EURASIP Journal on Wireless Communications and Networking

RESEARCH Open Access

# Defense against adversarial attacks in traffic sign images identification based on 5G

Fei Wu*, Limin Xiao*, Wenxue Yang and Jinbin Zhu

* Correspondence: woofei@buaa.edu.cn; xiaolm@buaa.edu.cn
School of Computer Science and Engineering, Beihang University, Beijing 100191, China

## Abstract

In the past decade, artificial intelligence and Internet of things (IoT) technology have been rapid development, gradually began to integrate with each other, especially in coming 5G era. Admittedly, image recognition is the key technology due to a huge number of video cameras integrated in intelligent IoT equipment, such as driverless cars. However, the rapidly growing body of research in adversarial machine learning has demonstrated that the deep learning architectures are vulnerable to adversarial examples. Thus, the raises questions about the security of intelligent Internet of thing (IoT) and trust sensitive areas. This emphasizes the urgent need for practical defense technology that can be deployed to real-time combat attacks at any time. Well-crafted small perturbations lead to the misclassification of legitimate images by neural networks, but not the human visual system. It is worth noting that many attack strategies are designed to disrupt image pixels in a visually imperceptible manner. Therefore, we propose a new defense method and take full advantage of 5G high-speed bandwidth and mobile edge computing (MEC) effectively. We use singular value decomposition (SVD) which is the optimal approximation of matrix in the sense of square loss to eliminate the perturbation. We have conducted extensive and large-scale experiments with German Traffic Sign Recognition Benchmark (GTSRB) datasets and the results show that adversarial attacks, such as Carlini-Wagner's $l_2$, Deepfool, and I-FSGM, can be better eliminated by the method and provide lower latency.

**Keywords:** Traffic signs, Adversary attacks, 5G, Defense, Deep learning

## 1 Introduction

In recent years, under the background of the continuous expansion of data scale and the great improvement of computing power, artificial intelligence, and IoT technology has developed rapidly. For example, deep learning has achieved far better performance than others in the fields of computer vision, speech recognition, and natural language processing which make humans want to integrate deep learning technology into the IoT equipment to make them capable of making decisions especially image classification and target tracking. However, its security problems are also constantly exposed in the rapid development, few people pay attention about that. In pattern recognition, the adversary adds carefully designed perturbation to the image to generate adversarial examples. Due to the linear nature of high-dimensional space, the influence of this small

© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

pixel change on the feature space is magnified, and finally, misleads the deep learning model to make a high-confidence misclassification. As shown in Fig. 1, the stop sign is identified as flowerpot by YOLO multiple object detector. The result is the applications of driverless vehicles etc. are faced with a serious security threat; up to now, adversarial defense is still a great obstacle to the popularization of artificial intelligence in the field of reliability.

In an unmanned system, vehicles obtain external road information through lidar, camera, optical detector, and so on. How to identify traffic signs quickly and accurately is a key point in the field of self-driving. A typical traffic sign image recognition system based on mobile edge computing (MEC) in the 5G network is shown in Fig. 2.

Coincidentally, we generate scale-invariant adversarial examples to test the safety and security of driverless vehicles which proved safety by contribution [1], but the result is thought provoking. We print the well-crafted scale-invariant adversarial examples out with a color printer and put it in the forward view of the driverless car. The system of driverless cars classifies it as a monitor or computer but not a traffic light. Changing the angle also produce the same result, as Fig. 3. The drawbacks of traditional driverless cars are obvious, the inducement is that the previous communication transmission rate cannot achieve real-time data transmission, analysis, and interaction. Not only that, but other intelligent IoT equipment also faces the same problem. But with the development of 5G remote communication technologies for vehicle-to-everything (V2X) communications, driverless car can upload live-traffic to MEC node and monitoring center of synchronization analysis and decision-making, the problem can be solved easily.

Due to the success of adversarial examples attack the DNN classifiers, they are increasingly being used as part of control pipelines in the real-world system such as driverless car [2, 3], UAVs [4, 5], and robots [6]. This threat has attracted people's attention, adversarial example begins in the digital domain and more recently extends to the physical domain [7, 8].

To eliminate the safety problems caused by scale-invariant adversarial samples in the field of self-driving. On the one hand, some scholars aim to improve the robustness of deep learning model to adversarial attack, the training set contains both real images and adversarial images which is called defensive distillation. On the other hand, other scholars try to preprocess the input image to eliminate the perturbation on adversarial examples. Shortly before, Yue Zhou et al. [9] proposed an OPV (overall probability

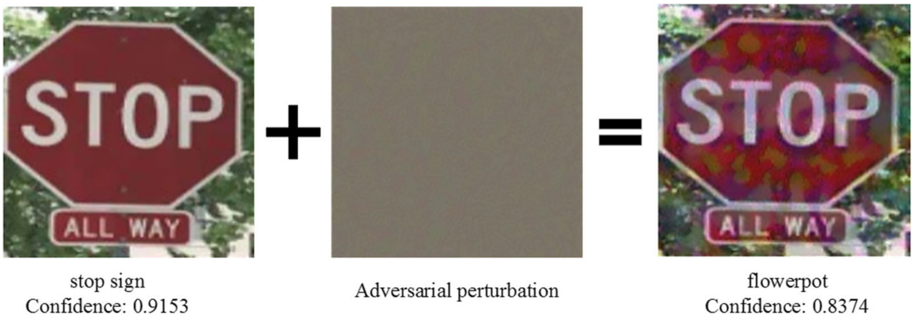

**Fig. 1** This is an adversarial example crafted for deep learning model (this figure shows an adversarial example generated by FGSM method)

