ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20230607000701
Original scientific paper

Defending Against Local Adversarial Attacks through Empirical Gradient Optimization

Boyang SUN, Xiaoxuan MA*, Hengyou WANG

Abstract: Deep neural networks (DNNs) are susceptible to adversarial attacks, including the recently introduced locally visible adversarial patch attack, which achieves a success rate exceeding 96%. These attacks pose significant challenges to DNN security. Various defense methods, such as adversarial training, robust attention modules, watermarking, and gradient smoothing, have been proposed to enhance empirical robustness against patch attacks. However, these methods often have limitations concerning patch location requirements, randomness, and their impact on recognition accuracy for clean images.To address these challenges, we propose a novel defense algorithm called Local Adversarial Attack Empirical Defense using Gradient Optimization (LAAGO). The algorithm incorporates a low-pass filter before noise suppression to effectively mitigate the interference of high-frequency noise on the classifier while preserving the low-frequency areas of the images. Additionally, it emphasizes the original target features by enhancing the image gradients. Extensive experimental results demonstrate that the proposed method improves defense performance by 3.69% for 80 × 80 noise patches (representing approximately 4% of the images), while experiencing only a negligible 0.3% accuracy drop on clean images. The LAAGO algorithm provides a robust defense mechanism against local adversarial attacks, overcoming the limitations of previous methods. Our approach leverages gradient optimization, noise suppression, and feature enhancement, resulting in significant improvements in defense performance while maintaining high accuracy for clean images. This work contributes to the advancement of defense strategies against emerging adversarial attacks, thereby enhancing the security and reliability of deep neural networks.

Keywords: adversarial attack; adversarial patch; deep learning; local gradient smoothing

1 INTRODUCTION

In recent years, with the continuous development of deep neural networks (DNNs), deep learning (DL) has accurately built more complex function models on datasets. It has gradually entered the stage of practical application and has made remarkable achievements in many areas, such as those in face recognition [1-3], image segmentation [4-9], and autonomous driving [10-12].

However, numerous experimental studies have demonstrated that introducing small, imperceptible changes [13-15] to human vision in the original input samples can lead a DNN to make misclassifications with high confidence. As the vulnerability of DL continues to be exposed, its security dramatically limits the scope of DL in practice and reduces the validity of research problems [16]. To further discover its faults, researchers mainly use two types of attack methods. The first is based on the gradient ascent algorithm, such as the FGSM [15], I-FGSM [17], and DI2-FGSM [18], which maximize the loss function and thus achieve the attack target. However, minor perturbations in visual inputs are highly susceptible to several factors, such as lighting, filming equipment, and angle. Thus, to cause the attack algorithm to work more efficiently, researchers proposed a second type of attack method that reduces interference from external factors [19] and generates adversarial patches by heavily modifying a few image pixels that can be arbitrarily applied in realistic scenarios. For example, the attacker may cover his face with designed glasses to deceive face recognition systems [20]. In autonomous driving applications, an attacker can add a noise patch with rectangular or circular patterns on top of traffic signs to cause misclassification. These adversarial attacks present a significant challenge to existing deep learning systems. Thus, improving the robustness of the given deep learning network has become a hot topic in recent years.

Recently, researchers have been actively exploring the phenomenon of local adversarial attacks and their impact on deep learning models. Local adversarial attacks refer to targeted methods that introduce small, imperceptible perturbations to input samples to deceive the models and cause misclassification or incorrect outputs. Deep learning models have achieved remarkable success in tasks such as image classification [21-24], object detection [25], natural language processing [26], and speech recognition. However, they have also exhibited a high sensitivity to input data. The theoretical background of local adversarial attacks is rooted in the vulnerability of deep learning models, where adversarial samples, despite their proximity to clean samples in the input space, yield drastically different outputs in the model's output space.

Attackers exploit this vulnerability by employing local adversarial attack techniques, manipulating input samples with tiny perturbations to deceive deep learning models. These perturbations can be either targeted, aiming to induce misclassification, or random, intended to trigger erroneous judgments by the models. Local adversarial attacks rely on two key concepts: adversarial samples and adversarial loss functions. Adversarial samples are generated by introducing small perturbations to original input samples, designed to deceive the deep learning models. Adversarial loss functions serve as objective functions for optimizing the adversarial samples, considering both the model's classification accuracy and the magnitude of perturbations, in order to find the most deceptive perturbation. The impact of local adversarial attacks on deep learning models is significant. They reveal the fragility of deep learning models, where minute perturbations can lead to misclassification. This poses a critical challenge to the security and reliability of deep learning models. The success of local adversarial attacks suggests the existence of vulnerabilities in the decision boundaries of deep learning models and their excessive sensitivity to small variations in input data.

To enhance the robustness of deep learning models, researchers have proposed various defense methods, including adversarial training, noise reduction, and input transformations. These methods aim to increase the models' resilience to adversarial samples and reduce the success rate of attacks. For example, an adversarial training approach was proposed in [27], and a robust attention

module was designed in [28] to improve the empirical robustness against patch attacks. In addition, the certified defense was first introduced into local adversarial attacks in [29]. However, these defense methods are not useful on large datasets. Thus, to extend them to large datasets, a derandomized smoothing structured ablation method was proposed in [30], which made the certified defense suitable for adversarial patches on the ImageNet dataset. To improve the performance of the derandomized certified defense, the authors introduced ViT in terms of time and accuracy in [31], but the certified accuracy was only 43.8%. Watermarking [32] and gradient smoothing [33] in empirical defense can improve the defense accuracy by approximately 20-30%.

In recent years, local gradient smoothing (LGS) has emerged as a classical defense method for addressing local adversarial attacks [33]. LGS effectively addresses the issue of adversarial patch interference on classifiers by employing filtering techniques to suppress local high-frequency noise. However, a significant drawback of this approach arises when clean examples are encountered, as it often leads to the unintended filtering of important object details, thereby impeding the classifier's ability to effectively identify the original objects.

Achieving a balance between defense accuracy and preserving image details has proven challenging in previous research, particularly when dealing with larger datasets like ImageNet. In this study, our objective is to extend the heuristic algorithm to the complex ImageNet dataset while preserving the integrity of original image details. Motivated by the work of Naseer et al. [33], we propose a novel method to enhance the defense efficiency of traditional empirical methods for clean images. This is achieved by integrating gradient boosting and gradient smoothing techniques. Our aim is to improve defense accuracy while minimizing the loss of crucial object details. The contributions of our work can be summarized as follows:

- To reduce the negative influence of traditional heuristic defense algorithms on clean images, we first introduce a low-pass filter that further processes the first-order gradient map before suppressing high-frequency noise. The low-pass filter smooths the high-frequency region of the image while better protecting the low-frequency region of the image, which can effectively improve the accuracy of filtering high-frequency noise. Furthermore, the texture features of the original objects are further highlighted by enhancing the gradient details of clean examples; thus, the accuracy of the classifier is improved very well.
- To defend the noise of the local adversarial patch, we design a gradient optimization strategy. The experimental results show that our proposed method can efficiently defend against local noise attacks.

This paper is organized as follows: Section 2 focuses on the analysis of work related to the adversarial attack algorithm. Section 3 elaborates on the specific principles and details of the LAAGO algorithm, and Section 4 analyses the effects of the LAAGO algorithm in detail through experiments. Conclusions are provided in Section 5.

