SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •
Special Focus on Cyber Security in the Era of Artificial Intelligence

July 2022, Vol. 65 170306:1–170306:12
https://doi.org/10.1007/s11432-021-3457-7

# Certified defense against patch attacks via mask-guided randomized smoothing

Kui ZHANG[1], Hang ZHOU[2], Huanyu BIAN[1], Weiming ZHANG[1*] & Nenghai YU[1]

[1]*School of Cyber Science and Technology, University of Science and Technology of China, Hefei* 230027*, China;*
[2]*School of Computer Science, Simon Fraser University, Burnaby* V5A 1S6*, Canada*

Received 31 August 2021/Revised 5 January 2022/Accepted 16 March 2022/Published online 23 June 2022

**Abstract** The adversarial patch is a practical and effective method that modifies a small region on an image, making DNNs fail to classify. Existing empirical defenses against adversarial patch attacks lack theoretical analysis and are vulnerable to adaptive attacks. To overcome such shortcomings, certified defenses that provide a guaranteed classification performance in the face of strong unknown adversarial attacks are proposed. However, on the one hand, existing certified defenses either have low clean accuracy or need specified architecture, which is not robust enough. On the other hand, they can only provide provable accuracy but ignore the relationship to the number of perturbations. In this paper, we propose a certified defense against patch attacks that provides both the provable radius and high classification accuracy. By adding Gaussian noises only on the patch region with a mask, we prove that a stronger certificate with high confidence can be achieved by randomized smoothing. Furthermore, we design a practical scheme based on joint voting to find the patch with a high probability and certify it effectively. Our defense achieves 86.4% clean accuracy and 71.8% certified accuracy on CIFAR-10 exceeding the maximum 60% certified accuracy of existing methods. The clean accuracy of 67.8% and the certified accuracy of 53.6% on ImageNet are better than the state-of-the-art method, whose certified accuracy is 26%.

**Keywords** certified defense, adversarial patch, patch localization, randomized smoothing, joint voting

**Citation** Zhang K, Zhou H, Bian H Y, et al. Certified defense against patch attacks via mask-guided randomized smoothing. Sci China Inf Sci, 2022, 65(7): 170306, https://doi.org/10.1007/s11432-021-3457-7

## 1 Introduction

Deep neural networks (DNNs) have been applied extensively in various professions because of their superior performance. However, researchers [1–6] have shown that DNNs are vulnerable to adversarial attacks that a small perturbation constrained by $\ell_0$, $\ell_2$, or $\ell_\infty$ norm makes DNNs fail to predict correctly. Among these attacks, patch-based methods that replace a small area of the image with the generated patch are more practical in the physical world. There emerged many adversarial patch attacks [7–9], and studies [10–14] put a malicious patch on the clothes through multiple transformations, making object detectors such as YOLOv3 [15] and faster R-CNN [16] fail to classify or detect objects. Besides, some state-of-the-art face recognition models cannot recognize correctly [17–19] when wearing an adversarial hat or adversarial glasses.

To mitigate the risks of the adversarial patch, many defenses against such attacks have been proposed according to heuristic observations. From the perspective of the image level, many researchers like [20, 21] used the gradient or saliency map to locate the harmful area and then reconstructed it to get an input that will not affect the output. Some other studies [22, 23] start from the robustness of the model itself by making adversarial patches with stronger attack performance to perform adversarial training to enhance the robustness of the model. However, the lack of theoretical proof for these defenses makes the DNNs vulnerable to adaptive attacks [24, 25]. To solve this problem, a series of provable defenses are proposed. Some studies [26] use linear relaxation to obtain provable robustness but cannot scale to large size datasets. More studies [27–30] designed specific network structures with the goal of reducing the

[*] Corresponding author (email: zhangwm@ustc.edu.cn)

impact of adversarial patches on clean features and gave difficult relaxation conditions for verification, but this also leads to low accuracy.

Apart from the annoyance of low accuracy, there is another issue not considered in the current approach. Existing defenses model patch attacks as a special case of $l_0$ adversarial attacks, which means that an adversary needs to only consider how many pixels to modify rather than the amount of perturbation. However, there are many real-world scenarios where adversarial pixels cannot be modified to their maximum value. Thus it is reasonable and necessary to give a certificate with a robust radius of modifications that makes the bound tighter for patches in different scenarios.

In this paper, we propose mask-guided randomized smoothing (MRS), a general certified defense against adversarial patch attacks, achieving high certified accuracy and clean accuracy. We leverage the randomized smoothing scheme to demonstrate that when a Gaussian mask is used to smooth the patch region, a stronger certificate related to the amount of perturbation can be realized.

With the proposed MRS, the adversarial patch with a relatively larger noise level can still be classified correctly. The contributions of the paper are summarized as three-fold. (i) We give theoretical proof of the mask-guided certification against patch attacks and show that local randomized smoothing can obtain the guarantee of a larger robust radius, which can defend against unrestricted patch attacks. (ii) We propose a robust adversarial patch localization algorithm that can effectively localize adversarial patches and improve the certified radius in the face of real adversarial samples. (iii) We evaluate our method on both CIFAR-10 and ImageNet, achieving a larger top-1 certified accuracy compared with state-of-the-art methods.

## 2 Related work

### 2.1 Adversarial patch attack

Due to the complexity of the physical world, methods that add small perturbations to the whole image are no longer applicable for real-world adversarial attacks on the recognition systems. The patch generated by replacing a small part of the image with a predesigned mask is the most frequently used attack in the real world, e.g., a small sticker on a road sign or a T-shirt can make a classifier go wrong. The process of generating an adversarial patch can be mathematically formulated as follows. We define the set of all possible regions $l_p$ of the patch as $L$ and define an operator $A(\boldsymbol{x}, \boldsymbol{x}_p, l_p)$ as an adversarial patch $\boldsymbol{x}_p$ placed at a random region $l_p$ on the image $\boldsymbol{x}$. The initial adversarial input $\boldsymbol{x}^0_{\mathrm{adv}}$ is given by replacing an arbitrary region of the input $\boldsymbol{x}$ with a binary mask $\boldsymbol{M}_{l_p}$, i.e.,

$$\boldsymbol{x}^0_{\mathrm{adv}} = \boldsymbol{M}_{l_p} \odot A + (\boldsymbol{J} - \boldsymbol{M}_{l_p}) \odot \boldsymbol{x}, \tag{1}$$

where $\odot$ denotes Hadamard product and $\boldsymbol{J}$ is an all-ones matrix. The final adversarial patch is generated through multiple iterations by gradient-based or optimization-based algorithms, i.e.,

$$\boldsymbol{x}^n_p = \underset{\boldsymbol{x}_p}{\arg\max}\, \mathcal{L}\left[f(A(\boldsymbol{x}, \boldsymbol{x}_p, l_p)), y\right], \quad l_p \in L, \tag{2}$$

where $y$ is the ground truth label of $\boldsymbol{x}$ and $\boldsymbol{x}^n_p$ is the final adversarial patch after $n$ iterations.

