BWA: Research on Adversarial Disturbance Space Based on Blind Watermarking and Color Space

Ziwei Xu[1](✉), Chunyang Ye[1,2], and Shuaipeng Dong[2]

[1] School of Cyberspace Sevurity, Hainan University, Haikou 570228, China
zwx2642@163.com

[2] School of Computer Science and Technology, Hainan University, Haikou 570228, China

Abstract. Effective generation of adversarial examples can help to improve the training of neural models to avoid adversarial example attacks. Watermark-based adversarial example generation methods regard watermark as a meaningful noise to perturb the neural models. Therefore, the resulting adversarial examples are more similar to the original images yet more difficult to defend. Existing Watermark-based adversarial example generation methods adopt the visible watermarking technology. This however may reduce the success rate of the attacks because the adversarial examples with visible watermarks can be easily perceptible by humans. To address this issue, we propose a novel approach to generate adversarial examples based on the combination of frequency domain and color space perturbation. In particular, we use wavelet transform to hide the watermark, making it invisible and introducing noises to the frequency of the images. We then select the Lab color space Similarity as an optimization scheme for perturbations control. Experimental results show that under the same dataset, the maximum attack success rate of the adversarial example generated by our algorithm can reach 98.56%. In addition, the generated adversarial examples are highly portable, the successful attacks on VGG, Resnet101, and Inception-v3 can reach more than 95%, and the color space perturbation optimization achieves an average RGB channel similarity of 97.22%.

Keywords: Adversarial examples · neural networks · blind watermarks

1 Background

Deep Neural Networks show excellent performance in different fields [1], such as image classification [2, 3], text analysis [4], speech recognition [5], to name a few. However, deep learning models are often vulnerable to well-designed adversarial attacks, resulting in immeasurable security problems. For example, the existence of adversarial samples threatens [6] driving face recognition and road signs. In terms of voice, it causes instruction recognition errors, speaker recognition errors, information leakage, and even unintelligible problems. To improve the training of the neural models to avoid adversarial sample attacks, effective generation of adversarial examples is needed.

J. C. Hung et al. (Eds.): IC 2023, LNEE 1045, pp. 678–688, 2023.
https://doi.org/10.1007/978-981-99-2287-1_95

Currently, numbers methods have been came up to create adversarial examples [7] demonstrated that it is possible to add subtle perturbations to images that are imperceptible to humans, thus misleading deep neural network image classifiers to make wrong classifications. Goodfellow observed that in high-dimensional spaces, the linear behavior of deep neural networks can be utilized by adversarial examples, and proposed a Fast Gradient Sign Attack method to effectively calculate adversarial perturbations and generate adversarial examples [8]. The Deepfool algorithm [9] generates the smallest normative adversarial perturbation by iterative pushing the images to the classification boundary. By limiting L∞, L2 and the L0 norm makes the perturbation imperceptible, the optimization-based method C&W [10] search for adversarial examples with smaller perturbations amplitude. Typically, researchers use the L2 norm limit to evaluate distortion [11] (as a measure of perceptual similarity), because the attack strategy tricks the classifier by adding noise. L2 similarity has a distinct feature: it is highly sensitive to sample illumination and viewpoint changes [12], so this metric is not optimal. Different from other attacks, image watermarking is added to the original image as a meaningful noise without affecting people's recognition of the image. However, the disadvantage is that the adding of watermarks into the original image makes the difference between images large, and exposing the visible watermark information are more likely to raise suspicions about images with a high security factor.

To solve this problem, we put forward an adversarial examples generation method based on blind watermarking. To hide the watermark in the image, we use blind watermark to add peturbations in the frequency domain. In particular, in the frequency domain, add a blind watermark to the image with a random number in the transform domain. Then, the image is converted back to the original domain so that the difference between them cannot be identified by human eyes. We also add tiny color perturbations in the Lab color space to further optimize the image with a better attack effect and similarity with the original image. In this way, we can generate stable and strong general perturbations without a large amount of data.