2 RELATED WORK

The current adversarial attacks on DNNs are broadly classified into two categories. The first category of digital attacks introduces tiny amounts of noise into the original image that is imperceptible to the human eye. The adversarial samples generated by modifying each image pixel by a small amount can lead a DNN to make misclassifications with high confidence. Currently, for digital attacks, existing defense algorithms such as JPEG compression [34], feature compression [35], median filtering, Gaussian filtering, and total variation minimization [36] are usually considered effective in specific situations. The second category of physical attacks is different from the first because it causes local high-frequency adversarial noise by modifying a few image pixels by large visible amounts. [37] and [19] Physical attack is not affected by realistic factors such as illumination and angle. However, few widely used defense algorithms for local adversarial attacks exist. The only algorithms that are commonly used for this purpose are the LGS algorithm [33] and the digital watermarking (DW) [32] algorithm. These two types of attack strategies and their corresponding defense algorithms are described in detail below.

2.1 Traditional Attacks

The adversarial example generation problem can be converted into a constrained optimization problem. Given a classifier $\mathbb{F}(y|x.)$, solve the following optimization problem to increase the attack success rate of the adversarial sample:

$$\begin{aligned} &\max_{x'} \mathbb{F}\left(y=\bar{y}|x'\right) \\ &\text{subject to}: |x-x'|_p \le \varepsilon \end{aligned} \tag{1}$$

$x \in \mathbb{R}^n$ is the original example of the input, $\bar{y}$ is the target label, and ε is the maximum value of the restricted perturbation. The attacker increases the target label $\bar{y}$ by entering an adversarial sample with noise δ, and the mathematical definition is formulated as follows:

$$x' = x + \delta \in \mathbb{R}^n \tag{2}$$

Such traditional adversarial attacks change most of the pixels in the image and thus allow the generation of adversarial examples, including the FGSM [14], I-FGSM [17], DI^2-FGSM [18] and PGD [13] algorithms, which fall into this category. Fig. 1 illustrates the attack application of the FGSM algorithm. However, on the defensive side, methods such as JPEG compression, total variance minimization and feature compression are often considered to be effective against such attacks, especially when ε is tiny.

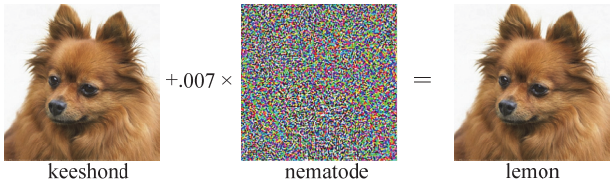


Figure 1 FGSM [15] Example of adversarial perturbation

2.2 Traditional Attack Defense

This section focuses on two algorithms, JPEG compression and total variation minimization (TVM), which defend against traditional attack strategies.

[34] conducted extensive research on the effectiveness of JPEG compression's defenses and demonstrated that JPEG compression using the discrete cosine transform (DCT) can effectively remove high-frequency components that are less important to human vision and retain more critical low-frequency components by using the following steps:

- The image is first split into 8 × 8 blocks using downsampling, and each block is processed separately throughout the subsequent compression process.
- Each color channel is converted from a small block of RGB to the YC_bC_r color space, where Y and C_bC_r represent the luminance and chrominance of the image, respectively.
- After discrete cosine transform, the images are reversible despite being in a damaged state. The frequency amplitude is finally quantized by measuring the data by dividing by a constant and rounding the result to the nearest integer.

The total variational minimization model is an image denoising method based on the idea of the mathematical variational method in which the image is brought into a smoothed state by calculating the minimization of the energy function (according to Eq. (3)) after it has been determined for the image. Guo et al. [36] considered smoothing adversarial examples through TVM, JPEG compression and image padding. The use of TVM was shown to be very effective in removing small perturbations due to its ability to detect and remove minor changes in the image.

$$\min TV(f) = \int_{\Omega} \sqrt{|\nabla f|^2}\, d_x d_y = \int_{\Omega} \sqrt{f_x^2 + f_y^2}\, d_x d_y \tag{3}$$

2.3 Local Adversarial Attack

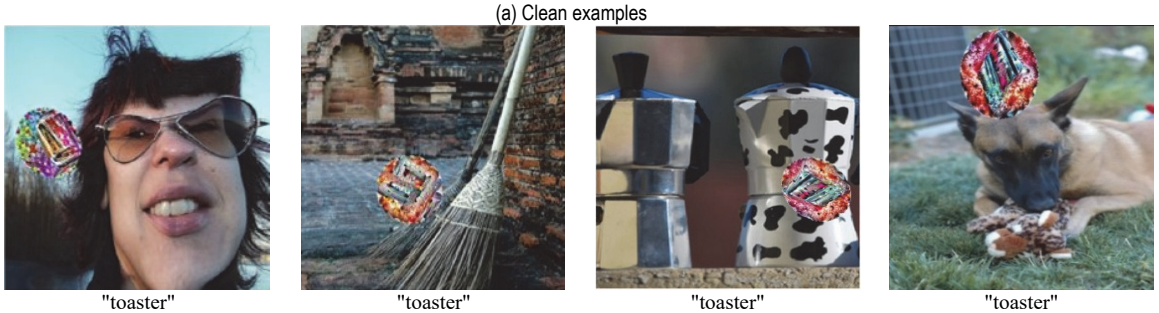
2.3.1 Adversarial Patch

Traditional adversarial examples will lose their offensive effect to some extent when encountering different lighting, scenarios, angles, and devices in the real world, and they cannot be directly applied in real-world attacks. Therefore, in 2017, Athalye et al. [38] introduced an Expectation over Transformation (EOT) attack to create adversarial examples that remain adversarial over a selected transformation distribution. As seen in Fig. 2, Brown et al. created a perturbation noise patch for real-world attacks and scenes independent based on the EOT attack, which acts randomly on a region in the image and causes the classifier to misclassify a target label. The principle is a patch operator proposed by Brown et al. $A(p,x,l,t)$, where p is the patch block, x is the original image, l is the location where the patch is placed on the image, and t defines patch operations such as rotation transformations. The possibility of applying the patch to the original image location l by means of operator A increases the target label $\bar{y}$ in the following optimization problem:

$$\bar{p} = \arg\max_{p} E_{x\sim X, t\sim T, l\sim L}\left[\log \Pr\left(\bar{y}\middle|A(p,x,l,t)\right)\right] \tag{4}$$

where X denotes the training image set, T denotes the distribution of all transformations, and L denotes the distribution of the positions in the image.

(a) Clean examples



(b) Adversarial examples

Figure 2 (a) and (b) show the original category of the image and the target category exhibited after being attacked, respectively

2.3.2 LaVAN

The LaVAN algorithm explores a perturbation that is visible but is restricted in position and does not obscure the primary target. The authors show that modifying 2% of the pixels to form a patch that can attack the state-of-the-art Inception v3, and the patch can be migrated as follows:

$$x' = (1-m)\odot x + m\odot\delta, \text{s.t.}, m\in\mathbb{R}^n \tag{5}$$

The network domain and image domain are mentioned in this paper. The network domain is ignored, while the image domain is used because it is relevant. We also performed an analysis based on the gradient that the network does not pay special attention to the patch region, which contradicts the assumptions in the adversarial patch. The results are displayed in Fig. 3.

