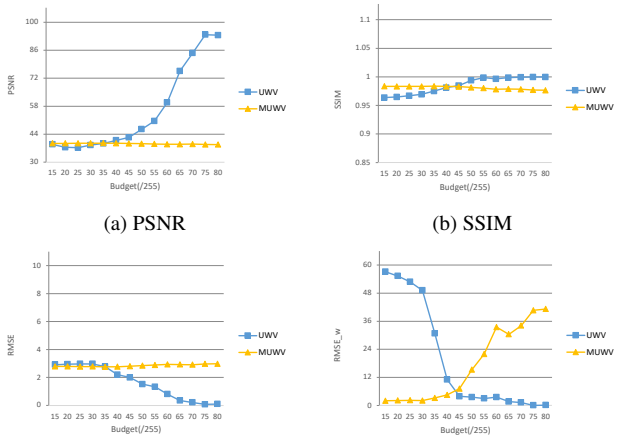
Moreover, our perturbations show a decrease in both RMSE and RMSE_w, which are measures of image error. This means that our perturbations are effective in reducing the error between the perturbed image and the original input.

In contrast, the random noise perturbation shows a decrease in the quality metrics and an increase in the error metrics, which shows that random noise is not an effective perturbation method for enhancing image quality.

Overall, our perturbation methods, UWV and MUWV, demonstrate their effectiveness in improving image quality and reducing error compared to the baseline and random noise perturbations.

Method	PSNR	SSIM	RMSE	RMSE_w
Clean	39.0806	0.9618	2.8560	59.2999
RN	39.0806	0.9618	2.8560	59.2999
UWV	**41.7725**	0.9832	10.1829	**2.1965**
MUWV	39.4947	**0.9842**	**2.5792**	2.7642

Table 1. Impact of 2 vaccines on WDNet with perturbations and random noises constrained by L_∞ norm bound 8/255. "Clean" represents the watermarked image without vaccines, and "RN" represents the watermarked image with random noise. Higher PSNR/SSIM or lower RMSE_w/RMSE_w values indicate better protection. The best results are shown in boldface.



(a) PSNR (b) SSIM

(c) RMSE (d) RMSE_w

Figure 4. The impact of watermark vaccine budgets on metrics. The x-axis shows the perturbation budgets ε/255 and the vertical axis shows the value of metrics.

4.3. Different Budget of Watermark Vaccine

The graphs in Figure 4 demonstrate the trade-off between the strength of the watermark perturbations and the resulting performance of the watermarking algorithm. The performance of a watermarking algorithm is typically evaluated based on various metrics, such as robustness, imperceptibility, and capacity.

In low-budget scenarios, MUWV shows superior performance compared to the other watermarking techniques, as demonstrated by a qualitative comparison and perturbation example shown in Figure 5, indicating that MUWV is more efficient at embedding watermarks with minimal perturbations and achieving good performance while minimizing the impact on image quality, which is particularly important in applications where image quality is a priority, such as in the field of digital art or in medical imaging.

However, as the budget increases, the performance of MUWV slightly declines. This may be due to the fact that a larger budget is applied only to the watermark image, resulting in damage to the watermark and therefore decreased performance. On the other hand, UWV's performance improves as the budget increases because it uses a global perturbation to the entire image. This makes UWV more suitable for applications where a higher level of robustness is required, such as in copyright protection for commercial images.

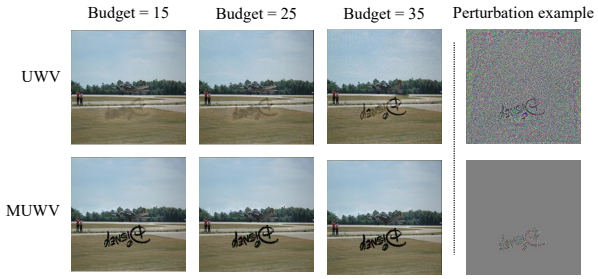


Figure 5. Qualitative comparison between UWV and MUWV in different budgets. For each row, we will display images infecting UWV/MUWV under a specific budget. Finally, we will demonstrate an example of perturbation from UWV/MUWV.