The paper has the following outstanding contributions: First, we propose a method of making adversarial examples based on blind watermarking. Compared with the existing watermarking method, our method has stronger aggressiveness and better concealment. Second, by adding tiny color perturbations in the Lab color space, stable and aggressive perturbations can be generated without a large amount of data. This increases the adaptability and robustness of our approach. Third, we conduct extensive experiments to evaluate our proposal. Through experiments, the adversarial examples generated by our algorithm under the same dataset have a maximum attack success rate of 98.56%. In the test of attacking Vgg, it is 5.6% higher than the existing methods on average, and in the test of Resnet101 and Inception-v3. In addition, the generated adversarial examples are highly portable, and the successful attacks on VGG, Resnet101, and Inception-v3 can reach more than 95%, and the color space perturbation optimization achieves an average RGB channel similarity of 97.22%. We also analyze various factors of adding watermark, and discuss the influence of attack rate and image similarity.

The overall organizational structure of the rest of this article is as follows. Section 2 presents the background and related work on adversarial examples. Section 3 introduces the adversarial sample based on blind watermark and its improved idea of adding color perturbation, and the fourth part details the proposed adversarial method, including the comparison of some indicators. The last section summarizes our work and looks forward to future research directions.

Subsequent paragraphs, however, are indented.

2 Related Work

2.1 Adversarial Attack

Gradient-based attack methods include Carlini and Wagner attack (C&W) [13], Deepfool [9], JSMA [14]. Most of these attacks target image classification. However, with the deepening of research, adversarial examples not only attack image classification, but also become more and more popular in other computer vision tasks. [16, 17] Sharif proposed to estimate the prediction score of the model gradient using finite differences. These iterative attacks estimate the gradient by sampling from the noise distribution around the feature points. While this approach is successful, it requires a lot of model queries. Adversarial Transformation Network (ATN) [18] propose an autoencoder-based network to create adversarial examples. [19] Gragnaniello adopted a GAN network from which to create adversarial examples. However, since the adversarial samples have no direct correspondence with the original images, the perturbations may be very obvious and fail to deceive the human eye. The improved method proposed by [20] adopts the AdvGAN generator based on the auto-encoder to obtain the maximum range perturbation from the perspective of the original image. There are novel research algorithms such as one-pixel-attack [21], which modifies a single pixel point by random check and optimization to attack; there are also patch-based adversarial examples. [22–24] all paste patches on the original image, and only update the parameters of the patch by improving the loss backpropagation. The innovation lies in how the loss function is designed. However, online iterative attack strategies limit their application scenarios, in order to generate adversarial sequences, the only downside is that they always require access to the model‘s weights during the attack.

2.2 Visible Watermarking Methods

Digital watermarking is a kind of protection information embedded in the carrier file by applying computer algorithm. Digital watermarks can guarantee the security of information and protect the copyright of works. [25] proposed a blind technique based on fast Walsh-Hadamard transform, SVD, key mapping and coefficient sorting, but its effect against geometric attacks such as rotation and shearing is relatively weak. Difference. The above algorithms have caused varying degrees of changes to the image data. [26] proposed a robust watermarking scheme that exploits the multi-resolution and multi-scale properties of nonsub exampled wavelet transforms to analyze the orientation features of a given image. [7] Jia proposed a new optimization algorithm, when adversarial watermarks are generated using evolutionary algorithms (BHE) in a black-box attack environment, the location of the watermark and the watermark are highly correlated with the transparency attribute. In terms of the similar distance from the original image, the addition of visible watermarks will cause suspicion. In order to avoid this situation, we propose blind watermark perturbations. Considering the security issues of the Internet cannot be ignored, add an invisible watermark to protect their Copyright, which enables copyright protection of images while conducting adversarial attacks with better robustness.

3 Methodology

3.1 Research Ideas

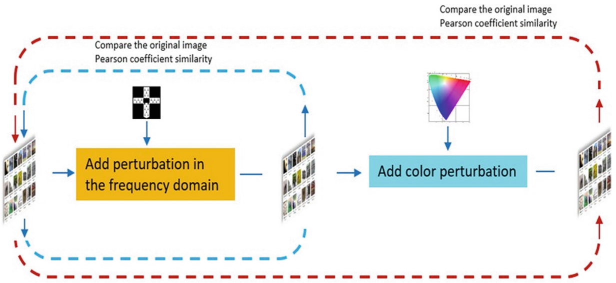
Compare the original image
Pearson coefficient similarity
Compare the original image
Pearson coefficient similarity
Add perturbation in the frequency domain
Add color perturbation

Fig. 1. The process of adversarial example making.

