Hindawi
Discrete Dynamics in Nature and Society
Volume 2024, Article ID 9753490, 1 page
https://doi.org/10.1155/2024/9753490

Retraction

Retracted: Adversarial Attacks Defense Method Based on Multiple Filtering and Image Rotation

Discrete Dynamics in Nature and Society

Received 23 January 2024; Accepted 23 January 2024; Published 24 January 2024

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] F. Li, X. Du, and L. Zhang, "Adversarial Attacks Defense Method Based on Multiple Filtering and Image Rotation," *Discrete Dynamics in Nature and Society*, vol. 2022, Article ID 6124895, 11 pages, 2022.

Hindawi
Discrete Dynamics in Nature and Society
Volume 2022, Article ID 6124895, 11 pages
https://doi.org/10.1155/2022/6124895

Research Article

Adversarial Attacks Defense Method Based on Multiple Filtering and Image Rotation

Feng Li, Xuehui Du, and Liu Zhang

Information Engineering University, Zhengzhou 450000, China

Correspondence should be addressed to Feng Li; fengli_edu@21cn.com

Received 21 October 2021; Revised 24 November 2021; Accepted 2 December 2021; Published 16 April 2022

Academic Editor: Ahmed Farouk

Adversarial examples in an image classification task cause neural networks to predict incorrect class labels with high confidence. Many applications related to image classification, such as self-driving and facial recognition, have been seriously threatened by adversarial attacks. One class of the existing defense methods is the preprocessing-based defense which transforms the inputs before feeding them to the system. These methods are independent of the classification models and have excellent defensive effects under oblivious attacks. An image filtering method is often used to evaluate the robustness of adversarial examples. However, filtering induces the loss of valuable features that reduce the classification accuracy and weakens the adversarial perturbation. Furthermore, the fixed filtering parameters cannot effectively defend against the adversarial attack. This paper proposes a novel defense method based on different filter parameters and randomly rotated filtered images. The output classification probabilities are statistically averaged, which keeps the classification accuracy while removing the perturbation. Experimental results show that the proposed method improves the defense capability of various models against diverse kinds of oblivious adversarial attacks. Under the adaptive attack, the transferability of the adversarial examples among different models is significantly reduced.

1. Introduction

Deep learning models, especially the Convolutional Neural Network (CNN), have been rapidly advanced in many tasks, from image classification to object detection, image title generation, speech recognition, and text recognition. Consequently, they have been widely applied in various practical applications such as a self-driving car, significantly reducing the image processing related-manual labour. However, the emergence of adversarial examples poses a severe security threat to applying deep learning models. Szegedy et al. [1] showed for the first time that neural networks could make false classification by adding a small amount of human-unobtrusive disturbance to images. Since then, a variety of attacks have been studied by researchers. The adversarial example attack methods of the deep neural network for the image classification task are divided into four categories: neighbourhood search-based attacks, gradient-based attacks, optimization-based attacks, and image transformation-based attacks. The neighbourhood search-based attack methods adopt the brute force search scheme to find adversarial examples in the neighbourhood of an attack-free image to deceive the classification model. For example, the single-pixel attack [2] generated adversarial examples by changing the value of a pixel in the image. To construct the loss function, the gradient-based attack method replaces the correct class label with the target class label. It generates adversarial examples in the direction of gradient descent to attack. This category includes L-BFGS (Limited Memory, Broyden, Fletcher, Goldfarb, Shanno) [1], Fast Gradient Sign Method (FGSM) [3], Iterative FGSM (I-FGSM) [4], and the more aggressive Project Gradient Descent (PGD) attack [5]. Moosavi et al. [6] found that the gradient-based attack method can generate the adversarial perturbation that has strong generality. Based on this, a Universal Adversarial Perturbation (UAP) attack was proposed. In gradient-based attacks, adversarial examples are further generated with less perturbation and more difficulty to detect by adding optimization methods.

DeepFool attack [7] pushed the sample across the nearest classification boundary and made the generated adversarial

perturbation smaller. Jacobian-based Saliency Map Attack (JSMA) [8] could trick the classification model by changing only a small number of pixels. Carlini and Wagner (C&W) attack [9] adopted multiple optimizations, achieving a better attack effect. Further, adversarial examples could be generated through the transformation of images. Xiao et al. [10] used spatial transformation to attack. In [11], an image processing approach was used to extract noise as an adversarial perturbation to generate adversarial examples. The existence of adversarial examples indicates that deep learning models are very fragile. Research on adversarial examples and defense methods can help improve the robustness and security, which is of great significance for the practical application of models.

The adversarial example can be regarded as the original picture plus the adversarial perturbation. Adversarial perturbation is intuitively reflected as a noise signal similar to a specific texture. Domain-adversarial training [12, 13] was proposed to improve the robustness of CNN by increasing shape bias. However, [14] showed no defense effect against adversarial example attacks. Therefore, domain-adversarial training, aimed to increase shape preference, did not enhance the model's defense against the attacks. Further studies for defense methods should consider the causes and attack methods of the adversarial example attacks.

The existence of an adversarial example comes from the deep linearity of the model, and the defense against adversarial example attack needs to prevent the attacker from using this deep linearity. Defense methods can be divided into two categories: model reinforcement methods and image preprocessing-based methods. Model reinforcement methods include adversarial training [3], Thermometer Code (TE) [15], Distillation [16], and PixelDefend [17]. These methods provide a good defense effect, but the training cost is high. Furthermore, it is highly dependent on the model and attack method. The image preprocessing-based defense adopted rotation, scaling, JPG compression [18], randomization [19], wavelet transform [20, 21], Principal Component Analysis (PCA) [22], and median blurring [23]. The preprocessing-based defense is independent of the deep learning model and dataset. Moreover, the defense can be achieved by destroying the gradient and the integrity of adversarial perturbation, which can improve the defense capability of any model. It can be used alone or in combination with other methods. However, the preprocessing-based defense still has two shortcomings: it weakens the adversarial perturbation and reduces the classification accuracy of the model for clear images. In addition, adaptive attacks have been shown to bypass such defenses. For example, Athalye et al. [24] proved that the obfuscated gradients generated by image preprocessing could be bypassed. Backward Pass Differentiable Approximation (BPDA) was proposed to attack defenses with one or more nondifferentiable components. Since the Expectation over Transformation (EOT) [25] method has been a standard technique for computing gradients of models with randomized components, a gradient-based attack can be used to attack such defense. The adaptive attack proposed by Tramer et al. [26] can bypass various defense methods based on image preprocessing.