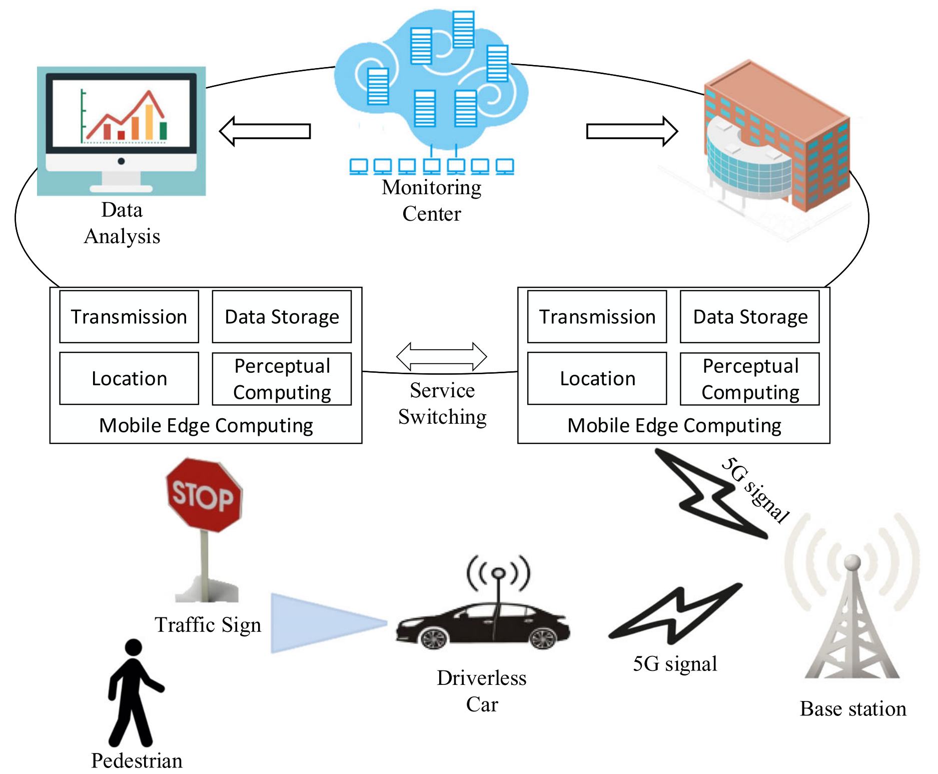

**Fig. 2** Typical traffic sign image recognition system based on the IoT and 5G (this figure shows the working mode of driverless vehicles based on 5G and mobile edge computing technology)

value) defense algorithm based on Markov chain to defend adversarial examples. The algorithm regards the image as a Markov chain composed of pixels, the value of each pixel is only related to the previous pixel, and the correlation of adjacent pixels is used as an index to judge the level of image cleanliness. By resetting the pixel value to purify the adversarial examples, deep neural network classifiers can achieve correct classification. Kejun Zhang et al. [10] proposed a general method for simultaneous detection of perturbation images and generate clean images using a single model. The RGB image is

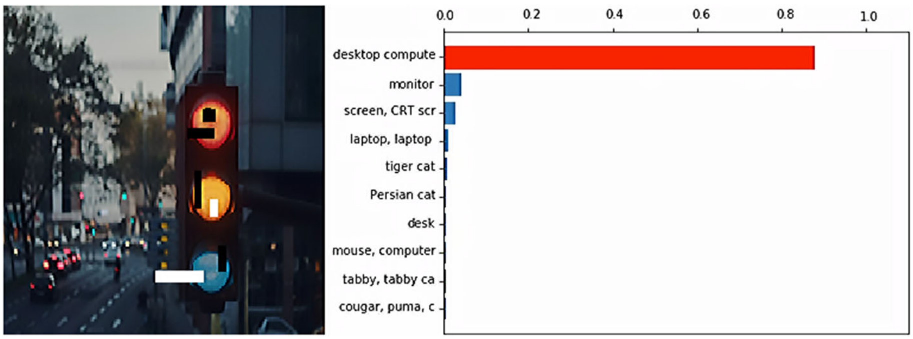

**Fig. 3** Error target identification by driverless car system (this figure shows the drawbacks of traditional local unmanned systems)

transformed into a YCrCb color space which is easier to extract edge information. Then Scharr operator is used to extract edge information to identify whether the image is clean or not. The image edge information matrix is transformed into a uniform size glcm (gray level co-occurrence matrix) and input into a deep neural network with separable depth convolution to complete training and detection. Songkui Chen et al. [11] proposed a new image denoising algorithm to solve the problem of serious loss of details in traditional denoising algorithms. The algorithm uses GANs to learn the image denoising process. The generated network takes the adversarial example as input and the denoised image as output. The discriminant network is used for distinguishing the detail loss of the denoised image and the real image. Based on maintaining good performance, the network achieves a better effect of detail retention. Li Yuancheng et al. [12] proposed a U-Net-based deep denoising neural network (UDDN) to remove perturbation from adversarial examples. The main idea is to take the acquisition of noise as the learning goal, then subtract the noise from the adversarial image to get a clean image. Nilaksh Das et al. proposed the use of JPEG compression converts different regions of the image into different compression coefficients to resist adversarial disturbance and achieve better results [13]. Chuan Guo et al. make comprehensive use of image transformation techniques, such as cropping-rescaling, bit-depth reduction, total variance minimization, and quilting, to construct a multilevel defense system and demonstrate the effectiveness of the method by experiments [14]. Cihang Xie et al. proposed a method of image random resizing and adversarial training to mitigate the influence of perturbation on images, and proved by experiments that randomization has little effect on the classification of normal images, but it can significantly increase the correct classification rate of adversarial examples [15].

### 1.1 Our contributions and impact

In this paper, we propose a new technique capable of effectively mitigating adversarial examples and prior knowledge about potential attacks is hardly required and consider the real situation of self-driving in 5G environment, the problem of adversarial attacks on object recognition can be effectively solved through singular value decomposition and 5G network, the process is as follows.

1. The high bandwidth in the 5G environment can send the road condition image captured by the unmanned vehicle back to the MEC node in real-time. At same time, the YOLO multiple object detector recognizes the image locally.
2. In MEC node, received images are implemented singular value decomposition and retain a certain proportion of singular values. Then, the image is reconstructed by increasing 10 singular values in turn and gets $N/10$ images ($N$ denote the number of retained singular values). The trained deep learning model is used to identify these images. The label with the largest number of occurrences and the highest average confidence is the final output.
3. The result is sent back the unmanned vehicle through the 5G network. This would help the unmanned vehicle to mitigate the misclassification caused by the adversarial examples.

