Towards Practical Certifiable Patch Defense with Vision Transformer

Zhaoyu Chen[1*] Bo Li[2*†] Jianghe Xu[2] Shuang Wu[2] Shouhong Ding[2] Wenqiang Zhang[1†]
[1]Academy for Engineering and Technology, Fudan University [2]Tencent Youtu Lab
{zhaoyuchen20,wqzhang}@fudan.edu.cn {libraboli,jankosxu,calvinwu,ericshding}@tencent.com

Abstract

Patch attacks, one of the most threatening forms of physical attack in adversarial examples, can lead networks to induce misclassification by modifying pixels arbitrarily in a continuous region. Certifiable patch defense can guarantee robustness that the classifier is not affected by patch attacks. Existing certifiable patch defenses sacrifice the clean accuracy of classifiers and only obtain a low certified accuracy on toy datasets. Furthermore, the clean and certified accuracy of these methods is still significantly lower than the accuracy of normal classification networks, which limits their application in practice. To move towards a practical certifiable patch defense, we introduce Vision Transformer (ViT) into the framework of Derandomized Smoothing (DS). Specifically, we propose a progressive smoothed image modeling task to train Vision Transformer, which can capture the more discriminable local context of an image while preserving the global semantic information. For efficient inference and deployment in the real world, we innovatively reconstruct the global self-attention structure of the original ViT into isolated band unit self-attention. On ImageNet, under 2% area patch attacks our method achieves 41.70% certified accuracy, a nearly 1-fold increase over the previous best method (26.00%). Simultaneously, our method achieves 78.58% clean accuracy, which is quite close to the normal ResNet-101 accuracy. Extensive experiments show that our method obtains state-of-the-art clean and certified accuracy with inferring efficiently on CIFAR-10 and ImageNet.

1. Introduction

Despite achieving outstanding performance on various computer vision tasks [18, 19, 22–26, 28, 34, 41, 42, 44, 45], deep neural networks (DNNs) are vulnerable and susceptive to adversarial examples [11, 14, 15], which are attached to a perturbation on images by adversaries [33]. Patch attack is one of the most threatening forms of adversarial examples, which can modify pixels arbitrarily in a continuous region, and can implement physical attacks on autonomous systems via their perception component. For example, putting stickers on traffic signals can make the model misprediction [10].

While several practical patch defenses are proposed [12, 30], they only obtain robustness against known attacks but not against more powerful attacks that may be developed in the future [5, 35]. Therefore, we focus on certifiable defense against patch attacks in this paper, which allows guaranteed robustness against all possible attacks for the given threat model. In recent years, this community has received great attention. For example, Chiang et al. propose the first certifiable defense against patch attacks by extending interval bound propagation (IBP) [5] on CIFAR10. Then later work introduces small receptive fields or randomized smoothing to improve certification on CIFAR10 and scale to ImageNet.

However, the accuracy gap between certifiable patch defense and normal model limits the practical application of these defense methods. For example, PatchGuard [39] can achieve 84.7% clean accuracy and 57.7% certified accuracy under 4×4 patches on CIFAR10. However, when PatchGuard [39] is extended to large-scale datasets, such as ImageNet, it can only obtain 54.6% clean accuracy and 26.0% certified accuracy under 2% patches, which is much lower than the normal ResNet-50 [13] (76.2%). Consequently, a breakthrough is needed urgently to narrow the gap and move towards practical certifiable patch defenses.

Recently, transformer [36] has achieved significant success in speech recognition and natural language processing. Inspired by this, Vision Transformer (ViT) [9] has been proposed and obtain potential performance in computer vision, such as image classification [1, 9], objects detection [4] and semantic segmentation [1]. ViT models the context between different patches and obtains long-range dependencies by self-attention. Compared with convolutional neural networks (CNNs), ViT has achieved promising performance, which has the potential to improve certification. Furthermore, Derandomize smoothing (DS) [21] is a classic certifiable patch defense based on randomized smoothing robustness schemes and provides high confident certified robustness by structured ablation. It can also be generalized to

*indicates equal contributions.
†indicates corresponding author.

other network architectures. Therefore, integrating ViT into DS is a potential certifiable patch defense. However, direct replacing the CNN structure in DS with ViT leads to trivial results: (1) the accuracy is still lower than normal classification networks; (2) the excessive inference time limits the application of the method in practice.

To address these issues and move towards practical certifiable patch defense, we propose an efficient certifiable patch defense with ViT to improve accuracy and inference efficiency. First, we introduce a progressive smoothed image modeling task to train ViT. Specifically, the training objective is to gradually recover the original image tokens based on the smoothed image bands. By gradually reconstructing, the base classifier can explicitly capture the local context of an image while preserving the global semantic information. Consequently, more discriminative local representations can be obtained through very limited image information (a thin smoothed image band), which improves the performance of the base classifier. Then, we renovate the global self-attention structure of the original ViT into the isolated band-unit self-attention. The input image is divided into bands and the self-attention in each band-like unit is calculated separately, which provides the feasibility for the parallel calculation of multiple bands. Finally, our method achieves 78.58% clean accuracy on ImageNet and 41.70% certified accuracy within efficient inference under 2% area patch attacks. The clean accuracy is quite close to the normal ResNet-101 accuracy. Extensive experiments demonstrate that our method obtains state-of-the-art clean and certified accuracy with inferring efficiently on CIFAR-10 and ImageNet. Our major contributions are as follows:

- We introduce ViT into certifiable patch defense and propose a progressive smoothed image modeling task, which lets the model capture more discriminable local context of an image while preserving the global semantic information.
- We renovate the global self-attention structure of the ViT into the isolated band-unit self-attention, which considerably accelerates inference.
- Experiments show that our method obtains state-of-the-art clean and certified accuracy with inferring efficiently on CIFAR-10 and ImageNet. Additionally, our method achieves 78.58% clean accuracy on ImageNet and 41.70% certified accuracy in efficient inference under 2% area patch attacks. The clean accuracy is quite close to the normal ResNet-101 accuracy.

