ViP: Unified Certified Detection and Recovery for Patch Attack with Vision Transformers

Junbo Li[1], Huan Zhang[2], and Cihang Xie[1]

[1] University of California Santa Cruz
[2] Carnegie Mellon University

Abstract. Patch attack, which introduces a perceptible but localized change to the input image, has gained significant momentum in recent years. In this paper, we present a unified framework to analyze certified patch defense tasks, including both *certified detection* and *certified recovery*, leveraging the recently emerged Vision Transformers (ViTs). In addition to the existing patch defense setting where only one patch is considered, we provide the very first study on developing certified detection against the *dual patch attack*, in which the attacker is allowed to adversarially manipulate pixels in two different regions.

By building upon the latest progress in self-supervised ViTs with masked image modeling (*i.e.*, masked autoencoder (MAE)), our method achieves state-of-the-art performance in both certified detection and certified recovery of adversarial patches. Regarding certified detection, we improve the performance by up to ~16% on ImageNet without training on a single adversarial patch, and for the first time, can also tackle the more challenging dual patch setting. Our method largely *closes the gap* between detection-based certified robustness and clean image accuracy. Regarding certified recovery, our approach improves certified accuracy by ~2% on ImageNet across all attack sizes, attaining the new state-of-the-art performance.

Keywords: Certified Defense, Patch Attacks, Vision Transformer

1 Introduction

Deep neural networks (DNNs) are vulnerable to adversarial attacks [9, 22]. Researchers have come up with various attacks to craft visually imperceptible adversarial examples that can lead to a model failing in a set of image recognition tasks, including classification [22], object detection [29], semantic segmentation [4, 29], *etc*. Among these attack methods, patch attack [2, 8, 13, 25, 30] considers arbitrarily modifying a small and continuous region in an image, which utilizes characteristics of physical objects. Due to arbitrary location and small size of the patch attack, it is more challenging to defend against such an attack. Existing empirical methods designed for defending against patch attacks [11, 20] reported ~70% robust accuracy on ImageNet [6]. However, if we consider a stronger attacker who is aware of the pre-processing step, the robustness of these defenses will severely drop to ~50% [3].

To fix such issues, another series of works focus on designing provable mechanisms, which aim to provide a provable defense against adversarial attacks. Specifically, on account of different levels of provable defense, there are usually two kinds of tasks: certified detection [10, 14, 17, 28] and certified recovery [15, 18, 21, 26] for adversarial patches. The former task is to detect whether an image was successfully attacked or not, while the latter one aims to classify an image correctly under any patch attacks smaller than a particular size. In general, certified recovery is considered as a much more challenging task than certified detection in the real-world scenario.

In certified detection, a small mask is applied on a clean image and slides from upper left to lower right (*i.e.*, acting like a convolution kernel). We have a partially occluded image for a patch mask applied on each position. All these different images are sent to a DNN. Finally, the original image can be certifiably detected for any patch attacks if all the output prediction results are strictly consistent. Related works include Minority Reports Defense [17], PatchGuard++ [28], and ScaleCert [10]. However, these methods are either computationally intractable for large-scale data or rely on CNNs with a small reception field to extract features, restricting their further applications. Recently, Huang *et al.* [14] introduces ViTs for certified detection and substantially improves performance, even for defending against larger patch attacks.

We find that methods for certified recovery based on randomized smoothing [5, 16] are similar to certified detection. In this setting, we first forward a small subset of an image each time and then make a majority voting for the outputs of all these small subsets. Appropriate geometry structures provide that a patch can only intersect with restricted small subsets. Therefore if the gap between the majority prediction and the sub-majority prediction is large enough, we can guarantee that the voting result will not change regardless of where the attacker put the patch attack.

The current state-of-the-art methods for these two tasks are both achieved with vision transformers [14, 21]. Interestingly, we find that these two tasks can actually be solved in a unified framework with vision transformer structures and a strategy of dropping patches. A patch attack can be certifiably detected if we drop a few patches each time and all the predictions are strictly consistent. In comparison, an image can be certifiably recovered if we drop many patches each time, and the gap between the majority voting predictions and the sub-majority voting predictions is big enough. Due to the similarity of the two tasks, we can use the same framework and network structure to solve these problems. Moreover, in real-world attack settings, we cannot restrict how many patches an attacker can use, so it is necessary to design general defense algorithms beyond single-patch attack.

In this work, we present ViP, a unified analysis framework for certified robustness including both certified detection certified recovery. Benefited by the recent progress in self-supervised vision transformers, especially the powerful masked autoencoder (MAE) [12], we achieve the state-of-the-art performance on all related tasks. By evaluating certified detection in a zero-shot manner, our method improves the certified detection rate by up to ~16% on ImageNet over the prior art [14]. Moreover, we develop the first theoretical guarantee for dual-patch attack detection. As a byproduct of improvement on single-patch attack, we successfully generalize the dual-patch detection to the large-scale dataset like ImageNet. In addition, our methods improves the certified accuracy by ~2% on all tasks in certified recovery compared to the state-of-the-art [21].