## 2 Background: adversarial attacks

In this part, we briefly explain the generation principle of adversarial samples and its applicable scenarios. Assuming there is an image classification C and an image $x \in X$, $X = [0, 1]^{H * W * C}$, thus an untargeted adversarial example $x^{'}$ can be defined as $x^{'} \in X$, $C(x) \neq C(x^{'})$ $d(x, x^{'}) \leq \rho$, where $d(,)$ is the dissimilarity function meanwhile $\rho \geq 0$. In practice, $d(,)$ often use Euclidean distance $d(x, x^{'}) = \| x - x^{'} \|_2$ or Chebyshev distance $d(x, x^{'}) = \| x - x^{'} \|_\infty$. Targeted attack is similar to an untargeted attack, but a specific target label needs to be specified, i.e., $C(x^{'}) = k$. It also has been proved that it is feasible to train an alternative model that is highly similar to the target model in practice once the target model is a DNN trained by gradient back propagation [16, 17].

There are four highly aggressive attacks, I-FGSM (iterative fast gradient sign method), Deepfool, Carlini and Wagner Attacks (C&W), and JSMA (Jacobian-based Saliency Map Attack), detailed information shown in Table 1.

### 2.1 Method generating adversarial example

In this section, we use four methods to generate adversarial images and briefly describe them. There are two main ways to generate adversarial attacks, single-step, which only perform gradient computation once, and iterative method, which perform multiple times. In theory, the perturbation of a single-step attack is weak, and some of them not cause misclassification. On the contrary, iterative attacks have strong aggression for a specific network but less transferability.

#### *2.1.1 Iterative fast gradient sign method attack (I-FGSM)*

This is the iterative version of the fast gradient sign method attack (FGSM) [18], which computing perturbations subject to an $l_\infty$ constraint. FGSM is led by minimizing loss function $J(x^*, y)$, normally the loss function is cross entropy [19].

$$x^* = x + \in * \, sign(\nabla_x J(x, y)) \tag{1}$$

Its iterative version is represented as follows. Usually, $\alpha = \epsilon / T$, $\epsilon$ is disturbance, $T$ refers to the number of iterations. The iterative version has stronger white-box attack ability, but the transferability of the attack sample is poor.

$$x_0^* = x, x_{t+1}^* = x_t^* + a * \, sign\left(\nabla_x J\left(x_t^*, y\right)\right) \tag{2}$$

#### *2.1.2 Carlini and Wagner attacks (C&W)*

C&W is an optimization-based attack, which adds a relaxation term to the perturbation minimization problem of model-based differentiable alternatives. This attack influence

**Table 1** Typical Iterative attack method

| Method | I-FGSM | C&W | Deepfool | JSMA |
|---|---|---|---|---|
| Category | White box | White box | White box | White box |
| Target | Target or non-target | Target or non-target | Target or non-target | Target or non-target |
| Specific | Image specific | Image specific | Image specific | Image specific |
| Perturbation | $l_\infty$ | $l_0$, $l_2$, $l_\infty$ | $l_2$, $l_\infty$ | $l_0$ |
| Learning | Iterative | Iterative | Iterative | Iterative |
| Strength | Strong | Strong | Strong | General |

$l_0$, $l_2$, and $l_\infty$ and it is highly aggressive. Such attack can be described as the following minimization problems [20].

$$\|x-x'\|_2+\lambda\ \max\left(-\kappa, Z\left(x'\right)_k-\ \max\left\{Z\left(x'\right)_{k'}: k'\neq k\right\}\right) \tag{3}$$

Where $\kappa$ controls the confidence with which an image is misclassified by the target model, and $Z$ (*) is the output from the logit layer.

#### *2.1.3 Deepfool*

Deepfool attack is introduced by Moosavi Dezfooli et al. [21], which includes the target version and the non-target version. The author proves an effective method to apply the minimum perturbations to the misclassification under the $l_2$ distance metric. The method performs iterative steps for the antagonistic direction of the gradient provided by the local linear approximation of the classifier until the hyperplane crossover is made. This process can be described as follows:

$$\rho_{adv}\left(\hat{k}\right)=E_x\frac{\Delta\left(x;\hat{k}\right)}{\|x\|_2} \tag{4}$$

$$\Delta\left(x;\hat{k}\right)\coloneqq\ \min_r\ \|r\|_2 subject\ to\, \hat{k}(x+r)\neq\hat{k}(x) \tag{5}$$

Where $\Delta(x;\hat{k})$ denote the robustness of $\hat{k}$ at point $x$, $r$ denotes the perturbation, $\hat{k}(x)$ is the estimated label, and $E_x$ is the expectation over the distribution of data.

#### *2.1.4 Jacobian-based saliency map attack (JSMA)*

Giving the saliency map computed by the model's Jacobian matrix, the attack tries to modify the most significant pixel at each iteration until the prediction has changed to the target class. This attack attempts to create an adversarial disturbance at $l_0$ distance metric [22]. Here, it needs to be explained that when calculating the gradient, the FGSM and Deepfool discussed earlier are obtained by deriving the loss function, while the forward derivative in JSMA is obtained by deriving the output of the last layer of the neural network. The specific calculation process of the forward derivative ∇F(X) is as follows. Where $j$ represents the corresponding output classification and $i$ represents the corresponding input characteristics.

$$\nabla \mathrm{F}(x)=\frac{\partial F(X)}{\partial X}=\left[\frac{\partial F_j(X)}{\partial x_i}\right]_{i\in 1\ldots M, j\in 1\ldots N} \tag{6}$$

$$\frac{\partial F_j(X)}{\partial x_i}=\left(W_{n+1,j}\cdot\frac{\partial H_n}{\partial X_i}\right)*\frac{\partial f_{n+1,j}}{\partial x_i}\left(W_{n+1,j}\cdot H_n+b_{n+1,j}\right) \tag{7}$$

According to the different disturbance modes (forward disturbance and reverse disturbance), two methods for calculating antagonistic salience maps are proposed.