2. Related Work

2.1. Patch Attacks

Patch attacks are one of the most threatening forms of physical attacks, in which adversaries can arbitrarily modify pixels within the small continuous region. GAP [3] first creates universal, robust, targeted adversarial image patches in the real world and causes a classifier to output any target class. Then LaVAN [17] shows that it is possible to learn visible and localized adversarial patches that cover only 2% of the pixels in the image and cause image classifiers to misclassify to arbitrary labels in the digital domain. Because patch attack can be implemented in the form of stickers in the physical world, it brings great harm to vision systems, such as object detection [16, 38] and visual tracking [8].

2.2. Certifiable Patch Defense

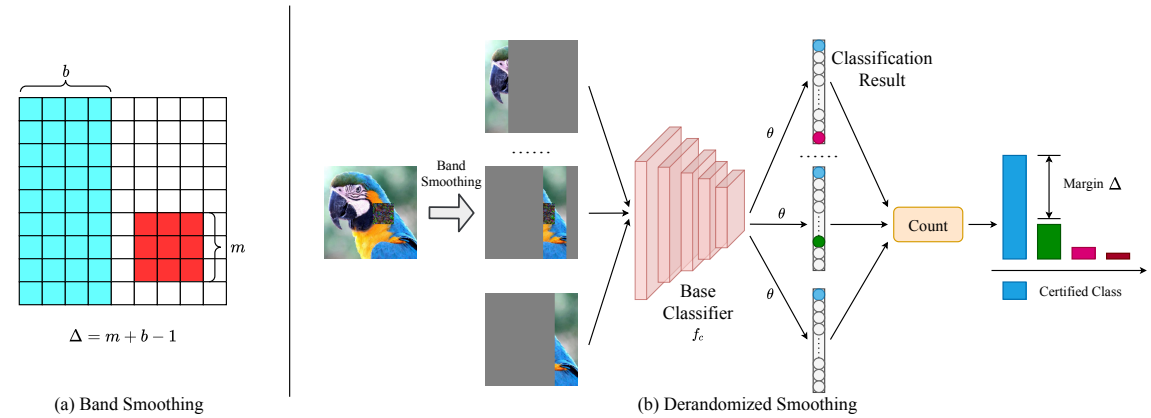
Several practical patch defenses were proposed such as digital watermark [12] and local gradient smoothing [30]. However, Chiang et al. [5] demonstrate that these defenses can be easily broken by white-box attacks which account for the pre-processing steps in the optimization procedure. It means that practical patch defenses only obtain robustness against known attacks but not against more powerful attacks that may be developed in the future.

Therefore, it is important to have guarantees of robustness in face of the worst-case adversarial patches. Recent work [27] focuses on certified defenses against patch attacks, which allow guaranteed robustness against all possible attacks for the given threat model. Chiang et al. [5] propose the first certifiable defense against patch attacks by extending interval bound propagation (IBP) on MNIST and CIFAR10. However, it is hard to scale to the ImageNet. Levine et al. propose Derandomized Smoothing (DS) [21], which trains a base classifier by smoothed images and a majority vote determines the final classification. This method provides significant accuracy improvement when compared to IBP on ImageNet but its inference is computationally expensive. Some work is based on using CNNs with the small receptive field, such as Clipped BagNet (CBN) [43], Patchguard [39] and BagCert [29].

2.3. Vision Transformer

Transformer [36] is the mainstream method in the natural language processing field, which captures long-range dependencies through self-attention and achieves state-of-the-art performance. Vision Transformer (ViT) [9] is the first work to achieve comparable results with traditional CNN architectures constructed only by self-attention blocks. It divides the image into a sequence of fixed-size patches and models the context between different patches and obtains long-range dependencies by multi-head self-attention.

The accuracy and certified robustness of the existing certifiable patch defenses are still not enough to be applied in practice. We introduce ViT into certifiable patch defense with the progressive smoothed image modeling task. With isolated band unit self-attention, our method achieves significant improvements in accuracy and inference efficiency,



(a) Band Smoothing (b) Derandomized Smoothing

Figure 1. Introduction of Derandomized Smoothing (DS). The red patch represents the adversarial patch and the blue band represents the retained image after the band smoothing in (a). (b) describes the pipeline of DS. First, DS smoothes the image in the band smoothing and obtains the smoothed images from different positions. Then the smoothed images are fed into the base classifier f_c and we obtain the classification result by the threshold θ. Finally, DS counts the result and applies Equation 2 to judge whether the image is certified.

Algorithm 1 Progressive Smoothed Image Modeling Task

Input: the image x, the label Y, the tokenizer Z, transformer encoder f, weighting factor λ, MLP head mlp and the number of stages N_s

Output: f, mlp

for $i \in [1,\ N_s]$ **do**
  Smooth images x and obtain smoothed images x_s
  Determine expected reconstructed images x_e
  Calculate visual tokens z with x via the tokenizer Z
  Calculate the output representation H_O with x_e via the transformer encoder f
  Calculate logits l with H_O via the MLP head mlp
  Select the reconstructed tokens H_R and corresponding visual tokens Z_R by Equation 3 and Equation 4
  Calculate the loss L by Equation 5 or Equation 6
  Update f and mlp through L backward
end for
return f, mlp

which enables practical certifiable patch defense.

3. Method

In this section, we first review the certified mechanism of DS. Second, we propose a progressive smoothed image modeling task to help ViT capture the more discriminable local context of an image while preserving the global semantic information. Finally, we propose the isolated band unit self-attention to accelerate inference and move towards practical certifiable patch defense.