2 Related Works

2.1 Certified Detection

McCoyd *et al.* [17] is the first work on certified detection for adversarial patches. Their certification is achieved by generating a prediction grid. However, this method is computationally infeasible on large-scale dataset like ImageNet. To reduce computational complexity, Xiang *et al*. [28] uses CNNs with small reception field and conducts masking on feature level. However, they still cannot get a good performance on ImageNet. The performance is restricted due to locality information. Moreover, Han *et al*. [10] proposes to only forward the top k of SIN, Superficial Important Neurons. Recently, Huang *et al*. [14] proposes to use vision transformer structures to do certified detection, which improves a lot on both performance and speed.

2.2 Certified Recovery

Earlier works on certified recovery include [3, 19], which rely on the bound of activation value. However they are infeasible to extend to large-scale datasets. Based on traditional randomized smoothing method [5, 16], Levine *et al*. [15] first proposes the (de)randomized smoothing method designed for patch attack, and scales to ImageNet. The follow-up work [21] changes to use vision transformers, but it remains unclear why vision transformers work better than CNNs. If assuming information of patch size is known, PatchCleanser in [27] designs a two-stage certification process that enjoys a much better recovery rate.

2.3 Vision Transformers

Application of self-attention blocks in vision transformers has achieved a huge success these years [7, 23]. Due to patchfying an image to be a token sequence, a vision transformer can accept almost arbitrary subparts of an image as the input. This greatly helps self-supervision, especially using masked-image-modeling strategy [1, 12, 24, 31]. These works demonstrate the great potential of vision transformers. Especially for the task of certified detection or recovery of adversarial patches, vision transformers are a better choice compared to traditional CNNs, because vision transformers intrinsically use patches as their inputs.

3 Certified Patch Defense

3.1 Problem Setup

We consider the L_0 patch attack in this work, which shares the same setting as many previous works. Specifically, for a classifier $F : \mathbb{R}^{C\times H\times W} \to \{1, \cdots, N\}$ and an image $x \in \mathbb{R}^{C\times H\times W}$ with channel C, height H and width W, the attacker can adversarially choose (a, b) as the upper left position, and arbitrarily change pixels within the corresponding square patch of size p. For an attack set $\mathcal{A}$ and $A \in \mathcal{A}$, denote $\mathcal{K}(A)$ to be the set of pixels that are changeable by A. Denote $\mathcal{A}_p$ to be the set of attacks that can arbitrarily change pixels inside a $p \times p$ square. So for any $A \in \mathcal{A}_p$, $\mathcal{K}(A)$ is a $p \times p$ square. Since an arbitrary rectangular can be covered by a larger square, here we only consider a square patch attack.

In our defense framework, the classifier F is based on base classifiers, and we denote these base classifiers as $f_1, \cdots, f_n$. Moreover, since the certification framework relies on mask strategy, here we make some additional notations. Suppose M is a $0-1$ mask matrix that shares the same height and width with image x, where masked pixels are 0 and others are 1. Denote $\mathcal{O}(M)$ to be the positions of pixels that are masked. For a classifier f, define $f_M(x) = f(M \bigodot x)$.

3.2 Certified Detection for Patch Defense

In this task, our goal is to decide whether an image is successfully attacked or not. This can be achieved by choosing base classifiers $f_1, \cdots, f_n$ and F properly. Our algorithms are based on the following theorem.

Definition 1. *For an attack set $\mathcal{A}$, the base classifiers $f_1, \cdots, f_n$ are called **compatible** to $\mathcal{A}$, if they satisfy that for any image $x \in \mathbb{R}^{C\times H\times W}$ and any attack $A \in \mathcal{A}$, there exists $1 \le i \le n$ such that $f_i(x) = f_i(A(x))$.*

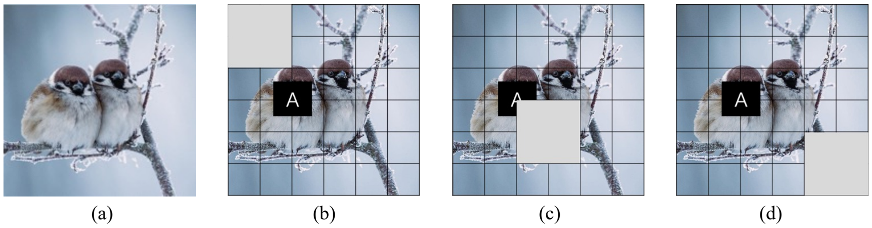
Theorem 1. *For an attack set $\mathcal{A}$, suppose the base classifiers $f_1, \cdots, f_n$ are compatible to $\mathcal{A}$. For an image $x \in \mathbb{R}^{C\times H\times W}$, an attack from $\mathcal{A}$ can be either certified detected or regarded as "harmless" if no warning is raised under the following definition of F:*

$$F(x) := \begin{cases} a, & \text{if } f_1(x) = \cdots f_n(x) = a \\ warning, & else \end{cases}$$