In this paper, we use localized patches discussed above as the threat model. Adversarial patch attack is mainly divided into two types according to the visual effect, one is a drastic change easily recognizable to the human eye, and the other is a small amount of modification difficult to detect.

**Cluttered noise.** Early adversarial patches are usually directly generated by gradient-based methods such as fast gradient sign method (FGSM) [2], project gradient descent (PGD) [6] without more optimizations. Patches generated by LaVAN [8] and the method proposed by Brown et al. [7] do not obscure the foreground and are robust to general affine transformations. To enable the patch to evade detection, Subramanya et al. [31] added an additional constraint to suppress the class activation values of patch locations. Gittings et al. [32] used deep image prior [33] to reconstruct adversarial examples that resemble the appearance of natural images. Yang et al. [9] optimized the positions and textures of a group of class-specific textures by reinforcement learning, successfully implementing a black-box patch attack. Although these methods can successfully execute attacks, the adversarial perturbations are still obvious compared to the context.

**Context-aware noise.** Context-aware noise refers to the perturbation that is inconspicuous for the human. Since sharp noise can be easily recognized by human eyes, increasing researchers focus on generating adversarial patches that are invisible to both human eyes and neural detectors. Fendley et al. [34] designed a semi-transparent patch added to the original image to balance obtrusiveness and attack success rate. Brunner et al. [35] used the traditional copy-move method to initialize the patch for an efficient black-box targeted attack. PS-GAN [36] generated visually natural adversarial patches correlated with the image context based on the generative adversarial network (GAN) framework. Luo et al. [37] proposed a GAN-based framework with multiple scales of generators and discriminators to generate adversarial patches consistent with contexts.

### 2.2 Adversarial patch defense

Defense against patch attack can be roughly divided into two types: empirical defense and certified defense. Empirical defense shows powerful defensive ability against adversarial examples, while the certified defense can not only defend the adversarial attacks effectively but also provide the theoretical lower bound of adversarial perturbation and the provable accuracy.

**Empirical defense.** The first defense against patch attacks is proposed by Hayes [20] who used the saliency map of the image to localize the patch. Based on the observation that the gradient of classification loss for the input is generally large and dense around the location of perturbed pixels and then utilized inpainting to remove the patch. Naseer et al. [21] computed the normalized first-order local image gradients and mapped them into the original image to suppress the adversarial patch. Wu et al. [22] tried to find the strongest patch attack and then mixed them into the training data for adversarial training to improve the robustness of the model against the adversarial patches. Though these defenses above show strong robustness against patch attack, they are vulnerable to adaptive attacks without an analytical guarantee.

**Certified defense.** Chiang et al. [26] designed the first provable defense against patch attacks via interval bound propagation (IBP) [38]. The range of influence on the output of the last layer is acquired by estimating the interval at which perturbed pixels affect the output of each layer to derive the provable robustness. Inspired by randomized smoothing [39], Levine and Feizi [40] tried to obtain certified robustness against sparse adversarial attacks by randomly ablating input features. Further, they proposed (de)randomized smoothing (DRS) [27] which selected a small block traversing the image to classify and then chose the majority vote of the outputs as the prediction. The prediction is considered provable only if the number of top-1 classes is larger by $2k$ than the number of top-2 classes where $k$ is the number of blocks that are affected by the patch. However, both the clean accuracy and the certified accuracy on ImageNet are fairly low due to the use of the ablated version of the image. BAGCERT [30] trained a model with a small receptive field that still utilized the majority vote of predictions of patch blocks to classify and used similar validation conditions as DRS, obtaining higher accuracy. BAGCERT can get the prediction of all patch blocks by a single forward propagation, so the inference speed is greatly improved compared with DRS. Zhang et al. [28] used Clipped BagNet (CBN) that the logits of BagNet were clipped to restrain the influence of the patch. MR [41] defense used a relatively coarse occlusion prediction method and was difficult to scale to large resolution images because the larger the occlusion and step size, the greater the accuracy loss and the longer the time. Xiang et al. [29] also used the BagNet model with a small receptive field to control the number of malicious features, and then masked the detected malicious features to derive provable robustness. Despite provable robustness, most of the methods perform poorly on both clean and certified accuracy that does not work well on ImageNet.

### 2.3 Basics of randomized smoothing

Randomized smoothing based techniques have been proved effective in many studies, and now we describe how randomized smoothing provides provable robustness for image classification. Given a base classifier $f$ mapping an input $\boldsymbol{x} \in \mathbb{R}^d$ to classes $\mathcal{Y}$, a smoothed classifier $g$ can be constructed from the base classifier $f$ by adding an isotropic Gaussian noise to $\boldsymbol{x}$. Specifically, for an input $\boldsymbol{x}$, the smoothed classifier $g$ returns the class assigned the largest probability when $\boldsymbol{x}$ is perturbed with Gaussian noise $\mathcal{N}\left(0, \sigma^2 \boldsymbol{I}\right)$ and then passed through $f$, i.e.,

$$g(\boldsymbol{x}) = \underset{c \in \mathcal{Y}}{\arg\max}\ \mathbb{P}(f(\boldsymbol{x}+\epsilon)=c),\ \text{where } \epsilon \sim \mathcal{N}\left(0, \sigma^2 \boldsymbol{I}\right). \tag{3}$$

The noise level $\sigma$ controls the trade-off between robustness and accuracy. When the lower bound on the probability of the predicted class and the upper bound on the probability of the remaining class are known, the classifier $g$ can be shown to be robust within the $\ell_2$ ball of the input by Neyman-Pearson lemma [42]. The upper and lower bounds of the probabilities can be calculated with high confidence using the Monte Carlo algorithm.