We add the watermark to the three color gamuts of RGB through wavelet transform and singular value decomposition, respectively, with the set random number seeds. In the transform domain algorithm, the multi-resolution characteristics of the transform domain and the inherent characteristics of the SVD singular value are fully utilized to enhance the invisibility and robustness of the watermark. Accordingly, we propose algorithm attack model (Fig. 1). We disguise the adversarial perturbation as a frequency-domain blind watermark. Our work is to generate adversarial images that cannot be classified correctly, the optimization or constraint of which is expressed by the following formula:

$$\begin{aligned} &minimize\ D(x, x+\theta) \\ &such\ that\ C(x+\delta)=t \\ &x+\delta \in [0,1]^n \end{aligned} \tag{1}$$

where: x ∈ Rm is a clean input, δ is the perturbation added, D is the distance metric, which measures the distance between the original image and the antagonistic sample, C is the classifier, t is the label of the misclassification of the adversarial example, $[0, 1]^n$ limits the perturbation between (0,1).

3.2 Algorithm Details

Using wavelet transform can improve the visual concealment and robustness of the watermark. Through the transformation, the features can be fully highlighted, localized analysis of temporal and spatial frequencies can be performed, and through a series of operations such as zooming, panning, etc., the purpose of refining the scale on the signal is achieved, and finally time subdivision is automatically realized at high frequencies, and frequency subdivision is automatically realized at low frequencies. Therefore, it will not miss every detail of the signal, and adding disturbances in the frequency domain can be more invisible. First complete the wavelet transform processing and singular value decomposition operations, we randomly add the watermark to the RGB color gamuts with the same random number seed, and add images to the wavelet frequency domain, and then improve the process of adding color disturbance (see Fig. 2). First of all, image scrambling is a mapping of two-dimensional space, and the function of scrambling is to change the arrangement and combination of the original images and the spatial correlation. For an image W of size N × N, use formula (1) to perform Arnold transform:

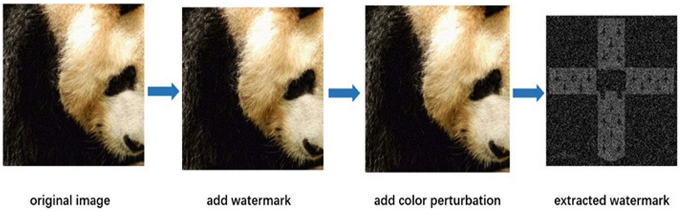

Fig. 2. From left to right, the images are the original image, the watermarked image, the image with added color disturbance, and the extracted watermark.

image W of size $N \times N$, use formula (1) to perform Arnold transform:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ k & k+1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} (modN) \tag{2}$$

In formula:

(x, y) is the pixel point of the original image, (x′, y′) is the pixel point of the new image after transformation, N is the image order, that is, the size of the image, generally considering a square image, k is an integer belonging to [1, N].

Denote the transformation matrix as W, thus, do the iterative procedure:

$$I_{xy}^{n+1} = WI_{xy}^{n}(modN), I_{xy}^{n} = (x, y), n = 0, 1, 2, ... \tag{3}$$

Singular value decomposition (SVD) in numerical analysis is a numerical algorithm that diagonalizes a matrix. From a linear algebra perspective, a grayscale image can be viewed as a non-negative matrix. The image W′ ∈ Rm × n, where R represents the real number domain:

$$W = U\Sigma V^T \tag{4}$$

Then perform first-level wavelet decomposition on the carrier image to obtain 3 high frequency subbands HH, LH, HL and 1 low frequency subband LL; divide the low frequency subband image into m blocks. Repeat the singular value decomposition and embed the watermark according to a certain intensity factor until all the watermark information is embedded, and then perform the inverse wavelet transform to obtain the watermarked image.

