Certified Patch Robustness via Smoothed Vision Transformers

Hadi Salman* MIT hady@mit.edu
Saachi Jain* MIT saachij@mit.edu
Eric Wong* MIT wongeric@mit.edu
Aleksander Mądry MIT madry@mit.edu

Abstract

Certified patch defenses can guarantee robustness of an image classifier to arbitrary changes within a bounded contiguous region. But, currently, this robustness comes at a cost of degraded standard accuracies and slower inference times. We demonstrate how using vision transformers enables significantly better certified patch robustness that is also more computationally efficient and does not incur a substantial drop in standard accuracy. These improvements stem from the inherent ability of the vision transformer to gracefully handle largely masked images.[1]

1. Introduction

High-stakes scenarios warrant the development of certifiably robust models that are *guaranteed* to be robust to a set of transformations. These techniques are beginning to find applications in real-world settings, such as verifying that aircraft controllers behave safely in the presence of approaching airplanes [19], and ensuring the stability of automotive systems to sensor noise [54].

We study robustness in the context of adversarial patches—a broad class of arbitrary changes contained within a small, contiguous region. Adversarial patches capture the essence of a range of maliciously designed physical objects such as adversarial glasses [44], stickers/graffiti [12], and clothing [55]. Researchers have used adversarial patches to fool image classifiers [4], manipulate object detectors [18, 24], and disrupt optical flow estimation [38].

Adversarial patch defenses can be tricky to evaluate—recent work broke several empirical defenses [1, 16, 35] with stronger adaptive attacks [6, 48]. This motivated *certified* defenses, which deliver provably robust models without having to rely on an empirical evaluation. However, certified guarantees tend to be modest and come at a cost: poor standard accuracy and slower inference times [25, 26, 56, 63]. For example, a top-performing, recently proposed method reduces standard accuracy by 30% and increases inference time by two orders of magnitude, while certifying only 13.9% robust accuracy on ImageNet against patches that take up 2% of the image [25]. These drawbacks are commonly accepted as the cost of certification, but severely limit the applicability of certified defenses. Does certified robustness really need to come at such a high price?

Our contributions

In this paper, we demonstrate how to leverage vision transformers (ViTs) [10] to create certified patch defenses that achieve significantly higher robustness guarantees than prior work. Moreover, we show that certified patch defenses with ViTs can actually maintain standard accuracy and inference times comparable to standard (non-robust) models. At its core, our methodology exploits the token-based nature of attention modules used in ViTs to gracefully handle the ablated images used in certified patch defenses. Specifically, we demonstrate the following:

Improved guarantees via smoothed vision transformers. We find that using ViTs as the backbone of the derandomized smoothing defense [25] enables significantly improved certified patch robustness. Indeed, this change alone boosts certified accuracy by up to 13% on ImageNet, and 5% on CIFAR-10 over similarly sized ResNets.

Standard accuracy comparable to that of standard architecures. We demonstrate that ViTs enable certified defenses with standard accuracies comparable to that of standard, non-robust models. In particular, our largest ViT improves state-of-the-art certified robustness on ImageNet while maintaining standard accuracy that is similar to that of a non-robust ResNet (>70%).

Faster inference. We modify the ViT architecture to drop unnecessary tokens, and reduce the smoothing process to

*Equal contribution.

[1] Our code is available at https://github.com/MadryLab/smoothed-vit.

pass over mostly redundant computation. These changes turn out to vastly speed up inference time for our smoothed ViTs. In our framework, a forward pass on ImageNet becomes up to two orders of magnitude faster than that of prior certified defenses, and is close in speed to a standard (non-robust) ResNet.

2. Certified patch defense with smoothing & transformers

Smoothing methods are a general class of certified defenses that combine the predictions of a classifier over many variations of an input to create predictions that are certifiably robust [7, 26]. One such method that obtains robustness to adversarial patches is derandomized smoothing [25], which aggregates a classifier's predictions on various *image ablations* that mask most of the image out.

These approaches typically use CNNs, a common default model for computer vision tasks, to evaluate the image ablations. The starting point of our approach is to ask: are convolutional architectures the right tool for this task? The crux of our methodology is to leverage vision transformers, which we demonstrate are more capable of gracefully handling the image ablations that arise in derandomized smoothing.

2.1. Preliminaries

Image ablations. Image ablations are variations of an image where all but a small portion of the image is masked out [25]. For example, a column ablation masks the entire image except for a column of a fixed width (see Figure 1 for an example). We focus primarily on column ablations and explore the more general block ablation in Appendix G.

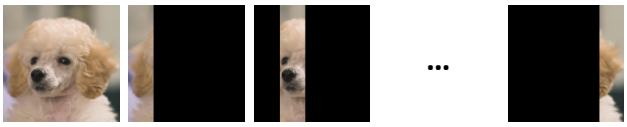


Figure 1. Examples of column ablations for the left-most image with column width 19px.

For a input $h \times w$ sized image $\mathbf{x}$, we denote by $\mathcal{S}_b(\mathbf{x})$ the set of all possible column ablations of width b. A column ablation can start at any position and wrap around the image, so there are w total ablations in $\mathcal{S}_b(\mathbf{x})$.