Proof. Suppose x is an image that satisfies

$$f_1(x) = \cdots = f_n(x) = a.$$

For any attack $A \in \mathcal{A}$, denote $U = unique(f_1(A(x)), \cdots, f_n(A(x)))$ to be the deduplication of the set $\{f_1(A(x)), \cdots, f_n(A(x))\}$, and $\#U$ to be the number of elements of U. If $\#U \ge 2$, then obviously x is attacked. If $\#U = 1$, suppose we have $f_1(A(x)) = \cdots = f_n(A(x)) = b$. However, there exist $1 \le i \le n$ such that $f_i(x) = f_i(A(x))$ because $f_1, \cdots, f_n$ are compatible to $\mathcal{A}$. Since $f_i(x) = a$ and $f_i(A(x)) = b$, we have $a = b$. So this attack fails to make any changes to the prediction result.



(a) (b) (c) (d)

Fig. 1. (a)-(d) is certified detection process. (a) is the original image. We slide a square mask from upper left to lower right. (b), (c) and (d) are three different positions. The black patch “A” is an adversarial attack. One of such gray square masks can fully cover the adversarial patch.

Our theorem is a more general version of previous certification [14, 17]. We have no restrictions of base classifiers constructions, and can be adapted to any attack set out of single-patch attack. For example, we can also solve dual-patch attack or more generally, sparse adversarial attack proposed in [16].

Base Classifiers Design Denote $\mathcal{A}_p$ to be the set of square patch attacks with size smaller than p. When attacks are restricted to $\mathcal{A}_p$, it is not hard to design the base classifiers. Fixed a known classifier f, we consider classifiers $\tilde{f} \in \mathcal{F} = \{f_M, M : 0\text{-}1mask\}$. So we only need to choose n masked areas $M_1, \cdots, M_n$ such that for any $A \in \mathcal{A}_p$, there exists $1 \leq i \leq n$ such that $M_i \bigodot x = M_i \bigodot A(x)$, which means $\mathcal{K}(A) \subset \mathcal{O}(M_i)$.

Therefore, it is sufficient to assign a M_i for every possible position of $p \times p$ area. Moreover, since we need $f_{M_1}(x) = \cdots = f_{M_n}(x)$, we should make $\mathcal{O}(M_i)$ as small as possible, so that the unmasked area is big enough to achieve consistency for all these classifiers. Based on these analysis and the principles of simplicity, we first set a proper size $m \geq p$ and a proper stride s. We can have two straight solutions. Firstly, we can slide the $m \times m$ square area with stride s from upper left to lower right. Each $m \times m$ area acts as $\mathcal{O}(M_i)$ for some M_i. Recent works based on CNNs have similar ideas [10, 17, 28]. However, the complexity increases quadratically as image size increases. So it is impractical to make a certification for high resolution like 224. Second, we can slide a band of width m with stride s from left to right. Each band acts as $\mathcal{O}(M_i)$ for some M_i. This is a linear-complexity algorithm that hasn't been explored before for certified detection.

Although the above ideas are restricted to use due to the high complexity, we find that vision transformers can perfectly solve this problem. Figure 1 and 2 (a)(b) show how we can take advantage of vision transformers for quadratic and linear complexity respectively. First, the natural square patch structure can act as the role of a sliding window with a big size and stride. For any $A \in \mathcal{A}_p$, it will only intersect with restricted patches. For example, if we use regular vision transformers with patch size 16, the attack A only influences r_p^2 patches, where $r_p = \lceil (p-1)/16 \rceil + 1$. Second, after patchifying the image to

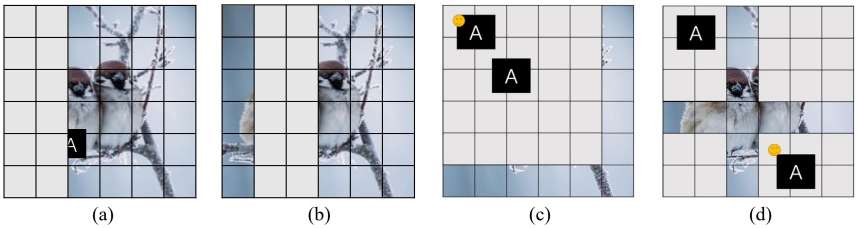


Fig. 2. (a) and (b) are the illustration of linear-complexity certification. We slide a band mask from left to right. At least one gray mask will fully cover the adversarial patch (Figure (b)). In (c) and (d), we can see that when two adversarial patches are close enough, we can certainly cover them with a reasonable bigger area. However, they can distribute arbitrarily far from each other in the image. Our generalized window solve this problem as shown in (d). Here we mark the start position of a generalized window with a smiley face.

be a sequence of patches, we can take the tokens that are not influenced by A, and drop others. This can additionally reduce complexity. Finally, with a large kernel size and stride, the certification process can be two-magnitude faster than previous methods. So we can certify ImageNet data with a high speed. This method using vision transformers with quadratic complexity is also illustrated in [14].

