## Appendix A Proof of Theorem 1

**Theorem 1.** Given a base classifier $f$ and its smoothed version $g$: $\mathbb{R}^d \to \mathcal{Y}$, an example $\mathbf{x}$, a masked noise distribution $\epsilon \sim \mathcal{N}\left(0, \sigma^2\mathbf{M}\right)$ with a binary mask $\mathbf{M} \in \mathbb{R}^d$ where $\mathbf{M}$ 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:

$$\mathbb{P}(f(\mathbf{x}+\varepsilon) = c_A) \geqslant \underline{p_A} \geqslant \overline{p_B} \geqslant \max_{c \neq c_A} \mathbb{P}(f(\mathbf{x}+\varepsilon) = c). \tag{A1}$$

Then we have:

$$g(\mathbf{x}+\delta) = c_A, \qquad \forall \|\delta\|_2 < R \tag{A2}$$

where

$$R = \frac{\sigma}{2}(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B})). \tag{A3}$$

*Proof.* Define two variables

$$\begin{aligned} X &:= \mathbf{x} + \epsilon = \mathcal{N}\left(\mathbf{x}, \sigma^2\mathbf{M}\right) \\ Y &:= \mathbf{x} + \delta + \epsilon = \mathcal{N}\left(\mathbf{x}+\delta, \sigma^2\mathbf{M}\right). \end{aligned}$$

We know from [1] [2] that if $\exists \beta$ satisfying

$$\Omega = \left\{\delta^T\mathbf{v} \leqslant \beta, \mathbf{v} \in \mathbb{R}^d\right\}$$

and

$$\mathbb{P}(f(X) = c_A) \geqslant \mathbb{P}(X \in \Omega),$$

then

$$\mathbb{P}(f(Y) = c_A) \geqslant \mathbb{P}(Y \in \Omega) \tag{A4}$$

in which the inequality sign is reversed and $c_A$ is replaced by $c_B$, and the lemma still holds.
Define the half-spaces

$$\begin{aligned} \mathbb{A} &= \left\{\mathbf{v} : \delta^T(\mathbf{v}-\mathbf{x}) \leqslant \sigma\|\delta\|_2\Phi^{-1}(\underline{p_A})\right\} \\ \mathbb{B} &= \left\{\mathbf{v} : \delta^T(\mathbf{v}-\mathbf{x}) \geqslant \sigma\|\delta\|_2\Phi^{-1}(1-\overline{p_B})\right\}, \end{aligned}$$

then we have

$$\begin{aligned} \mathbb{P}(X \in A) &= \mathbb{P}\left(\delta^T(X-\mathbf{x}) \leqslant \sigma\|\delta\|_2\Phi^{-1}\left(\underline{p_A}\right)\right) \\ &= \mathbb{P}\left(\delta^T\mathcal{N}\left(0, \sigma^2\mathbf{M}\right) \leqslant \sigma\|\delta\|_2\Phi^{-1}\left(\underline{p_A}\right)\right) \\ &= \mathbb{P}\left(\sigma\|\mathbf{M}\odot\delta\|_2\mathbb{Q} \leqslant \sigma\|\delta\|_2\Phi^{-1}\left(\underline{p_A}\right)\right) \\ &= \Phi\left(\frac{\|\delta\|_2\Phi^{-1}\left(\underline{p_A}\right)}{\|\mathbf{M}\odot\delta\|_2}\right), \mathbb{Q} \sim \mathcal{N}(0,1) \end{aligned} \tag{A5}$$

$$\begin{aligned} \mathbb{P}(X \in B) &= \mathbb{P}\left(\delta^T(X-\mathbf{x}) \geqslant \sigma\|\delta\|_2\Phi^{-1}\left(1-\overline{p_B}\right)\right) \\ &= \mathbb{P}\left(\delta^T\mathcal{N}\left(0, \sigma^2\mathbf{M}\right) \geqslant \sigma\|\delta\|_2\Phi^{-1}\left(1-\overline{p_B}\right)\right) \\ &= \mathbb{P}\left(\sigma\|\mathbf{M}\odot\delta\|_2\mathbb{Q} \geqslant \sigma\|\delta\|_2\Phi^{-1}\left(1-\overline{p_B}\right)\right) \\ &= \Phi\left(\frac{\|\delta\|_2\Phi^{-1}\left(\overline{p_B}\right)}{\|\mathbf{M}\odot\delta\|_2}\right), \end{aligned} \tag{A6}$$