3.1. Preliminaries

Smoothing in derandomized smoothing means to keep a part of the continuous image and smooth other parts of the image. For example, band smoothing means smoothing the entire image except for a band of a fixed-width b, as shown in Figure 1 (a). DS trains the base classifier with smoothed images. For an input image $x \in R^{c\times h\times w}$, let the base classifier be expressed as $f_c(x, b, p, \theta)$, where x is the input images, b is the width of the band, p is the position of the retained band, θ is the threshold for voting and c is the class label. For each class c, $f_c(x, b, p, \theta)$ is 1 if its logits of the class c is greater than the threshold θ, otherwise it is 0. To calculate certified robustness, DS counts the number of bands on which the base classifier is applied to each class.

$$\forall c, \quad n_c(p) = \sum_{p=1}^{w} f_c(x, b, p, \theta). \tag{1}$$

An image is certified only if the statistics of the highest class (e.g. the label c) are greater than a margin than the next highest class c'. The shape of the adversarial patch is supposed to be $m \times m$. The number of intersections between the band and this patch is at most $\Delta = m + b - 1$. Therefore, an image is certified when Δ satisfies the following conditions:

$$n_c(x) > \max_{c' \neq c} n_{c'}(x) + 2\Delta. \tag{2}$$

When the threshold θ is determined, the highest class has been guaranteed not to be affected by the adversarial patch. Therefore, we define **clean accuracy** as the accuracy that the classification is correct after voting. **Certified accuracy** is the accuracy that the classification is correct and Δ satisfies Equation 2 after voting.

3.2. Progressive Smoothed Image Modeling

Since the base classifier can only use very limited information (such as bands), we need the base classifier to have

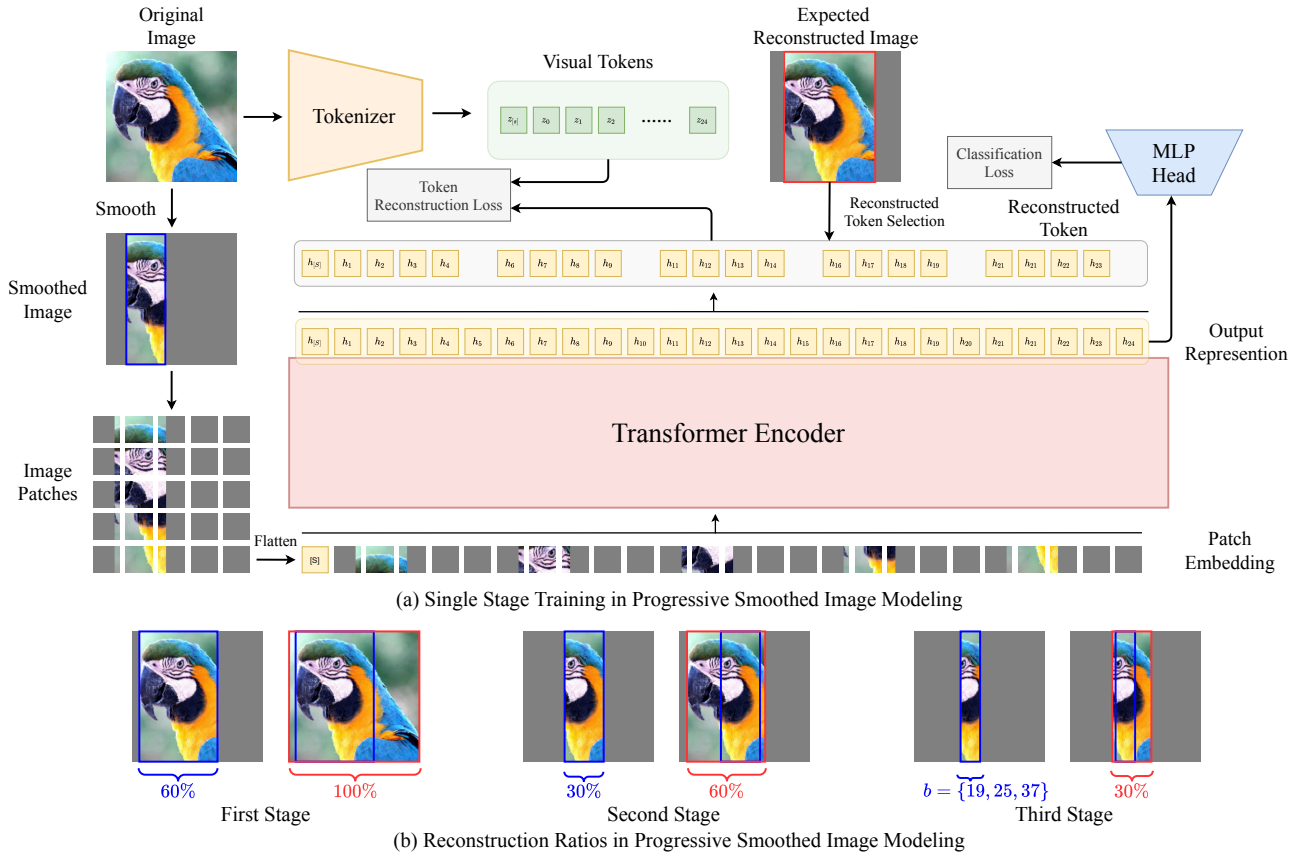


Figure 2. Introduction of Progressive Smoothed Image Modeling. (a) describes a single stage smooth training in progressive smoothed image modeling. We expect smoothed images to be reconstructed as expected reconstruction images. (b) describes reconstructed ratios in multi-stage training. The blue boxes represent the band after smoothing and the red boxes represent the expected reconstructed band.

the ability to better capture discriminable features. Specifically, in natural language processing, the masked language modeling (MLM) training paradigm (like BERT [7]) has been proved to be effective in learning more discriminative features and improving the performance of the model. Inspired by MLM, we propose a smoothed image modeling task to train ViT.

However, unlike languages which are human-created signals with dense semantic correlations between words, as natural signals, the visual content of different parts in an image has a high degree of freedom. Therefore, it is very difficult to use a band with a width of b ($b \ll h, w$) to recover the full-scale image tokens in one stage. Hence, we use a multistage smoothed image modeling task called progressive smoothed image modeling to train the base classifier, as shown in Figure 2. By gradually reconstructing the smoothed image parts, the base classifier can explicitly capture the local context of an image while preserving the global semantic information. Consequently, more discriminative local representations can be obtained through very limited image information, which improves the performance of the base classifier.