$$S(X,t)[i] = \left\{ \begin{array}{l} 0,\ \ if\,\dfrac{\partial F_t(X)}{\partial X_i} < 0\ \ or\ \ \displaystyle\sum_{j\neq t}\dfrac{\partial F_j(X)}{\partial X_i} > 0 \\ \left(\dfrac{\partial F_j(X)}{\partial X_i}\right)\left|\displaystyle\sum_{j\neq t}\dfrac{\partial F_j(X)}{\partial X_i}\right| otherwise \end{array} \right\} \tag{8}$$

$$S(X,t)[i] = \left\{ \begin{array}{l} 0,\ \ if\,\dfrac{\partial F_t(X)}{\partial X_i} > 0\ \ or\ \ \displaystyle\sum_{j\neq t}\dfrac{\partial F_j(X)}{\partial X_i} < 0 \\ \left(\dfrac{\partial F_j(X)}{\partial X_i}\right)\left|\displaystyle\sum_{j\neq t}\dfrac{\partial F_j(X)}{\partial X_i}\right| otherwise \end{array} \right\} \tag{9}$$

According to the characteristics obtained from the antagonistic salience map, disturbance (forward disturbance or reverse disturbance) can be added to it. If the added disturbance is not enough to change the classification results, the disturbed samples can be used to repeat the above process.

## 3 Methodology

The basic idea of defending adversarial samples is to eliminate or destroy the negligible perturbation of the input before being identified by the target model. The infinitesimal $\eta$ difference between adversarial examples $X^{ADV}$ and clean images, $X$ can be expressed as follows.

$$\left\|X^{ADV} - X\right\| = \eta \tag{10}$$

As mentioned earlier, the adversarial sample only adds a slight perturbation to the normal image. These perturbations hardly affect the human visual system, but intelligent IoT equipment cannot work properly. Therefore, we try to perform singular value decomposition on the adversarial examples to eliminate or filter out certain parts of adversarial perturbation to restore the correct decision of the neural network model.

### 3.1 Theoretical background

For each $A \in C^{m*n}$, singular values $\delta_1$, $\delta_2$, ..., $\delta_r$ of a matrix is unique, it describes the distribution characteristics of matrices. The matrix A is regarded as a linear transformation, which maps the points of $m$-dimensional space to $n$-dimensional space. After singular value decomposition, the transformation is divided into three parts, $U$, $\varDelta$, and $V$, where $U$ and $V$ are standard orthogonal matrices. Orthogonal transformation can reduce the correlation of image data, obtain the overall characteristics of the image, and help to represent the original image with less data, which is very meaningful for image analysis, storage, and image transmission.

If $A$ is a digital image, then $A$ can be regarded as two-dimensional time-frequency information, the singular value decomposition formula of $A$ is expressed as follows:

$$A = UDV^H = U\begin{bmatrix} \nabla & 0 \\ 0 & 0 \end{bmatrix} V^H = \sum_{i=1}^{r} \delta_i u_i v_i^H \tag{11}$$

To make it more intuitive, we graphically show the decomposition process on 2 × 2 matrixes in Fig. 4. Where $u_i$ and $v_i$ are column vectors of $U$ and $V$, $\delta_i$ is a non-zero singular value of $A$. The image can be regarded as the result of the superposition of $r$ subgraphs. After singular value decomposition, the texture and geometric information of the image are concentrated in $U$, $V$, while the singular value in $\varDelta$ represents the energy information of the image. Taking the RGB image as an example, the process of singular value decomposition is shown in Fig. 5.

At the same time, the singular value of matrix has the following properties:

Property 1 The singular value is stable. Assuming $A$, $B \in C^{m*n}$, the singular values of $A$ and $B$ are $\boldsymbol{\lambda_1} \geq \boldsymbol{\lambda_2} \geq \ldots \geq \boldsymbol{\lambda_p}$ and $\boldsymbol{\tau_1} \geq \boldsymbol{\tau_2} \geq \ldots \geq \boldsymbol{\tau_p}$ ($p = min\,(m, n)$), there is $|\boldsymbol{\lambda_i} - \boldsymbol{\tau_i}| \leq ||A - B||_{\mathbf{2}}$. This property indicates that singular values of the image do not change much after passing through the SVD for the image with color change, noise interference and so on.

Property 2 Singular values are proportional invariant. The singular values of matrix $A$ and $kA$ are $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ and $\tau_1 \geq \tau_2 \geq \ldots \geq \tau_p \geq 0$, there is $|\mathrm{a}|\,(\lambda_1, \lambda_2, \ldots, \lambda_p) = (\tau_1, \tau_2, \ldots, \tau_p)$. This property indicates that the normalization process to eliminate the effect of amplitude on feature extraction will not change the relative size of the singular value.

Property 3 Singular values have rotation invariance. If $P$ is a unitary matrix, then the singular value of matrix $P$A is the same as that of matrix $A$, i.e., $|AA^H - \sigma^j I| = |PA(PA)^H - \sigma^j I| = 0$.

Property 4 Matrix approximation. Assuming $A \in C^{m*n}$, $rank(A) = r \geq s$, if $\varDelta_s = diag\,(\delta_1, \delta_2, \ldots, \delta_s)$, $A_s = \sum_{i=1}^{s} \delta_i u_i v_i^H$, $rank(A_s) = rank\,(\Delta_s) = s$. There is the following conclusion:

$$\|A - A_s\|_F = \min\left\{||A - B||_F | B \in C^{m*}\right\} \tag{12}$$

The above formula shows that in the sense of $F$-norm, matrix $A_s$ is the best approximation of matrix $A$ in space $C_s^{m*n}$. Therefore, according to the need to retain $s(s < r)$ singular values greater than a certain threshold and discard the remaining $r - s$ singular