The filtering method can take advantage of the fact that the adversarial perturbation is mostly high-frequency noise whose intensity is relatively small compared with the original image. A median filter, Gaussian filter, and bilateral filter [27] were commonly used to filter out the high-frequency components, reducing the influence of the adversarial perturbation. According to [4], Gaussian filtering defended part of the samples against attack in sacrificing the significant classification accuracy. In low-pass filters such as mean filter [28] and Gaussian filter, the high-frequency components can be effectively removed, retaining the structured information of the image. However, the useful edge information is weakened, reducing the classification ability of the model. Bilateral filtering is an advanced filter that not only eliminates the noise but also effectively retains the edge information. In such a way, the visual quality of the image is enhanced while resisting adversarial attacks.

The accuracy degradation makes the filtering method infeasible to defend against the adversarial attack. In order to overcome this problem, this paper proposes to use multiple bilateral filters combined with image rotation as an effective and efficient defense method (Figure 1). In a nutshell, most of the existing methods have the following issues:

(1) Model reinforcement methods suffer from the limitation of high training cost

(2) Model reinforcement methods are highly dependent on the model and attack method

(3) Most image preprocessing-based weakens the adversarial perturbation and reduces the classification accuracy of the model for clear images

(4) Filtering methods' accuracy degrades in defending against adversarial attacks

Therefore, it is need of the hour to develop a more secure defense method.

In order to address above mentioned issues, this work contributes to improving the defense procedure in the following aspects:

(i) It proposed a two-stage filtering method to maintain better image quality.

(ii) It proposed using prefiltering to reduce high-frequency components in an image. It is followed by using a larger size of the filter to further reduce the adversarial perturbation components.

(iii) The use of random parameters increases the uncertainty of the defense method and the difficulty of the attack.

(iv) It proposed rotating the filtered images before classification.

(v) It adjusted the parameters of the bilateral filter and statistically averaged multiple classification probability.

Adversarial example attacks can be divided into the white-box attack, black-box attack, and gray-box attack. The white-box attack refers to the attack designed after the attacker has full knowledge of the parameters and structure

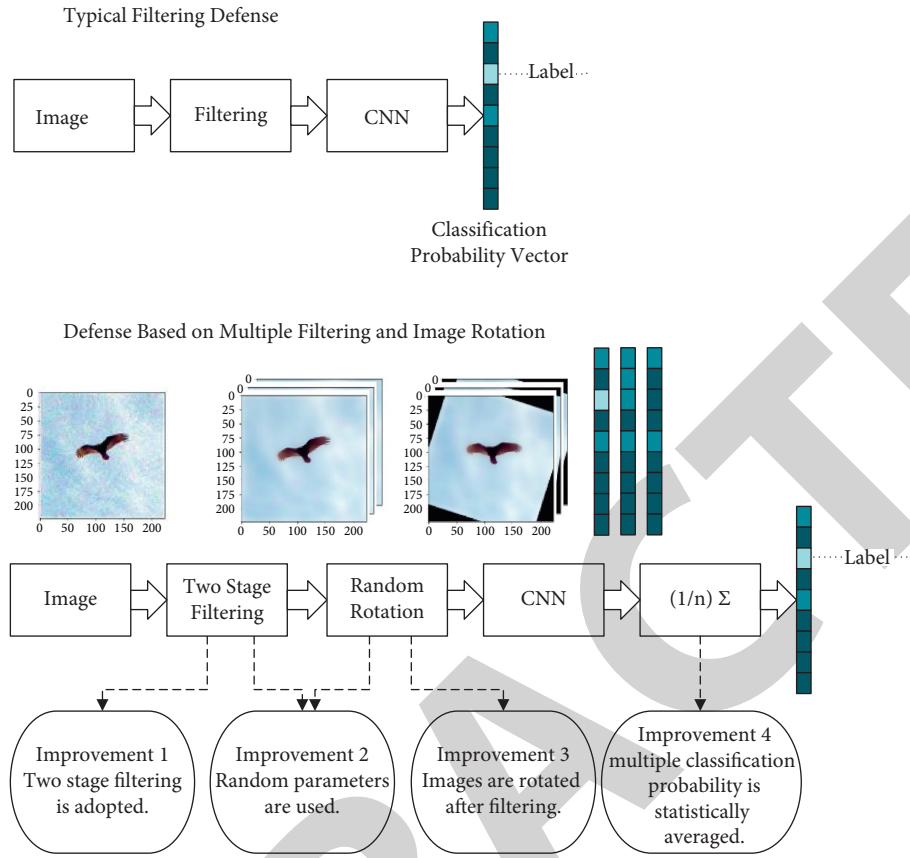

Figure 1: The improved defense procedure by the proposed method in four aspects.

of the target model. In contrast, the black-box attack usually refers to the attack designed when the attacker can only access the output of the target model but cannot obtain the internal information of the model. The gray box attack is between white-box and black-box attacks, and only part of the model and defense approach is known. In the proposed method, random filters and image rotation are used, which lead to a random gradient. Both gray and white box attacks were used to test the defense performance, respectively. The gray-box attack method used in this paper, called oblivious attack, is that the attacker does not know the defense method and only attacks according to the classification model. The white-box attack method used in this paper, called adaptive attack, is that attacks are conducted according to both the classification model and defense method.

The experimental results show that the proposed method achieves excellent defensive efficacy against all kinds of oblivious attacks while only slightly sacrificing the image classification accuracy of some models. Furthermore, the proposed method reduces the transferability of the adversarial examples under adaptive attack.

The rest of the paper is organized as follows: Section 2 describes the proposed defense algorithm. Section 3 presents the results of experiments in different scenarios and provides critical analysis. Finally, the paper is concluded in Section 4.