In ViT, an image is spilt into a sequence of patches as inputs. Formally, we need to flatten a image $x \in R^{c \times h \times w}$ into $(N = hw/p^2)$ patches $x^p \in R^{N \times p^2 c}$, where the shape of the image x is (h, w), the number of channels is c, and (p, p) is the shape of patches (e.g. $p = 16$). The patches $\{x_i^p\}_{i=1}^N$ are projected to obtain the patch embeddings $\{Ex_i^p\}_{i=1}^N$, where $E \in R^{p^2 c \times d}$ and d is the embedding dimension. Like Bert [7], we concatenate class token $E_{[s]}$ to patch embeddings Ex_i^p. Simultaneously, in order to encode position information, we need to add 1D learnable position embeddings E_{pos} to patch embeddings Ex_i^p. Then, the input vector $H_I = [E_{[s]}, Ex_i^p, ..., Ex_N^p] + E_{pos}$ is fed to transformer and $H_O = [h_{[s]}, h_i, ..., h_N]$ is used as the output representation for the image patches with respect to x.

Here, we first introduce single stage training in progressive smoothed image modeling, as illustrated in Figure 2 (a). BERT-based training has been explored in vision tasks. The difficulty is that it is non-trivial to recover tokens in computer vision. For accelerating convergence, we introduce a tokenizer as the supervision of reconstruction. There are two types of supervision for reconstruction: VAE and distillation. For VAE, we use pre-trained VAE [31] for su-

pervision. For distillation, we use the output of pre-trained ViT [9] for supervision. As given a smoothed image x_s, we split it into N image patches $\{x_{s_i}^p\}_{i=1}^N$ and obtain N visual tokens $\{z_i\}_{i=1}^N$. Centered on the band of x_s, we select the reconstructed band and generate the expected reconstructed image. According to the expected reconstructed image, we have a reconstructed token selection to obtain reconstructed tokens. The band in the expected reconstructed image produces a band mask $\{M_i^b\}_{i=1}^N$ corresponding to the patch x_i^p which needs constructing. Hence, the reconstructed tokens and corresponding visual tokens are rephrased as:

$$H_R = \{h_i : M_i^b = 1\}_{i=1}^N. \tag{3}$$

$$Z_R = \{z_i : M_i^b = 1\}_{i=1}^N. \tag{4}$$

The objective of Smoothed Training is to simultaneously minimize the classification loss and token reconstruction loss. For VAE, the total loss can be expressed as:

$$\min \underbrace{CE(l,\, Y)}_{\text{classification loss}} + \lambda \cdot \underbrace{CE(Z_R,\, H_R)}_{\text{token reconstruction loss}} . \tag{5}$$

For distillation, the total loss can be expressed as:

$$\min \underbrace{CE(l,\, Y)}_{\text{classification loss}} + \lambda \cdot \underbrace{||Z_R - H_R||_2}_{\text{token reconstruction loss}} . \tag{6}$$

Here, l is the output logits after passing MLP head and Y is the label of x. $\lambda = 1000$ balances the gradients between token reconstruction loss and classification loss.

Figure 2 (b) shows the reconstruction ratio varies within each stage. The blue boxes represent the band after smoothing and the red boxes represent the expected reconstructed band. In the first stage, we randomly smooth the approximately 40% image. The remaining 60% of the images are used to reconstruct the whole image. In the second stage, we smooth away 70% of the images and utilize the remaining 30% of the bands, to reconstruct 60% of the patches within a neighborhood centered on the 30% band, including the 30% band. In the last stage, only the band with width b is reserved, and all other parts are smoothed. The band with width b is used to reconstruct 30% of the patches in the neighborhood centered on the band. Our method greatly narrows the accuracy gap between DS and normal model by progressive smoothed image modeling, making it possible to achieve certifiable patch defense in practice.

3.3. Isolated Band Unit Self-attention

To achieve practical certifiable patch defense, high accuracy needs to be combined with efficient inference. With progressive smoothed image modeling, DS improves the clean and certified accuracy, however, it requires hundreds of inferences for various smoothed images, which limits its application in practice.

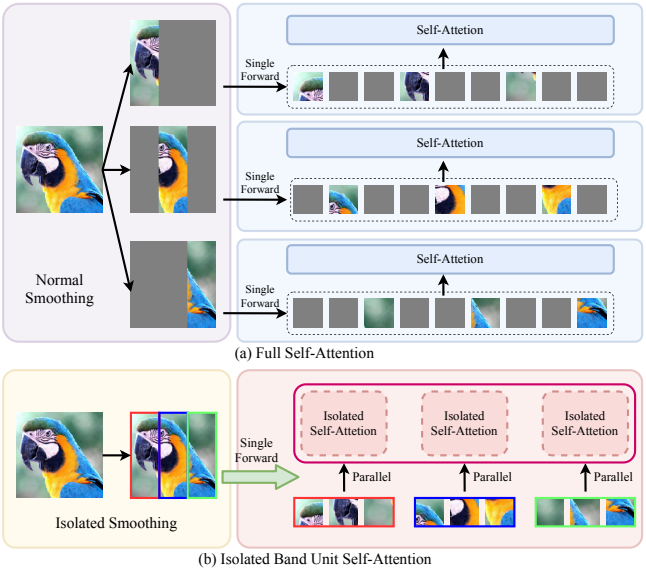


Figure 3. Introduction of Isolated Band Unit Self-attention. (a) describes the normal training that smoothed parts are redundant and unnecessary to calculate. (b) introduces the isolated band unit self-attention that smoothed parts are dropped and self-attention is only calculated within parallel windows.

Smoothed parts of the smoothed image introduce redundant information and invalid calculation. As shown in Figure 3 (a), normal smoothing utilizes the whole smoothed image but its calculation is unnecessary for smoothed parts. The long-range dependencies of ViT use this redundant information, which harms their accuracy and introduces extra calculation cost. Furthermore, it is inefficient to calculate one forward calculation for every smoothed image.