$$\begin{aligned} \mathbb{P}(Y \in A) &= \mathbb{P}\left(\delta^T(Y-\mathbf{x}) \leqslant \sigma\|\delta\|_2\Phi^{-1}\left(\underline{p_A}\right)\right) \\ &= \mathbb{P}\left(\delta^T\mathcal{N}\left(0, \sigma^2\mathbf{M}\right) + \|\delta\|_2^2 \leqslant \sigma\|\delta\|_2\Phi^{-1}\left(\underline{p_A}\right)\right) \\ &= \mathbb{P}\left(\sigma\|\mathbf{M}\odot\delta\|_2\mathbb{Q} \leqslant \sigma\|\delta\|_2\Phi^{-1}\left(\underline{p_A}\right) - \|\delta\|_2^2\right) \\ &= \mathbb{P}\left(\mathbb{Q} \leqslant \frac{\sigma\|\delta\|_2\Phi^{-1}\left(\underline{p_A}\right) - \|\delta\|_2^2}{\sigma\|\mathbf{M}\odot\delta\|_2}\right) \\ &= \Phi\left(\frac{\sigma\|\delta\|_2\Phi^{-1}\left(\underline{p_A}\right) - \|\delta\|_2^2}{\sigma\|\mathbf{M}\odot\delta\|_2}\right), \end{aligned} \tag{A7}$$

$$\begin{aligned}
\mathbb{P}(Y \in B) &= \mathbb{P}\left(\delta^T(Y-\mathbf{x}) \geqslant \sigma\|\delta\|_2\Phi^{-1}\left(1-\overline{p_B}\right)\right) \\
&= \mathbb{P}\left(\delta^T\mathcal{N}\left(0,\sigma^2\mathbf{M}\right) + |\delta\|_2^2 \geqslant \sigma\|\delta\|_2\Phi^{-1}\left(1-\overline{p_B}\right)\right) \\
&= \mathbb{P}\left(\sigma\|\mathbf{M}\odot\delta\|_2\mathbb{Q} \geqslant \sigma\|\delta\|_2\Phi^{-1}\left(1-\overline{p_B}\right) - \|\delta\|_2^2\right) \\
&= \mathbb{P}\left(\mathbb{Q} \geqslant \frac{\sigma\|\delta\|_2\Phi^{-1}\left(1-\overline{p_B}\right) - \|\delta\|_2^2}{\sigma\|\mathbf{M}\odot\delta\|_2}\right) \\
&= \mathbb{P}\left(\mathbb{Q} \leqslant \frac{\sigma\|\delta\|_2\Phi^{-1}\left(\overline{p_B}\right) + \|\delta\|_2^2}{\sigma\|\mathbf{M}\odot\delta\|_2}\right) \\
&= \Phi\left(\frac{\sigma\|\delta\|_2\Phi^{-1}\left(\overline{p_B}\right) + \|\delta\|_2^2}{\sigma\|\mathbf{M}\odot\delta\|_2}\right).
\end{aligned} \tag{A8}$$

The mask $\mathbf{M}$ covers the patch, so we have

$$|\delta\|_2 = \|\mathbf{M}\odot\delta\|_2, \tag{A9}$$

thus we have

$$\mathbb{P}(X \in A) = \Phi\left(\Phi^{-1}(\underline{p_A})\right) = \underline{p_A} \tag{A10}$$

$$\mathbb{P}(X \in B) = \Phi\left(\Phi^{-1}(\overline{p_B})\right) = \overline{p_B} \tag{A11}$$

$$\mathbb{P}(f(X) = c_A) \geqslant \mathbb{P}(X \in A) = \underline{p_A} \tag{A12}$$

$$\mathbb{P}(f(X) = c_B) \leqslant \mathbb{P}(X \in B) = \overline{p_B}. \tag{A13}$$

From A4, we have

$$\mathbb{P}(f(Y) = c_A) \geqslant \mathbb{P}(Y \in A) \tag{A14}$$

$$\mathbb{P}(f(Y) = c_B) \leqslant \mathbb{P}(Y \in B). \tag{A15}$$

If we want to get $\mathbb{P}(f(Y) = c_A) > \mathbb{P}(f(Y) = c_B)$, just let $\mathbb{P}(Y \in A) > \mathbb{P}(Y \in B)$, we can get that $\mathbb{P}(Y \in A) > \mathbb{P}(Y \in B)$ if and only if

$$R < \frac{\sigma}{2}(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B})). \tag{A16}$$