2. The Proposed Defense Algorithm

The image filtering-based adversarial attack defense methods are independent of the deep learning model and the attack types. However, the image quality could be decreased, which potentially reduces the accuracy of the classification model. How to maintain the image quality while filtering the adversarial perturbation is the key to this approach.

2.1. Defense Framework. In order to reduce the influence of adversarial perturbation, the defense method reformulates single image classification as multiple image classification through two-stage filtering and image rotation. The multiple corresponding classification probability vectors are then averaged, and the index of the maximum probability is considered the true class label of the image.

As shown in Figure 2, the defense method consists of four parts:

(1) The input image is prefiltered to remove the high-frequency component. The first layer of CNN usually takes a small convolution kernel, and high-frequency signals will greatly impact classification. Furthermore, adversarial perturbation contains high-frequency components. Thus, filtering out high-frequency components can effectively reduce the impact of adversarial perturbation.

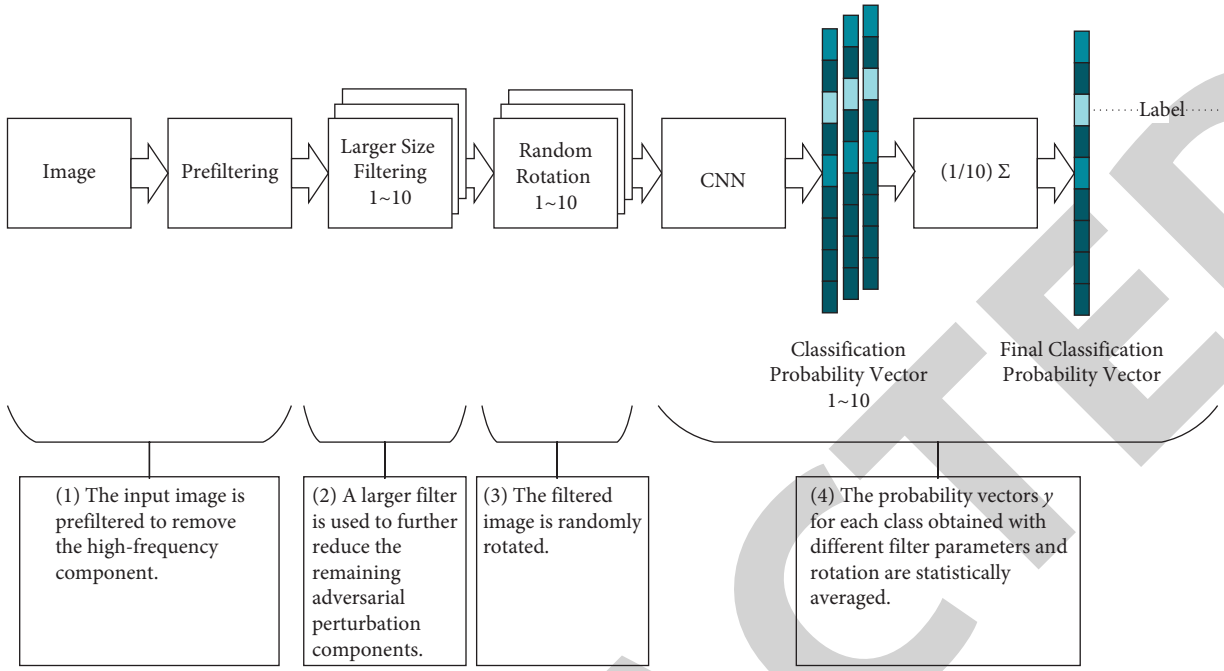

Figure 2: Defense framework.

(2) A larger filter is used to further reduce the remaining adversarial perturbation components. Since the high-frequency component of the adversarial perturbation has been reduced during the prefiltering process, the remaining perturbation has a relatively small color variance. It would be more efficient to use a larger filter for the remaining perturbation. As the intensity of adversarial perturbation is small, a small color variance in filtering can effectively reduce the adversarial perturbation and maintain the image quality. This filtering method mainly reduces the chroma stripes in the dark area of the image with a little effect on the image brightness. According to [29], human vision is less sensitive to chroma than brightness, as shown in Figure 3. Thus, the visual quality of the image can be maintained by filtering.

(3) The filtered image is randomly rotated before being classified by the model. Although the adversarial perturbation is weakened to a large extent after filtering, it is not entirely disappeared. The rotation can reduce the match between the model and adversarial perturbation [30, 31].

(4) The probability vectors y_i for each class obtained with different filter parameters and rotation are statistically averaged. Figure 4 shows that the image classification probability varies significantly for different filter parameters when single filtering is used. When a standard deviation (Std) is small (Std < 50), the class probability of the adversarial examples changes dramatically, inducing unstable classification results. When Std is large (Std > 300), filtering causes blurred images, and the probability of the true class of almost all clear images and adversarial examples is close to 0,

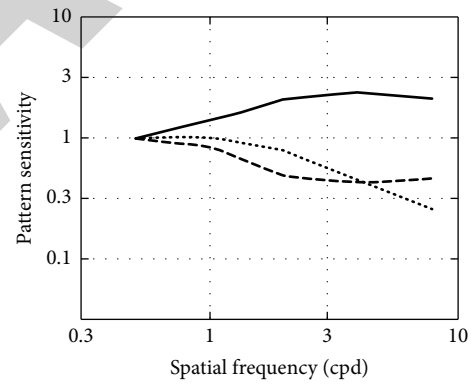

Figure 3: Color and pattern sensitivity functions [29]. The solid curve shows the sensitivities of the white-black, the dotted curve shows the red-green, and the dashed curve shows the blue-yellow. The white-black sensitivity is the highest among the three at any certain spatial frequency.

which is easy to misclassify. When Std is moderate (50 < Std < 300), the probability of true class is high and suitable for classification. It is defined here as the high confidence interval. However, the probability of the true class of some images fluctuates within this range. Thus, by averaging results from different Stds in the high confidence interval, the model's more stable classification performance can be obtained.