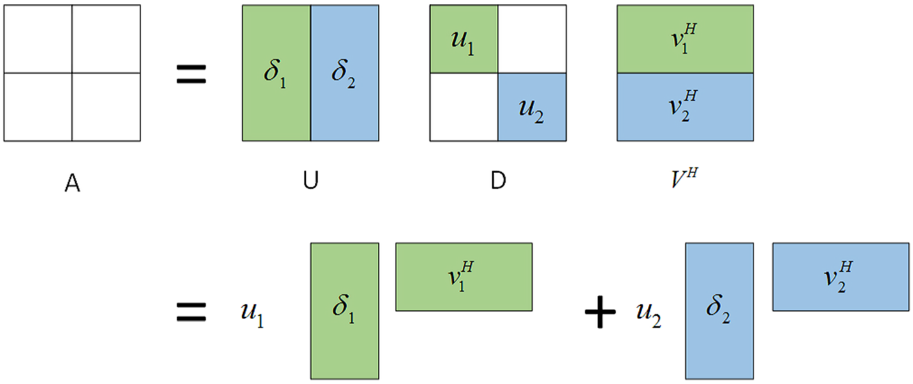

**Fig. 4** Singular value decomposition process of 2 × 2 matrix (this figure shows the singular value decomposition of a 2 × 2 matrix)

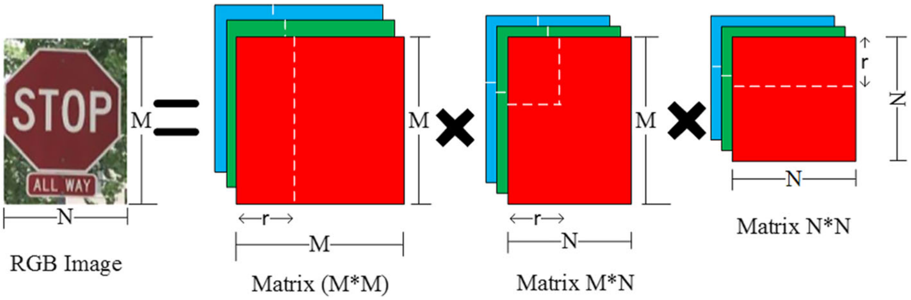

**Fig. 5** Singular value decomposition of pixel matrix (the "*r*" in the figure represents the retained singular values and corresponding eigenvectors)

values to ensure the approximation of the two matrixes in a certain sense. This property can be used for matrix reduction, and data compression also provides a basis for finding the "essential information" retained in the adversarial samples and removing the perturbation.

### 3.2 Defense method

The road condition information is captured by the vehicle camera for target identification and uploaded to MEC node through 5G. Then the MEC node synchronizes the analysis of signposts and road conditions using the SVD methodology mentioned above, the result immediately returned to the corresponding unmanned car and helps it to make the correct decision as shown in Fig. 6. This would significantly decrease the security risk. The threshold we set retain first $n$ singular values $\{\delta_1, \delta_2, \ldots, \delta_n\}$ $(\delta_1 \geq \delta_2 \geq \ldots \geq \delta_n)$ and satisfy the equality relation $0.6\sum_{i=1}^{s}\delta_i \leq \sum_{i=1}^{n}\delta_i \leq 0.7\sum_{i=1}^{s}\delta_i$.

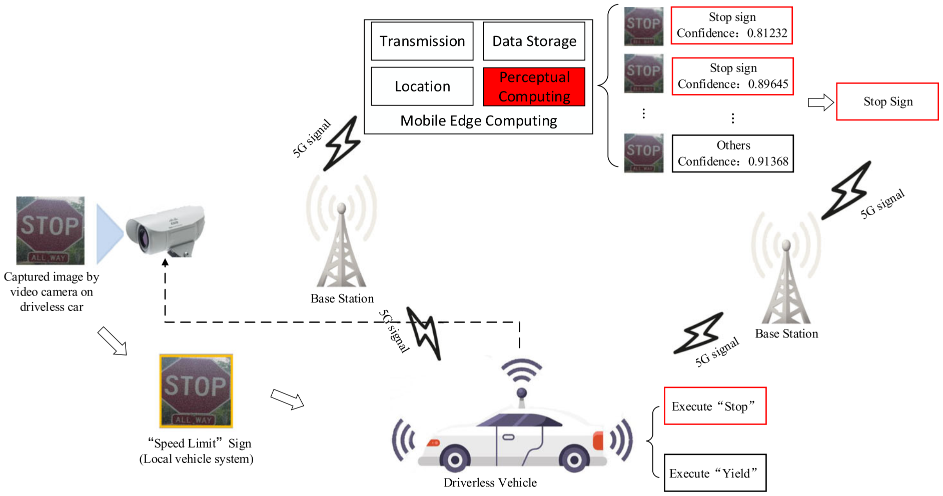

**Fig. 6** The pipeline of our defense model (this figure shows the perturbed "stop sign" image uploads to the MEC node and processed by singular value decomposition. Then the MEC node sends the correct decision to the corresponding driverless car and helps it execute the right instruction)

## 4 Results and discussion

### 4.1 Experimental setup

Specifically, we choose German Traffic Sign Recognition Benchmark dataset [23] to perform the experiment, the deep neural network classifier we choose google Inception-v4 [24] and achieve 96.8% accuracy on test set. This model is named "target model" and the target model combine defense method is named "defense model".

With respect to adversarial example generation, we explain as follows. For FGSM attack, the $\varepsilon$ value is a key parameter. Namely, the size of $\varepsilon$ directly affects the success rate of adversarial example generation. The larger $\varepsilon$ is probably introducing obviously perturbation and easily being discovered. So, we craft the adversarial examples with appropriate different $\varepsilon$ values than 2/255 for ImageNet images. For Deepfool attack, we use the default settings to generate Deepfool examples. The algorithm alters the image more unimpressive than FGSM, but the generated adversarial examples still fool the DNN classifiers successfully. For C&W attack, it is an optimization-based algorithm and its aim is to seek out perturbation as small as possible. In other word, it can find closer adversarial examples than the other attack techniques. For instance, C&W $l_2$ attack craft adversarial examples with much lower distortion than FGSM and $\kappa$ we set respectively 2.0 in the experiment. For JSMA attack, we use the default setting to generate corresponding adversarial examples. It should be noted that JSMA is perceptible perturbations, which limits the number of altered pixels but not the amplitude of the pixels. This may cause that generated adversarial examples are easy to spot, but we still keep them in our experiment.

With respect to the defense model, it consists of the original networks we list above, and there add two layers. For the SVD layer, singular value decomposition is performed on the pixel matrix of the image. In line with the equivalence relationship, $0.6\sum_{i=1}^{s}\delta_i \leq \sum_{i=1}^{n}\delta_i \leq 0.7\sum_{i=1}^{s}\delta_i$ remains about the first 60% of singular values in descending order and reconstruct the images according to the gradient.