Certified Detection for Dual-Patch Attack Dual-patch attack is considered to be challenging for certified defense. In this setting, the attacker is allowed to attack two arbitrary patches. Naturally, we hope to use what we do in the defense of a single-patch attack: finding some i such that $\mathcal{K}(A) \subset \mathcal{O}(M_i)$. It is natural when two adversarial patches are close to each other enough. However, what if they are far from each other since we allow the attacker to choose patches arbitrarily? What kind of masks should we choose to make the base classifiers compatible to the attack set?

Does our theoretical guarantee fail completely? The answer is no. Actually, we can still certify dual-patch attack cases with slight modification. The first key is to modify the topological structure of an image. For image size $H \times W$, we define the generalized window as follows:

Definition 2 (Generalized Window). *For any $1 \leq a \leq H, 1 \leq b \leq W$, a generalized window M of size (m_1, m_2) starting from position (a, b) is defined to be $M = \{((a+i)\%H, (b+j)\%W)\}_{i=1}^{m_1}{}_{j=1}^{m_2}$. Here % means taking the remainder.*

Obviously, when a, b are small, generalized windows are just the same as the regular square windows. With generalized windows, we can say the left and right sides of an image are connected, and the upper and lower sides are also connected. Our certification is based on generalized windows.

Theorem 2. *For an even number q and a square grid of size $q \times q$, any two sub-areas of size $p_n \times p_n$ can be covered by a generalized window of size $(p_n + q/2, p_n + q/2)$.*

Actually the geometric understanding is quite straightforward. Please refer to Figure 2. We leave the formal proof of Theorem 2 to appendix. Now we can do certification for dual-patch cases totally same as single-patch case. Here q is the patch number typically set to be $224/16 = 14$, also denote p_s is the patch size. Denote $\mathcal{A}_p \times \mathcal{A}_p$ to be the set of dual-patch attacks of size p. By Theorem 2, for any $A \in \mathcal{A}_p \times \mathcal{A}_p$, there exists a generalized window M of size $((p_n + q/2)p_s, (p_n + q/2)p_s)$, where $p_n = \lceil (p-1)/p_s \rceil + 1$, that fully covers areas influenced by A. So let $\{M_i\}_{i=1}^{q^2}$ satisfy that $\{\mathcal{O}(M_i)\}_{i=1}^{q^2}$ are all the generalized windows of size $((p_n + q/2)p_s, (p_n + q/2)p_s)$, which are exact combinations of some patches. Then define $f_i = f_{M_i}$ for $1 \leq i \leq n$ and a known classifier f. Theorem 1 provides that we can certified detect every successful attack from $\mathcal{A}_p \times \mathcal{A}_p$ if the predictions of f_i are consistent.

3.3 Certified Recovery for Patch Defense

This task is more challenging because we aim to directly make a correct prediction no matter an adversarial attack is successful or not. This task also highly relies on the choices of $f_1, \cdots, f_n$ and F.

Definition 3. *For an attack set $\mathcal{A}$, the base classifiers $f_1, \cdots, f_n$ are called **m-compatible** to $\mathcal{A}$, if*

$$\sup_{x \in \mathbb{R}^{C \times H \times W}} \sum_{i=1}^{n} \mathbb{I}\left\{ f_i(x) \neq f_i(A(x)) \right\} = m.$$

Theorem 3. *For an attack set $\mathcal{A}$, suppose the base classifiers $f_1, \cdots, f_n$ are m-compatible to $\mathcal{A}$. For an image x, and all labels $\mathcal{Y} = \{1, \cdots, N\}$, denote $n_j(x)$ to be the number of base classifiers that return label $j \in \mathcal{Y}$. Also, denote $\{n_{i_j}(x)\}_{j=1}^{N}$ to be the descending sort of $\{n_j(x)\}_{j=1}^{N}$. Define*

$$F(x) = \arg\max_{1 \leq j \leq N} n_j(x).$$

If $n_{i_1}(x) > n_{i_2}(x) + 2m$, then $F(x) = F(A(x))$ for any $A \in \mathcal{A}$