2.2. Defense Algorithm. The proposed defense algorithm consists of the following three steps, as shown in Figure 5:

Step 1: the prefiltering is mainly to reduce high-frequency perturbation

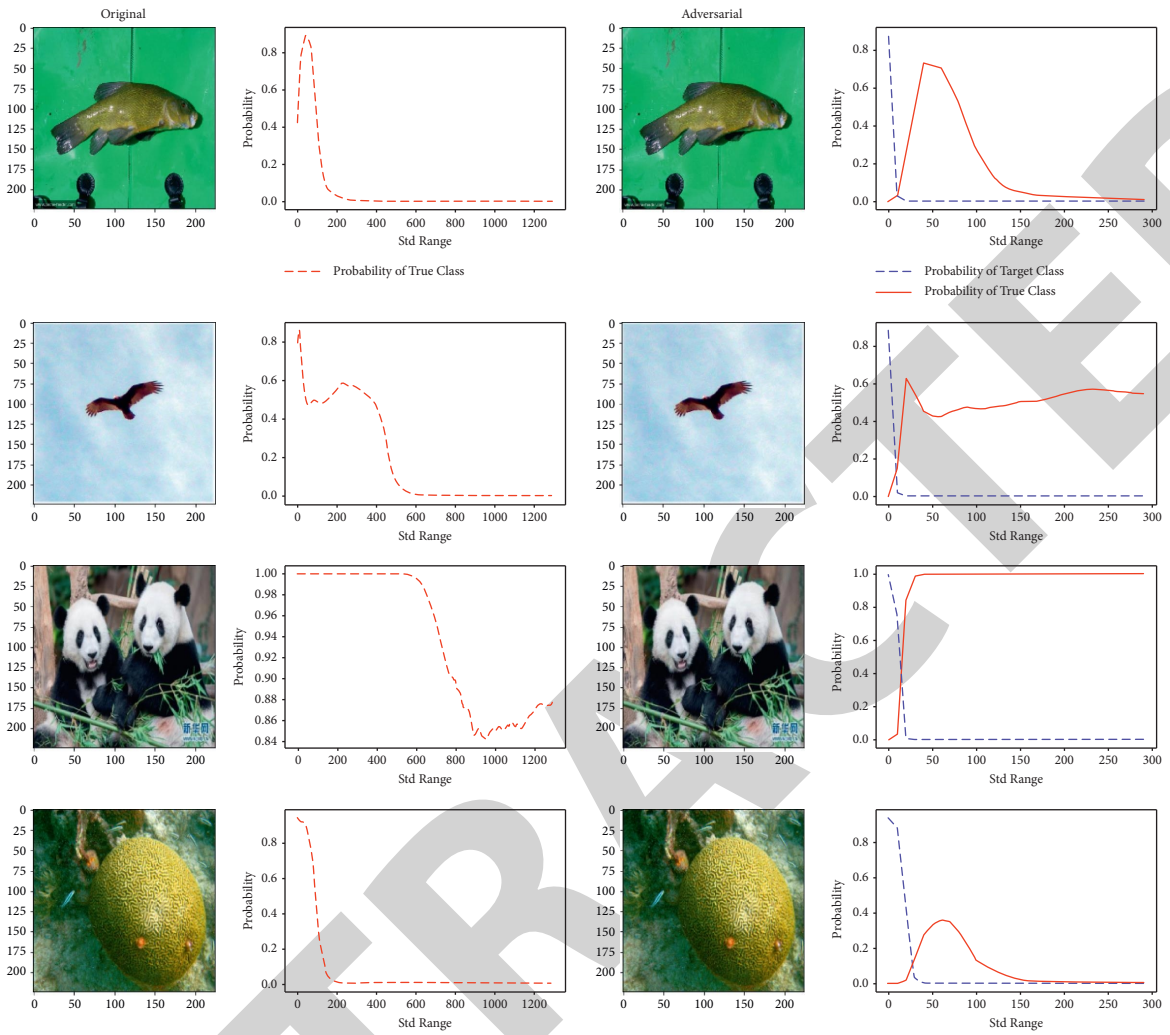

Figure 4: The image classification results vary significantly with different filter variances (std: standard deviation).

Step 2: a cyclic body aims to obtain the classification probability given by the model under different filtering and rotation parameters

Step 3: the statistical average of the classification probability in step 2 is obtained to determine the final classification label

The detailed process for each step is described as follows.

Let us define the original image as img, the real class label of the image as C(img), and the classification model as *model*. Then, the attacker generates an adversarial example img' based on *model* and img.

A bilateral filter is defined as a function of

$$x_f = \text{bilateralFilter}(\text{input}, \text{size}, \text{std1}, \text{std2}), \quad (1)$$

where input and size represent the input image and the filter size, respectively. std1 and std2 represent the standard deviations for the color and space domains, respectively.

Step 1: first, the input image is prefiltered to reduce the high-frequency components in the image, where bilateral filters with sizes 3 and 5 are applied successively, $std1 = std2 = 20$. Note that slightly higher classification accuracy is obtained when only the filter size = 3 is used for prefiltering, but it is more sensitive to noise:

$$\begin{aligned} x &= \text{bilateralFilter}(\text{input}, 3, 20, 20), \\ x &= \text{bilateralFilter}(x, 5, 20, 20). \end{aligned} \quad (2)$$

Step 2: the filtered images with a larger kernel size of 19 are rotated. Then, the classification model computes the probability vector y_i where $i = 1, 2, 3, \ldots, 9, 10$ as follows:

$$\begin{aligned} &\text{std1} = 1, \\ &\text{std2} = 20i + rb_i,\ rb_i \in (-10, 10), \\ &-10^\circ < rc_i < 10^\circ, \end{aligned} \quad (3)$$

where rb_i and rc_i are random parameters. The input image x is filtered as follows:

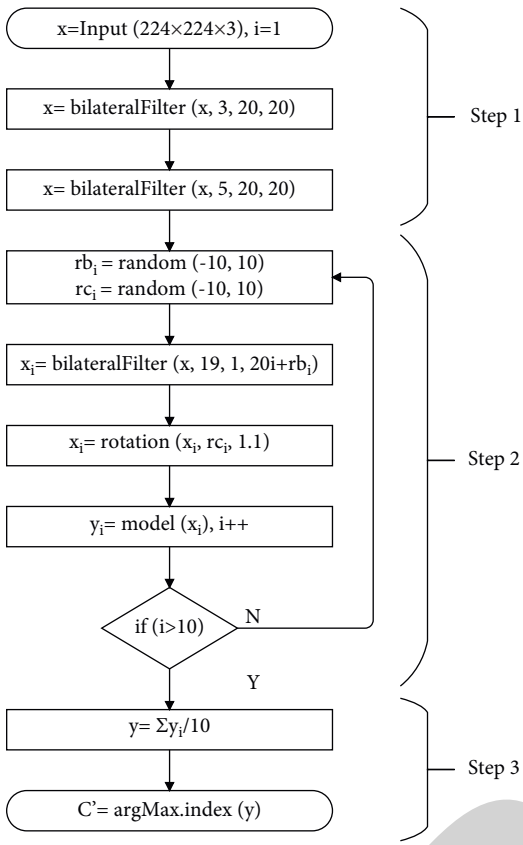

Figure 5: Defense algorithm.

$$x_i = \text{bilateralFilter}(x, 19, \text{std1}, \text{std2}). \quad (4)$$

Then, the filtered image is resized 1.1 times (to reduce the influence of the zero paddings) and rotated (the center of rotation is any position in the image):

$$x_i = \text{rotation}(x_i, rc_i, 1.1). \quad (5)$$

The classification model predicts the probability y_i:

$$y_i = \text{model}(x_i). \quad (6)$$

Step 3: the probabilities y_i are averaged to obtain y:

$$y = \frac{1}{10}\sum_{i=1}^{10} y_i. \quad (7)$$

Then, the class label of the image is determined as C' through:

$$C' = \arg\max.\text{index}(y). \quad (8)$$

3. Experimental Results

The employed dataset X consists of 1000 images randomly selected from the ILSVRC2012 validation set. Adversarial attack methods: I-FGSM [4], FGSM [3], Deep Fool [6], C&W [9], and EOT-PGD (PGD [5] attack embedded with EOT [25]) were used to attack, and the target models are AlexNet [32], VGG (Visual Geometry Group) [33], Inception [34], and ResNet (Residual Neural Network) [35]. In the oblivious attacks (Experiments 1 and 2), only the target model and the input image are utilized to generate adversarial examples using a specified attack method. In the adaptive attacks (Experiment 3), the gradient was calculated by the EOT method, and the model was attacked with PGD. The comparative experiments were conducted on the environment with PyTorch and OpenCV.

We measured the performance of the proposed method and other methods in terms of classification accuracy [36–38]. Classification accuracy is defined as the ratio of the number of correct predictions to the total number of input samples. Mathematically, it can be defined as follows:

$$\text{accuracy} = \frac{\text{number of correct predictions}}{\text{total number of samples}}. \quad (9)$$

Accuracy can also be calculated in terms of positives and negatives from the confusion matrix as follows:

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}, \quad (10)$$

where TP = true positives, TN = true negatives, FP = false positives, and FN = false negatives.

Experiment 1. (Effectiveness of defense against I-FGSM classic attack: I-FGSM is used to attack, and the effectiveness of defense against adversarial example attacks is evaluated. For all the images in dataset X, the target labels were randomly selected, and I-FGSM was used to generate the targeted adversarial examples. The evaluation results are shown in Table 1.

Bold numbers in Table 1 show that, for clear images, the top-1 accuracy of our method for some models is higher than that without defense. The proposed method achieves a classification accuracy increase of 16.23% and 6.48% over bilateral filtering and median blurring in the clear images. Due to multiple filtering in a high confidence interval and the combination with image rotation, the classification accuracy is maintained. It can keep the image salience to a large extent when resisting adversarial attacks. As shown in Figure 6, the accuracy of our method increases by more than 10% on average over the other image processing-based methods [18, 19], providing better defense efficacy.

Experiment 2. Effectiveness of defense against other oblivious attacks. The adopted oblivious attacks mainly include strong attacks based on gradient and optimization. The performance of the proposed defense method is evaluated against FGSM, DeepFool, and C&W attack methods on dataset X1 (200 images randomly selected from dataset X). The experimental results are shown in Table 2. For clear images, the proposed defense improves the classification accuracy of Inception, ResNet-50, and ResNet-152. The experimental results show that the proposed defense algorithm is efficient against various oblivious attacks. In the case of ResNet-50, the classification accuracy was only 1% lower

Table 1: Top-1 accuracy (%) under I-FGSM attack ($\epsilon = 0.007$, $n = 100$).

Target model	Without defense		With (11, 100, 100) bilateral filter		With size 3 median blur		Proposed method	
	Clear	Attack	Clear	Attack	Clear	Attack	Clear	Attack
AlexNet	**52.2**	0.4	37.0	31.3	48.5	30.4	**53.4**	50.7
VGG-16	70.8	0.3	46.8	42.5	61.4	51.0	66.6	63.3
VGG-19	71.3	0.3	48.0	42.9	63.0	53.6	69.4	65.1
Inception	**69.8**	8.4	57.7	54.2	64.1	53.5	**70.4**	67.7
ResNet-50	**73.5**	0.0	61.3	55.9	69.8	64.5	**75.5**	73.3
ResNet-152	**76.8**	0.2	65.3	61.3	73.6	67.8	**78.2**	77.3

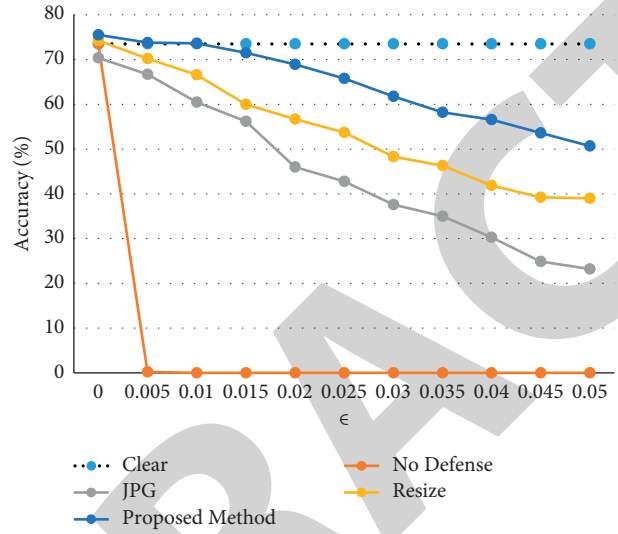

Figure 6: The comparison of our work with JPG compression [18] and randomization [19] (ResNet-50 top-1 accuracy under I-FGSM attack).