## Appendix B Pseudocode for Practical Localization

**Algorithm B1** Localization based on Joint voting

**Require:** Base classifier $f$, adversarial image $\mathbf{x}$, block set $\mathbb{B}^{r\times r}$, cross-shaped filter $\mathbb{T}$, threshold $\tau$.
**Ensure:** Patch location set $\mathbb{L}$ or Abstain.

initial $\mathbb{L} \leftarrow \varnothing$, $\mathbb{C}^{r\times r} \leftarrow \mathbf{0}$, score set $\mathbb{S} \leftarrow \varnothing$ and $\texttt{pred} \leftarrow \mathbf{0}$
**for** $b^{(i,j)}$ in $\mathbb{B}$ **do**
  $\mathbf{x}^{(i,j)} \leftarrow \textsc{Inpaint}(\mathbf{x}, b^{(i,j)})$
  $\texttt{pred}(i,j) \leftarrow f(\mathbf{x}^{(i,j)})$
**end for**
$\hat{l}_A \leftarrow$ top indices in pred
**for** $l^{(i,j)}$ in pred **do**
  **if** $l^{(i,j)} \neq \hat{l}_A$ **then**
    $\mathbb{C}[i,j] \leftarrow 1$
  **end if**
**end for**
**for** $\mathbb{C}[p,q] \in \mathbb{C}$ **do**
  $\mathbb{S}[p,q] \leftarrow \textsc{Sum}(\mathbb{T}_{\mathbb{C}[p,q]} \odot \mathbb{C})$
**end for**
$s_A \leftarrow \max(\mathbb{S})$
**for** $\mathbb{S}[m,n] \in \mathbb{S}$ **do**
  **if** $\mathbb{S}[m,n] = s_A$ **then**
    $\mathbb{L} \leftarrow \mathbb{L} \cup \mathbb{B}[m,n]$
  **end if**
**end for**
**for** $\mathbb{L}[m,n] \in \mathbb{L}$ **do**
  **for** $\mathbb{C}[p,q] \in \mathbb{C}$ **do**
    **if** $|p-m| < \tau \wedge |q-n| < \tau$ **then**
      $\mathbb{L} \leftarrow \mathbb{L} \cup \mathbb{B}[p,q]$
    **end if**
  **end for**
**end for**
**return** $\mathbb{L}$

The pseudocode of localization is described in algorithm B1. Function $\textsc{Inpaint}(\mathbf{x}, b_i)$ reconstructs block $b_i$ using its neighbor pixels. The filter is centered on the currently calculated block and takes out the values in the filter. Different from the certification, the process of the prediction will perform a binomial test on the number of $c_A$ and $c_B$ to ensure that the majority voted category is correct with high probability.

# Appendix C Other Results

## Appendix C.1 More comparisons with DRS

For DRS, the certified condition is $Counts[c] > Counts[c_{next}] + 2 \times AffectedBands$, which implies that the larger the patch is, the harder the condition is reached. So we compare the performance to Column smoothing of DRS when assuming the patch size is $10 \times 10$ and $15 \times 15$ on CIFAR-10 datasets. Table C1 shows that our method is much better than DRS on large patches.

**Table C1** Certified accuracy of our method and DRS (Column smoothing) on CIFAR-10 when the patch size is $10 \times 10$ and $15 \times 15$. We use 100-accuracy for the $10 \times 10$ patch and 225-accuracy for the $15 \times 15$ patch.

| | $(10 \times 10)$ | $(15 \times 15)$ |
|---|---|---|
| DRS | 29.0 | 1.9 |
| Ours | 60.0 | 32.7 |

## Appendix C.2 Accuracy of Localization

From Appendix A, we only need to ensure that the localization can completely cover the patch to calculate a credible certificate, so the localization accuracy (LA) is defined as the ratio of images where the patch is completely covered to the total test images. The coverage area is defined as the ratio of the mask covering the patch to the image, and we count the average coverage area (ACA) of the successfully localized images to help measure the effectiveness of the localization algorithm. The larger the image resolution, the more accurate the localization, in agreement with the experimental part. And as long as the localization algorithm provides the mask, the certificate provided by MRS can be obtained.

**Table C2** The localization accuracy (LA) and average coverage area (ACA) on CIFAR-10 and ImageNet. The larger the LA and the smaller the ACA the better the localization algorithm performs.

| | LA | ACA |
|---|---|---|
| CIFAR-10 | 96.2 | 12.2 |
| ImageNet | 99.8 | 11.7 |

**References**

1 Neyman J, Pearson E S. On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society A, 1933, 231: 289-337

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