Proof. Our goal is to prove that if $n_{i_1}(x) > n_{i_2}(x) + 2m$, then for any $A \in \mathcal{A}$, $n_{i_1}(A(x)) = \max_{1\leq j\leq n} n_j(A(x))$. By definition, we have

$$\begin{aligned}
n_{i_1}(A(x)) &= \sum_{i=1}^{n} \mathbb{I}\{f_i(A(x)) = i_1\} \\
&= \sum_{i=1}^{n} \mathbb{I}\{f_i(x) = i_1\} + \sum_{i=1}^{n} (\mathbb{I}\{f_i(A(x)) = i_1\} - \mathbb{I}\{f_i(x) = i_1\}) \\
&= n_{i_1}(x) + \\
&\quad \sum_{i=1}^{n} (\mathbb{I}\{f_i(A(x)) = i_1, f_i(x) \neq i_1\} - \mathbb{I}\{f_i(A(x) \neq i_1, f_i(x) = i_1)\}) \\
&\geq n_{i_1}(x) - \sum_{i=1}^{n} \mathbb{I}\{f_i(A(x)) \neq i_1, f_i(x) = i_1\} \\
&\geq n_{i_1}(x) - \sum_{i=1}^{n} \mathbb{I}\{f_i(A(x)) \neq f_i(x)\} \\
&\geq n_{i_1}(x) - \sup_{x\in\mathbb{R}^{C\times H\times W}} \sum_{i=1}^{n} \mathbb{I}\{f_i(A(x)) \neq f_i(x)\} = n_{i_1}(x) - m.
\end{aligned}$$

Due to $n_{i_1}(x) > n_{i_2}(x) + 2m$, this gives for any $j \neq i_1$,

$$n_{i_1}(A(x)) \geq n_{i_1}(x) - m > n_{i_2}(x) + m \geq n_j(x) + m.$$

Moreover,

$$\begin{aligned}
&n_{i_1}(A(x)) > n_j(x) + m = \sum_{i=1}^{n} \mathbb{I}\{f_i(x) = j\} + m \\
&= \sum_{i=1}^{n} \mathbb{I}\{f_i(A(x)) = j\} + \sum_{i=1}^{n} (\mathbb{I}\{f_i(x) = j\} - \mathbb{I}\{f_i(A(x)) = j\}) + m \\
&= n_j(A(x)) + \\
&\sum_{i=1}^{n} \mathbb{I}\{f_i(x) = j, f_i(A(x)) \neq j\} - \sum_{i=1}^{n} \mathbb{I}\{f_i(x) \neq j, f_i(A(x)) = j\} + m \\
&\geq n_j(A(x)) - \sum_{i=1}^{n} \mathbb{I}\{f_i(x) \neq j, f_i(A(x)) = j\} + m \\
&\geq n_j(A(x)) - \sup_{x\in\mathbb{R}^{C\times H\times W}} \sum_{i=1}^{n} \mathbb{I}\{f_i(x) \neq j, f_i(A(x)) = j\} + m = n_j(A(x)).
\end{aligned}$$

Hence, for any $j \neq i_1$, we have

$$n_{i_1}(A(x)) > n_j(A(x)).$$

Therefore, for any $A \in \mathcal{A}$, we have $n_{i_1}(A(x)) = \max_{1\leq j\leq n} n_j(A(x))$ and $f(x) = f(A(x))$.

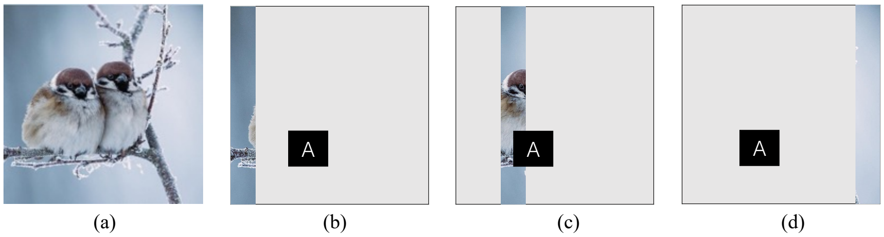


Fig. 3. Certification process using derandomized smoothing. (a) is the original full image. We slide the band of size $(224, w)$ from left to right with stride s. (b), (c), (d) are three positions. The black patch "A" is an adversarial patch with size p. The band in (c) intersect with the adversarial patch, and the adversarial patch only intersects bands close to (c).

Similar to Theorem 1, Theorem 3 is also very flexible. Actually for any base classifiers, at least we have $f_1, \cdots, f_n$ is $m-$compatible for $m = n$. So all we need to do is to make m small enough.

Base Classifiers Design When it comes to square patch attack set $\mathcal{A}_p$, we also choose base classifiers based on the mask strategy. The goal is different from the first task. Previously, we want a consistency so we keep as many as possible unmasked area. Now we need the gap between n_{i_1} and n_{i_2} is bigger than the 'inconsistency' of predictions between clean image and attacked image, which means we need to minimize the influence of any attack $A \in \mathcal{A}_p$. Therefore, we need to mask as large as possible areas. Notice that there exists a trade-off since masking more area will reduce m in Theorem 3, but decrease the accuracy of a single classifier.