## 3 Method

In this section, we describe the proposed certified defense against adversarial patch attack based on randomized smoothing [39]. We first calculate and prove the theoretical lower bound of the certificate when the patch location is known, and then give a practical method approaching the lower bound with high probability when the input is arbitrary images.

### 3.1 Robustness guarantee against patch attack

One advantage of the randomized smoothing technique is that it is model agnostic. Thus it can be extended to many different large models, but it yields a small provable radius. To extend its radius to meet the needs of defending against patch attacks, we use a mask to guide the smoothing. When $\underline{p_A}$ and $\overline{p_B}$ are obtained, we prove the theoretical expression of the robust radius against patch attack under the guidance of mask $\boldsymbol{M}$, where $\underline{p_A}$ is the lower bound on the probability of the class $c_A$ with the highest output probability and $\overline{p_B}$ is the upper bound on the probability of the class $c_B$ with the second-highest probability.

**Theorem 1.** Given a base classifier $f$ and its smoothed version $g$: $\mathbb{R}^d \to \mathcal{Y}$ defined in (3), an image $\boldsymbol{x}$, a masked noise distribution $\epsilon \sim \mathcal{N}\left(0, \sigma^2 \boldsymbol{M}\right)$ with a binary mask $\boldsymbol{M} \in \mathbb{R}^d$ where $\boldsymbol{M}$ is all zero except that the selected part is one, $c_A, c_B \in \mathcal{Y}$ and $\underline{p_A}, \overline{p_B} \in [0,1]$ that satisfy

$$\mathbb{P}(f(\boldsymbol{x}+\varepsilon)=c_A) \geqslant \underline{p_A} \geqslant \overline{p_B} \geqslant \max_{c \neq c_A} \mathbb{P}(f(\boldsymbol{x}+\varepsilon)=c). \tag{4}$$

Then we have

$$g(\boldsymbol{x}+\delta)=c_A, \qquad \forall \|\delta\|_2 < R, \tag{5}$$

where

$$R=\frac{\sigma}{2}(\Phi^{-1}(\underline{p_A})-\Phi^{-1}(\overline{p_B})). \tag{6}$$

*Proof.* See Appendix A.

We show that the theoretical upper bound of the robust radius can be obtained by estimating $\underline{p_A}$ and $\overline{p_B}$, and will further elaborate on the changes brought by MRS in the following corollary. When adding local noises following Gaussian distribution on the patch, the adversarial point moves back to the decision region of correct image distribution with a high probability. Theorem 1 proves that the adversarial patch cannot change the output of the smooth classifier when the modification of the patch is smaller than the provable radius $R$. By sampling multiple times, we can use majority vote to obtain reliable predictions and use hypothesis testing to estimate $\underline{p_A}$ and $\overline{p_B}$. The detailed MRS certification algorithm is given in Algorithm 1. Function LowerConfBound($\texttt{counts}[\hat{c}], n, 1-\alpha$) obtains lower confidence interval $\underline{p_A}$ with probability at least $1-\alpha$ over the randomness for $\texttt{counts}[\hat{c}] \sim \text{Binomial}(n, p_A)$.

Below we exemplify the change in radius and accuracy after adding local Gaussian noise to the patch region and demonstrate its effectiveness.

**Corollary 1.** Considering a linear classifier with two classes $f(\boldsymbol{x})=\operatorname{sign}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b)$ and the smoothed classifier $g$, also superimposing a masked noise on the input, we can derive $p_A=\Phi(\frac{|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b|}{\sigma\|\boldsymbol{M}\odot\boldsymbol{w}\|_2})$ and the provable radius $R=\frac{|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b|}{\|\boldsymbol{M}\odot\boldsymbol{w}\|_2}$.

*Proof.* Take $p_A=\mathbb{P}(g(\boldsymbol{x})=1)$ as an example,

$$\begin{aligned}
& p_A=\mathbb{P}(g(\boldsymbol{x})=1)=\mathbb{P}(f(\boldsymbol{x}+\epsilon)=1) \\
\Leftrightarrow\ & p_A=\mathbb{P}(\operatorname{sign}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+\boldsymbol{w}^{\mathrm{T}}\epsilon+b)=1) \\
\Leftrightarrow\ & p_A=\mathbb{P}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+\boldsymbol{w}^{\mathrm{T}}\epsilon+b \geqslant 0)
\end{aligned}$$

**Algorithm 1** Mask-guided randomized smoothing certification

```
Require: Base classifier $f$, standard deviation of Gaussian noise $\sigma$, image $\boldsymbol{x}$, location mask $\boldsymbol{M}$, sample times $n$ and $n_0$, $\alpha$.
Ensure: ABSTAIN or predicted label $\hat{c}$, certified radius $r$.
for $i = 1$ to $n_0$ do
    Sample noise $\varepsilon_i \sim \mathcal{N}\left(0, \sigma^2 \boldsymbol{M}\right)$;
    $\texttt{output}_i \leftarrow f(\boldsymbol{x} + \varepsilon_i)$;
    Append $\texttt{output}_i$ in $\texttt{counts}_{n_0}$;
end for
$\hat{c} \leftarrow$ top indice in $\texttt{counts}_{n_0}$;
for $j = 1$ to $n$ do
    Sample noise $\varepsilon_j \sim \mathcal{N}\left(0, \sigma^2 \boldsymbol{M}\right)$;
    $\texttt{output}_j \leftarrow f(\boldsymbol{x} + \varepsilon_j)$;
    Append $\texttt{output}_j$ in $\texttt{counts}_n$;
end for
$\underline{p_A} \leftarrow$ LowerConfBound($\texttt{counts}_n[\hat{c}], n, 1 - \alpha$);
if $\underline{p_A} > \frac{1}{2}$ then
    return $\hat{c}$, radius $\sigma\Phi^{-1}(\underline{p_A})$;
else
    return ABSTAIN;
end if
```

$$
\begin{aligned}
&= \mathbb{P}\left(\sigma \left\|\boldsymbol{M} \odot \boldsymbol{w}\right\|_2 \mathbb{Q} \geqslant -\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} - b\right) \\
&= \mathbb{P}\left(\mathbb{Q} \leqslant \frac{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b}{\sigma \left\|\boldsymbol{M} \odot \boldsymbol{w}\right\|_2}\right) \\
&= \Phi\left(\frac{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b}{\sigma \left\|\boldsymbol{M} \odot \boldsymbol{w}\right\|_2}\right), \quad \mathbb{Q} \sim \mathcal{N}\left(0, 1\right).
\end{aligned} \tag{7}
$$