Therefore, we innovatively renovate the global self-attention structure of the original ViT into isolated band unit self-attention. Specifically, the input image is divided into bands by sliding windows, and the self-attention in each band-like unit is calculated separately, which provides the feasibility for the parallel calculation of multiple bands. As shown in Figure 3 (b), we choose patches of each band by parallel sliding windows and infer multiple bands within one forward calculation. In the isolated band unit self-attention, a window is a band and isolated self-attention is only calculated within the window.

Compared to normal smoothing, ViT splits x_s into hw/p^2 patches and we only select the $N = hb/p^2$ patches within windows for fine-tuning and inferring. The time complexity of the whole image input of self-attention operation is $O\left(N^2d + Nd^2\right)$, where the first term is the complexity of attention operation and another is the complexity of fully-connected operation. Compared with the input of the whole image, the isolated self-attention can reduce the

Table 1. The parameters of models on ImageNet.

Model	BagNet33	ViT-s	ResNet50	ResNext101	ViT-B
Parameters	18M	22M	26M	88.79M	86M

Table 2. Clean and certified accuracy compared with state-of-the-art certifiable patch defenses on CIFAR10.

Method		Clean Accuracy (%)	Certified Accuracy (%)	
			2×2	4×4
Baseline	CBN	84.20	44.20	9.30
	DS	83.90	68.90	56.20
	PG	84.70	69.20	57.70
	BagCert	86.00	73.33	64.90
Smooth Model	ViT-S	80.40	61.50	51.78
	ECViT-S	87.56	73.82	65.10
	ResNext101	85.34	69.32	60.68
	ViT-B	91.28	78.10	70.78
	ECViT-B	**93.48**	**82.80**	**76.38**

Table 3. Clean and certified accuracy compared with state-of-the-art certifiable patch defenses on ILSVRC2012.

Method	Clean Accuracy (%)	Certified Accuracy (%)		
		1% pixels	2% pixels	3% pixels
CBN	49.50	13.40	7.10	3.10
PG (1%)	55.10	32.30	-	-
PG (2%)	54.60	26.00	26.00	-
PG (3%)	54.10	19.70	19.70	19.70
DS	64.67	30.14	24.70	20.88
BagCert	46.00	-	23.00	-
ECViT-S(b=37)	69.88	35.03	29.74	25.74
ECViT-B(b=37)	**78.58**	**47.39**	**41.70**	**37.26**

amount of calculation to $\frac{1}{w/b}$ of the former. Furthermore, it has $\lceil \frac{w}{b} \rceil$ adjacent windows for inference at the same time, where the shape of windows is (h, b). Therefore, we change the original forward calculation from w times to b $(w \gg b)$ times. Therefore, it is possible to deploy a certifiable patch defense in real systems through efficient inference.

4. Experiments

We conduct extensive experiments on CIFAR10 [20] and ImageNet [6]. First, we compare our method with the state-of-the-art certifiable patch defenses on both datasets. In addition, we conduct ablation studies to investigate the factors that affect clean and certified accuracy.

4.1. Experimental Setup

In our experiments, we use Pytorch for the implementation and train on NVIDIA Tesla V100 GPUs. We

Table 4. Clean and certified accuracy compared with smooth models on ILSVRC2012.

Method	Clean Accuracy (%)	Certified Accuracy (%)			Inference Time (s)
		1% pixels	2% pixels	3% pixels	
ResNet50(b=19)	62.03	29.03	23.53	19.77	69.00
ResNet50(b=25)	64.67	30.14	24.70	20.88	
ResNet50(b=37)	67.60	27.15	21.86	18.14	
ViT-S(b=19)	63.88	33.08	27.78	23.84	87.13
ViT-S(b=25)	66.49	33.90	28.59	24.57	
ViT-S(b=37)	69.01	33.37	28.07	24.15	
ECViT-S(b=19)	64.69	34.38	28.85	24.74	9.66
ECViT-S(b=25)	67.14	**35.57**	**30.06**	**25.98**	13.20
ECViT-S(b=37)	**69.88**	35.03	29.74	25.74	23.54
ResNext101(b=19)	69.36	40.74	34.97	30.58	567.75
ResNext101(b=25)	71.96	41.86	36.03	32.17	
ResNext101(b=37)	74.89	42.79	36.69	33.24	
ViT-B(b=19)	66.92	34.65	28.71	24.80	136.75
ViT-B(b=25)	70.57	36.72	31.13	26.80	
ViT-B(b=37)	74.68	37.61	31.88	27.44	
ECViT-B(b=19)	73.49	46.83	40.72	36.29	16.63
ECViT-B(b=25)	75.30	46.56	40.79	36.21	22.72
ECViT-B(b=37)	**78.58**	**47.39**	**41.70**	**37.26**	40.50

choose different networks as base classifiers, such as ResNet50 [13], ResNext101-32x8d (ResNext101) [40], ViT-S/16-224 (ViT-S) and ViT-B/16-224 (ViT-B) [9]. In the smooth model, the methods directly apply the corresponding backbone into DS and are all fine-tuned from the ImageNet pre-trained model [37]. The parameters of models on ImageNet are shown in Table 1. We train the models in 120 epochs on ImageNet and 600 epochs on CIFAR10.

The parameters of our Efficient Certifiable ViT (ECViT) is the same as the ViT. We train ECViT from the same ViT pre-trained models. For exmaple, ECViT-B is based on ViT-B and ECViT-S is based on ViT-S. For CIFAR10, we train ECViT 150 epochs for each stage during the progressive smoothed image modeling task and fine-tune 150 epochs in isolated band unit self-attention. For ImageNet, we train ECViT 30 epochs for each stage during the progressive smoothed image modeling task and fine-tune 30 epochs in isolated band unit self-attention.

We report clean and certified accuracy compared ECViT with Interval Bound Propagation (IBP) [5], Derandomized Smoothing (DS) [21], Clipped BagNet (CBN) [43], Patchguard [39] (PG) and BagCert [29]. Here, DS is based on the band smoothing and Patchguard is based on the mask BagNet [2]. For each stage and fine-tuning in the progressive smoothed image modeling or smoothed models, we set the optimizer to AdamW, the loss function to be a cross-entropy loss, the batch size to 512, the warm-up epoch to 5, the learning rate to be 2e-5, the threshold θ to 0.2, and the weight decay to be 1e-8. ECViT is trained on CIFAR10 and ImageNet for a total of 600 and 120 epochs, consistent with

Table 5. Ablation study of training for different stages and tokenizers on ILSVRC2012 .