Derandomized smoothing. Derandomized smoothing [25] is a popular approach for certified patch defenses that constructs a *smoothed classifier* comprising of two main components: (1) a *base classifier*, and (2) a set of image ablations used to smooth the base classifier. Then, the resulting smoothed classifier returns the most frequent prediction of the base classifier over the ablation set $\mathcal{S}_b(\mathbf{x})$. Specifically, for an input image $\mathbf{x}$, ablation set $\mathcal{S}_b(\mathbf{x})$, and a base classifier f, a smoothed classifier g is defined as:

$$g(\mathbf{x}) = \arg\max_c n_c(\mathbf{x}) \tag{1}$$

where

$$n_c(\mathbf{x}) = \sum_{\mathbf{x}' \in S_b(\mathbf{x})} \mathbb{I}\{f(\mathbf{x}') = c\}$$

denotes the number of image ablations that were classified as class c. We refer to the fraction of images that the smoothed classifier correctly classifies as *standard accuracy*.

A smoothed classifier is *certifiably robust* for an input image if the number of ablations for the most frequent class exceeds the second most frequent class by a large enough margin. Intuitively, a large margin makes it impossible for an adversarial patch to change the prediction of a smoothed classifier since a patch can only affect a limited number of ablations.

Specifically, let Δ be the maximum number of ablations in the ablation set $S_b(\mathbf{x})$ that an adversarial patch can simultaneously intersect (e.g., for column ablations of size b, an $m \times m$ patch can intersect with at most $\Delta = m + b - 1$ ablations). Then, a smoothed classifier is certifiably robust on an input $\mathbf{x}$ if it is the case that for the predicted class c:

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

If this threshold is met, the most frequent class is guaranteed to not change even if an adversarial patch compromises every ablation it intersects. We denote the fraction of predictions by the smooth classifier that are both correct and certifiably robust (according to Equation 2) as *certified accuracy*.

Vision transformers. A key component of our approach is the vision transformer (ViT) architecture [10]. In contrast to convolutional architecures, ViTs use self-attention layers instead of convolutional layers as their primary building block and are inspired by the success of self-attention in natural language processing [49]. ViTs process images in three main stages:

1. *Tokenization:* The ViTs split the image into $p \times p$ patches. Each patch is then embedded into a positionally encoded *token*.

2. *Self-Attention:* The set of tokens are then passed through a series of multi-headed self-attention layers [49].

3. *Classification head:* The resulting representation is fed into a fully connected layer to make predictions for classification.

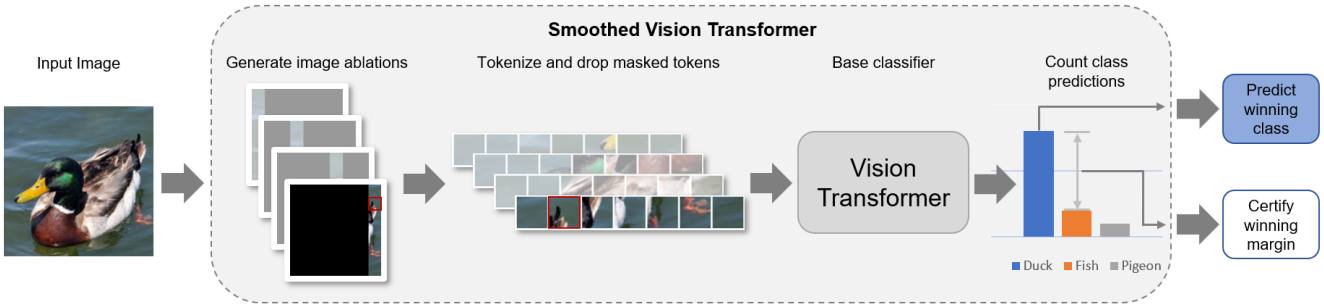


Figure 2. Illustration of the smoothed vision transformer. For a given image, we first generate a set of ablations. We encode each ablation into tokens, and drop fully masked tokens. The remaining tokens for each ablation are then fed into a vision transformer, which predicts a class label for each ablation. We predict the class with the most predictions over all the ablations, and use the margin to the second-place class for robustness certification.

Recent works have investigated whether ViTs can improve robustness in various settings. ViTs initially appeared to be more robust than CNNs to natural and adversarial perturbations [36]. However, recent work showed that this might not be the case [2]. In Appendix B we demonstrate that standard ViTs and CNNs can both be easily broken using simple patch attacks.

2.2. Smoothed vision transformers

Two central properties of vision transformers make ViTs particularly appealing for processing the image ablations that arise in derandomized smoothing. Firstly, unlike CNNs, ViTs process images as sets of tokens. ViTs thus have the natural capability to simply drop unnecessary tokens from the input and "ignore" large regions of the image, which can greatly speed up the processing of image ablations.

Moreover, unlike convolutions which operate locally, the self-attention mechanism in ViTs shares information *globally* at every layer [49]. Thus, one would expect ViTs to be better suited for classifying image ablations, as they can dynamically attend to the small, unmasked region. In contrast, a CNN must gradually build up its receptive field over multiple layers and process masked-out pixels.

Guided by these intuitions, our methodology leverages the ViT architecture as the base classifier for processing the image ablations used in derandomized smoothing. We first demonstrate that these *smoothed vision transformers* enable substantially improved robustness guarantees, without losing much standard accuracy (Section 3). We then modify the ViT architecture and smoothing procedure to drastically speed up the cost of inference of a smoothed ViT (Section 4). We present an overview of our approach in Figure 2.