Similarly, we can prove that

$$
p_A = \Phi\left(\frac{-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} - b}{\sigma \left\|\boldsymbol{M} \odot \boldsymbol{w}\right\|_2}\right), \tag{8}
$$

when $p_A = \mathbb{P}(g(\boldsymbol{x}) = -1)$, so we have

$$
p_A = \Phi\left(\frac{\left|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b\right|}{\sigma \left\|\boldsymbol{M} \odot \boldsymbol{w}\right\|_2}\right). \tag{9}
$$

In a two-class classifier, $p_A = 1 - p_B$. According to Theorem 1, we have

$$
R = \sigma\Phi^{-1}(p_A) = \frac{\left|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b\right|}{\left\|\boldsymbol{M} \odot \boldsymbol{w}\right\|_2}. \tag{10}
$$

Based on Corollary 1, the mask $\boldsymbol{M}$ that limits the size of the perturbation region controls the trade-off between the accuracy and the size of the provable radius, and this also holds for deeper and larger networks by similar extrapolation. For a larger multi-class nonlinear classifier, the impact of $\sigma$ on the $\underline{p_A}$ is small when the M keeps small. This is consistent with our intuitive observation that when all the elements of the $\boldsymbol{M}$ are one, the certified radius degenerates to [39], with a corresponding decrease in the lower bound on the probability of the predicted class.

### 3.2 Robust localization based on joint voting

To evaluate the effectiveness of MRS, we propose a robust localization algorithm to simulate the smoothing of adversarial patches at unknown random locations. To obtain a larger radius, the location of the patch should be positioned as precisely as possible. Based on Section 2, sharp noise with tight topology can be easily detected by higher-order filters, while context-aware noise is hard to localize. To cope with all possible situations, we propose a robust localization algorithm by joint voting to ensure that patches can be safely located.

The general framework of the methodology is shown in Figure 1. The input image $\boldsymbol{x}$ is occluded by an $s \times s$ square block from left to right and top to bottom with the stride of one pixel. For each position of the block, we leverage an inpainting [43] algorithm to reconstruct it and pass the restored image through $f$ to get a prediction $l \in \{c_0, c_1, \ldots, c_n\}$. All predictions make up the prediction map. Then we count

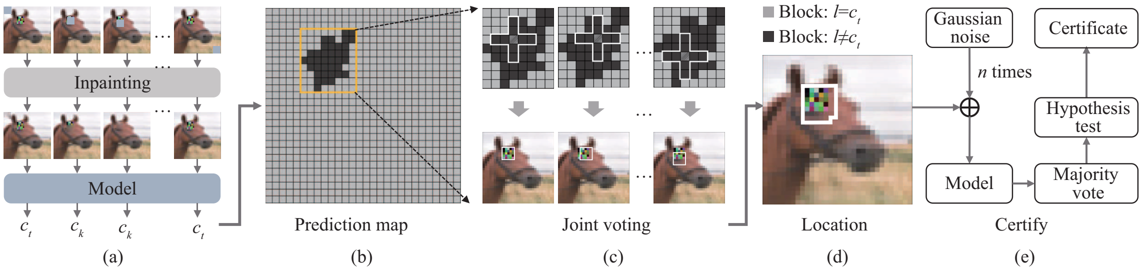

**Figure 1** Overview of MRS. (a) We first use a sliding kernel to repair the image to get a prediction map composed of the outputs; (b) light gray represents target attack class, dark gray represents non-attack class; (c) the joint voting in the four directions of the candidate area is represented by a white cross-shaped template; (d) the region with the highest score in the joint voting determines the final location; (e) the certificate is calculated by estimating the exact lower bound of the probability through hypothesis testing.

the number of each class,

$$n_k = \sum_{k \in \{c_0, c_1, \ldots, c_n\}} \text{COUNT}[l = c_k]. \tag{11}$$

The class $c_t$ with the maximum value $n_t$ is identified as the attacked class. On the one hand, an image that can be predicted correctly has a very low probability of being predicted incorrectly after masking and recovery, and the error rate drops lower after the training data is augmented with local transformations. On the other hand, if an adversarial patch attacks the model successfully, it still succeeds with a high probability when the inpainted region does not contain the original patch. This ensures that $n_t$ is the attacked class.

Commonly, the prediction of a restored adversarial image is different from that of the clean image except for the case that the patch is partially or fully inpainted. According to the positional relation between the candidate block and the adversarial patch, we divide them into the following three cases. The first case is that the region with the label incorrectly labeled does not contain adversarial pixels. These regions are relatively few and sparsely distributed across the image. The second case is that the region with the label incorrectly labeled contains partial adversarial pixels. This is the most common case that there is no guarantee that an attacker still has a successful attack with a random portion of the patch. The third case is that the inpainted region covers the actual patch region. Thus the predicted label remains the same as that of the inpainted clean image.

Based on the above analysis, we propose a robust algorithm for locating the patch. Denote $r_i$ as the $i$-th square region of the image and denote $b_i$ as the input that the $r_i$ is inpainted. After obtaining predictions of all $b_i$, we use a candidate block $r_i$ centered cross-shaped filter $\mathbb{T}$ to count the number of blocks incorrectly labeled in the template. More formally, denoting $\mathbb{C} = \{r_i \mid g(b_i) \neq c_t\}$, the score of each candidate region $r_i$ is

$$\text{score}_i = |\mathbb{T} \cap \mathbb{C}|\,. \tag{12}$$

We mark the $r_i$ with the largest score as the adversarial patch region. When there are two or more candidate regions in the image with scores greater than a threshold $\tau$ and no connectivity, inpainting cannot help the image to be correctly predicted, and we will directly discard this image without participating in the certification. Template T accepts the votes from its four directions into a final vote to determine the malicious region. If the template T shrinks its four directions to contain only the square region of the central block, the localization phase can be easily defeated. To ensure that the localization area can cover the patch, we mark the candidate areas that are within 3 pixels from the middle block as patch areas as well. The white cross-shaped box in Figure 1 represents the template, which means the number of blocks in $\mathbb{C}$ that overlap $m$ rows or $m$ columns horizontally and vertically with the $s \times s$ candidate in the center. Through exhaustive search and filtering, the algorithm can effectively locate the adversarial patch while excluding the interference caused by candidate blocks in clean areas of the image.