Network	Band Size	Stages	Clean Accuracy (%)	Certified Accuracy (%)		
				1%	2%	3%
Distllation	b=19	one_stage	72.01	43.49	37.64	33.31
		two_stage	72.78	44.87	39.01	34.63
		three_stage	72.93	45.78	40.03	35.61
	b=25	one_stage	74.48	44.14	38.39	33.87
		two_stage	75.09	45.63	39.91	35.54
		three_stage	75.12	46.63	40.93	36.49
	b=37	one_stage	78.05	44.89	39.24	34.60
		two_stage	77.98	45.54	39.94	35.45
		three_stage	78.05	46.46	40.84	36.39
VAE	b=19	one_stage	72.60	43.91	38.00	33.60
		two_stage	73.30	45.50	39.69	35.12
		three_stage	73.49	46.83	40.72	36.29
	b=25	one_stage	75.15	45.50	39.56	35.18
		two_stage	75.48	46.34	40.49	35.99
		three_stage	75.30	46.56	40.79	36.21
	b=37	one_stage	77.95	44.49	38.78	34.28
		two_stage	78.40	46.53	40.73	36.36
		three_stage	78.58	47.36	41.70	37.26

smoothed models.

4.2. Certification on CIFAR10

Following the setting of the previous work [21], we evaluate clean and certified accuracy on 5,000 images of the CIFAR10 validation set. We select two patch sizes, including 2×2 and 4×4. In the experiment, the images are all up-sampled from 32 to 224, and for the band smoothing, the band size b is fixed to 4 on the original size of 32. Table 2 shows clean and certified accuracy against patches of different sizes. Experiments show that ECViT can effectively improve smoothed ViT and achieve state-of-the-art accuracy. For ViT-S, under the same structure, ECViT-S improves clean accuracy by 7.16% and certified accuracy by $\sim$13%. Our best ECViT-B still has 82.80% certified accuracy under 2×2 patches with 93.48% clean accuracy.

4.3. Certification on ImageNet

We evaluate our proposed ECViT on ILSVRC2012 validation set [32]. ILSVRC2012 validation set has 50,000 images and we test clean accuracy and certified accuracy on the whole 50,000 images. Following the setting of previous work [21, 39], we select three different patch sizes, including 1% (23×23), 2% (32×32) and 3% (39×39). Table 3 shows clean and certified accuracy compared state-of-the-art methods on ILSVRC2012. ECViT-B surpasses BagCert's clean and certified accuracy by 7.48% and 9.47% and achieves state-of-the-art clean and certified accuracy in comparison to the previous state-of-the-art methods.

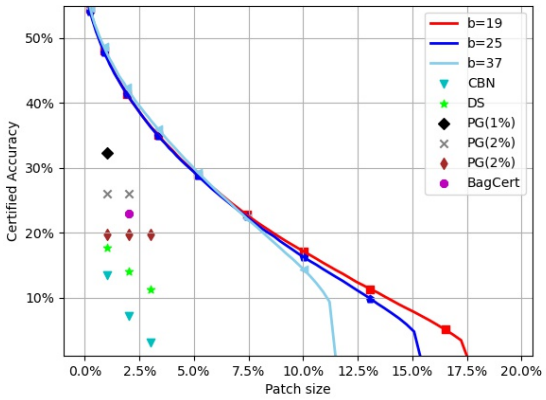


Figure 4. Certified accuracy under different patch sizes on ILSVRC2012. Patch size represents the proportion of adversarial patch in the image.

Table 4 shows clean and certified accuracy compared other smooth models on ILSVRC2012. Smooth models compared to other than ECViT are based on different backbones in DS. Inference time is calculated when the batch size of the image is 1024 to complete the vote in seconds. To simulate the inference in practice, inference time contains the time of the complete process, including data pre-processing and inference. In the case of band size $b = 37$, our ECViT-B improves the clean accuracy by 3.9% compared to ViT-B. The certified accuracy is improved by about 10% in different patch sizes, and it speeds up 4-8x times for ViT-B in the inference time. In the smoothing model, we achieve the fastest inference efficiency. Our method achieves state-of-the-art certification on ILSVRC2012 while maintaining 78.58% clean accuracy, which is very close to the normal ResNet-101.

4.4. Ablation Study

In this section, we mainly focus on studying the effect of the number of stages, different patch sizes, and tokenizers on the clean and certified accuracy.

Number of Stages. To verify the effectiveness of multi-stage progressive smoothed image modeling, we study the effect of different stages on accuracy. Specifically, taking two_stage as an example, two_stage represents the direct fine-tuning after the progressive smoothed image modeling of the first two stages. Among them, the total training epoch is the same. The meanings of one_stage and $three\_stage$ are similar with two_stage. Table 5 shows the training of different stages and tokenizers on ILSVRC2012 validation set for ECViT-B. Table 6 reflects the different stages of ablation experiments on CIFAR10. Clean and certified accuracy basically increases with the increase of training stages. Experiments show that the progressive smoothed image mod-

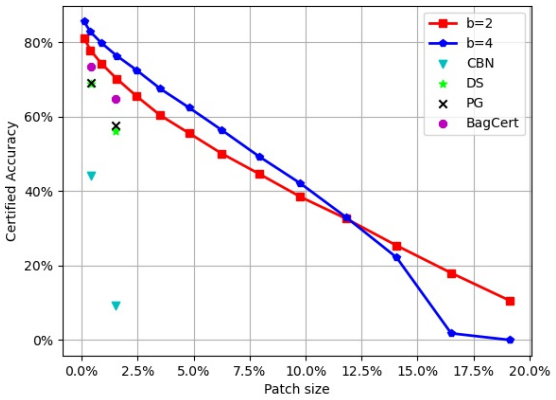


Figure 5. Certified accuracy under different patch sizes on CIFAR10. Patch size represents the proportion of adversarial patch in the image.