2.4 Defending Against Local Adversarial Attack

2.4.1 Digital Watermarking Defense

Hayes et al. [32] proposed both nonblind and blind defense strategies to address the challenges presented by local adversarial attacks. One of the nonblind defenses refers to Alexandru Telea [39] in the process of image restoration, where the defender needs to know the location of the adversarial patch in advance during reconstruction. In practice, however, the location of local adversarial attacks is usually randomly placed, and the threat of attack is disarmed as soon as the defender knows the location of the adversarial patch. Local adversarial attacks can shift the classifier's attention from the original object to the adversarial high-frequency noise. In the blind defense image restoration process, the experimental hypothesis is more realistic in that the defender is only known to the noisy image, and the authors use the attention mechanism to first find the location of the high-frequency noise using the saliency map. Then, they further process the region to achieve noise suppression before inputting it into the classifier.

The advantage of digital watermarking (DW) is that the location of the adversarial mask can be found efficiently using the saliency map. However, this advantage is also a drawback when defending clean examples because the saliency map provides the location of the original object with high probability, and clean examples can significantly degrade the classifier's recognition performance after processing. In a report on the performance of blind defense, only 400 randomly selected images were tested using VGG19 [40], and the accuracy dropped by 12% on the clean examples, making it difficult to guarantee the stability of the classifier in practical applications.

(a) Padlock, 90.22% (b) Dumbbell, 92.08%

(c) Soccer Ball, 91.81% (d) Brass, 92.23%

Figure 3 A demonstration of the high accuracy of the target classes presented by the images after a patch attack of size 42 × 42 (approximately 2% of the image)

2.4.2 Local Gradient Smoothing Algorithm

The adversarial patch essentially introduces a local high-frequency noise that will be given priority attention by the model, thus making the classifier output the target label $\bar{y}$. Since the accuracy of DW on clean examples is drastically reduced, Muzammal Naseer et al. [33] proposed a novel defense mechanism, the local gradient smoothing algorithm, which can be summarized as a scaled normalized gradient map that is processed and remapped onto the original image to suppress high-frequency regions and is described as follows:

First, the image first-order gradient is defined as follows:

$$\left|\nabla X(a,b)\right| = \sqrt{\left(\frac{\partial x}{\partial a}\right)^2 + \left(\frac{\partial x}{\partial b}\right)^2} \tag{6}$$

where a and b correspond to the horizontal and vertical direction gradients, respectively. Subjecting the gradients to the normalization operation reduces the impact of the defense algorithm on the original image and improves the image recognition success rate. Second, Muzammal Naseer et al. [33] designed a block-wise approach to dividing the gradient magnitude map into k overlapping blocks of the same size (τ) and filtering out the regions that are most likely to be noisy according to the threshold (γ). The specific definition is as follows:

$$g'_{h,w} = \varpi\left(g(x), h, w, \tau, o\right) \in \mathbb{R}^{\tau},$$
$$G_{h,w} = \begin{cases} g'_{h,w}, & if \ \frac{1}{\left|g'_{h,w}\right|}\sum_i \sum_j g'_{h,w}(i,j) > \gamma \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

Finally, as shown in Fig. 4, the filtered local high-frequency noise is suppressed and mapped back to the original image by using the following suppression formula:

$$\Gamma(x) = x \odot \left(1 - \gamma * G_{h,w}\right) \tag{8}$$

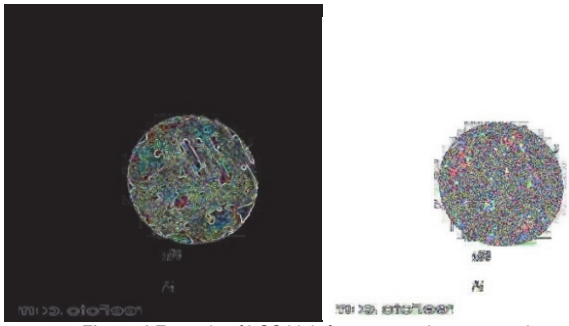


Figure 4 Example of LGS high-frequency noise suppression

In the experiments using the Inception v3 model, Muzammal Naseer et al. [33] achieved a defense accuracy of 67.49% against an adversarial patch attack of approximately 4% with a 5.59% drop in recognition accuracy for clean samples after they were processed with the traditional LGS algorithm. This result is a slight

improvement compared to the digital watermark DW, but it still requires further development.

The accuracy of traditional LGS defense methods still needs to be optimized. Additionally, the LGS does a high degree of damage to the DNN, which makes it difficult to achieve usable accuracies. We propose a local adversarial attack empirical defense algorithm using gradient optimization (LAAGO) to solve these issues.

3 LAAGO ALGORITHM

In the following section, the implementation principle of the proposed local adversarial attacks empirical defense algorithm using gradient optimization, also called LAAGO, is described in terms of both algorithmic flow and algorithmic principle.

3.1 LAAGO Algorithm Flow

In the traditional LGS algorithm, high-frequency noise is suppressed to enable the classifier $\mathbb{F}$ to accurately identify the true class y of the input image x. In the experiments on the LGS algorithm, all of the adversarial patches are randomly placed in the image's edge region so that the patch cannot obscure the original target location. However, due to the random location of the confrontation patches in practical applications, there is a high probability of obscuring some important features of the target objects. To ensure that some original objects with missing features still have high recognition accuracy, we introduce a low-pass filter and gradient enhancement to improve the stability of the algorithm on clean samples. The flow chart of our proposed LAAGO algorithm is shown in Fig. 5.