### 3.3 Security analysis of localization

When considering white-box adaptive attackers, the voting and discarding mechanisms guarantee that the localization algorithm will not be defeated. The searching and voting mechanism will put the adversary in a dilemma. Specifically, to circumvent the localization, the adversary needs to generate square patches that satisfy the following condition: every row and every column of the patch and their combination must attack successfully at the same time. It is inherently very difficult to generate a single row or column of adversarial pixels with attack performance. In the one-pixel attack [44], though a single adversarial pixel could be found using the differential evolution algorithm, it is very hard to find a satisfactory solution if the pixels are restricted to a very small area. This is a contradictory optimization problem that would put the attacker in a dilemma of making only a part of the patch or the whole exist offensively. The only failure case is that there are a large number of adjacent areas with the same incorrect label after being inpainted, because the samples are distributed on the decision boundary, which is inherently not robust.

There is an implicit assumption in our localization algorithm that the prediction remains essentially the same after a tiny region of the image being inpainted. When there are two or more candidate regions in the image with scores greater than a certain threshold and no connectivity, the localization algorithm can choose to return ABSTAIN. This is an optional operation. The discarding operation has little impact on the certification because a well-trained classifier is robust to such deterministic and benign inpainting changes and only a sparse number of positions in very few images are misclassified after being inpainted, as proved in recent work [45]. In the experiments, the number of samples that make the localization algorithm return ABSTAIN is small.

## 4 Experiments

### 4.1 Setup

**Datasets.** We conduct experiments on CIFAR-10 [46] and ImageNet [47]. The images in ImageNet are resized to $224 \times 224$, and the images in CIFAR-10 are with the size of $32 \times 32$. Following [39], for both datasets, we use the full training set and randomly sample 500 images from the test set for testing. The accuracy of the test set of CIFAR-10 and ImageNet is 97.6% and 75.2%, respectively.

**Metrics.** We define certified $R$-accuracy as the accuracy of correct predictions when the certified radius $r$ of a test image is larger than a threshold $R$, and the mask of localization completely covers the adversarial patch. Different from the existing defense evaluation criteria, we can better measure the defense performance under patches with various interference levels to approximate a more realistic situation by modulating $R$, the noise radius, as another measurement.

**Implement details.** For CIFAR-10, we train the model with the augmentation of $\sigma = 7$. For ImageNet, we use a pretrained model with $\sigma = 8$ to warm up for training with $\sigma = 257$, which is effective in improving clean accuracy and convergence speed. The training process takes about 1 hour using CIFAR-10 on a single NVIDIA RTX 2080 Ti GPU and takes about 3 days using ImageNet on 4 NVIDIA RTX 2080 Ti GPUs.

To evaluate the proposed defense in a realistic adversarial environment, we use PGD [6] with $\ell_\infty = 255/255$ to generate the adversarial patch in 80 random locations for our base classifier and select the one that causes the greatest classification loss.

### 4.2 Main results

**Comparison with state-of-the-art methods.** Figures 2(a) and (d) show the certified $R$-accuracy curve on CIFAR-10 and ImageNet. The certified 25-accuracy on CIFAR-10 exceeds 70% while empirical accuracy, which only considers the prediction is correct, can reach 80%. Despite the certified radius being the theoretical upper bound, in practice, an adversarial patch can be successfully defended even if the $\ell_2$ distance is greater than this upper bound. The results on ImageNet are generally consistent with those on CIFAR-10. The only difference is that due to the higher resolution of ImageNet, which provides more semantic information, the localization is more accurate, and the decrease rate in provable accuracy is smaller compared to that of CIFAR-10.

We compare clean accuracy and certified accuracy of MRS with several certified patch defenses [27–30]. Clean accuracy refers to the accuracy of the defense on clean images. CBN [28], MASK-DS [29], MASK-BN [29], BAGCERT [30], and DRS [27] are tested on clean images while our method is tested on the generated adversarial samples with the localization to obtain an approximate certified accuracy due to the high computational cost. Table 1 shows that both certified accuracy and clean accuracy on CIFAR-10 and ImageNet greatly exceed the best current results, especially on ImageNet that the clean accuracy is