### 4.2 Result of clean image test

We first seek confirmation whether these defense operations have an impact on the recognition rate of the normal image. We see that the combined operation of SVD has little effect on the recognition accuracy of normal images. The accuracy of the defense model to normal images is reduced by an average of 1.746%. The processed normal image is shown in Fig. 7.

### 4.3 Singular value analysis

Before testing the defense model, we first analyze the influence of the size of the singular value on the image reconstruction and design a comparison of the difference in singular values between normal images and adversarial samples. As shown in Fig. 8, the image reconstructed by the larger singular values in the front has a high similarity with the original image, but continue to add the singular value, the change is no longer obvious. Whereafter, we compared the singular values of normal images with the corresponding adversarial example and randomly sample the singular values of 18 locations, plot them in Fig. 9. It shows that the singular values in the middle and tail of the adversarial samples have a large offset when compared with the normal images. We

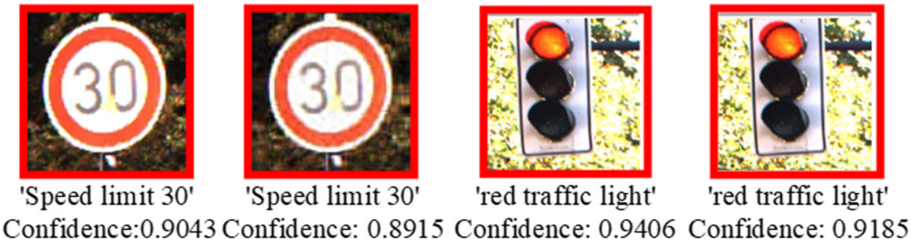

**Fig. 7** Classification comparison of normal images and their processing images (this figure shows the impact of the proposed method on clean traffic sign images. The left is original captured image, the right is processed image)

speculate that the use of these relatively "pure" singular values to reconstruct the image is helpful to the correct classification of the adversarial examples.

### 4.4 Result of different attack scenario experiment

In this section, we evaluate the proposed method on different attack scenarios. Considering the actual situation, we perform the poster-printing attacks and sticker attacks. In the poster-printing attacks scenario, we refer to the experiment of Kurakin et al. [25] and generate adversarial examples by methods mentioned in section 2. From Table 2, we can see that the defense model can effectively mitigate adversarial effects for poster-printing attacks. Especially when defense strategy combined with adversarial training, they perform very well on all attacks, adversarial examples achieve a very high attack rate on the target model, but hardly pass the defense model.

In the sticker attack scenario, we try to generate the physical perturbations in the form of stickers. Resembling graffiti or art on the signpost to misclassify the target

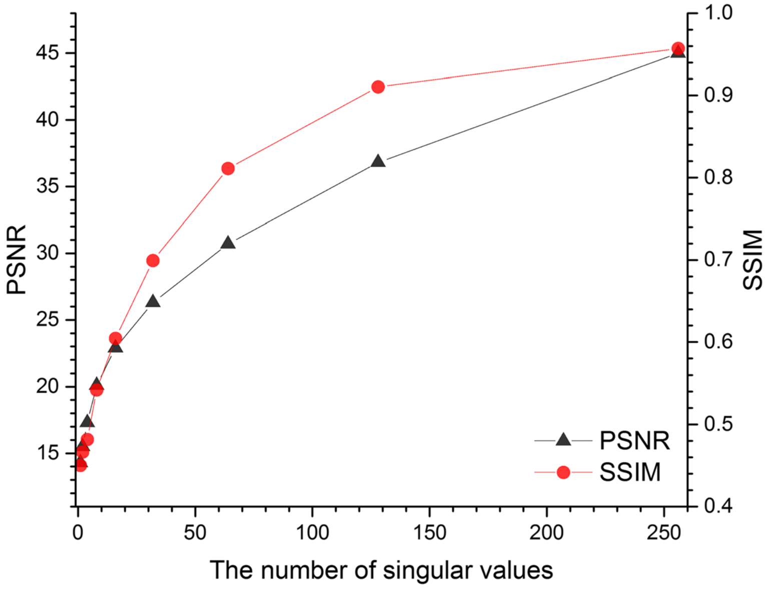

**Fig. 8** The relationship of PSNR and SSIM with the number increasement of singular values (this figure shows the singular value of the first 60% to 70% can demonstrate the main information of the image. This means the SSIM value should exceed 0.9 or the PSNR value exceed 40)

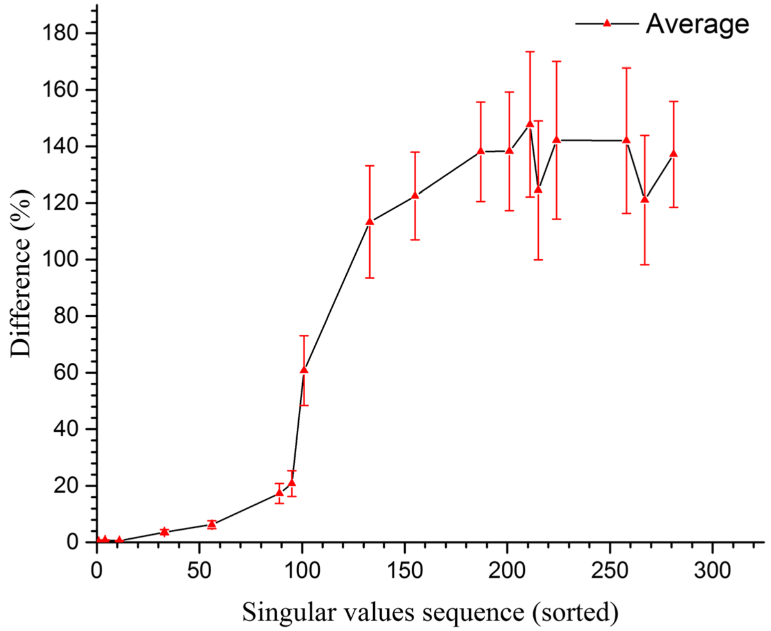

**Fig. 9** The difference of singular values between normal images and adversarial examples (this figure shows the larger singular value at the top of the sort is less likely to be affected by adversarial perturbation, but it is not true for the little one)

model, the result shown in Table 3. The sticker camouflage graffiti attack and art attacks both have aggressive adversarial attacks, but they can be easily mitigated by defense model which includes 5G technology and image singular value decomposition.