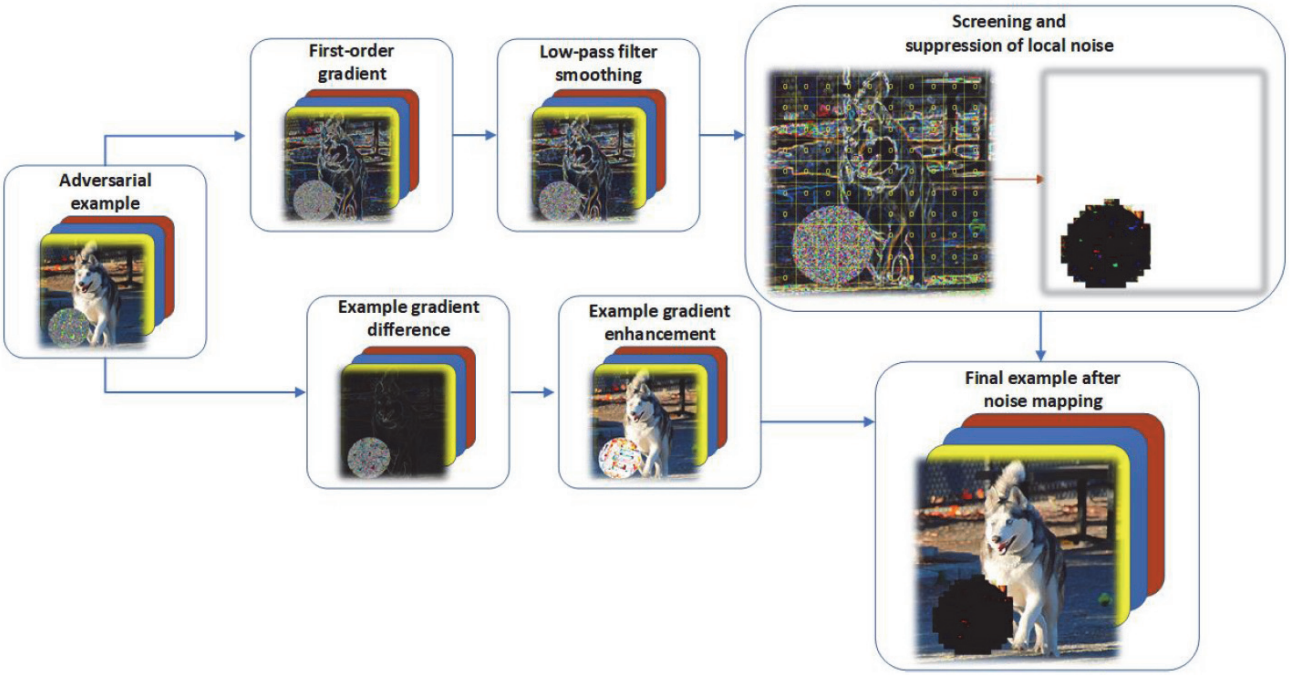


Figure 5 Illustration of the LAAGO algorithm. In the first half of the workflow, we use a low-pass filter to smooth the first-order gradient map. The processed gradient map is then divided into overlapping blocks of equal size, and local noise is suppressed using gradient descent after filtering out the most likely noisy regions according to a threshold. In the second half of the workflow, we extract the gradient of the original image and enhance it according to the enhancement factor. Finally, the suppressed local adversarial noise is mapped to the gradient-enhanced original image to complete the defense.

In Fig. 5, the LAAGO algorithm uses a low-pass smoothing filter to effectively reduce the interference of localized adversarial noise before screening local antagonistic noise in the first-order gradient map. Before mapping the suppressed local adversarial noise, we perform gradient enhancement on the unprocessed adversarial examples to help focus the DNN's attention on the contour texture features of the original object. The suppressed local adversarial noise is then mapped onto the gradient-enhanced adversarial example to obtain the final example. The improved principle, which is shown in Figure 5, is discussed in more depth in the next subsection. The LAAGO algorithm pseudocode is provided in Algorithm 1.

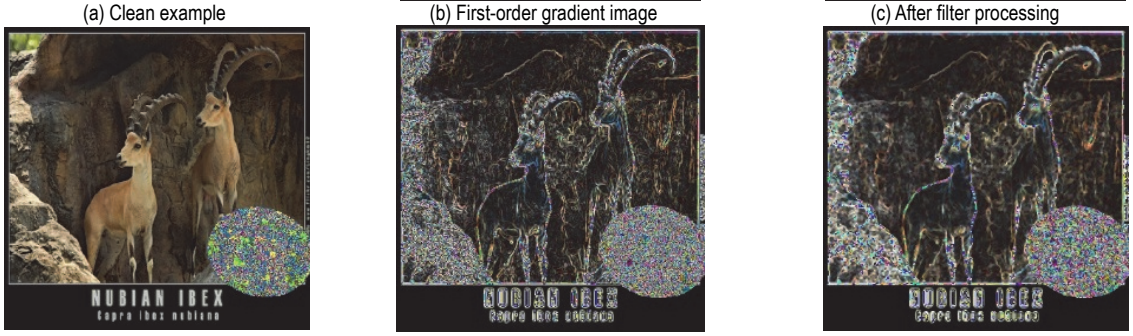
3.2 Principle of the LAAGO Algorithm

3.2.1 Introduction of Low-Pass Filters

After obtaining the image's first-order gradient, the conventional LGS algorithm is used to normalize the image's overall first-order gradient map to ensure that the subsequent search for high-frequency noise locations is accurate. In realistic scenarios, images acquired by different devices in different environments are easily affected by defocusing and noise pollution, among other factors. To reduce the impact of image noise on the classification, we use a low-pass filter instead of the normalization operation in the original text. The low-pass filter smooths the high-frequency region of the image while better protecting the low-frequency region of the image, which can effectively alleviate the problem of biased results caused by screening high-frequency noise.

As many filters have low-pass properties, we chose a nonlinear median filter for image processing to better match the subsequent gradient enhancement. The median filter can play the role of both noise removal and image edge protection in some special cases and has a low impact on the original image's edges, as shown in Fig. 6.

(a) Clean example (b) First-order gradient image (c) After filter processing



(d) Adversarial example (e) First-order gradient image (f) After filter processing

Figure 6 (b) and (e) are the first-order gradient maps of (a) and the adversarial example (d), respectively. (c) and (f) are the corresponding images after low-pass filtering. It is easy to see that the edge features of the images are better preserved, while the high-frequency noise points are altered

Algorithm 1: A local adversarial attack empirical defense algorithm using gradient optimization (LAAGO)

Input: *image* x ,
screening threshold γ ,
smoothing factor λ ,
block size τ ,
Number of chunks k ,
Gradient enhancement factor θ

begin
$\hat{x} \longleftarrow$ *the gradient map of the image* x .
$\hat{x}' \longleftarrow$ *after using low-pass filters.*
$g'_{h,w} \longleftarrow \hat{x}\ .\ unfold\ (k, k-\tau)$
for $G_{h,w} \in g'_{h,w}$ **do**
$G_{h,w}\begin{cases} g'_{h,w} & \leftarrow \left(g'_{h,w} \cdot sum(\ )/\hat{x}.prod\right) \geq \gamma \\ 0 & \leftarrow otherwise \end{cases}$
$\Phi(x) \longleftarrow 1-\lambda * G_{h,w}$
for $k \in x_c$ **do**
for $i, j \in x_h, x_w$ **do**
$|g_x| \leftarrow \left(x[i+1, j, k] - x[i, j, k]\right)$
$|g_y| \leftarrow \left(x[i, j+1, k] - x[i, j, k]\right)$
$x' \longleftarrow x + \theta * \left(|g_x| + |g_y|\right)$
$\Gamma(x) \longleftarrow \Phi(x) \odot x'$
Result: $\Gamma(x)$

The first-order gradient image after low-pass filter processing is divided into k overlapping blocks of the same size (τ) after they are processed using Eq. (7). One of the gradient optimization strategies is to suppress local adversarial noise by using gradient descent after filtering it out. Gradient descent is performed after filtering out the regions that are most likely to be noisy according to the following threshold (γ):

$$\Phi(x) = 1 - \lambda * G_{h,w} \tag{9}$$

3.2.2 Sample Gradient Enhancement

The second strategy of gradient optimization is to use gradient enhancement in the original image because we first apply gradient enhancement to the input image x , which can improve essential features such as the original object's contour texture and thus improve the recognition accuracy of the classifier. The specific details are as follows:

If the image is simply viewed as a two-dimensional function $f(x, y)$, the formula for finding the rate of change of the grayscale image, is defined as follows:

$$\begin{aligned} \frac{\partial f(x,y)}{\partial x} &= \lim_{\delta \to 0} \frac{f(x+\delta, y) - f(x,y)}{\delta} \\ \frac{\partial f(x,y)}{\partial y} &= \lim_{\delta \to 0} \frac{f(x, y+\delta) - f(x,y)}{\delta} \end{aligned} \tag{10}$$