In practice, it is important to find the right balance between the strength of the perturbations and the resulting performance of the watermarking algorithm. The goal is to minimize the impact on image quality while still preventing watermark removal. Therefore, it is necessary to carefully evaluate the trade-offs between the different evaluation indicators and select the most suitable watermarking technique for the specific application.

5. Conclusion and Discussion

Watermarking techniques have been subject to attacks by malicious actors who attempt to remove or alter the watermark, making it difficult to protect the content from infringement. Existing watermark removal networks have become increasingly sophisticated, making it more challenging to protect the watermark from elimination. To address this challenge, we propose the use of universality in our watermarking frameworks, UWV and MUWV. This concept allows the watermark to be protected from elimination by existing watermark removal networks, ensuring the integrity and ownership of the content.

In our experiments, we found that the use of UWV may impact the visual quality of the image, which could affect its value and appeal to potential users. Therefore, we developed MUWV, which only perturbs the watermark part of the image, while leaving the rest of the image untouched. This results in an image that retains a higher visual quality, making it more appealing to potential users while still having a stronger ability to resist watermark removal by existing networks.

References

[1] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424, 2000. 1

[2] Gordon W Braudaway. Protecting publicly-available images with an invisible image watermark. In *Proceedings of international conference on image processing*, volume 1, pages 524–527. IEEE, 1997. 1

[3] Zhiyi Cao, Shaozhang Niu, Jiwei Zhang, and Xinyi Wang. Generative adversarial networks model for visible watermark removal. *IET Image Processing*, 13(10):1783–1789, 2019. 3

[4] Xinyun Chen, Wenxiao Wang, Chris Bender, Yiming Ding, Ruoxi Jia, Bo Li, and Dawn Song. Refit: a unified watermark removal framework for deep learning systems with limited data. In *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security*, pages 321–335, 2021. 1

[5] Danni Cheng, Xiang Li, Wei-Hong Li, Chan Lu, Fake Li, Hua Zhao, and Wei-Shi Zheng. Large-scale visible watermark detection and removal with deep convolutional networks. In *Pattern Recognition and Computer Vision: First Chinese Conference, PRCV 2018, Guangzhou, China, November 23-26, 2018, Proceedings, Part III 1*, pages 27–40. Springer, 2018. 1, 3

[6] Ingemar Cox, Matthew Miller, Jeffrey Bloom, Jessica Fridrich, and Ton Kalker. *Digital watermarking and steganography*. Morgan kaufmann, 2007. 1

[7] Xiaodong Cun and Chi-Man Pun. Split then refine: stacked attention-guided resunets for blind single image visible watermark removal. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1184–1192, 2021. 3

[8] Tali Dekel, Michael Rubinstein, Ce Liu, and William T Freeman. On the effectiveness of visible watermarks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2146–2154, 2017. 2

[9] Yosef Gandelsman, Assaf Shocher, and Michal Irani. " double-dip": unsupervised image decomposition via coupled deep-image-priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11026–11035, 2019. 2

[10] Mahmoud Hassaballah and Ali Ismail Awad. *Deep learning in computer vision: principles and applications*. CRC Press, 2020. 3

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[12] Jan Hendrik Metzen, Mummadi Chaithanya Kumar, Thomas Brox, and Volker Fischer. Universal adversarial perturbations against semantic image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 2755–2764, 2017. 2

[13] Chun-Hsiang Huang and Ja-Ling Wu. Attacking visible watermarking schemes. *IEEE transactions on multimedia*, 6(1):16–30, 2004. 2

[14] Tejas Jambhale and H Abdul Gaffar. A deep learning approach to invisible watermarking for copyright protection. In *Inventive Communication and Computational Technologies: Proceedings of ICICCT 2021*, pages 493–503. Springer, 2022. 1