Setup. We focus primarily on the column smoothing setting and defer block smoothing results to Appendix G. We consider the CIFAR-10 [22] and ImageNet [9] datasets, and perform our analysis on three sizes of vision transformers—ViT-Tiny (ViT-T), ViT-Small (ViT-S), and ViT-Base (ViT-B) models [10, 51]. We compare to residual networks of similar size—ResNet-18, ResNet-50 [17], and Wide ResNet-101-2 [60], respectively. Further details of our experimental setup are in Appendix A. Further experiments exploring data-augmentation are in Appendix C.

3. Improving certified and standard accuracies with ViTs

Recall that even though certified patch defenses can guarantee robustness to patch attacks, this robustness typically does not come for free. Indeed, certified patch defenses tend to have substantially lower standard accuracy when compared to typical (non-robust) models, while delivering a fairly limited degree of (certified) robustness.

In this section, we show how to use ViTs to substantially improve both standard and certified accuracies for certified patch defenses. To this end, we first empirically demonstrate that ViTs are a more suitable architecture than traditional convolutional networks for classifying the image ablations used in derandomized smoothing (Section 3.1). Specifically, this change in architecture alone yields models with significantly improved standard and certified accuracies. We then show how a careful selection of smoothing parameters can enable smoothed ViTs to have even higher standard accuracies that are comparable to typical (non-robust) models, without sacrificing much certified performance (Section 3.2).

Our ImageNet and CIFAR-10 results are summarized in Table 1 and Table 2, respectively. We further include the inference time to evaluate a batch of images, using the modifications described in Section 4. See Appendix H for extended tables covering a wider range of experiments.

Table 1. Summary of our ImageNet results and comparisons to certified patch defenses from the literature: Clipped Bagnet (CBN), BAGCERT, Derandomized Smoothing (DS), and PatchGuard (PG). Time refers to the inference time for a batch of 1024 images, b is the ablation size, and s is the ablation stride. An extended version is in Appendix H.

Standard and Certified Accuracy on ImageNet (%)					
	Standard	1% pixels	2% pixels	3% pixels	Time (sec)
Baselines					
Standard ResNet-50	76.1	—	—	—	0.67
WRN-101-2	78.9	—	—	—	3.1
ViT-S	79.9	—	—	—	0.4
ViT-B	81.8	—	—	—	0.95
CBN [63]	49.5	13.4	7.1	3.1	3.05
BAGCERT [32]‡	45.3	—	22.9	—	8.60
DS [25]*	44.4	17.7	14.0	11.2	149.5
PG [56]†	55.1†	32.3†	26.0†	19.7†	3.05
Smoothed models					
ResNet-50 (b = 19)	51.5	22.8	18.3	15.3	149.5
ViT-S (b = 19)	**63.5**	**36.8**	**31.6**	**27.9**	**14.0**
WRN-101-2 (b = 19)	61.4	33.3	28.1	24.1	694.5
ViT-B (b = 19)	69.3	**43.8**	**38.3**	**34.3**	31.5
ViT-B (b = 37)	**73.2**	43.0	38.2	34.1	58.7
ViT-B (b = 19, s = 10)	68.3	36.9	36.9	31.4	**3.2**

Table 2. Summary of our CIFAR-10 results and comparisons to certified patch defenses from the literature: Clipped Bagnet (CBN), Derandomized Smoothing (DS), and PatchGuard (PG). Here, b is the column ablation size out of 32 pixels. An extended version is in Appendix H.

Standard and Certified Accuracy on CIFAR-10 (%)			
	Standard	2×2	4×4
Baselines			
CBN [63]	84.2	44.2	9.3
DS [25]*	83.9	68.9	56.2
PG [56]†	84.7†	69.2†	57.7†
Smoothed models			
ResNet-50 (b = 4)	86.4	71.6	59.0
ViT-S (b = 4)	**88.4**	**75.0**	**63.8**
WRN-101-2 (b = 4)	88.2	73.9	62.0
ViT-B (b = 4)	**90.8**	**78.1**	**67.6**

3.1. ViTs outperform ResNets on image ablations.

We first isolate the effect of using a ViT instead of a ResNet as the base classifier for derandomized smoothing. Specifically, we keep all smoothing parameters fixed and only vary the base classifier. Following [25], we use column ablations of width $b = 4$ for CIFAR-10 and $b = 19$ for ImageNet for both training and certification.

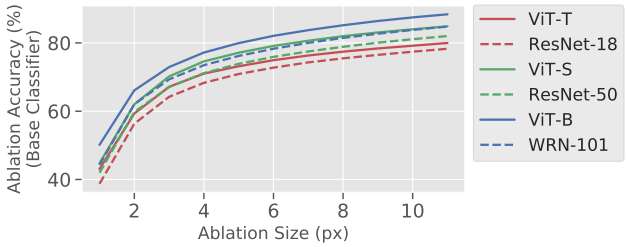
Ablation accuracy. The performance of derandomized smoothing entirely depends on whether the base classifier can accurately classify ablated images. We thus measure the accuracy of ViTs and ResNets at classifying column ablated images across a range of evaluation ablation sizes as shown in Figure 3. We find that ViTs are significantly more accurate on these ablations than comparably sized ResNets. For example, on ImageNet, ViT-S has up to 12% higher accuracy on ablations than ResNet-50.