Table 6. Ablation study of training for different stages on CIFAR10. Clean and certified accuracy basically increases with the increase of training stages.

Networks	Band Size	Stages	Clean Accuracy (%)	Certified Accuracy (%)	
				2×2	4×4
ECViT-S	b=2	one_stage	77.94	66.36	57.86
		two_stage	77.84	66.56	56.80
		three_stage	**80.50**	**69.06**	**59.58**
	b=4	one_stage	86.56	71.26	62.60
		two_stage	86.30	71.36	62.38
		three_stage	**87.56**	**73.82**	**65.10**
ECViT-B	b=2	one_stage	85.60	76.42	68.66
		two_stage	86.06	77.20	69.06
		three_stage	**86.40**	**77.86**	**70.20**
	b=4	one_stage	92.46	81.54	74.56
		two_stage	93.14	82.40	75.98
		three_stage	**93.48**	**82.80**	**76.38**

eling task allows the base classifier to explicitly capture the local context of an image while preserving the global semantic information. Consequently, more discriminative local representations can be obtained through a very limited image information (a smoothed thin image bands), which improves the performance of the base classifier.

Patch Sizes. The size of adversarial patches will greatly affect the certification. Figure 4 and Figure 5 respectively reflect the certified accuracy variation on ImageNet and CIFAR10 for ECViT-B. The certified accuracy decreases as the patch size becomes larger. When the patch size reaches 10%, ECViT-B still has a certified accuracy of $\sim 17.00\%$ and $\sim 39.00\%$ on ImageNet and CIFAR10.

Tokenizer. In order to verify that progressive smoothed image modeling is also effective for other tokenizers, we conduct the following experiments on ECViT-B. Table 7 illustrates the comparison of different tokenizers and other network architectures. We can see that compared to ViT-B, the distilled tokenizer has a significant improvement, but it is still lower than the VAE tokenizer. This also verifies that our method can be adapted to different tokenizers.

Table 7. Clean and certified accuracy compared with other tokenizers on ILSVRC2012. The VAE is better than the distilled.

Band size	Smooth Model	Clean Accuracy (%)	Certified Accuracy (%)		
			1% pixels	2% pixels	3% pixels
b=19	ViT-B(b=19)	66.92	34.65	28.71	24.80
	Ours(distilled)	72.93	45.78	40.03	35.61
	Ours(vae)	**73.49**	**46.83**	**40.72**	**36.29**
b=25	ViT-B(b=25)	70.57	36.72	31.13	26.80
	Ours(distilled)	75.12	46.63	40.93	36.49
	Ours(vae)	**75.30**	**46.56**	**40.79**	**36.21**
b=37	ViT-B(b=37)	74.68	37.61	31.88	27.44
	Ours(distilled)	78.05	46.46	40.84	36.39
	Ours(vae)	**78.58**	**47.39**	**41.70**	**37.26**

5. Conclusions

Certifiable patch defenses allow guaranteed robustness against all possible attacks for the given threat model. Existing certifiable patch defenses sacrifice the clean accuracy of classifiers and only obtain a low certified accuracy on toy datasets, which limits their applications in practice. To move towards practical certifiable patch defense, we introduce ViT into the framework of DS. With the progressive smoothed image modeling and isolated band unit self-attention, our ECViT obtains state-of-the-art accuracy and efficient inference efficiency on CIFAR-10 and ImageNet. In our work, we hope to provide a new perspective on how a suitable network architecture can improve the upper certification of patch defenses in practice.

Limitations and broader impacts. In spite of the limitation that adversarial patches are assumed to be square in shape rather than rectangular or irregular in shape, we believe that ECViT introduces a practical certifiable patch defense in order to assist physical systems in mitigating the threat of patch attacks.

Acknowledgements

This work was supported by National Natural Science Foundation of China (No.62072112), National Key R&D Program of China (2020AAA0108301), Scientific and Technological Innovation Action Plan of Shanghai Science and Technology Committee (No.20511103102), Fudan University-CIOMP Joint Fund (No. FC2019-005), Double First-class Construction Fund (No. XM03211178).

References

[1] Hangbo Bao, Li Dong, and Furu Wei. Beit: BERT pre-training of image transformers. *CoRR*, abs/2106.08254, 2021. 1

[2] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 6

[3] Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch, 2017. 2

[4] Zhiyang Chen, Yousong Zhu, Chaoyang Zhao, Guosheng Hu, Wei Zeng, Jinqiao Wang, and Ming Tang. DPT: deformable patch-based transformer for visual recognition. *CoRR*, abs/2107.14467, 2021. 1

[5] Ping-Yeh Chiang, Renkun Ni, Ahmed Abdelkader, Chen Zhu, Christoph Studer, and Tom Goldstein. Certified defenses for adversarial patches. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 1, 2, 6

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009. 6

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. 4

[8] Li Ding, Yongwei Wang, Kaiwen Yuan, Minyang Jiang, Ping Wang, Hua Huang, and Z. Jane Wang. Towards universal physical attacks on single object tracking. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 1236–1245. AAAI Press, 2021. 2

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 1, 2, 5, 6

[10] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1625–1634. IEEE Computer Society, 2018. 1

[11] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 1

[12] Jamie Hayes. On visible adversarial perturbations & digital watermarking. In *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1597–1604. Computer Vision Foundation / IEEE Computer Society, 2018. 1, 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 1, 6

[14] Hao Huang, Yongtao Wang, Zhaoyu Chen, Yuheng Li, Zhi Tang, Wei Chu, Jingdong Chen, Weisi Lin, and Kai-Kuang Ma. Cmua-watermark: A cross-model universal adversarial watermark for combating deepfakes. *CoRR*, abs/2105.10872, 2021. 1

[15] Hao Huang, Yongtao Wang, Zhaoyu Chen, Zhi Tang, Wenqiang Zhang, and Kai-Kuang Ma. Rpattack: Refined patch attack on general object detectors. In *2021 IEEE International Conference on Multimedia and Expo, ICME 2021, Shenzhen, China, July 5-9, 2021*, pages 1–6. IEEE, 2021. 1