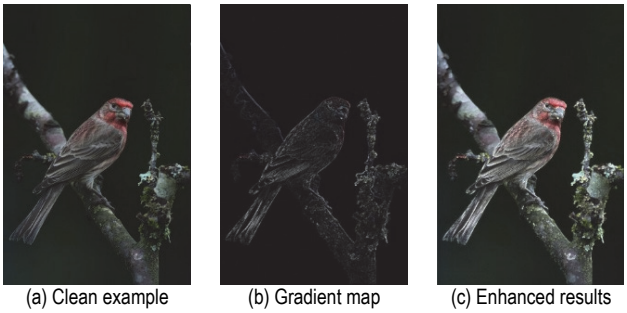
Because the image is discretized by pixels and is a noncontinuous two-dimensional function, δ cannot be infinitely small, and for an unnormalized image, its smallest δ is the pixel with a gray value of 1. Thus, the formula is as follows ($\delta = 1$):

$$\begin{aligned} \frac{\partial f(x,y)}{\partial x} &= f(x+1, y) - f(x,y) = g_x \\ \frac{\partial f(x,y)}{\partial y} &= f(x, y+1) - f(x,y) = g_y \end{aligned} \tag{11}$$

g_x and g_y represent the gradient of the image in the x and y directions at point (x, y), respectively, and they can be equated to the difference between two adjacent pixels. As shown in Fig. 7, to reduce the computational

effort, we use the absolute values of g_x and g_y and weight them to obtain their approximate gradient values and thus achieve gradient enhancement. The formula for this process is expressed as follows:

$$x' = x + \theta\left(\left|g_x\right| + \left|g_y\right|\right) \tag{12}$$



(a) Clean example (b) Gradient map (c) Enhanced results

Figure 7 (c) is obtained by adding a gradient (b) with weight θ to the original image (a)

The parameter θ controls the degree of gradient enhancement, which ultimately maps the suppressed local noise to the gradient-enhanced image and inputs it into the model:

$$\Gamma(x) = x' \odot \Phi(x) \tag{13}$$

As shown in Fig. 8, when using the LAAGO algorithm to process the adversarial examples with local noise, the position of the adversarial patch is not explicitly limited. Moreover, there is a high probability that the patch will obscure some of the original object's significant details. However, after low-pass filtering and gradient enhancement, the contours of the original object are enhanced, while the local adversarial noise is suppressed, as shown in Fig. 8. The LAAGO algorithm minimizes the impact of the local adversarial noise and the defense algorithm on the original image, and the effectiveness of the algorithm will be further verified by the experimental results in Section 4.

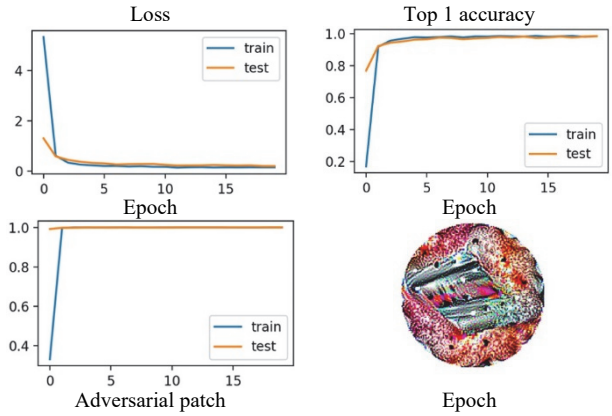
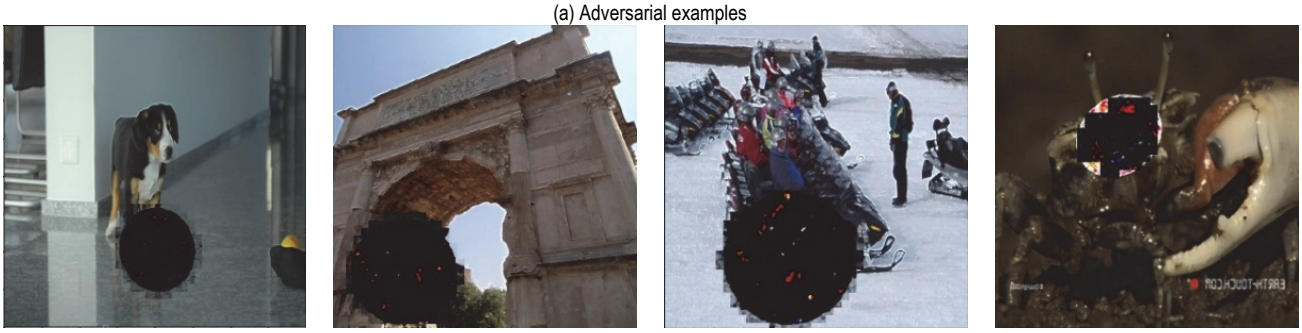


Figure 9 Attack success rate of the adversarial dataset and the adversarial patch training process

(a) Adversarial examples



(b) After LAAGO processing

Figure 8 (a) Adversarial examples with local patch adversarial noise. (b) The contour texture feature of the original object is enhanced after processing by the LAAGO algorithm

4 EXPERIMENTAL RESULTS AND ANALYSIS

4.1 Experimental Setup and Preparation

To better compare the defensive effectiveness of the traditional LGS and LAAGO algorithms, we follow the Inception v3 model from the original traditional LGS algorithm [41] when defending against adversarial patches. ImageNet2012 [42] is the dataset chosen for this experiment, and all of the attacks in our experiments were conducted in a white-box setting. 1000 adversarial examples were selected as the experimental dataset in the traditional LGS, and all of the patches were chosen to be placed in the edge region of the image to prevent the patches from obscuring the essential details in the image, which does not occur in actual applications. To obtain more

accurate results, when faced with an adversarial patch attack, we were confronted with patches that were completely randomly placed within the image area. We ran several iterations of attack optimization per image and terminated the optimization early if the classifier made misclassifications with a confidence above or equal to 99%. The results for 2000 adversarial examples misclassified as a toaster with confidence above 96% are reported. For the LaVAN attack, we iterated over 1000 images of size 299 × 299 with a confidence level greater than the 90% hostile dataset.

We used a low-pass filter instead of the image normalization operation, and the hyperparameters for screening the high-frequency threshold (λ) needed to be reselected. Fig. 10 shows that the optimal threshold λ was selected after averaging 30 experiments in the interval from 0.0 to 1.2.

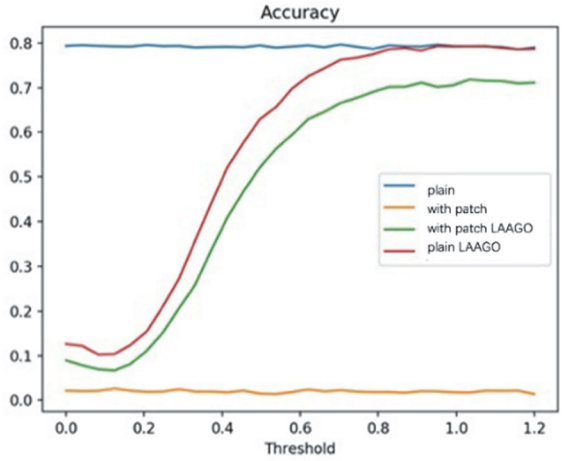


Figure 10 Parameter threshold (λ) influence graph

Specifically, the experiments were compared with the traditional LGS algorithm [33], JPEG compression [34], total image variance minimization (TVM) [36], Gaussian filtering (GF), median filtering (MF), and $\mathcal{L}_0$ gradient smoothing [43]. All of the experiments were conducted on a desktop Windows PC equipped with an Intel i7-10700k octa-core CPU clocked at 4.20 GHz and 32 GB of RAM.