Certified patch robustness. We next measure the effect of improved ablation accuracy on certified accuracy. We find that using a ViT as the base classifier in derandomized smoothing substantially boosts certified accuracy compared

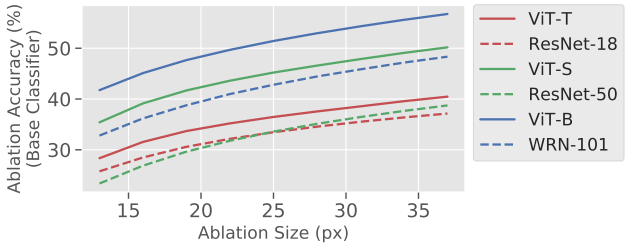
*We found that ResNets could achieve a significantly higher certified accuracy than was reported by [25] if we use early stopping-based model selection. We elaborate further in Appendix A.

†The PatchGuard defense uses a specific mask size that guarantees robustness to patches smaller than the mask, and provides no guarantees for larger patches. In this table, we report their best results: each patch size corresponds to a separate model that achieves 0% certified accuracy against larger patches. Comparisons across the individual models can be found in Appendix H.

‡No code was available, so we extracted the numbers from the paper.

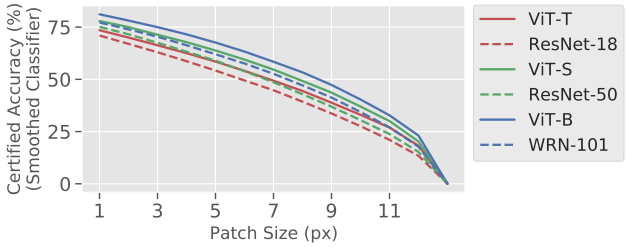


(a) CIFAR-10

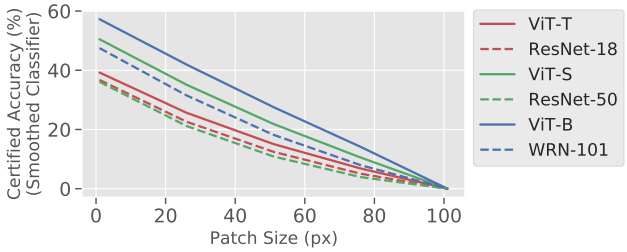


(b) ImageNet

Figure 3. Accuracies on column-ablated images for models on CIFAR-10 and ImageNet. The models were trained on column ablations of width $b = 19$ for ImageNet and $b = 4$ for CIFAR-10, and evaluated on a range of ablation sizes. ViTs outperform ResNets on image ablations by a sizeable margin.



(a) CIFAR-10



(b) ImageNet

Figure 4. Certified accuracies for ViT and ResNet models on CIFAR-10 and ImageNet for various adversarial patch sizes. Certification was performed using a fixed ablation of size $b = 4$ for CIFAR-10 and $b = 19$ for ImageNet (as in [25]).

to ResNets across a range of model sizes and adversarial patch sizes, as shown in Figure 4. For example, against 32×32 adversarial patches on ImageNet (2% of the image), a smoothed ViT-S improves certified accuracy by 14% over a smoothed ResNet-50, while the larger ViT-B reaches a certified accuracy of 39%—well above the highest reported baseline of 26% [56].

Standard accuracy. We further find that smoothed ViTs can mitigate the precipitous drop in standard accuracy observed in previously proposed certified defenses, particularly so for larger architectures and datasets. Indeed, the smoothed ViT-B remains 69% accurate on ImageNet—14.2% higher standard accuracy than that of the best performing prior work (Table 1). A full comparison between the performance of smoothed models and their non-robust counterparts can be found in Appendix H.

3.2. Ablation size matters

In the previous section, we fixed the width of column ablations at $b = 19$ for derandomized smoothing on ImageNet, following [25]. We now demonstrate that properly choosing the ablation size can improve the standard accuracy even further—by 4% on ImageNet—without sacrificing certified performance.

Specifically, we take ImageNet models trained on column ablations with width $b = 19$, and change the smoothing procedure to use a different width at *test* time. We report the resulting standard and certified accuracies in Figure 5, and defer additional experiments on changing the ablation size during training to Appendix D.1.

Although [25] found a steep trade-off between certified and standard accuracy in CIFAR-10 (which we verify in Appendix D.2), we find this to not be the case for ImageNet for either CNNs or ViTs. We can thus substantially increase the ablation size to improve standard accuracy *without* significantly dropping certified performance as shown in Figure 5. For example, increasing the width of column ablations to $b = 37$ improves the standard accuracy of the smoothed ViT-B model by nearly 4% to 73% while maintaining a 38% certified accuracy against 32×32 patches. In addition to being 12% higher than the standard accuracy of the best performing prior work, this model's standard accuracy is only 3% lower than that of a *non-robust* ResNet-50.

Thus, using smoothed ViTs, we can achieve state-of-the-art certified robustness to patch attacks in the ImageNet setting while attaining standard accuracies that are more comparable to those of non-robust ResNets.

4. Faster inference with ViTs

Derandomized smoothing with column ablations is an expensive operation, especially for large images. Indeed, an image with $h \times w$ pixels has w column ablations, so the forward pass of smoothed model is w times slower than a normal forward pass—*two orders of magnitude* slower on ImageNet.