Table 2: Top-1 accuracy (%) under FGSM, DeepFool, and C&W attacks.

Target model	Clear		FGSM		DeepFool		C&W	
	No defense	Our method	No defense	Our method	No defense	Our method	No defense	Proposed method
AlexNet	58.5	57.0	20.0	55.0	0.0	56.5	0.0	54.0
VGG-16	76.0	71.0	20.0	61.5	0.0	61.0	0.0	65.0
VGG-19	76.0	71.5	20.0	63.0	0.0	66.0	0.0	71.5
Inception	**71.0**	**73.5**	31.5	71.0	0.0	72.5	0.0	71.0
ResNet-50	**79.0**	**82.0**	26.5	78.0	0.0	79.5	0.0	81.0
ResNet-152	**79.0**	**80.0**	37.0	78.5	0.0	76.5	0.0	79.0

against FGSM attacks than clear images and even higher against DeepFool and C&W attacks. In addition, DeepFool and C&W attacks were significantly stronger than FGSM attacks with no defense, while the proposed defense method is the opposite. This is due to the adversarial perturbation, which is considered as noise in the image processing-based defense method, defended by denoising. The smaller the noise, the better the defense effect. Since the DeepFool and C&W attacks generate less adversarial perturbation, the defense approach is more effective against both attacks.

Figure 7 depicts the images in each step of the proposed method. The adversarial perturbation is more easily detected by the naked eye for uniform color and homogeneous regions. Bilateral filtering can remove most of the adversarial perturbation. Still, the residual part is difficult to eliminate while maintaining the visual quality of the image. An image rotation can further reduce its influence on classification. The bilateral filtering accords with the characteristics of the human visual system by denoising noises while fully retaining the useful features for image classification. The removed noise is the part of the image with no clear semantic meaning, and thus the filtered image is clearer. Alternatively, the edges of the object in the image can be retained so that the accuracy of image classification accuracy is maintained.

Experiment 3. (Effectiveness of the defense against adaptive attack: the performance of the proposed defense method is evaluated against adaptive attack on dataset X1.

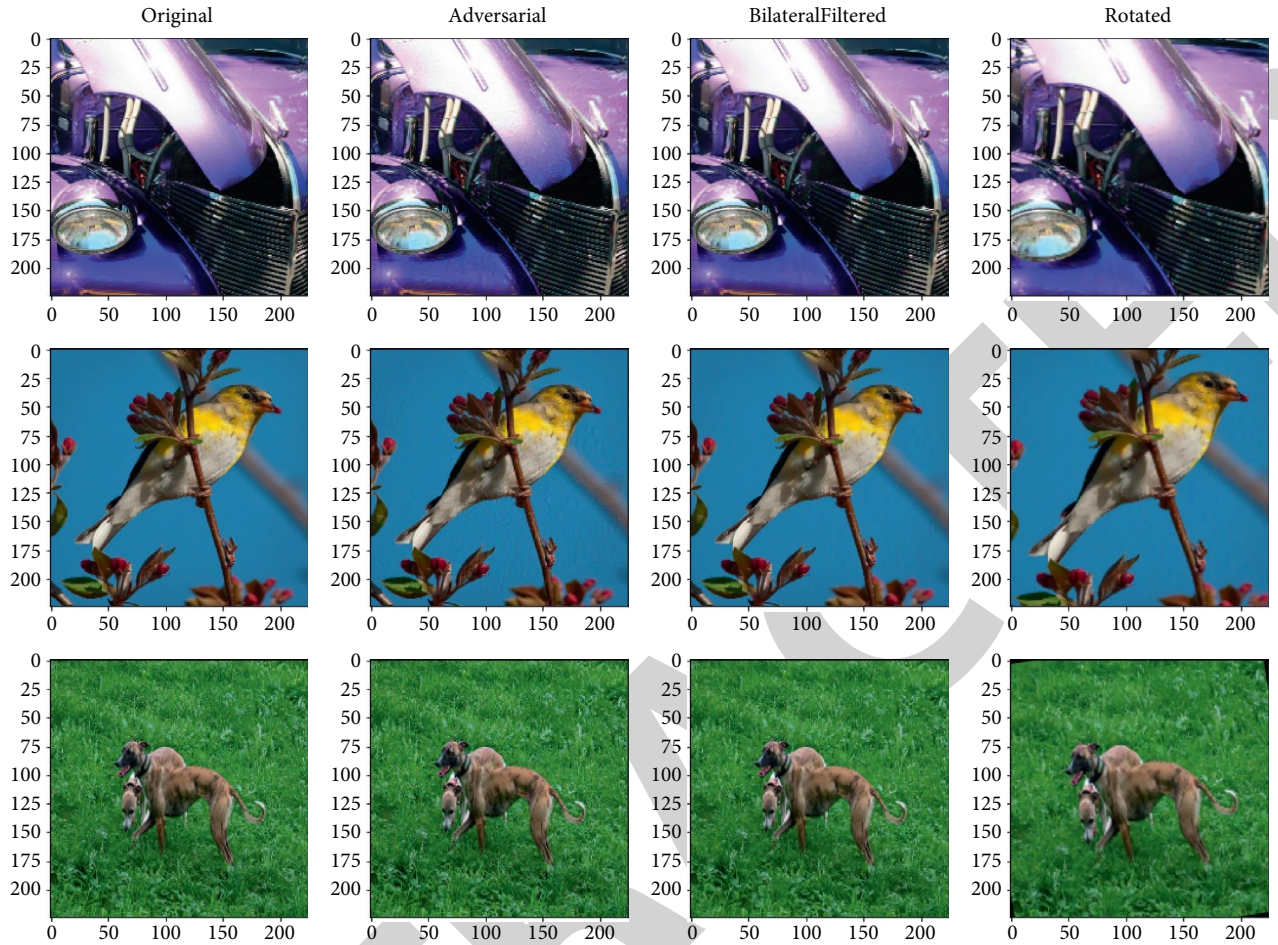

Figure 7: The images in each step of the proposed method (model = AlexNet, attack = C&W).