3.3 Loss Function

The parameter settings of adding watermark have different effects. When the depth is greater, the attack ability is stronger, but the picture changes will be blurred. The perturbation in the color space will not affect the perturbation in the frequency domain, and the watermark can still be extracted. Lab is designed based on people's perception of color, more specifically, it is perceptually uniform, if the number (the three channels of L, a, b) changes in the same magnitude, then it gives people a visual effect.

$$\begin{aligned} \Delta E_{00} &= \sqrt{\left(\frac{\Delta L'}{k_L S_L}\right)^2 + \left(\frac{\Delta C'}{k_C S_C}\right)^2 + \left(\frac{\Delta H'}{k_H S_H}\right)^2 + \Delta R}, \\ \Delta R &= R_T\left(\frac{\Delta C'}{k_C S_C}\right)\left(\frac{\Delta H'}{k_H S_H}\right) \end{aligned} \tag{5}$$

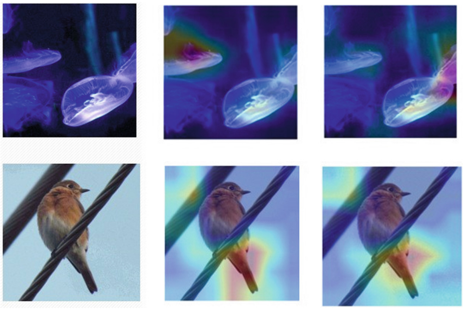

Fig. 3. The first column: original image. The second column: thermal map shown in the original image. The third column: thermal map of the image subjected to BWA attack. It shows that the feature regions that neural networks pay more attention to are not global.

4 Evaluation

4.1 Comparing the Results of Attacking Different Networks with Other Methods

Compared with the watermarking algorithm, our BWM algorithm has better concealment. Figure 3 shows the comparison of the results of our method attacking different networks. Our method is compared with adv-watermark, one-pixel, FGSM, advGan methods respectively. The dataset adopts ImageNet datasets. We can see that the advgan method has the best effect and achieves the highest attack rate. In Resnet101 and Inception-v3, BWA'rate reaches the highest.

Table 1. Attack rate of different networks.

Adv-methods	VGG	Resnet101	Inception-v3
Adv-watermark	0.934	0.873	0.921
One-pixel	0.935	0.920	0.876
FGSM	0.827	0.881	0.955
AdvGan	0.973	0.923	0.970
BWA	0.954	0.952	0.976

The random number used for S_1 watermark encryption, S_2 is the random number for adding the watermark to the picture, mod_1, mod_2 are used for the divisor of the embedding algorithm. In theory, and the larger the divisor, the stronger the robustness, but the greater the distortion of the output image (see Table 1). Figure 4 shows the impact of different parameters on the attack power, the larger the mod_1, the greater the distortion of the picture, and the color space will have a certain blur and change.

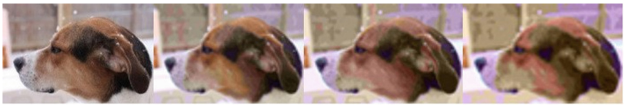

Fig. 4. The parameters of the last three images are shown in Error! Reference source not found**.** With the increase of mod1, the color does not change much, and the texture is gradually blurred. It can be seen that the embedding depth of watermark first affects the texture features.

Table 2. The effect of different parameters.

Parameters	Example	Attack rate	Similarity (R G B)		
s1, s2	mod1,mod2		R	G	B
5539,3336	56,25	0.950	0.9665	0.9899	0.9800
5539,3336	79,25	0.954	0.9487	0.9844	0.9702
5539,3336	108,25	0.954	0.9265	0.9755	0.9526
5539,3336	175,25	0.967	0.9379	0.8945	0.9673

The same analysis can be analyzed as shown in Table 2. When S_1, S_2, and mod_2 are set to 8399, 5536, and 25, respectively, under the constant random number setting, as mod_1 becomes larger, the attack rate can also increase to a certain extent (see Fig. 5).

4.2 The Effect of Watermark Times

The number of watermark additions also affects the attack capability. The more times the watermark is added, the greater the disturbance added in the frequency domain space, and the greater the impact on the original image (see Table 4)

Table 3. The effect of different parameters.

Parameters	Example	Attack rate	Similarity (R G B)		
s1, s2	mod1,mod2		R	G	B
8539,5536	176,25	0.963	0.8607	0.9432	0.9036
8539,5536	258,25	0.958	0.7846	0.8949	0.8286
8539,5536	296,25	0.977	0.7108	0.8892	0.7814
8539,5536	336,25	0.983	0.6670	0.8497	0.7312

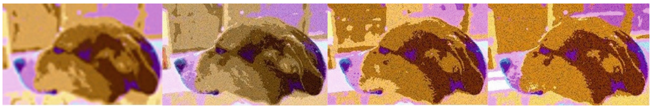

Fig. 5. The parameters of the next four pictures are shown in Error! Reference source not found. Slight changes in color and texture can be clearly seen, and high frequency features are destroyed, and the attack rate increases at the expense of the similarity of the images.