To address this, we first modify the ViT architecture

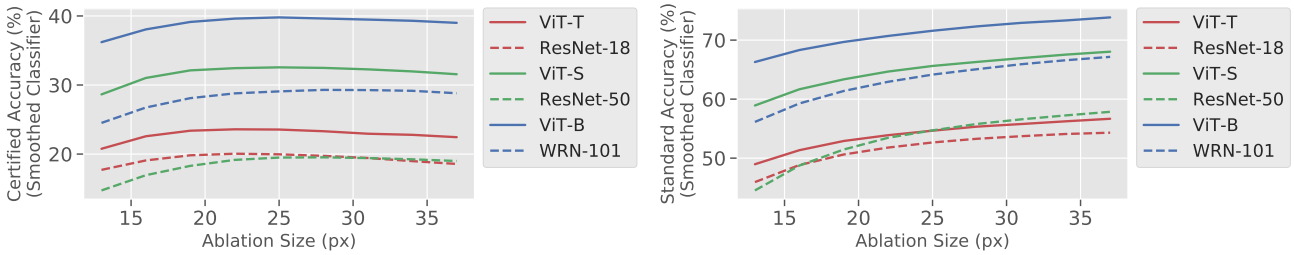


Figure 5. Certified (left) and standard (right) accuracies for a collection of smoothed models trained with a fixed ablation size $b = 19$ on ImageNet, and evaluated with varying ablation sizes. Certified accuracy remains stable across a range of ablation sizes, while standard accuracy substantially improves with larger ablations.

to avoid unnecessary computation on masked pixels (Section 4.1). We then demonstrate that reducing the number of ablations via striding offers further speed up (Section 4.2). These two (complementary) modifications vastly improve the inference time for smoothed ViTs, making them comparable in speed to standard (non-robust) convolutional architectures.

4.1. Dropping masked tokens

Recall that the first operation in a ViT is to split and encode the input image as a set of *tokens*, where each token corresponds to a patch in the image. However, for image ablations, a large number of these tokens correspond to fully masked regions of the image.

Our strategy is to pass only the *subset* of tokens that contain an unmasked part of the original image, thus avoiding computation on fully masked tokens. Specifically, given an image ablation, we alter the ViT architecture to do the following steps:

1. Positionally encode the entire ablated image into a set of tokens.
2. Drop any tokens that correspond to a *fully* masked region of the input.
3. Pass the remaining tokens through the self-attention layers.

The algorithm for dropping masked tokens is described in Algorithm 1, and the overall inference procedure for a smoothed ViT is summarized in Algorithm 2. As one would expect, since the positional encoding maintains the spatial information of the remaining tokens, the ViT's accuracy on image ablations barely changes when we drop the fully masked tokens. We defer a detailed analysis of this phenomenon to Appendix E.

Computational complexity. We now provide an informal summary of the computational complexity of this procedure, and defer a formal asymptotic analysis to Appendix E.1. After tokenization, the bulk of a ViT consists of two main operation types:

- *Attention operators*, which have costs that scale quadratically with the number of tokens but linearly in the hidden dimension.
- *Fully-connected operators*, which have costs that scale linearly with the number of tokens but quadratically in the hidden dimension.

Reducing the number of tokens thus directly reduces the cost of attention and fully connected operators at a quadratic

Algorithm 1 Mechanism for processing an image ablation $\mathbf{z} \in \mathbb{R}^{3\times h\times w}$ with mask $\mathbf{m}$ using a ViT with tokens of size $p \times p$ while dropping masked tokens. The ViT is decomposed into a positional encoder E and attention layers V.

function ProcessAblation($\mathbf{z}, \mathbf{m}$)
  $\mathcal{T} = \{\}$ *Initialize set of tokens for an ablation*
  for $i, j \in [h/p] \times [w/p]$ **do**
    if not $\mathbf{m}_{ip:(i+1)p, jp:(j+1)p} = \mathbf{0}$ **then**
      $\mathcal{T} = \mathcal{T} \cup E(\mathbf{z}_{ip:(i+1)p, jp:(j+1)p}, i, j)$
    end if
  end for
  return $V(\mathcal{T})$
end function

Algorithm 2 Forward pass for a smoothed ViT on an input image $\mathbf{x}$ for k classes and ablation set $\mathcal{S}(\mathbf{x})$, where $\mathbf{z}, \mathbf{m} \in \mathcal{S}(\mathbf{x})$ are the image ablations $\mathbf{z}$ and the corresponding mask $\mathbf{m}$.

function SmoothedViT($\mathbf{x}$)
  $c_i = 0$ for $i \in [k]$ // *Initialize counts to zero*
  for $\mathbf{z}, \mathbf{m} \in \mathcal{S}(\mathbf{x})$ **do**
    y = ProcessAblation($\mathbf{z}, \mathbf{m}$)
    $c_y = c_y + 1$ // *Update counts*
  end for
  return $\arg\max_y c_y$
end function

Table 3. Multiplicative speed up of inference for a smoothed ViT with dropped tokens over a smoothed ResNet, measured over a batch of 1024 images with $b = 19$.

	ResNet-18	ResNet-50	WRN-101
ViT-T	**5.85x**	21.96x	101.99x
ViT-S	2.85x	**10.68x**	49.62x
ViT-B	1.26x	4.75x	**22.04x**

and linear rate, respectively. For a small number of tokens, the linear scaling from the fully-connected operators tends to dominate. The cost of processing column ablations thus scales linearly with the width of the column, which we empirically validate in Figure 6. Further details about how we time these models can be found in Appendix A.4.