4.2 Experimental Results

4.2.1 Defense against the Adversarial Patch

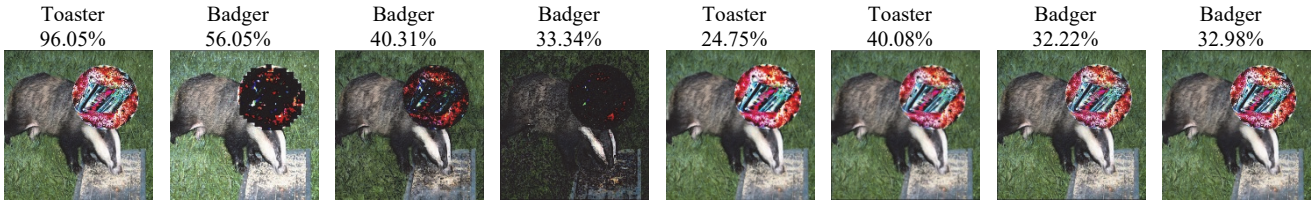
As described in this section, the size of the images was set to 400 × 400 when testing the Inception v3 model, and the adversarial patch sizes, which included 70 × 70 (~3% image), 80 × 80 (~4% image), and 120 × 120 (~10% image), were used to verify the proposed method. The processing results and display of each type of algorithm are shown in Tab. 1 and Fig. 11.

Research has shown that DL networks mainly focus on object texture features when recognizing objects [44]. Therefore, to further remove the attack properties of the adversarial patch, we improved the $\mathcal{L}_0$ gradient smoothing algorithm. After filtering out the high-frequency noise using Eq. (7), the $\mathcal{L}_0$ gradient smoothing algorithm was used in the local high-frequency noise region, which caused the algorithm to eliminate the texture information associated with the adversarial patch. However, it can be concluded from the experimental results (Tab. 1) that this defense does not achieve the desired results because even though the adversarial patch has been invalidated, the high-frequency noise still interferes with the model's ability to make correct judgments. Therefore, similar to the LAAGO algorithm, the defense against local adversarial noise needs to directly suppress the abnormally high-frequency noise while increasing the original object's gradient strength.

To evaluate the performance of the proposed method, JPEG compression, TVM denoising, Median filters, and Gaussian filters were used for experimental comparison. The results show that these traditional defense methods are less effective against patching attacks. DW and LGS are more effective than traditional defense methods, but they obscure some detailed information when processing clean examples, resulting in poor recognition in the absence of attacks. The experimental results demonstrate the advantages of the LAAGO algorithm over the other defense methods. LGS reduces the recognition accuracy of clean examples from 78.24% to 72.65%, a decrease of 5.59%, but the LAAGO algorithm only loses 0.3% accuracy.

Table 1 Experimental results of various algorithms for defending against Adversarial patch

	No Attack	70 × 70 noise patch covering ~3% of image	80 × 80 noise patch covering ~4% of image	120 × 120 noise patch covering ~10% of image
No. defense	78.24%	11.95%	8.65%	6.79%
LGS [λ = 3.7]	72.65%	68.59%	68.15%	67.49%
LGS [λ = 2.3]	73.30%	61.10%	60.60%	60.02%
LGS [λ = 1.7]	73.95%	50.00%	48.85%	48.50%
JPEG [quality = 60]	72.76%	25.23%	11.45%	5.60%
JPEG [quality = 30]	71.73%	33.13%	22.65%	12.60%
DW	53.27%	66.29%	64.83%	62.40%
TVM [weights = 20]	72.89%	3.78%	2.17%	1.02%
MF [window = 5]	71.75%	29.20%	29.10%	27.15%
GF [window = 3]	75.10%	16.00%	15.65%	15.40%
$\mathcal{L}_0$ [lambda = $3e^{-2}$]	3.20%	1.50%	1.20%	0.40%
*** LAAGO**	**77.94%**	**71.93%**	**71.84%**	**70.14%**



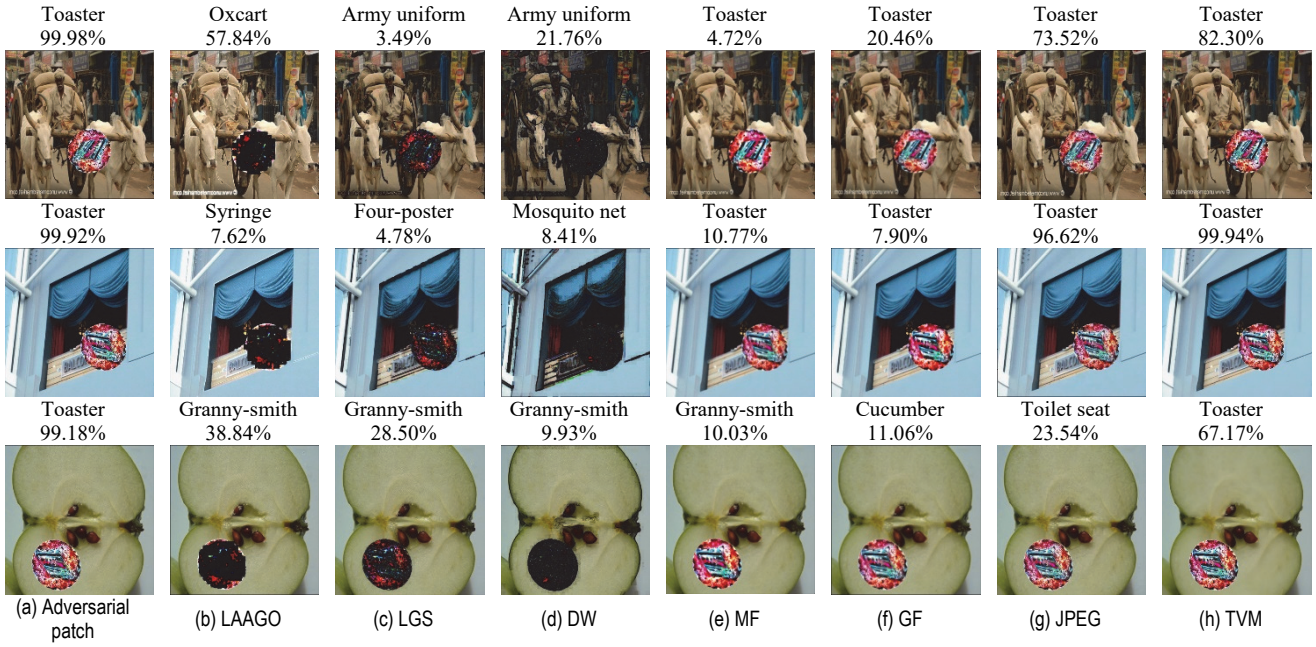


Figure 11 The Inception v3 confidence scores are shown on the example images. The images in the figure used the adversarial patch attacks, where (a) is the original adversarial sample, and (b) used the LAAGO algorithm. Columns c, d, e, f, g, and h list the results after LGS with λ equal to 2.3; DW, MF, GF, JPEG with quality 30; and TVM with weights equal to 10 processing, respectively

4.2.2 Defense Against LaVAN Attacks

LaVAN has been shown to enable attacks in any position on an image with patch sizes smaller than the adversarial patch. The defense accuracy of our algorithm generalizes surprisingly well to other attacks. We used images with a size of 299 × 299 as input to the Inception v3 model and experimented with a 42 × 42 size LaVAN patch (~2% image). For the experiments, the LaVAN patch was fixed at the position (210, 210). The experimental results are shown in Tab. 2 and Fig. 12.