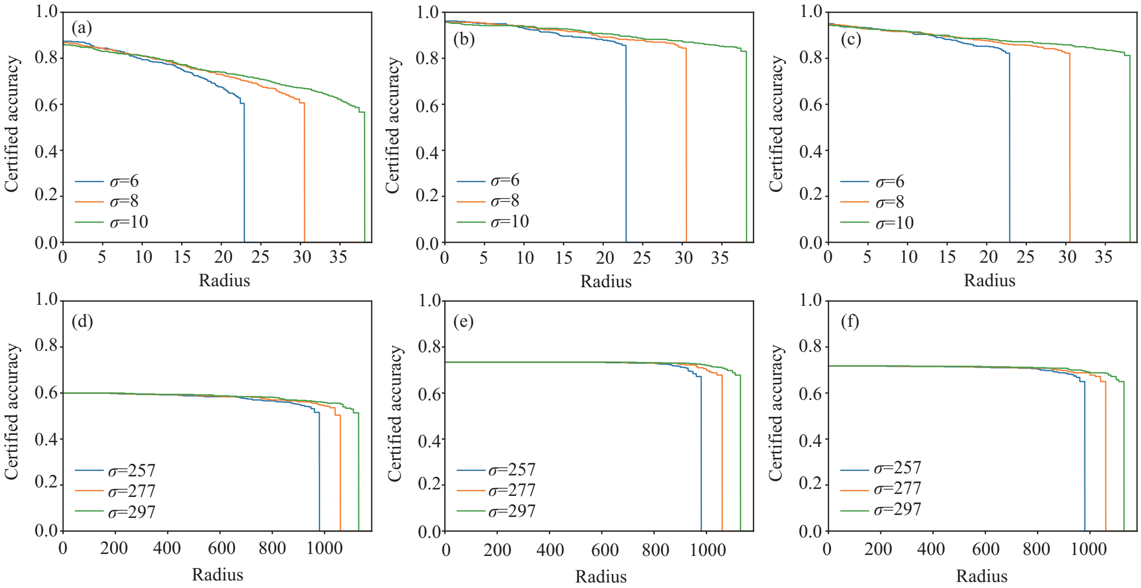

**Figure 2** $R$-accuracy curves of different noise level $\sigma$. (a), (b), and (c) are results on CIFAR-10, (d), (e), and (f) are results on ImageNet. (a) and (d) are the practical defense results on the real adversarial example set and (b), (c), (e), and (f) are theoretical results of patches at two different locations on the image.

**Table 1** Clean accuracy versus certified accuracy of different methods on CIFAR-10 and ImageNet

| | CIFAR-10 | | ImageNet | |
|---|---|---|---|---|
| | Clean accuracy | Certified accuracy | Clean accuracy | Certified accuracy |
| CBN [28] | 4.2 | 9.3 | 49.5 | 7.1 |
| DRS [27] | 83.9 | 56.3 | 43.1 | 14.5 |
| MASK-BN [29] | 83.9 | 47.3 | 54.6 | 26.0 |
| MASK-DS [29] | 84.6 | 57.7 | 43.6 | 16.0 |
| BAGCERT [30] | 86.0 | 60.0 | 45.3 | 22.9 |
| Ours | **86.4** | **71.8** | **67.8** | **53.6** |

67.8% and the certified 1024-accuracy is nearly 27% better than the state-of-the-art results. Experiments comparing with DRS at more sizes are provided in Appendix C.1.

**Theoretically optimal results.** To get a near theoretically optimal certified accuracy, the mask should be set to the same size and same location as the adversarial patch. Since the impact on classification accuracy may vary when patches are placed in different locations, we use two different settings to evaluate. In the first setting, we set the position of the smoothing mask in the center of the image, which ensures the coverage of the important region of the foreground object. In the second setting, we use Grad-cam [48] to predict the most significant region for network classification and then select the region with the highest activation value for the $s \times s$ size as the mask location. In both of settings, we used $10^5$ samples to verify so that we can obtain a certified radius of up to $4\sigma$ with 99.9% confidence.

Figures 2(b) and (c) show the $R$-accuracy curves under three different $\sigma$ values with different settings on CIFAR-10. The horizontal axis represents the $\ell_2$ norm of perturbations. Notice that for a $5 \times 5$ patch, when the radius reaches 25, the adversary is allowed to perform unrestricted attacks where the pixel value can be modified from 0 to 255. We show that the provable accuracy is still around 88% when the robust radius reaches 25, outperforming the state-of-the-art theoretical results. The curves in these two figures do not change significantly, suggesting that the smoothing classifier is not sensitive to smoothing regions. Regardless of where the patch is in the image, as long as it is smoothed, the classifier can be unaffected through the certification process.

Figures 2(e) and (f) are results on ImageNet. For a $32 \times 32$ patch, the proposed classifier can defend against unrestricted attacks when the radius reaches 1024. We show that the certified 1024-accuracy is around 61.8%, achieving the best theoretical results to date. The reason that all the curves on the figure are flat before reaching the upper bound of the radius is that the smoothed image will obtain a very high correct confidence through the smoothed classifier $g$ so that the certified radius is close to the upper

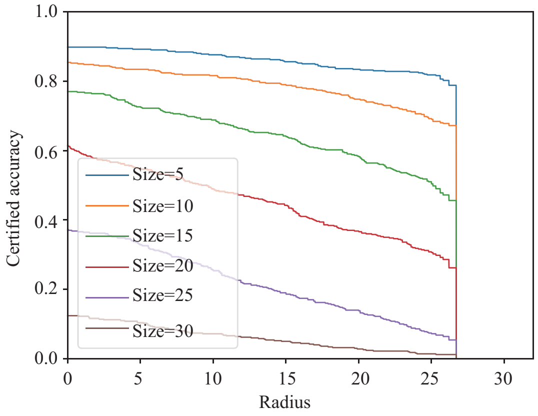

**Figure 3** Certified $R$-accuracy using masks with the size from $5 \times 5$ to $30 \times 30$ on CIFAR-10.

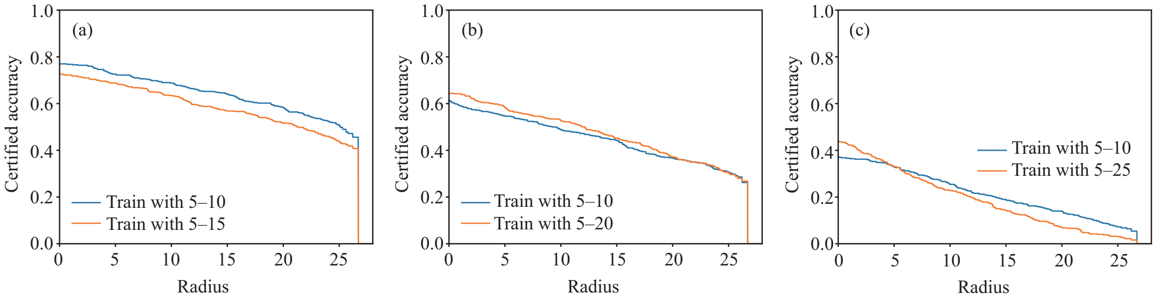

**Figure 4** The certification results for training using different size masks. The size of the mask used for certification: (a) $15 \times 15$, (b) $20 \times 20$, and (c) $25 \times 25$.

bound for most of the input images.

In conclusion, results on a large-scale dataset such as ImageNet are consistent with that of CIFAR-10. As shown in Figure 2, the noise level does not seem to affect the $R$-accuracy much, which will be detailed in the following.