Because the mask is big here, we also need to use generalized windows otherwise there will be only limited areas to choose. However, square masks do not work in recovery, we show this through an example. Considering patch attack that can change about 32×32 pixels of a 224×224 image (approximately 2%). For $m \geq 32$, there are $[(m - 32 + 1)/s]^2$ masks that can fully cover a 32×32 attack. This means for stride s, about

$$\left(\frac{224}{s}\right)^2 - \left(\frac{m-31}{s}\right)^2$$

base classifiers could be influenced by 32×32 attack. Suppose m is approximately 200, then the above term is approximately $20000/s^2$. But the number of total base classifiers is only $224^2/s^2 \approx 50000/s^2$. So it is hard to make the gap larger than $40000/s^2$. Therefore, if we use mask-based classifiers, we should consider choices out of squares. Actually rectangular mask whose either height or width equals to the original image works. That's what recent works focusing on (de)randomized smoothing did. Figure 3 shows this process. Let $M_1, \cdots, M_n$ be generalized windows of size (H, m) (or similarly, (m, W)) and stride s. In this case, the masked area only slides along one dimension instead of two. Each $p \times p$ square is fully covered by $\lceil (m-p+1)/s \rceil$ of them. Finally, we let $f_i = f_{M_i}$ for $1 \leq i \leq n$. When it comes to vision transformer backbones, we only take patches that are not fully covered, which bring less complexity with reduced input tokens.

3.4 Similarities of Certified Detection and Certified Recovery

Recently, both tasks achieve the state-of-the-art result using vision transformer structures, but it remains unclear why vision transformers work well. From the above analysis, we can find that there does exist many similarities in these two tasks. In both, we find a masked area and slide them around the image. After obtaining predictions using masked part, we either analyze consistency or vote. Since our model either does zero-shot certification or slightly finetunes with target size of unmasked area, the training recipe acts an important role. The recent progress of self-supervised vision transformers inspires us to choose vision transformers pretrained with the mask-image-modeling (MIM) method as our backbone, to additionally generate base classifiers in both tasks.

4 Results

In this section, we compare ViP and previous methods for various certified robustness tasks on ImageNet, which is challenging for most of the previous certification methods. Our methods with masked autoencoder (MAE) achieves state-of-the-art performance on all related tasks.

4.1 Certified Detection

Single-Patch Detection We first evaluate our certified detection methods on single-patch detection in a zero-shot manner.Results are shown in Table 1. We range the possible influenced patch number from 2×2 to 8×8 following [14]. We also compare with methods based on CNNs [10, 28] on pixel level. Results are shown in Table 2.

Both DeiT and MAE surpass results in [14] a lot. Although DeiT and the original ViT model in [14] shares similar accuracy on clean images, we find that DeiT has a very big improvement about $3 \sim 11\%$ over the original ViT model. This illustrates that data augmentation in training helps a lot. Additionally, MAE makes further improvement, about $6 \sim 16\%$ compared to [14]. At the first time, detection-based certified robustness is approaching the clean accuracy, even with a zero-shot manner.

We also show the results of linear-complexity certified detection in the last line of Table 1. This new linear time detection algorithm produces slightly worst results compared to our original quadratic time algorithm, yet being much faster. Furthermore, our linear time algorithm outperforms the quadratic algorithm in

Table 1. Results of single-patch certified detection on ImageNet. Here 2×2 means that an adversarial patch attack can at most influence 2×2 square patches of size 16×16. So 2×2 actually corresponds to patch attack of size not larger than 17×17 in the original 224×224 image. The same is true for 3×3 to 8×8. For example, 3×3 corresponds to patch attack of size $18 \times 18 \sim 33 \times 33$, and so on.

methods	complexity	speed	clean	certified robustness						
			acc.	2×2	3×3	4×4	5×5	6×6	7×7	8×8
PatchVeto [14]	quadratic	0.92s	81.8	72.0	67.2	61.9	56.4	50.5	44.1	37.1
ViP_DeiT_base	quadratic	0.92s	81.9	75.0	71.4	67.4	63.2	58.6	53.8	48.4
ViP_MAE_base	quadratic	0.92s	83.7	**77.7**	**74.6**	**70.9**	**67.2**	**62.9**	**58.4**	**53.4**
ViP_MAE_base	linear	0.11s	83.7	74.5	70.6	66.2	61.6	56.8	52.1	46.7

[9] with an improvement about $3 \sim 9\%$. Our certification time for one image against 2×2 attack is reduced from 0.92 seconds to 0.11 seconds.

We then compare on pixel level. Here we choose patch size from $\{24, 32, 40\}$, which corresponds to $1 \sim 3\%$ pixels respectively.