### 4.5 Result of comparison experiment

This section, we conduct a comparison with the methods based on generative adversarial networks [26], JPEG compression [12] and image hybrid transformation which include cropping, TVM, and quilting (hereinafter referred to as CTQ) [13], result as shown in Table 4. The target model we choose *Inceptionv4*, and every method we set 5G, 4G, and offline model. Compared with other methods, we can clearly see the SVD+5G has certain advantages than other models. Importantly, the proposed method

**Table 2** Top-1 classification accuracy under the poster-printing attack scenario (M means meter, ° means angel. The former is the accuracy of the target model, the latter is the accuracy of the defense model)

| Setting | I-FGSM | C&W | Deepfool | JSMA |
|---|---|---|---|---|
| 5M 0° | 13.73%/82.91% | 5.91%/90.21% | 22.41%/82.61% | 15.71%/90.66% |
| 5M 15° | 11.27%/81.75% | 6.06%/84.82% | 23.04%/81.25% | 23.22%/89.34% |
| 5M 30° | 12.98%/85.96% | 10.31%/88.18% | 30.96%/87.15% | 26.65%/88.52% |
| 10M 0° | 12.96%/87.65% | 9.58%/87.52% | 25.63%/88.64% | 18.96%/90.72% |
| 10M 15° | 10.62%/89.51% | 8.69%/82.65% | 27.12%/89.61% | 23.41%/89.26% |
| 10M 30° | 6.09%/85.63% | 13.37%/85.31% | 29.59%/90.13% | 25.67%/89.17% |
| 20M 0° | 16.64%/87.69% | 14.24%/84.93% | 20.76%/87.69% | 21.19%/88.93% |
| 30M 0° | 27.21%/90.39% | 12.17%/87.45% | 19.36%/88.42% | 21.39%/89.93% |

**Table 3** Top 1 classification accuracy under the graffiti attack and art attacks scenario (%)

| Models | Graffiti | Art |
|---|---|---|
| 5M 0° | 43.73%/89.78% | 36.23%/90.21% |
| 5M 15° | 41.27%/87.65% | 56.73%/91.56% |
| 5M 30° | 28.98%/86.42% | 57.49%/93.41% |
| 10M 0° | 46.96%/85.74% | 23.58%/92.37% |
| 10M 15° | 10.62%/86.93% | 28.43%/93.20% |
| 10M 30° | 36.42%/86.29% | 20.81%/90.79% |
| 20M 0° | 38.96%/88.04% | 31.65%/89.67% |
| 30M 0° | 40.29%/89.45% | 24.68%/91.35% |

hardly requires prior knowledge and suit for deployed in MEC node to support driverless cars and make it safer. Meanwhile, SVD+5G takes the least time due to the powerful 5G technology.

### 4.6 Discussion and limitation

In general, the proposed method can work better when combine with 5G and mobile edge computing to improve the security of the driverless car. It is necessary to pay attention to external factors, such as weather and communication interference in the practical application. One scenario is, in extreme cases, when the edge computing node is unable to mitigate the adversarial samples by the proposed method, the unmanned vehicle should be informed to switch emergency mode to find the nearest safe location to park. Another scenario is when 5G signal cannot be used due to signal interference in a specific area, the unmanned vehicle should switch to 4G or 3G in a timely manner and restore the communication connection as soon as possible. During this period, the unmanned vehicle should also slow down or park in a nearest safe area. For a variety of reasons, our simulation experiment and data analysis are limited, but we will continue to do further research. All in all, the goal of the method is to minimize the impact of adversarial samples on driverless vehicles.

**Table 4** The comparison result of the proposed method and other methods

| | Accuracy | Recall | Runtime |
|---|---|---|---|
| JPEG + 5g | 75.63% | 88.91% | 8.21 ms |
| JPEG + 4g | 74.85% | 87.86% | 31.51 ms |
| JPEG offline | 67.88% | 84.13% | 15.47 ms |
| APE-GAN+ 5G | 60.64% | 75.96% | 35.14 ms |
| APE-GAN+ 4G | 60.15% | 76.82% | 297.51 ms |
| APE-GAN offline | 58.64% | 75.76% | 42.36 ms |
| CTQ + 5G | 71.52% | 87.56% | 18.51 ms |
| CTQ + 4G | 72.17% | 88.62% | 90.31 ms |
| CTQ offline | 70.19% | 85.73% | 29.83 ms |
| SVD + 5G | 87.58% | 96.17% | 7.12 ms |
| SVD + 4G | 87.85% | 95.45% | 33.19 ms |
| SVD offline | 85.69% | 94.37% | 15.75 ms |

## 5 Conclusion

In this paper, we propose a singular value decomposition-based fast mechanism to mitigate adversarial examples in the physical world, especially in the field of automatic drive. We simulate the unmanned system and conduct experiments to verify the effectiveness of our defense methods by using google *Inception v4* deep neural networks for training, against different attack methods and assume a variety of attack scenarios. The experimental results clearly indicate that the processed images by defense model can effectively eliminate perturbation for the adversarial attack on signpost for misleading driverless cars. retaining the larger singular values in the pixel matrix can help to better find the "essential information" hidden in the image which is able to improve the robustness of the defense model. Meanwhile, the combine method of SVD+5G will greatly increase the cost of the adversary. In addition, the proposed defense model has strong practicability and easily to be deployed in the MEC node to make unmanned equipment in the coverage area free from adversarial attacks. In future work, we hope to extend the experiment to different IoT scenarios and guarantee the secure development and stability of intelligent IoT in the coming 5G era.