Table 4. The effect of watermark times ($s_1 = 8399$, $s_2 = 5536$, $mod_1 = 258$, $mod_2 = 25$).

Times	Similarity			Attack rate
	R	G	B	
1	0.8607	0.9432	0.9036	0.958
2	0.7846	0.8949	0.8226	0.962
3	0.7108	0.8892	0.7814	0.963
4	0.6670	0.8497	0.7321	0.963

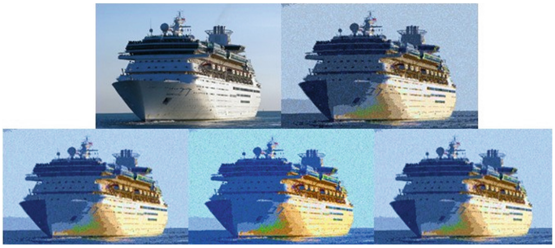

Fig. 6. For the original image, you can refer to the picture on the left in the first row, the pictures are watermarked 2, 3, and 4 times in sequence.

4.3 The Effect of Color Space Perturbation on Watermark Extraction

Before and after the watermark is embedded, the human eye cannot directly perceive the existence of the watermark, which has good concealment of the watermark. We added color perturbation to make the adversarial example image more closer to the initial image (see Fig. 6). Compared with the original watermarking attack, adding color space perturbation has a higher similarity, the results of color perturbation are shown in Table 5. The extracted watermark is a grayscale image (see Figs. 7 and 8).

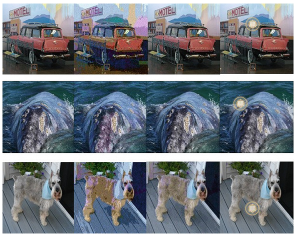

Fig. 7. The first column is the original picture, the middle column is the picture that has been watermarked once, the third column is the picture after adding color perturbation, and the fourth column is the confrontation sample of the Adwatermark algorithm.

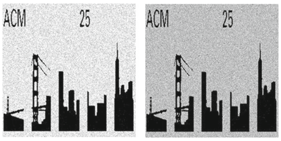

Fig. 8. The left picture: extracted watermark.

The right picture: watermark extracted after adding color disturbance.

Table 5. The effect of color space perturbation on watermark extraction (s1 = 8399, s2 = 5536 mod1 = 176, mod2 = 25).

Method	Similarity R	G	B	Attack rate
No color perturbation	0.8607	0.9432	0.9036	0.964
add color perturbation	0.9489	0.9824	0.9853	0.976

5 Conclusion

Our paper proposes a black-box attack, which adds perturbation in the form of blind water- mark in the wavelet frequency domain, and converts it to RGB space with a certain attack ability. We introduced the Lab color space for optimization, and added color disturbance in the lab color gamut. The results demonstrate that our method can successfully attack several networks. We hope that more researchers will pay attention to adversarial attacks and defenses in the field of neural networks in the future.

Acknowledgment. This work was supported in part by the Key Research and Development Program of Hainan Province under grant No. ZDYF2020008, ZDYF2020008, the Natural Science Foundation of Hainan Province under the grant No. 2019RCO88, 2019CXTD400, and grants from State Key Laboratory of Marine Resource Utilization in South China Sea and Key Laboratory of Big Data and Smart Services of Hainan Province.