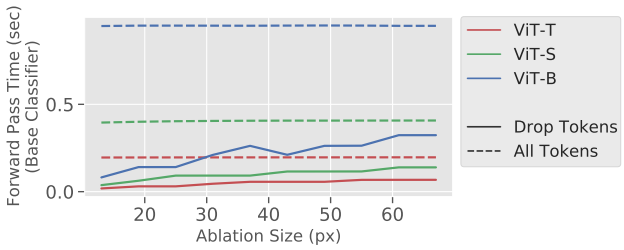


Figure 6. The average time to compute a forward pass for ViTs on 1024 column ablated images with varying ablation sizes, with and without dropping masked tokens. The cost of processing a full image without dropping masked tokens corresponds to the maximum ablation size $b = 224$.

4.2. Empirical speed-up for smoothed ViTs

Smoothed classifiers must process a large number of image ablations in order to make predictions and certify robustness. Consequently, using our ViT (with dropped tokens) as the base classifier for derandomized smoothing directly speeds up inference time. In this section, we explore how much faster smoothed ViTs are in practice.

We first measure the number of images per second that smoothed ViTs and smoothed ResNets can process. We use column ablations of size $b = 19$ on ImageNet, following [25]. In Table 3 that describes our results, we find speedups of 5-22x for smoothed ViTs over smoothed ResNets of similar size, with larger architectures showing greater gains. Notably, using our largest ViT (ViT-B) as the base classifier is 1.25x faster than using a ResNet-18, despite being 8x larger in parameter count. Dropping masked tokens thus substantially speeds up inference time for smoothed ViTs, to the point where using a large ViT is comparable in speed to using a small ResNet.

Strided ablations. We now consider a complementary means of speeding up smoothed classifiers: directly reducing the size of the ablation set via *strided* ablations. Specifically, instead of using every possible ablation, we can subsample every s-th ablation for a given stride s. Striding can reduce the total number of ablations (and consequently speed up inference) by a factor of s, *without* substantially hurting standard or certified accuracy (Table 1). We study this in more detail in Appendix F.

Strided ablations, in conjunction with the dropped tokens optimization from Section 4.1, lead to smoothed ViTs having inference times comparable to standard (non-robust) models. For example, when using stride $s = 10$ and dropping masked tokens, a smoothed ViT-S is only 2x slower than a single inference step of a standard ResNet-50, while a smoothed ViT-B is only 5x slower. We report the inference time of these models, along with their standard and certified accuracies, in Table 1.

5. Related work

Certified defenses. An extensive body of research has studied the development of certified or provable defenses to adversarial perturbations. This line of research largely falls into one of three categories: tighter or exact verifiers [11, 21, 31, 46, 57], convex relaxation-based defenses [14, 15, 33, 37, 43, 50, 52, 53, 62], and smoothing-based defenses [7, 23, 25, 27, 28, 41, 42, 58]. In the case of patches, the earliest certified defense used an instance of convex relaxation (interval bounds) to derive provable guarantees to adversarial patch [6]. Subsequent work [26] focused on randomized smoothing. This approach smooths classifiers over random noise, but tend to be extremely expensive to use (4-5 orders of magnitudes slower than a standard, non-robust model) [7, 26]. Recently, [29] proposed a variant based on randomized cropping that performs similarly to [25] but with better guarantees under worse-case patch transformations.

Deterministic smoothing. To mitigate the expensive inference times of randomized smoothing, [25] proposed derandomized smoothing, which used a finite set of ablations to smooth a base classifier. This substantially reduced the computational requirements of smoothing, but is still two orders of magnitude slower than standard models. Several other defenses, including Clipped BagNet [63], BAGCERT [32], and PatchGuard [56], rely on restricting the model's receptive field. These approaches are faster than derandomized smoothing, but have other limitations. Clipped BagNet (CBN) has substantially weaker robustness guarantees than derandomized smoothing. BAGCERT achieves higher robustness guarantees than CBN, but lower standard accuracy. PatchGuard has further higher but *brittle* guarantees: a defended model is optimally defended

against a specific patch size, and achieves no robustness at all against patches that are even slightly larger than the one considered.

Empirical methods: attacks and defenses. Another line of work studies empirical approaches for generating adversarial patches and designing empirical defenses. Adversarial patches have been developed for downstream tasks such as image classification [20], object detection [5, 13, 30], and facial recognition [3, 44, 45]. Several of these attacks work in the physical domain [4, 5, 13], and can successfully target tasks such as traffic sign recognition [5, 13]. Heuristic defenses to these attacks include watermarking [16] and gradient smoothing [35]; however, these defenses were shown to be vulnerable adaptive attacks [6]. More recently, [39] proposed an adversarial training approach and [34] proposed a robust attention module to improve empirical robustness to patch attacks.

Vision transformers. Our work leverages the vision transformer (ViT) architecture [10], which adapts the popular attention-based model from the language setting [49] to the vision setting. Recent work [47] has released more efficient training methods as well as pre-trained ViTs that have made these architectures more accessible to the wider research community.

6. Conclusion

We demonstrate how applying visual transformers (ViTs) within the smoothing framework leads to significantly improved certified robustness to adversarial patches while maintaining standard accuracies that are on par with regular (non-robust) models. Further, we put forth changes to the ViT architecture and the corresponding smoothing procedure that greatly speed up the resulting inference times over previous smoothing approaches by up to two orders of magnitude—they end up being only 2-5x slower than that of a regular ResNet. We believe that these improvements finally establish models that are certifiably robust to adversarial patches as a viable alternative to standard (non-robust) models.