[15] Xiaojun Jia, Xingxing Wei, Xiaochun Cao, and Xiaoguang Han. Adv-watermark: A novel watermark perturbation for adversarial examples. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1579–1587, 2020. 2

[16] Pei Jiang, Shiwen He, Hufei Yu, and Yaoxue Zhang. Two-stage visible watermark removal architecture based on deep learning. *IET Image Processing*, 14(15):3819–3828, 2020. 1

[17] Poonam Kadian, Shiafali M Arora, and Nidhi Arora. Robust digital watermarking techniques for copyright protection of digital data: A survey. *Wireless Personal Communications*, 118:3225–3249, 2021. 1

[18] Xiang Li, Chan Lu, Danni Cheng, Wei-Hong Li, Mei Cao, Bo Liu, Jiechao Ma, and Wei-Shi Zheng. Towards photo-realistic visible watermark removal with conditional generative adversarial networks. In *Image and Graphics: 10th International Conference, ICIG 2019, Beijing, China, August 23–25, 2019, Proceedings, Part I 10*, pages 345–356. Springer, 2019. 3

[19] Jing Liang, Li Niu, Fengjun Guo, Teng Long, and Liqing Zhang. Visible watermark removal via self-calibrated localization and background refinement. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4426–4434, 2021. 1, 3, 5

[20] Aishan Liu, Xianglong Liu, Jiaxin Fan, Yuqing Ma, Anlan Zhang, Huiyuan Xie, and Dacheng Tao. Perceptual-sensitive gan for generating adversarial patches. In *Proceedings of the AAAI conference on artificial intelligence*, 2019. 2

[21] Aishan Liu, Jiakai Wang, Xianglong Liu, Bowen Cao, Chongzhi Zhang, and Hang Yu. Bias-based universal adversarial patch attack for automatic check-out. In *ECCV*, 2020. 2

[22] Xinwei Liu, Jian Liu, Yang Bai, Jindong Gu, Tao Chen, Xiaojun Jia, and Xiaochun Cao. Watermark vaccine: Adversarial attacks to prevent watermark removal. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIV*, pages 1–17. Springer, 2022. 1, 2, 4, 5

[23] Yang Liu, Zhen Zhu, and Xiang Bai. Wdnet: Watermark-decomposition network for visible watermark removal. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3685–3693, 2021. 2, 3, 4, 5

[24] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017. 2

[25] Jaesik Park, Yu-Wing Tai, and In So Kweon. Identigram/watermark removal using cross-channel correlation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 446–453. IEEE, 2012. 2

[26] Hector Santoyo-Garcia, Eduardo Fragoso-Navarro, Rogelio Reyes-Reyes, Gabriel Sanchez-Perez, Mariko Nakano-

Miyatake, and Hector Perez-Meana. An automatic visible watermark detection method using total variation. In *2017 5th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–5. IEEE, 2017. 1

[27] Shiyu Tang, Ruihao Gong, Yan Wang, Aishan Liu, Jiakai Wang, Xinyun Chen, Fengwei Yu, Xianglong Liu, Dawn Song, Alan Yuille, et al. Robustart: Benchmarking robustness on architecture design and training techniques. *arXiv preprint arXiv:2109.05211*, 2021. 2

[28] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, Eftychios Protopapadakis, et al. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018, 2018. 3

[29] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1369–1378, 2017. 2

[30] Chaoning Zhang, Philipp Benz, Chenguo Lin, Adil Karjauv, Jing Wu, and In So Kweon. A survey on universal adversarial attack. *arXiv preprint arXiv:2103.01498*, 2021. 2

[31] Chongzhi Zhang, Aishan Liu, Xianglong Liu, Yitao Xu, Hang Yu, Yuqing Ma, and Tianlin Li. Interpreting and improving adversarial robustness of deep neural networks with neuron sensitivity. *IEEE Transactions on Image Processing*, 2021. 2

[32] Zhengxia Zou, Sen Lei, Tianyang Shi, Zhenwei Shi, and Jieping Ye. Deep adversarial decomposition: A unified framework for separating superimposed images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12806–12816, 2020. 2