## 4.3 Ablation study

**Certified $R$-accuracy of mask with different sizes.** Figure 3 shows the $R$-accuracy curves of different mask sizes on CIFAR-10. When assuming the patch size is $30 \times 30$, i.e., 87.9% of the image area, the 0-accuracy has dropped to nearly 10% which amounts to a random guess for CIFAR-10 with a total of 10 classes. This is to be expected, as the perturbations with a large $\sigma$ almost swamp the entire image, the classification is equivalent to a prediction of random noise. When the area of the patch reaches 9.7% of the image, 25-accuracy is still better than state-of-the-art methods against a 2.4% pixel patch. To ensure that the adversary can make arbitrary changes in the patch region without changing the classification results, values of Gaussian noise should be much larger than clean pixel values to the smoothing region. When the smoothed region is small, the neural network learns this pattern well, but when the smoothed region is large, the clean pixel values are too small compared to the noise values, and the clean pixels become "noise" for the random noise, which the neural network cannot fit at all.

**Influence of training methods.** From a generalization perspective, using an augmentation strategy that matches the testing time to train the model allows the test accuracy to be as close as possible to the training accuracy. However, we need to add noise of different sizes during certification, and the larger the mask size, the harder it is to converge the training, and the corresponding clean accuracy will decrease. To adapt the model to different size masks, we used random square noise with the width from 5 to 10, 5 to 15, 5 to 20, 5 to 25 respectively as the augmentation during the training. Figure 4 shows the $R$-accuracy results for several different training methods. As the area of the mask used for certification increases, although the model with matching strategy will gain some advantages at small radii, the 25-accuracy still decreases due to low clean accuracy. In general, it is the best choice for the certification to use a small mask as the augmentation.

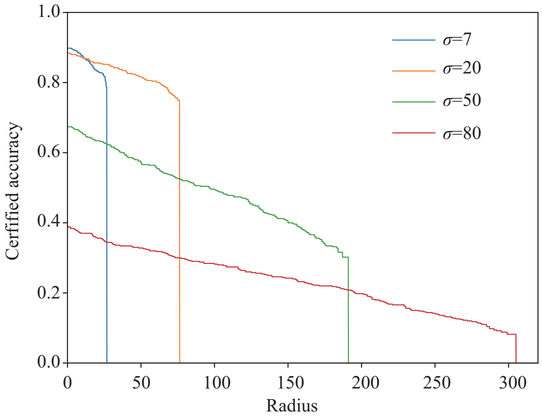

**Figure 5** Effect of $\sigma$ value in different scales ($\sigma = 7, 20, 50, 80$).

**Impact of $\sigma$.** We know from the above discussion that when $\sigma$ is certain, the smaller the smoothing area, the closer the provable $R$-accuracy will be for clean accuracy. However, when the smoothing area is fixed, the smaller the $\sigma$ is not the better. To investigate this issue, we perform experiments on the CIFAR-10 with a fixed mask size of $5 \times 5$ and compare the certifiable $R$-accuracy at different $\sigma$ values. The generalization performance of the network far exceeds expectations, and a network trained using $\sigma = 7$ as the augmentation can perform without degradation at more than three times the noise level greater than itself. Figure 5 shows that even though $\sigma = 80$ that 11 times larger than $\sigma = 7$, 25-accuracy still does not drop to 10%. The certification phase does not need to strictly match the augmentation strategy in training, which allows us to choose an augmentation strategy with fast convergence and high clean accuracy within a wide range. Although this powerful generalization performance is greatly reduced as the area of smoothed region is increased, we can still increase the $\sigma$ value to accommodate more modifications.

### 4.4 Discussion

**Certified radius adjustment in different cases.** The adversarial patch is usually modeled as an $l_0$ attack, which means one pixel can be modified up to 255. But in practice, pixel values of the patch region are not all zeros in general. Moreover, recent research [37] has been devoted to making the patch blend in with the surrounding environment which further reduces the noise value. In these cases, the range of the total amount of noise can be estimated from the pixels around the location where the patch is positioned. Thus the proposed method will further improve the certified accuracy compared to other methods by adaptively adjusting for the "worst case".

**Limitation.** The fact that our proposed defense method using strong external constraints including localization and smoothing contributes to the significant performance improvement over the state-of-the-art methods. Yet this operation brings about further computational cost than other methods, e.g., 14.6 s per CIFAR-10 image for inference. In fact, the proposed patch localization algorithm can be combined with any other provable adversarial defense method to boost their verification conditions to achieve even better performance, and our smoothing-based approach is only a practical example for the experiment.

## 5 Conclusion

In this work, we have introduced a certified defense against adversarial patch attacks that takes the $\ell_2$ norm of modifications in account. We provide the theoretical analysis and the proof of the certification based on randomized smoothing, and design a practical validation method on the real adversarial example set that the adversarial patches can be robustly located and efficiently verified. Numerous experiments demonstrate that our method exceeds the current methods. A more accurate and effective localization method will get higher accuracy and speed, which will be the future work.

**Acknowledgements** This work was supported in part by National Natural Science Foundation of China (Grant Nos. U20B2047, 62072421, 62002334, 62102386, 62121002), Exploration Fund Project of University of Science and Technology of China (Grant No. YD348000-
2001), and Fundamental Research Funds for the Central Universities (Grant No. WK2100000011).

**Supporting information** Appendixes A–C. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

### References

1 Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks. In: Proceedings of the 2nd International Conference on Learning Representations, 2014

2 Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: Proceedings of the 3rd International Conference on Learning Representations, 2015

3 Moosavi-Dezfooli S M, Fawzi A, Frossard P. Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016. 2574–2582

4 Carlini N, Wagner D. Towards evaluating the robustness of neural Networks. In: Proceedings of IEEE Symposium on Security and Privacy, 2017. 39–57

5 Chen P Y, Zhang H, Sharma Y, et al. Zoo: zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, 2017. 15–26

6 Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks. In: Proceedings of the 6th International Conference on Learning Representations, 2018

7 Brown T B, Mané D, Roy A, et al. Adversarial patch. 2017. ArXiv:1712.09665

8 Karmon D, Zoran D, Goldberg Y. LaVAN: localized and visible adversarial noise. In: Proceedings of the 35th International Conference on Machine Learning, 2018. 2507–2515