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
2. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Adv. Neural Inform. Process. Syst. **25**(2012)
3. Kurakin, A., Goodfellow, I., Bengio, S., et al.: Adversarial examples in the physical world. In: ICLR Workshop (2016)
4. Collobert, R., Weston, J.: A unified architecture for natural language processing: Deep neural networks with multitask learning. In: Proceedings of the 25th International Conference on Machine Learning, pp. 160−167 (2008)
5. Hinton, G., et al.: Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process. Mag. 29(6), 8297 (2012)
6. He, W., Wei, J., Chen, X., Carlini, N., Song, D.: Adversarial example defenses: ensembles of weak defenses are not strong (2017). https://arxiv.org/abs/1706.04701
7. Jia, X., Wei, X., Cao, X., Han, X.: Adv-watermark: a novel watermark perturbation for adversarial examples (2020). https://arxiv.org/abs/2008.01919
8. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: 2017 IEEE Symposium on Security and Privacy (sp), pp. 3957. IEEE (2017).https://arxiv.org/abs/1412.6572
9. Moosavi-Dezfooli, S-M., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2574–2582 (2016). https://doi.org/10.1109/CVPR.2016.282
10. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Berkay Celik, Z., Swami, A.: The limitations of deep learning in adversarial settings (2015). https://arxiv.org/abs/1511.07528
11. Gu, S., Rigazio, L.: Towards deep neural network architectures robust to adversarial examples. In: ICLR Computerence (2015)
12. Johnson, J., Alahi, A., Li, F-F.: Perceptual losses for real-time style transfer and super-resolution (2016). https://arxiv.org/abs/1603.08155
13. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP)
14. Croce, F., Hein, M.: Sparse and imperceivable adversarial attacks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4724–4732 (2019)
15. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks (2017). https://arxiv.org/abs/1706.06083
16. Engstrom, L., Tran, B., Tsipras, D., Schmidt, L., Madry, A.: Exploring the landscape of spatial robustness (2017). https://arxiv.org/abs/1712.02779
17. Sharif, M., Bauer, L., Reiter, M.K.: On the suitability of lp-norms for creating and preventing adversarial examples (2018). https://arxiv.org/abs/1802.09653
18. Eykholt, K., et al.: Robust physical-world attacks on deep learning models (2017). https://arxiv.org/abs/1707.08945

19. Gragnaniello, D., Marra, F., Poggi, G., Verdoliva, L.: Perceptual quality-preserving black-box attack against deep learning image classifiers (2019). https://arxiv.org/abs/1902.07776
20. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks (2015a). https://arxiv.org/abs/1506.01497
21. Su, J., Vargas, D.V., Sakurai, K.: One pixel attack for fooling deep neural networks. IEEE Trans. Evol. Comput. **23**(5), 828841 (2019). https://doi.org/10.1109/tevc.2019.2890858
22. Brown, T.B., Mane, D., Roy, A., Abadi, M., Gilmer, J.: Adversarial patch. arXiv preprint arXiv:1712.09665 (2017a)
23. Lee, M., Kolter, Z.: On physical adversarial patches for object detection. arXiv preprint arXiv:1906.11897 (2019b)
24. Thys, S., Van Ranst, W., Goedeme, T.: Fooling automated surveillance cameras: adversarial patches to attack person detection (2019b). https://arxiv.org/abs/1904.08653
25. Khanam, T., Dhar, P.K., Kowsar, S., Kim, J-M.: SVD-based image watermarking using the fast walsh-hadamard transform, key mapping, and coefficient ordering for ownership protection. Symmetry **12**(1), 52, (2019). https://doi.org/10.3390/sym12010052
26. Zhao, J., Xu, W., Zhang, S., Fan, S., Zhang, W.: A strong robust zero-watermarking scheme based on shearlets high ability for capturing directional features. Math. Probl. Eng. **2016** (2016). https://doi.org/10.1155/2016/2643263
27. Jiang, F., Gao, T., Li, De.: A robust zero-watermarking algorithm for color image based on tensor mode expansion. Multim Tools Appl. **79**(11), 75997614 (2020). https://doi.org/10.1007/s11042-019-08459-3
28. Liu, X., Yang, H., Liu, Z., Song, L., Li, H., Chen, J.: Dpatch: an adversarial patch attack on object detectors. (2018a). https://arxiv.org/abs/1806.02299
29. Ye, M., Luo, J., Zheng, G., Xiao, C., Wang, T., Ma, F.: Medat- tacker: exploring black-box adversarial attacks on risk prediction models in healthcare (2021). https://arxiv.org/abs/2112.06063
30. Zheng, X., Fan, Y., Wu, B., Zhang, Y., Wang, J., Pan, S.: Robust physical-world attacks on face recognition (2021). https://arxiv.org/abs/2109.09320
31. Tram‘er, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., McDaniel, P.: Ensemble adversarial training: attacks and defenses (2017). https://arxiv.org/abs/1705.07204
32. Sharif, M., Bhagavatula, S., Bauer, L., Reiter, M.K.: A general frame work for adversarial examples with objectives. ACM Trans. Privacy Secur.**22**(3), 130 (2019b)