[16] Lifeng Huang, Chengying Gao, Yuyin Zhou, Cihang Xie, Alan L. Yuille, Changqing Zou, and Ning Liu. Universal physical camouflage attacks on object detectors. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 717–726. Computer Vision Foundation / IEEE, 2020. 2

[17] Danny Karmon, Daniel Zoran, and Yoav Goldberg. Lavan: Localized and visible adversarial noise. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2512–2520. PMLR, 2018. 2

[18] Xiangtao Kong, Xina Liu, Jinjin Gu, Yu Qiao, and Chao Dong. Reflash dropout in image super-resolution. *arXiv preprint arXiv:2112.12089*, 2021. 1

[19] Xiangtao Kong, Hengyuan Zhao, Yu Qiao, and Chao Dong. Classsr: A general framework to accelerate super-resolution networks by data characteristic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12016–12025, June 2021. 1

[20] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images, 2009. 6

[21] Alexander Levine and Soheil Feizi. (de)randomized smoothing for certifiable defense against patch attacks. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in*

Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020. 1, 2, 6, 7

[22] Bo Li, Zhengxing Sun, and Yuqi Guo. Supervae: Superpixelwise variational autoencoder for salient object detection. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 8569–8576. AAAI Press, 2019. 1

[23] Bo Li, Zhengxing Sun, Qian Li, Yunjie Wu, and Anqi Hu. Group-wise deep object co-segmentation with co-attention recurrent neural network. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 8518–8527. IEEE, 2019. 1

[24] Bo Li, Zhengxing Sun, Lv Tang, and Anqi Hu. Two-breal net: Two-branch network for real-time salient object detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pages 1662–1666. IEEE, 2019. 1

[25] Bo Li, Zhengxing Sun, Lv Tang, Yunhan Sun, and Jinlong Shi. Detecting robust co-saliency with recurrent co-attention neural network. In Sarit Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 818–825. ijcai.org, 2019. 1

[26] Bo Li, Zhengxing Sun, Quan Wang, and Qian Li. Co-saliency detection based on hierarchical consistency. In Laurent Amsaleg, Benoit Huet, Martha A. Larson, Guillaume Gravier, Hayley Hung, Chong-Wah Ngo, and Wei Tsang Ooi, editors, *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, pages 1392–1400. ACM, 2019. 1

[27] Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, Jilin Li, and Feiyue Huang. Detecting adversarial patch attacks through global-local consistency. In Dawn Song, Dacheng Tao, Alan L. Yuille, Anima Anandkumar, Aishan Liu, Xinyun Chen, Yingwei Li, Chaowei Xiao, Xun Yang, and Xianglong Liu, editors, *ADVM '21: Proceedings of the 1st International Workshop on Adversarial Learning for Multimedia, Virtual Event, China, 20 October 2021*, pages 35–41. ACM, 2021. 2

[28] Siao Liu, Zhaoyu Chen, Wei Li, Jiwei Zhu, Jiafeng Wang, Wenqiang Zhang, and Zhongxue Gan. Efficient universal shuffle attack for visual object tracking. *arXiv preprint arXiv:2203.06898*, 2022. 1

[29] Jan Hendrik Metzen and Maksym Yatsura. Efficient certified defenses against patch attacks on image classifiers. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 2, 6

[30] Muzammal Naseer, Salman Khan, and Fatih Porikli. Local gradients smoothing: Defense against localized adversarial attacks. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2019, Waikoloa Village, HI, USA, January 7-11, 2019*, pages 1300–1307. IEEE, 2019. 1, 2

[31] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021. 4

[32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, 2015. 7

[33] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. 1

[34] Lv Tang, Bo Li, Yijie Zhong, Shouhong Ding, and Mofei Song. Disentangled high quality salient object detection. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 3560–3570. IEEE, 2021. 1

[35] Florian Tramèr, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 1

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. 1, 2

[37] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019. 6

[38] Zuxuan Wu, Ser-Nam Lim, Larry S. Davis, and Tom Goldstein. Making an invisibility cloak: Real world adversarial attacks on object detectors. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IV*, volume 12349 of *Lecture Notes in Computer Science*, pages 1–17. Springer, 2020. 2

[39] Chong Xiang, Arjun Nitin Bhagoji, Vikash Sehwag, and Prateek Mittal. Patchguard: Provable defense against adversarial patches using masks on small receptive fields. *CoRR*, abs/2005.10884, 2020. 1, 2, 6, 7

[40] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Hon-*

olulu, HI, USA, July 21-26, 2017, pages 5987–5995. IEEE Computer Society, 2017. 6

[41] Jie Zhang, Chen Chen, Bo Li, Lingjuan Lyu, Shuang Wu, Jianghe Xu, Shouhong Ding, and Chao Wu. A practical data-free approach to one-shot federated learning with heterogeneity. *arXiv preprint arXiv:2112.12371*, 2021. 1

[42] Jie Zhang, Lei Zhang, Gang Li, and Chao Wu. Adversarial examples for good: Adversarial examples guided imbalanced learning. *arXiv preprint arXiv:2201.12356*, 2022. 1

[43] Zhanyuan Zhang, Benson Yuan, Michael McCoyd, and David A. Wagner. Clipped bagnet: Defending against sticker attacks with clipped bag-of-features. In *2020 IEEE Security and Privacy Workshops, SP Workshops, San Francisco, CA, USA, May 21, 2020*, pages 55–61. IEEE, 2020. 2, 6

[44] Hengyuan Zhao, Xiangtao Kong, Jingwen He, Yu Qiao, and Chao Dong. Efficient image super-resolution using pixel attention. In *European Conference on Computer Vision*, pages 56–72. Springer, 2020. 1

[45] Yijie Zhong, Bo Li, Lv Tang, Hao Tang, and Shouhong Ding. Highly efficient natural image matting. *CoRR*, abs/2110.12748, 2021. 1