The EOT-PGD is selected as an adaptive attack, and the randomization defense and the proposed defense are used as the defense methods. Because the defense method with random parameters will result in a random gradient, the classical gradient-based attack method cannot use a random gradient to attack. The EOT method is used to get the gradient, and then the PGD is used for the attack model (AlexNet) to generate adversarial examples. As shown in Figure 8, when the randomization defense [19] is used, the solid orange line represents the accuracy of using AlexNet for classification. With the increase of the perturbation norm, the accuracy decreases rapidly. The orange dashed line indicates the classification accuracy using ResNet-101. The two curves almost coincide, showing the strong transferability of the adversarial example. When using the proposed defense method, the blue solid and dashed lines are clearly separated, and the transferability of the adversarial example is significantly reduced. According to [39], the ordinary targetless adversarial samples have strong transferability between models. Tables 3 and 4 summarize the classification accuracy for the randomization defense and the proposed defense. The first column and the first row represent the models used for attack and classification, respectively. The adversarial examples against randomization defense have strong transferability, while the proposed method greatly weakens the transferability of adversarial examples for all the models.

The proposed method takes advantage of the feature that the amplitude of adversarial perturbation is significantly smaller than the original image signal. In other words, the adversarial perturbation is regarded as a noise signal and filtered out by the image prefiltering process. Both the original image and the adversarial perturbation are weakened, and a larger size filter reduces the adversarial perturbation further. The random rotation reduces the matching degree of the model and the adversarial perturbation. Finally, a stable prediction of the image class can be obtained by statistical averaging. Under oblivious attacks, our method maintains the classification accuracy over the compared methods, which reveals an excellent defense efficacy against diverse types of attacks. Furthermore, due to the nonlinear transformation, under adaptive attacks, the transferability of the adversarial examples among models is weakened.

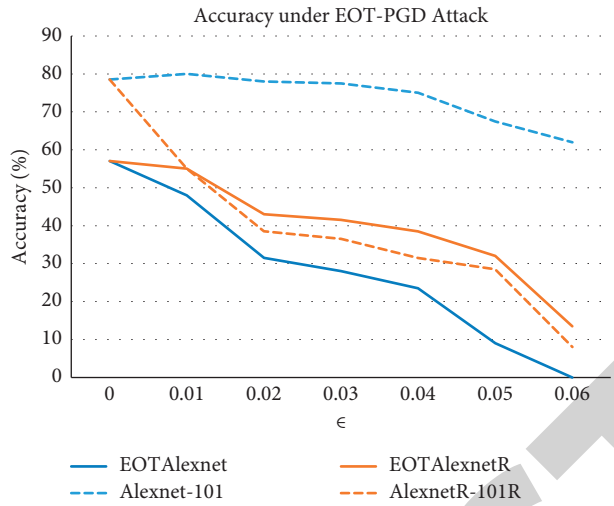

Figure 8: Transferability test (blue: our defense method and orange: randomization defense). The AlexNet model was attacked with EOT-PGD under different defenses to generate an adversarial example. Then, the adversarial examples were classified by using the AlexNet and ResNet-101 (solid and dashed lines represent the AlexNet and the ResNet-101, respectively).

Table 3: The classification accuracy (randomization defense).

Attack	Classify			
	AlexNet (%)	VGG-19 (%)	Inception (%)	ResNet-50 (%)
AlexNet	9	6	7.5	8
VGG-19	8.5	9.5	7.5	8.5
Inception	11	9.5	12	9.5
ResNet-50	16.5	15.5	19	20

Table 4: The classification accuracy (our defense).

Attack	Classify			
	AlexNet (%)	VGG-19 (%)	Inception (%)	ResNet-50 (%)
AlexNet	9	63.5	63.5	65
VGG-19	40.5	13.5	63	63.5
Inception	34.5	46.5	18	57
ResNet-50	30.5	35.5	60.5	20

4. Conclusions

This paper proposed a novel defense method based on filtering and image rotation against adversarial example attacks. The proposed method enhanced the capability of bilateral filtering-based defense against adversarial attacks from four aspects and entirely used the high confidence interval of the model under bilateral filtering. It could obtain promising defensive efficacy against the oblivious attack and effectively reduce the transferability of adversarial examples under adaptive attack. Experimental results prove the university of the proposed defense method due to the vast existence of the high confidence interval in deep learning-based image classification models. Future works will include investigating how the proposed defense method is collaborated with other image-understanding related tasks to make deep learning approaches for image data analysis safer and more widely used.

Data Availability

This article includes all the data. Further data can be requested from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding this article.

References

[1] C. Szegedy, W. Zaremba, I. Sutskever et al., "Intriguing properties of neural networks," 2013, https://arxiv.org/abs/1312.6199.

[2] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 5, pp. 828–841, 2019.

[3] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015*, San Diego, CA, USA, May 2015.

[4] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," in *Proceedings of the 5th International Conference on Learning Representations, ICLR 2017*, Toulon, France, April 2017.

[5] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2017, https://arxiv.org/abs/1706.06083.

[6] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Institute of Electrical and Electronics Engineers Inc, Honolulu, HI, USA, July 2017.

[7] S. M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: a simple and accurate method to fool deep neural networks," in *Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, IEEE Computer Society, Las Vegas, NV, USA, June 2016.

[8] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. Berkay Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *Proceedings of the 1st IEEE European Symposium on Security and Privacy, EURO SP 2016*, Institute of Electrical and Electronics Engineers Inc, Saarbruecken, Germany, March 2016.