9 Yang C L, Kortylewski A, Xie C, et al. Patchattack: a black-box texture-based attack with reinforcement learning. In: Proceedings of the 16th European Conference on Computer Vision, 2020. 681–698

10 Li Y, Bian X, Lyu S. Attacking object detectors via imperceptible patches on background. 2018. ArXiv:1809.05966

11 Lee M, Kolter J Z. On physical adversarial patches for object detection. 2019. ArXiv:1906.11897

12 Wu Z, Lim S N, Davis L, et al. Making an invisibility cloak: real world adversarial attacks on object detectors. In: Proceedings of the 16th European Conference on Computer Vision, 2020. 1–17

13 Xu K, Zhang G, Liu S, et al. Adversarial T-shirt! Evading person detectors in a physical world. In: Proceedings of the 16th European Conference on Computer Vision, 2020. 665–681

14 Saha A, Subramanya A, Patil K, et al. Role of spatial context in adversarial robustness for object detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2020. 784–785

15 Redmon J, Farhadi A. YOLOv3: an incremental improvement. 2018. ArXiv:1804.02767

16 Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell, 2017, 39: 1137–1149

17 Pautov M, Melnikov G, Kaziakhmedov E, et al. On adversarial patches: real-world attack on ArcFace-100 face recognition system. In: Proceedings of International Multi-Conference on Engineering, Computer and Information Sciences, 2019. 391–396

18 Komkov S A, Petiushko A. Advhat: real-world adversarial attack on arcface face id system. In: Proceedings of the 25th International Conference on Pattern Recognition, 2021. 819–826

19 Yang X, Wei F, Zhang H, et al. Design and interpretation of universal adversarial patches in face detection. In: Proceedings of the 16th European Conference on Computer Vision, 2020. 174–191

20 Hayes J. On visible adversarial perturbations & digital watermarking. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018. 1597–1604

21 Naseer M, Khan S, Porikli F. Local gradients smoothing: defense against localized adversarial attacks. In: Proceedings of IEEE Winter Conference on Applications of Computer Vision, 2019. 1300–1307

22 Wu T, Tong L, Vorobeychik Y. Defending against physically realizable attacks on image classification. In: Proceedings of the 8th International Conference on Learning Representations, 2020

23 Rao S, Stutz D, Schiele B. Adversarial training against location-optimized adversarial patches. In: Proceedings of European Conference on Computer Vision Workshops, 2020. 429–448

24 Athalye A, Carlini N, Wagner D A. Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples. In: Proceedings of the 35th International Conference on Machine Learning, 2018. 274–283

25 Carlini N, Athalye A, Papernot N, et al. On evaluating adversarial robustness. 2019. ArXiv:1902.06705

26 Chiang P Y, Ni R, Abdelkader A, et al. Certified defenses for adversarial patches. In: Proceedings of the 8th International Conference on Learning Representations, 2020

27 Levine A, Feizi S. (De) randomized smoothing for certifiable defense against patch attacks. In: Proceedings of Advances in Neural Information Processing Systems, 2020

28 Zhang Z, Yuan B, McCoyd M, et al. Clipped bagNet: defending against sticker attacks with clipped bag-of-features. In: Proceedings of IEEE Security and Privacy Workshops, 2020. 55–61

29 Xiang C, Bhagoji A N, Sehwag V, et al. Patchguard: a provably robust defense against adversarial patches via small receptive fields and masking. In: Proceedings of the 30th USENIX Security Symposium, 2021

30 Metzen J H, Yatsura M. Efficient certified defenses against patch attacks on image classifiers. In: Proceedings of the 9th International Conference on Learning Representations, 2021

31 Subramanya A, Pillai V, Pirsiavash H. Fooling network interpretation in image classification. In: Proceedings of IEEE International Conference on Computer Vision, 2019. 2020–2029

32 Gittings T, Schneider S, Collomosse J. Robust synthesis of adversarial visual examples using a deep image prior. In: Proceedings of the 30th British Machine Vision Conference, 2019

33 Ulyanov D, Vedaldi A S, Lempitsky V. Deep image prior. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018. 9446–9454

34 Fendley N, Lennon M, Wang I, et al. Jacks of all trades, masters of none: addressing distributional shift and obtrusiveness via transparent patch attacks. In: Proceedings of European Conference on Computer Vision Workshops, 2020. 105–119

35 Brunner T, Diehl F, Knoll A. Copy and paste: a simple but effective initialization method for black-box adversarial attacks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019

36 Liu A, Liu X, Fan J, et al. Perceptual-sensitive GAN for generating adversarial patches. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence, 2019. 1028–1035

37 Luo J, Bai T, Zhao J, et al. Generating adversarial yet inconspicuous patches with a single image. In: Proceedings of the 35th AAAI Conference on Artificial Intelligence, 2021. 15837–15838

38 Gowal S, Stanforth R. Scalable verified training for provably robust image classification. In: Proceedings of IEEE International Conference on Computer Vision, 2019. 4841–4850

39 Cohen J, Rosenfeld E, Kolter Z. Certified adversarial robustness via randomized smoothing. In: Proceedings of the 36th International Conference on Machine Learning, 2019. 1310–1320

40 Levine A, Feizi S. Robustness certificates for sparse adversarial attacks by randomized ablation. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence, 2020. 4585–4593

41 McCoyd M, Park W, Chen S, et al. Minority reports defense: defending against adversarial patches. In: Proceedings of Applied Cryptography and Network Security Workshops, 2020. 564–582

42 Neyman J, Pearson E S. On the problem of the most efficient tests of statistical hypotheses. Phil Trans R Soc Lond A, 1933, 231: 289–337

43 Telea A. An image inpainting technique based on the fast marching method. J Graphics Tools, 2004, 9: 23–34

44 Su J, Vargas D V, Sakurai K. One pixel attack for fooling deep neural networks. IEEE Trans Evol Comput, 2019, 23: 828–841

45 Black S, Keshavarz S, Souvenir R. Evaluation of image inpainting for classification and retrieval. In: Proceedings of IEEE Winter Conference on Applications of Computer Vision, 2020. 1060–1069

46 Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images. 2009

47 Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database. In: Proceedings of IEEE conference on Computer Vision and Pattern Recognition, 2009. 248–255

48 Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. Int J Comput Vis, 2020, 128: 336–359