Table 2. Results of single-patch certified detection compared to previous CNN-based methods on ImageNet. MRD [17] cannot scale to ImageNet.

methods	1% pixels		2% pixels		3% pixels	
	acc	rob	acc	rob	acc	rob
MRD [17]	-	-	-	-	-	-
PatchGuard++ [28]	61.8	36.3	61.6	33.9	61.5	31.1
ScaleCert [10]	62.8	60.4	58.5	55.4	56.4	52.8
ViP_MAE_base	**83.66**	**74.56**	**83.66**	**74.56**	**83.66**	**70.9**

Dual-Patch Detection This task of detecting attacks with two patches is much more challenging and was never demonstrated in prior works. We are the first to make this practical on large-scale dataset like ImageNet as far as we know. Table 3 compares different training recipe with certification in our framework. Conclusions are similar as the single-patch case. MAE-based model improves about 20% under two 2×2 adversarial patches. This also illustrates the effect of MIM-based pretraining methods; when the adversarial patch sizes are bigger, the detection rate is lower however the MAE-based model is consistently better. The low detection rate is reasonable since two big patches like 6×6 can actually cover most of the key information of an image. Note that these results are also obtained *without additional training* and we can potentially further improve performance by finetuning on target image size.

Table 3. Results of dual-patch certified detection on ImageNet. This setting is much more challenging and is not handled by existing works. Here 2 × 2 means each of these two patches can at most influence 2 × 2 square patches of size 16 × 16. So are 3 × 3 to 6 × 6.

model	clean	robustness				
	acc.	2×2	3×3	4×4	5×5	6×6
ViP_ViT_base	81.8	22.2	12.5	4.3	0.6	0.03
ViP_DeiT_base	81.9	35.5	26.4	16.6	6.3	0.4
ViP_MAE_base	83.66	**42.0**	**33.3**	**23.9**	**14.1**	**4.8**

4.2 Certified Recovery

Finally, we evaluate different certified recovery methods compared with previous CNN-based methods (Figure 2) and current state-of-the-art method [21] when adversarial patch size is unknown. Following the same setting in [21], we choose same training parameters and test with same width of unmasked area. In detail, for width $w \in \{19, 25, 37\}$, we randomly choose an unmasked area of size $(224, w)$. After patch embedding, we drop tokens that are fully masked. We train for 30 epochs, using SGD optimizer with momentum 0.9, fixed learning rate 1e-3 for batch size 256, with a weight decay 1e-4. We only use random resized crop, horizontal flip and color jitter for data augmentation. Our method using MAE_base all achieve a better performance with about 2% improvement over all different width and stride.

Moreover, we test the influence of fintuning epoch and model size for width 19. If we train for longer epochs like 60, the recovery rate can additionally gain for about 1%. So we have not achieved the limit. Also, we test MAE_large with 30-epoch finetuning. This can further give about 5% improvement compared with MAE_base, and nearly 10% improvement compared with DeiT_base. Compared with about 2% improvement over MAE_base and 5% improvement over DeiT on clean accuracy, we find that certified robustness benefits more from larger model.

Table 4. Results of certified recovery compared to previous CNN-based methods on ImageNet. Here we use ViP_MAE_base with width 19 and stride 1.

methods	1% pixels		2% pixels		3% pixels	
	acc	rob	acc	rob	acc	rob
DS [15]	44.4	17.7	44.4	14.0	44.4	11.2
PatchGuard [26]	55.1	32.3	54.6	26.0	54.1	19.7
ViP_MAE_base	**70.4**	**45.0**	**70.4**	**40.3**	**70.4**	**35.6**

Table 5. Results of certified recovery.

method	epoch	width	stride	clean acc.	adversarial robustness		
					24 × 24	32 × 32	40 × 40
ViP_DeiT_base [21]	30	19	10	68.3	36.9	36.9	31.4
			5	69.0	40.6	37.7	32.0
			1	69.3	43.8	38.3	34.3
		25	10	70.3	40.9	35.2	29.8
			5	70.8	41.6	36.0	33.0
			1	72.1	44.0	38.8	34.8
		37	10	72.6	41.3	36.1	30.8
			5	73.1	41.9	36.4	33.5
			1	73.2	43.0	38.2	34.1
ViP_MAE_base	30	19	10	69.5	38.8	38.8	33.1
			5	70.1	42.5	39.6	33.7
			1	70.4	45.0	40.3	35.6
		25	10	71.1	42.8	37.3	31.6
			5	72.3	43.7	37.8	34.9
			1	72.5	45.6	40.9	36.2
		37	10	75.3	43.9	38.3	32.8
			5	75.7	44.5	38.7	32.9
			1	75.8	45.1	40.4	35.7
ViP_MAE_base	60	19	10	69.9	39.8	39.8	34.2
			5	70.4	43.5	40.6	34.8
			1	70.8	46.0	41.4	36.7
ViP_MAE_large	30	19	10	73.6	44.6	44.6	38.6
			5	74.1	48.3	45.4	39.3

5 Conclusion

In this work, we propose a unified analysis framework for certified robustness tasks including both certified detection and certified recovery. For L_0 patch attack, these two tasks both rely on the choices of the masks. Our work illustrate the great potential of using recent progress on vision transformers, especially cutting-edge self-supervised masked-image-modeling methods, to promote patch defense. With our defense framework, certified robustness for both tasks are approaching the clean accuracy. Compared with earlier works which only have about $20 \sim 30\%$ certified robustness, we make a progress on making certification practical in real world.