[9] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proceedings of the 2017 IEEE Symposium on Security and Privacy, SP 2017*, Institute of Electrical and Electronics Engineers Inc, San Jose, CA, USA, May 2017.

[10] C. Xiao, J.-Y. Zhu, B. Li, W. He, M. Liu, and D. Song, "Spatially transformed adversarial examples," in *Proceedings of the 6th International Conference on Learning Representations, ICLR 2018*, Vancouver, Canada, April 2018.

[11] A. Agarwal, M. Vatsa, R. Singh, and N. K. Ratha, "Noise is inside me! generating adversarial perturbations with noise derived from natural filters," in *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPRW 2020*, IEEE Computer Society, Seattle, WA, USA, June 2020.

[12] F. Brochu, "Increasing shape bias in ImageNet-trained networks using transfer learning and domain-adversarial methods," 2019, https://arxiv.org/abs/1907.12892.

[13] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness," in *Proceedings of the 7th International Conference on Learning Representations, ICLR 2019*, New Orleans, LA, USA, May 2019.

[14] K. T. Co, L. Muñoz-González, L. Kanthan, B. Glocker, and E. C. Lupu, "Universal Adversarial perturbations to understand robustness of texture vs. shape-biased training," 2019.

[15] J. Buckman, A. Roy, C. Raffel, and I. Goodfellow, "Thermometer encoding: one hot way to resist adversarial examples," in *Proceedings of the 6th International Conference on Learning Representations, ICLR 2018*, Vancouver, Canada, April 2018.

[16] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *Proceedings of the 2016 IEEE Symposium on Security and Privacy, SP 2016*, Institute of Electrical and Electronics Engineers Inc, San Jose, CA, USA, May 2016.

[17] Y. Song, T. Kim, S. Nowozin, S. Ermon, and N. Kushman, "PixelDefend: leveraging generative models to understand and defend against adversarial examples," in *Proceedings of the 6th International Conference on Learning Representations, ICLR 2018*, Vancouver, Canada, April 2018.

[18] G. K. Dziugaite, Z. Ghahramani, and D. M. Roy, "A study of the effect of JPG compression on adversarial images," 2016, https://arxiv.org/abs/1608.00853.

[19] C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. Yuille, "Mitigating adversarial effects through randomization," in *Proceedings of the 6th International Conference on Learning Representations, ICLR 2018*, Vancouver, Canada, April 2018.

[20] P. Gupta and E. R. CIIdefense, "Defeating adversarial attacks by fusing class-specific image inpainting and image denoising," in *Proceedings of the 17th IEEE/CVF International Conference on Computer Vision, ICCV 2019*, Institute of Electrical and Electronics Engineers Inc, Seoul, Republic of Korea, October 2019.

[21] A. Prakash, N. Moran, S. Garber, A. DiLillo, and J. Storer, "Deflecting adversarial attacks with pixel deflection," in *Proceedings of the 31st Meeting of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2018*, IEEE Computer Society, Salt Lake City, UT, USA, June 2018.

[22] A. N. Bhagoji, D. Cullina, C. Sitawarin, and P. Mittal, "Enhancing robustness of machine learning systems via data transformations," in *Proceedings of the 52nd Annual Conference on Information Sciences and Systems, CISS 2018*, Institute of Electrical and Electronics Engineers Inc, Princeton, NJ, USA, March 2018.

[23] M. Osadchy, J. Hernandez-Castro, S. Gibson, O. Dunkelman, and D. Perez-Cabo, "No bot expects the DeepCAPTCHA! introducing immutable adversarial examples, with applications to CAPTCHA generation," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 11, pp. 2640–2653, 2017.

[24] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, International Machine Learning Society (IMLS), Stockholm, Sweden, July 2018.

[25] A. Anish, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, International Machine Learning Society (IMLS), Stockholm, Sweden, July 2018.

[26] F. Tramer, N. Carlini, W. Brendel, and A. Madry, "On adaptive attacks to adversarial example defenses," December 2020, https://arxiv.org/abs/2002.08347.

[27] S. Paris and F. Durand, "A fast approximation of the bilateral filter using a signal processing approach," in *Proceedings of the European conference on computer vision*, pp. 568–580, Springer, Berlin, Germany, 2006.

[28] B. Liang, H. Li, M. Su, X. Li, W. Shi, and X. Wang, "Detecting adversarial image examples in deep neural networks with adaptive noise reduction," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 1, pp. 72–85, 2021.

[29] A. B. Poirson and B. A. Wandell, "Pattern-color separable pathways predict sensitivity to simple colored patterns," *Vision Research*, vol. 36, no. 4, pp. 515–526, 1996.

[30] D. D. Thang and T. Matsui, "Image transformation can make neural networks more robust against adversarial examples," 2019, https://arxiv.org/pdf/1901.03037.pdf.

[31] S. Tian, G. Yang, and Y. Cai, "Detecting adversarial examples through image transformation," in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, AAAI press, New Orleans, LA, USA, February 2018.

[32] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proceedings of the 26th Annual Conference on Neural Information Processing Systems 2012, NIPS*, Lake Tahoe, NV, USA, December 2012.

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015*, San Diego, CA, USA, May 2015.

[34] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, IEEE Computer Society, Boston, MA, USA, June 2015.

[35] L. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, IEEE Computer Society, Las Vegas, NV, USA, June 2016.

[36] G. Kumar, K. Thakur, and M. R. Ayyagari, "MLEsIDSs: machine learning-based ensembles for intrusion detection systems—a review," *The Journal of Supercomputing*, vol. 76, pp. 8938–8971, 2020.

[37] G. Kocher and G. Kumar, "Machine learning and deep learning methods for intrusion detection systems: recent developments and challenges," *Soft Computing*, vol. 25, no. 15, pp. 9731–9763, 2021.

[38] K. Thakur, H. Alqahtani, and G. Kumar, "An intelligent algorithmically generated domain detection system," *Computers & Electrical Engineering*, vol. 92, Article ID 107129, 2021.

[39] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," in *Proceedings of the 5th International Conference on Learning Representations, ICLR 2017*, Toulon, France, April 2017.