Limitations. Similarly to other certified defenses, our method specifically focuses on patch attacks and does not guarantee robustness to attacks that fall outside of this threat model. Furthermore, although our approach is vastly faster than other smoothed models, smoothed ViTs are still slightly slower than standard (non-robust) models. Finally, the standard accuracy of our models may suffer if the predictive signal in an image comes only from a small region of the image, as that region might not be present in many image ablations.

Potential negative impact. A possible negative impact of our work is that it might instill overconfidence in the model. At test time, our robustness guarantees ensure that the prediction is stable but might be not necessarily correct, leading to a false sense of confidence. Additionally, users may erroneously extrapolate other forms of robustness from our guarantees of patch robustness. The guarantees presented in this paper are *robustness* guarantees and not *correctness* guarantees, in the sense that our models can guarantee that a prediction is stable if a certain region of the image is manipulate, but it cannot guarantee that the prediction will be correct. Therefore, we encourage users to be aware of these subtleties before using our technique.

7. Acknowledgements

Work supported in part by the NSF grants CCF-1553428 and CNS-1815221, and Open Philanthropy. This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001120C0015.

Research was sponsored by the United States Air Force Research Laboratory and the United States Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the United States Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

[1] Mitali Bafna, Jack Murtagh, and Nikhil Vyas. Thwarting adversarial examples: An l_0-robustsparse fourier transform. *arXiv preprint arXiv:1812.05013*, 2018. 1

[2] Yutong Bai, Jieru Mei, Alan Yuille, and Cihang Xie. Are transformers more robust than cnns? *arXiv preprint arXiv:2111.05464*, 2021. 3, 14

[3] Avishek Joey Bose and Parham Aarabi. Adversarial attacks on face detectors using neural net based constrained optimization. In *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*, 2018. 8

[4] Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch, 2018. 1, 8

[5] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Polo Chau. Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 52–68. Springer, 2018. 8

[6] Ping-yeh Chiang, Renkun Ni, Ahmed Abdelkader, Chen Zhu, Christoph Studor, and Tom Goldstein. Certified defenses for adversarial patches. *arXiv preprint arXiv:2003.06693*, 2020. 1, 7, 8

[7] Jeremy M Cohen, Elan Rosenfeld, and J Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning (ICML)*, 2019. 2, 7

[8] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical data augmentation with no separate search. *arXiv preprint arXiv:1909.13719*, 2(4):7, 2019. 14

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition (CVPR)*, 2009. 3

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. 1, 2, 3, 8, 12

[11] Rüdiger Ehlers. Formal verification of piece-wise linear feed-forward neural networks. In *Automated Technology for Verification and Analysis*, 2017. 7

[12] Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. Robust physical-world attacks on machine learning models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[13] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Florian Tramer, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Physical adversarial examples for object detectors. *CoRR*, 2018. 8

[14] Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. On the effectiveness of interval bound propagation for training verifiably robust models. 2018. 7

[15] Sven Gowal, Krishnamurthy Dj Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. Scalable verified training for provably robust image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 7

[16] Jamie Hayes. On visible adversarial perturbations & digital watermarking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1597–1604, 2018. 1, 8

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3, 12

[18] Shahar Hoory, Tzvika Shapira, Asaf Shabtai, and Yuval Elovici. Dynamic adversarial patch for evading object detection models. *arXiv preprint arXiv:2010.13070*, 2020. 1

[19] Kyle D Julian and Mykel J Kochenderfer. Guaranteeing safety for neural network-based aircraft collision avoidance systems. In *2019 IEEE/AIAA 38th Digital Avionics Systems Conference (DASC)*, pages 1–10. IEEE, 2019. 1

[20] Danny Karmon, Daniel Zoran, and Yoav Goldberg. Lavan: Localized and visible adversarial noise. In *International Conference on Machine Learning*, pages 2507–2515. PMLR, 2018. 8

[21] Guy Katz, Clark Barrett, David Dill, Kyle Julian, and Mykel Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks. In *International Conference on Computer Aided Verification*, 2017. 7

[22] Alex Krizhevsky. Learning multiple layers of features from tiny images. In *Technical report*, 2009. 3, 12

[23] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *Symposium on Security and Privacy (SP)*, 2019. 7

[24] Mark Lee and Zico Kolter. On physical adversarial patches for object detection. *arXiv preprint arXiv:1906.11897*, 2019. 1

[25] Alexander Levine and Soheil Feizi. (de) randomized smoothing for certifiable defense against patch attacks. *arXiv preprint arXiv:2002.10733*, 2020. 1, 2, 4, 5, 7, 12, 13, 15, 18, 20, 21

[26] Alexander Levine and Soheil Feizi. Robustness certificates for sparse adversarial attacks by randomized ablation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4585–4593, 2020. 1, 2, 7, 13

[27] Alexander Levine and Soheil Feizi. Wasserstein smoothing: Certified robustness against wasserstein adversarial attacks. In *International Conference on Artificial Intelligence and Statistics*, pages 3938–3947. PMLR, 2020. 7

[28] Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Certified adversarial robustness with additive noise. *arXiv preprint arXiv:1809.03113*, 2018. 7