### Abbreviations

IoT: Internet of things; MEC: Mobile edge computing; SVD: Singular value decomposition; GTSRB: German traffic sign recognition benchmark; V2X: Vehicle-to-everything; DNN: Deep neural networks; OPV: Overall probability value; glcm: Gray level co-occurrence matrix; UDDN: U-Net-based deep denoising neural network; I-FGSM: Iterative fast gradient sign method attack; C&W: Carlini and Wagner attacks; JSMA: Jacobian-based Saliency Map Attack; CTQ: Cropping, TVM and quilting; SSIM: Structural similarity index measure; PSNR: Peak signal to noise ratio

### Acknowledgements

Thanks are due to Wenxue Yang for assistance with the experiments and to Limin Xiao for valuable discussion and suggestions. Thanks to Jinbin Zhu for collecting the experimental data. In addition, the authors would like to thank the anonymous reviewers for their valuable and constructive criticisms that are very helpful to improve the quality of this paper.

### Authors' contributions

FW contributes to the main design ideas for this defense method. WY is responsible for coding and experimental testing. Jinbin Zhu is responsible for the statistics and mapping of experimental data. Limin Xiao is responsible for supervision and project administration. The authors read and approved the final manuscript.

### Funding

This research was supported by the fund of the National key R&D Program of China under Grant NO. 2017YFB1010000 and the fund of the State Key Laboratory of Software Development Environment under Grant NO. SKLSDE-2020ZX-15.

### Availability of data and materials

The datasets we use during the current study are available in the GTSRB repository, http://benchmark.ini.rub.de/?section=gtsrb&subsection=dataset#Downloads

The iterative fast gradient sign method adversarial examples generated by open-source code on the github, https://github.com/1Konny/FGSM

The Carlini and Wagner adversarial examples generated by open-source code on the github, https://github.com/toby-lyf/adv-attack.

The Deepfool method adversarial examples generated by open-source code on the github, https://github.com/toby-lyf/adv-attack.

The Jacobian-based Saliency Map Attack adversarial examples generated by open-source code on the github, https://github.com/RobertoFalconi/BlackBoxAttackDNN.

### Competing interests

The authors declare that they have no competing interests.

Received: 25 April 2020 Accepted: 11 August 2020
Published online: 09 September 2020

### References

1. J Lu et al. "NO need to worry about adversarial examples in object detection in autonomous vehicles." (2017).
2. A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The kitti vision benchmark suite. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pages 3354–3361. IEEE (2012).
3. T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971, (2015).

4. H. Bou-Ammar, H. Voos, and W. Ertel. Controller design for quadrotor uavs using reinforcement learning. In Control Applications (CCA), 2010 IEEE International Conference on, pages 2130–2135. IEEE, (2010).
5. C. Mostegel, M. Rumpler, F. Fraundorfer, and H. Bischof. Uav-based autonomous image acquisition with multi-view stereo quality assurance by confidence prediction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 1–10, (2016).
6. F. Zhang, J. Leitner, M. Milford, B. Upcroft, and P. Corke. Towards vision-based deep reinforcement learning for robotic motion control. arXiv preprint arXiv:1511.03791, 2015.
7. A. Athalye and I. Sutskever. Synthesizing robust adversarial examples. arXiv preprint arXiv:1707.07397, (2017).
8. M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter. Accessorize to a crime: real and stealthy attacks on state-ofthe-art face recognition. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pages 1528–1540. ACM, (2016).
9. Y. Zhou, X. Hu, L. Wang, S. Duan, Y. Chen, Markov chain based efficient defense against adversarial examples in computer vision. IEEE Access **7**, 5695–5706 (2018)
10. K. Zhang, Y. Liang, J. Zhang, Z. Wang, X. Li, No one can escape: A general approach to detect tampered and generated image. IEEE Access **7**, 129494–129503 (2019)
11. Chen, S., Shi, D., Sadiq, M., & Zhu, M. Image denoising via generative adversarial networks with detail loss. In Proceedings of the 2019 2nd International Conference on Information Science and Systems, pp. 261-265. ACM, Jeju Island (2019).
12. Y. Li, Y. Wang, Defense against adversarial attacks in deep learning. Appl. Sci. **9**(1), 76 (2019)
13. N Das, M Shanbhogue, ST Chen, F Hohman, S Li, L Chen, ... & DH Chau. Shield: Fast, practical defense and vaccination for deep learning using jpeg compression. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 196-204. ACM, London (2018).
14. C Guo, et al. "Countering adversarial images using input transformations." arXiv preprint arXiv:1711.00117 (2017).
15. Xie, C, et al. "Mitigating adversarial effects through randomization." arXiv preprint arXiv:1711.01991 (2017).
16. N Papernot, P McDaniel, I Goodfellow, S Jha, ZB Celik, and A Swami. Practical black-box attacks against machine learning. In Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, pp.506–519. ACM, Abu Dhabi (2017).
17. C Szegedy, W Zaremba, I Sutskever, J Bruna, D Erhan, I Goodfellow & R Fergus. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013).
18. Goodfellow, I. J., Shlens, J., & Szegedy, C. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014).
19. A Kurakin, I Goodfellow, and S Bengio. 2016. Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533 (2016).
20. N Carlini, & D Wagner. Towards evaluating the robustness of neural networks. In 2017 IEEE Symposium on Security and Privacy (SP), pp. 39-57. IEEE, San Jose (2017).
21. S. M Moosavi-Dezfooli, A Fawzi, & P Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2574-2582. (2016).
22. N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami. The limitations of deep learning in adversarial settings. In Security and Privacy (EuroS&P), 2016 IEEE European Symposium on, pages 372–387. IEEE, Las Vegas (2016).
23. J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel. Man vs. computer: benchmarking machine learning algorithms for traffic sign recognition. Neural Networks, (2012).
24. C Szegedy, S Ioffe, V Vanhoucke, & A Alemi. Inception-ResNet and the impact of residual connections on learning. arXiv preprint arXiv:1602.07261 (2016).
25. A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533, (2016).
26. S Shen et al. "Ape-gan: adversarial perturbation elimination with GAN." arXiv pre-print arXiv:1707.05474 (2017).

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Submit your manuscript to a SpringerOpen journal and benefit from:**

► Convenient online submission
► Rigorous peer review
► Open access: articles freely available online
► High visibility within the field
► Retaining the copyright to your article

**Submit your next manuscript at ► springeropen.com**