Tab. 2 demonstrates the relative effectiveness of traditional denoising algorithms such as JPEG compression and MF due to the absence of an extensive range of high-frequency noise in LaVAN attacks. However, LAAGO differs in that we not only perform gradient suppression within the noise region but also highlight the target features through gradient enhancement. The advantage of LAAGO over LGS and DW in filtering local adversarial patches is shown in Fig. 12. LAAGO effectively protects the features of the original target to improve the classification accuracy. Therefore, in the face of LaVAN attacks, which are more camouflaged, LAAGO still demonstrates excellent defense performance with 72.10% defense effectiveness.

Table 2 Experimental results of various algorithms for defending against LaVAN

	No Defense	*LAAGO	LGS [$\lambda = 2.3$]	DW [window = 3]	MF [quality = 10]	JPEG [weights = 10]	TVM
LaVAN	0%	72.10%	68.50%	65.50%	60.90%	51.20%	11.40%

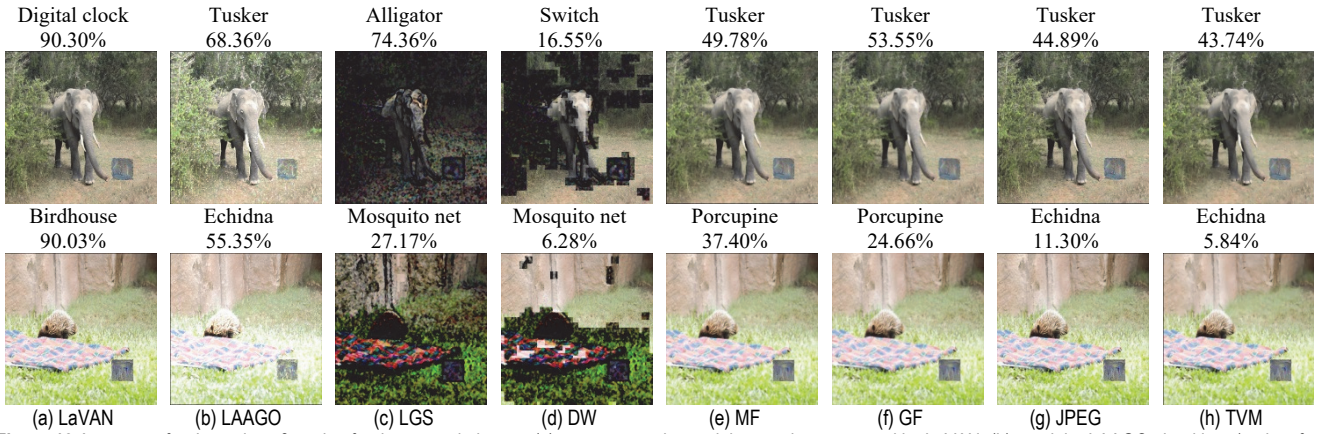


Figure 12 Accuracy after Inception v3 testing for the example image. (a) represents adversarial examples generated by LaVAN, (b) used the LAAGO algorithm. (c, d, e, f, g, h) show the results after LGS with lambda equal to 2.3; DW, MF, GF, JPEG with quality 30; and TVM with weights equal to 10 processing, respectively

5 CONCLUSION

In this paper, we aim to enhance the defense against local adversarial attacks by investigating the attack properties in the gradient domain. Traditional heuristic algorithms, although commonly used, can have a significant impact on clean images, making them challenging to implement in real-world applications. This

limitation may restrict the practicality of these algorithms when applied to scenarios where preserving the integrity of clean images is crucial.

To address these challenges, we introduce a two-fold approach. First, we apply a low-pass filter to smooth the high-frequency region of the image, aiming to minimize the influence of external factors such as environmental conditions, lighting, and equipment. By reducing high-frequency noise, we mitigate errors that may arise from screening high-frequency components. Secondly, we propose the LAAGO algorithm, which enhances the gradient texture details of the original object. By enhancing the DNN's attention towards the original object, we effectively improve the classification accuracy and algorithmic robustness of the classifier. This approach shows promise in strengthening the defense against local adversarial attacks.

Although LAAGO has demonstrated remarkable performance in experiments, it is important to acknowledge its limitations in the face of the continuous evolution and improvement of adversarial attacks. Heuristic algorithms may not be effective against new or targeted attacks, potentially leading to reduced defense effectiveness or failure when encountering new adversarial samples in the future. Furthermore, LAAGO may have weaknesses when dealing with attacks that do not exhibit obvious gradient changes.

Future research efforts will focus on extending the LAAGO algorithm to address these limitations. Specifically, our goal is to enhance the algorithm's accuracy in detecting high-frequency noise and dynamically adjust filtering thresholds, smoothing coefficients, and enhancement factors for images of different sizes. By implementing end-to-end defense strategies in practical applications, we aim to improve the overall effectiveness and applicability of the proposed defense techniques.

Acknowledgements

Funding: This work is supported in part by the National Key Research and Development Program of China (No. 2020YFB2103604).

6 REFERENCES

[1] Parkhi, O., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. *BMVC 2015 - Proceedings of the British machine vision conference*, 1-12.

[2] Marriott, R. T., Romdhani, S., & Chen, L. (2021). A 3d gan for improved large-pose facial recognition. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13445-13455.

[3] Huang, Z., Zhang, J., & Shan, H. (2021). When age-invariant face recognition meets face age synthesis: A multi-task learning framework. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7282-7291.

[4] Li, L., Zhou, T., Wang, W., Li, J., & Yang, Y. (2022). Deep hierarchical semantic segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1246-1257.

[5] Ozkahraman, M. & Livatyali, H. (2022). Artificial Intelligence in Foreign Object Classification in Fenceless Robotic Work Cells Using 2-D Safety Cameras. *Tehnicki vjesnik-Technical Gazette*, *29*(5), 1491-1498. https://doi.org/10.17559/TV-20211122150850

[6] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. *Proceedings of the IEEE international conference on computer vision*, 2961-2969.

[7] Ertuğrul, D. Ç. & Abdullah, S. A. (2022). A Decision-Making Tool for Early Detection of Breast Cancer on Mammographic Images. *Tehnicki vjesnik-Technical Gazette*, *29*(5), 1528-1536. https://doi.org/10.17559/TV-20211221131838

[8] Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on Pattern Analysis and Machine Intelligence*, *39*(12), 2481-2495. https://doi.org/10.1109/TPAMI.2016.2644615

[9] Cao, Y. (2021). Image and Graph Restoration Dependent on Generative Adversarial Network Algorithm. *Tehnicki vjesnik-Technical Gazette*, *28*(6), 1820-1824. https://doi.org/10.17559/TV-20210413111850

[10] Huval, B., Wang, T., Tandon, S., Kiske, J., Song, W., Pazhayampallil, J., Andriluka, M., Rajpurkar, P., Migimatsu, T., Cheng-Yue, R., Mujica, F., Coates, A., & Ng, A. Y. (2015). An empirical evaluation of deep learning on highway driving. arXiv preprint arXiv:1504.01716. https://doi.org/10.48550/arXiv.1504.01716

[11] Badue, C., Guidolini, R., Carneiro, R. V., Azevedo, P., Cardoso, V. B., Forechi, A., Jesus, L., Berriel, R., Paixão, T. M., Mutz, F., de Paula Veronese, L., Oliveira-Santos, T., & de Souza, A. F. (2021). Self-driving cars: A survey. *Expert Systems with Applications*, *165*, 113816. https://doi.org/10.1016/j.eswa.2020.113816

[12] Simon, J., Trojanová, M., Hošovský, A., & Sárosi, J. (2021). Neural Network Driven Automated Guided Vehicle Platform Development for Industry 4.0 Environment. *Tehnicki vjesnik-Technical Gazette*, *28*(6), 1936-1942. https://doi.org/10.17559/TV-20200727095821

[13] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083. https://doi.org/10.48550/arXiv.1706.06083

[14] Moosavi-Dezfooli, S. M., Fawzi, A., & Frossard, P. (2016). Deepfool: A simple and accurate method to fool deep neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2574-2582.