[29] Wan-Yi Lin, Fatemeh Sheikholeslami, jinghao shi, Leslie Rice, and J Zico Kolter. Certified robustness against adversarial patch attacks via randomized cropping. In *ICML 2021 Workshop on Adversarial Machine Learning*, 2021. 7

[30] Xin Liu, Huanrui Yang, Ziwei Liu, Linghao Song, Hai Li, and Yiran Chen. Dpatch: An adversarial patch attack on object detectors. *arXiv preprint*

arXiv:1806.02299, 2018. 8

[31] Alessio Lomuscio and Lalit Maganti. An approach to reachability analysis for feed-forward relu neural networks. In *ArXiv preprint arXiv:1706.07351*, 2017. 7

[32] Jan Hendrik Metzen and Maksym Yatsura. Efficient certified defenses against patch attacks on image classifiers. In *International Conference on Learning Representations*, 2021. 4, 7

[33] Matthew Mirman, Timon Gehr, and Martin Vechev. Differentiable abstract interpretation for provably robust neural networks. In *International Conference on Machine Learning (ICML)*, 2018. 7

[34] Norman Mu and David Wagner. Defending against adversarial patches with robust self-attention. 8

[35] Muzammal Naseer, Salman Khan, and Fatih Porikli. Local gradients smoothing: Defense against localized adversarial attacks. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1300–1307. IEEE, 2019. 1, 8

[36] Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *arXiv preprint arXiv:2105.10497*, 2021. 3, 14

[37] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2018. 7

[38] Anurag Ranjan, Joel Janai, Andreas Geiger, and Michael J Black. Attacking optical flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2404–2413, 2019. 1

[39] Sukrut Rao, David Stutz, and Bernt Schiele. Adversarial training against location-optimized adversarial patches. *arXiv preprint arXiv:2005.02313*, 2020. 8

[40] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. In *International Journal of Computer Vision (IJCV)*, 2015. 12

[41] Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 7

[42] Hadi Salman, Mingjie Sun, Greg Yang, Ashish Kapoor, and J Zico Kolter. Denoised smoothing: A provable defense for pretrained classifiers. *Advances in Neural Information Processing Systems*, 33, 2020. 7

[43] Hadi Salman, Greg Yang, Huan Zhang, Cho-Jui Hsieh, and Pengchuan Zhang. A convex relaxation barrier to tight robustness verification of neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 7

[44] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28, 2016*, pages 1528–1540, 2016. 1, 8

[45] Simen Thys, Wiebe Van Ranst, and Toon Goedemé. Fooling automated surveillance cameras: adversarial patches to attack person detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 8

[46] Vincent Tjeng, Kai Xiao, and Russ Tedrake. Evaluating robustness of neural networks with mixed integer programming. In *International Conference on Learning Representations (ICLR)*, 2019. 7

[47] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020. 8, 12

[48] Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *arXiv preprint arXiv:2002.08347*, 2020. 1

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 2, 3, 8

[50] Tsui-Wei Weng, Huan Zhang, Hongge Chen, Zhao Song, Cho-Jui Hsieh, Duane Boning, Inderjit S Dhillon, and Luca Daniel. Towards fast computation of certified robustness for ReLU networks. In *International Conference on Machine Learning (ICML)*, 2018. 7

[51] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019. 3, 12

[52] Eric Wong and J Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning (ICML)*, 2018. 7

[53] Eric Wong, Frank Schmidt, Jan Hendrik Metzen, and Zico Kolter. Scaling provable adversarial defenses. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 7

[54] Eric Wong, Tim Schneider, Joerg Schmitt, Frank R Schmidt, and J Zico Kolter. Neural network virtual sensors for fuel injection quantities with provable performance specifications. In *2020 IEEE Intelligent*

Vehicles Symposium (IV), pages 1753–1758. IEEE, 2020. 1

[55] Zuxuan Wu, Ser-Nam Lim, Larry S Davis, and Tom Goldstein. Making an invisibility cloak: Real world adversarial attacks on object detectors. In *European Conference on Computer Vision*, pages 1–17. Springer, 2020. 1

[56] Chong Xiang, Arjun Nitin Bhagoji, Vikash Sehwag, and Prateek Mittal. Patchguard: A provably robust defense against adversarial patches via small receptive fields and masking. In *30th {USENIX} Security Symposium ({USENIX} Security 21)*, 2021. 1, 4, 5, 7, 20, 21

[57] Kai Y. Xiao, Vincent Tjeng, Nur Muhammad Shafiullah, and Aleksander Madry. Training for faster adversarial robustness verification via inducing ReLU stability. In *International Conference on Learning Representations (ICLR)*, 2019. 7

[58] Greg Yang, Tony Duan, J. Edward Hu, Hadi Salman, Ilya Razenshteyn, and Jerry Li. Randomized smoothing of all shapes and sizes, 2020. 7

[59] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019. 14

[60] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 3, 12

[61] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 14

[62] Huan Zhang, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, and Luca Daniel. Efficient neural network robustness certification with general activation functions. *arXiv preprint arXiv:1811.00866*, 2018. 7

[63] Zhanyuan Zhang, Benson Yuan, Michael McCoyd, and David Wagner. Clipped bagnet: defending against sticker attacks with clipped bag-of-features. In *2020 IEEE Security and Privacy Workshops (SPW)*, 2020. 1, 4, 7, 20, 21