Acknowledgement

This work is supported by a gift from Open Philanthropy, TPU Research Cloud (TRC) program, and Google Cloud Research Credits program.

References

1. Bao, H., Dong, L., Wei, F.: Beit: Bert pre-training of image transformers. In: ICLR (2022)
2. Brown, T.B., Mané, D., Roy, A., Abadi, M., Gilmer, J.: Adversarial patch. arXiv preprint arXiv:1712.09665 (2017)
3. Chiang, P.Y., Ni, R., Abdelkader, A., Zhu, C., Studer, C., Goldstein, T.: Certified defenses for adversarial patches. In: ICLR (2020)
4. Cisse, M.M., Adi, Y., Neverova, N., Keshet, J.: Houdini: Fooling deep structured visual and speech recognition models with adversarial examples. In: NeurIPS (2017)
5. Cohen, J., Rosenfeld, E., Kolter, Z.: Certified adversarial robustness via randomized smoothing. In: ICML (2019)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
8. Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., Song, D.: Robust physical-world attacks on deep learning visual classification. In: CVPR (2018)
9. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: ICLR (2015)
10. Han, H., Xu, K., Hu, X., Chen, X., Liang, L., Du, Z., Guo, Q., Wang, Y., Chen, Y.: Scalecert: Scalable certified defense against adversarial patches with sparse superficial layers. In: NeurIPS (2021)
11. Hayes, J.: On visible adversarial perturbations & digital watermarking. In: CVPR Workshops (2018)
12. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR (2021)
13. Huang, L., Gao, C., Zhou, Y., Xie, C., Yuille, A.L., Zou, C., Liu, N.: Universal physical camouflage attacks on object detectors. In: CVPR (2020)
14. Huang, Y., Li, Y.: Zero-shot certified defense against adversarial patches with vision transformers. arXiv preprint arXiv:2111.10481 (2021)
15. Levine, A., Feizi, S.: (de) randomized smoothing for certifiable defense against patch attacks. In: NeurIPS (2020)
16. Levine, A., Feizi, S.: Robustness certificates for sparse adversarial attacks by randomized ablation. In: AAAI (2020)
17. McCoyd, M., Park, W., Chen, S., Shah, N., Roggenkemper, R., Hwang, M., Liu, J.X., Wagner, D.: Minority reports defense: Defending against adversarial patches. In: ACNS (2020)
18. Metzen, J.H., Yatsura, M.: Efficient certified defenses against patch attacks on image classifiers. In: ICLR (2021)
19. Mirman, M., Gehr, T., Vechev, M.: Differentiable abstract interpretation for provably robust neural networks. In: ICML (2018)
20. Naseer, M., Khan, S., Porikli, F.: Local gradients smoothing: Defense against localized adversarial attacks. In: WACV (2019)
21. Salman, H., Jain, S., Wong, E., Madry, A.: Certified patch robustness via smoothed vision transformers. In: CVPR (2022)

22. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. In: ICLR (2014)
23. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: ICML (2021)
24. Wei, C., Fan, H., Xie, S., Wu, C.Y., Yuille, A., Feichtenhofer, C.: Masked feature prediction for self-supervised visual pre-training. In: CVPR (2022)
25. Wu, Z., Lim, S.N., Davis, L.S., Goldstein, T.: Making an invisibility cloak: Real world adversarial attacks on object detectors. In: ECCV (2020)
26. Xiang, C., Bhagoji, A.N., Sehwag, V., Mittal, P.: Patchguard: A provably robust defense against adversarial patches via small receptive fields and masking. In: USENIX Security Symposium (2021)
27. Xiang, C., Mahloujifar, S., Mittal, P.: Patchcleanser: Certifiably robust defense against adversarial patches for any image classifier. In: USENIX Security Symposium (2022)
28. Xiang, C., Mittal, P.: Patchguard++: Efficient provable attack detection against adversarial patches. arXiv preprint arXiv:2104.12609 (2021)
29. Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L., Yuille, A.: Adversarial examples for semantic segmentation and object detection. In: ICCV (2017)
30. Yang, C., Kortylewski, A., Xie, C., Cao, Y., Yuille, A.: Patchattack: A black-box texture-based attack with reinforcement learning. In: ECCV (2020)
31. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenizer. In: ICLR (2022)