[15] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572. https://doi.org/10.48550/arXiv.1412.6572

[16] Akhtar, N., Mian, A., Kardan, N., & Shah, M. (2021). Advances in adversarial attacks and defenses in computer vision: A survey. *IEEE Access*, *9*, 155161-155196. https://doi.org/10.1109/ACCESS.2021.3127960

[17] Kurakin, A., Goodfellow, I. J., & Bengio, S. (2018). Adversarial examples in the physical world. *Artificial intelligence safety and security*, 99-112.

[18] Xie, C., Zhang, Z., Zhou, Y., Bai, S., Wang, J., Ren, Z., & Yuille, A. L. (2019). Improving transferability of adversarial examples with input diversity. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2730-2739.

[19] Brown, T. B., Mané, D., Roy, A., Abadi, M., & Gilmer, J. (2017). Adversarial patch. arXiv preprint arXiv:1712.09665. https://doi.org/10.48550/arXiv.1712.09665

[20] Sharif, M., Bhagavatula, S., Bauer, L., & Reiter, M. K. (2016). Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 1528-1540.

[21] Asriny, D. M. & Jayadi, R. (2023). Transfer learning VGG16 for classification orange fruit images. *Journal of System and Management Sciences*, *13*(1), 206-217. https://doi.org/10.33168/JSMS.2023.0112
[22] Low, M. X., Yap, T. T. V., Soo, W. K., Ng, H., Goh, V. T., Chin, J. J., & Kuek, T. Y. (2022). Comparison of label encoding and evidence counting for malware classification. *Journal of System and Management Sciences*, *12*(6), 17-30. https://doi.org/10.33168/JSMS.2022.0602
[23] Nayak, S., Panigrahi, C. R., Pati, B., Nanda, S., & Hsieh, M. (2022). Comparative Analysis of HAR Datasets Using Classification Algorithms. *Computer Science and Information Systems*, *19*(1), 47-63. https://doi.org/10.2298/CSIS201221043N
[24] Kim, K., Lee, J., Lim, H., Oh, S. W., & Han, Y. (2022). Deep RNN-Based Network Traffic Classification Scheme in Edge Computing System. *Computer Science and Information Systems*, *19*(1), 165-184. https://doi.org/10.2298/CSIS200424038K
[25] Khow, Z. J., Goh, K. O. M., Tee, C., & Law, C. Y. (2022). A YOVO5 based real-time helmet and mask detection system. *Journal of Logistics, Informatics and Service Science*, *9*(3), 97-111. https://doi.org/10.33168/LISS.2022.0308
[26] Hong, L. C., Tee, C., & Goh, M. K. O. (2022). Activities of daily living recognition using deep learning approaches. *Journal of Logistics, Informatics and Service Science*, *9*(4), 129-148. https://doi.org/10.33168/LISS.2022.0410
[27] Rao, S., Stutz, D., & Schiele, B. (2020). Adversarial training against location-optimized adversarial patches. *Computer Vision–ECCV 2020 workshops*, 429-448.
[28] Mu, N. & Wagner, D. (2021). Defending against adversarial patches with robust self-attention. *ICML 2021 workshop on uncertainty and robustness in deep learning*.
[29] Chiang, P. Y., Ni, R., Abdelkader, A., Zhu, C., Studer, C., & Goldstein, T. (2020). Certified defenses for adversarial patches. arXiv preprint arXiv:2003.06693. https://doi.org/10.48550/arXiv.2003.06693
[30] Levine, A. & Feizi, S. (2020). (De) Randomized smoothing for certifiable defense against patch attacks. *Advances in Neural Information Processing Systems*, *33*, 6465-6475.
[31] Salman, H., Jain, S., Wong, E., & Madry, A. (2022). Certified patch robustness via smoothed vision transformers. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15137-15147.
[32] Hayes, J. (2018). On visible adversarial perturbations & digital watermarking. *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 1597-1604.
[33] Naseer, M., Khan, S., & Porikli, F. (2019). Local gradients smoothing: Defense against localized adversarial attacks. *2019 IEEE winter conference on applications of computer vision*, 1300-1307.
[34] Dziugaite, G. K., Ghahramani, Z., & Roy, D. M. (2016). A study of the effect of jpg compression on adversarial images. arXiv preprint arXiv:1608.00853. https://doi.org/10.48550/arXiv.1608.00853
[35] Xu, W., Evans, D., & Qi, Y. (2017). Feature squeezing: Detecting adversarial examples in deep neural networks. arXiv preprint arXiv:1704.01155. https://doi.org/10.48550/arXiv.1704.01155
[36] Guo, C., Rana, M., Cisse, M., & Van Der Maaten, L. (2017). Countering adversarial images using input transformations. arXiv preprint arXiv:1711.00117. https://doi.org/10.48550/arXiv.1711.00117
[37] Karmon, D., Zoran, D., & Goldberg, Y. (2018). Lavan: Localized and visible adversarial noise. *International Conference on Machine Learning*, 2507-2515.
[38] Athalye, A., Engstrom, L., Ilyas, A., & Kwok, K. (2018). Synthesizing robust adversarial examples. *International conference on machine learning*, 284-293.
[39] Telea, A. (2004). An image inpainting technique based on the fast marching method. *Journal of Graphics Tools*, *9*(1), 23-34. https://doi.org/10.1080/10867651.2004.10487596
[40] Simonyan, K. & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. https://doi.org/10.48550/arXiv.1409.1556
[41] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818-2826.
[42] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*, 211-252. https://doi.org/10.1007/s11263-015-0816-y
[43] Xu, L., Lu, C., Xu, Y., & Jia, J. (2011). Image smoothing via L 0 gradient minimization. *Proceedings of the 2011 SIGGRAPH Asia conference*, 1-12.
[44] Zhang, T. & Zhu, Z. (2019). Interpreting adversarially trained convolutional neural networks. *International conference on machine learning*, 7502-7511.

Contact information:

Boyang SUN, Postgraduate
School of Electrical and Information Engineering,
Beijing University of Civil Engineering and Architecture,
Beijing 100044, China
E-mail: sunboyang0716@gmail.com

Xiaoxuan MA, Associate Professor
(Corresponding author)
School of Electrical and Information Engineering,
Beijing University of Civil Engineering and Architecture,
Beijing 100044, China
E-mail: maxiaoxuan@bucea.edu.cn

Hengyou WANG, Professor
School of Science,
Beijing University of Civil Engineering and Architecture,
Beijing 100044, China
E-mail: wanghengyou@bucea.edu.cn