A. Experimental setup

A.1. Models and architectures

We use three sizes of vision transformers—ViT-Tiny (ViT-T), ViT-Small (ViT-S), and ViT-Base (ViT-B) models [10, 51] and compare to to residual networks of similar (or larger) size—ResNet-18, ResNet-50 [17], and Wide ResNet-101-2 [60], respectively. These architectures and their corresponding number of parameters are summarized in Table 4. The table also reports the standard accuracy and the robust accuracy of pretrained versions.

Table 4. A collection of neural network architectures we use in our paper.

Architecture	ViT-T	ResNet-18	ViT-S	ResNet-50	ViT-B	WRN-101-2
Params	5M	12M	22M	26M	86M	126M
Standard Accuracy (%)	72.2	69.8	79.9	76.1	81.8	78.9

We use the same architectures for both ImageNet and CIFAR-10 models, and finetune our smoothed models from publicly released checkpoints pretrained on ImageNet. All our CIFAR-10 experiments are thus conducted on up-sampled CIFAR-10 images of size 224×224.

A.2. Datasets

We use two datasets:

1. CIFAR [22] https://paperswithcode.com/dataset/cifar-10.
2. ImageNet [40], with a custom (research, non-commercial) license, as found here https://paperswithcode.com/dataset/imagenet.

A.3. Training parameters

Derandomized smoothing requires that the base classifier predict well on image ablations. A standard technique for derandomized smoothing methods is to directly train the base classifier on image ablations [25]. Thus, unless otherwise stated, in each epoch we randomly apply a column ablation of fixed width to each image of the training set.

To facilitate training of the base classifiers, we start from pretrained ResNets[2] and ViT architectures[3] (see Table 4) and fine-tune as follows:

ImageNet. We train for 30 epochs using SGD of fixed learning rate of 10^{-3}, a batch size of 256, a weight-decay of 10^{-4}, a momentum of 0.9, and with column ablations of fixed width $b = 19$. For data-augmentation, we use random resized crop, random horizontal flip, and color jitter. We then apply column ablations.

CIFAR-10. We train for 30 epochs using SGD with a step learning rate of 10^{-2} that drops every 10 epochs by a factor of 10, a batch size of 128, a weight-decay of 5×10^{-4}, a momentum of 0.9, and with column ablations of fixed width $b = 4$. We only use random horizontal flip for data-augmentation, after which we apply column ablations. We then upsample all CIFAR-10 images to 224×224 (on GPU).

Training time. Training is relatively fast, with our largest ImageNet model (WRN-101-2) finishing in roughly two days on one NVIDIA V100 GPU. The smaller models such as ViT-T or ResNet-18 finish training in only a few hours.

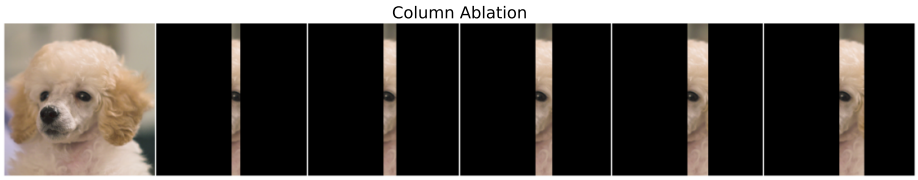
A.4. Compute and timing experiments

We use an internal cluster containing NVIDIA 1080-TI, 2080-TI, V100, and A100 GPUs. Scalability and timing experiments were performed on an A100 and averaged over 50 trials. When performing scalability experiments, we do not include data loading time or the time to move the input to the GPU.

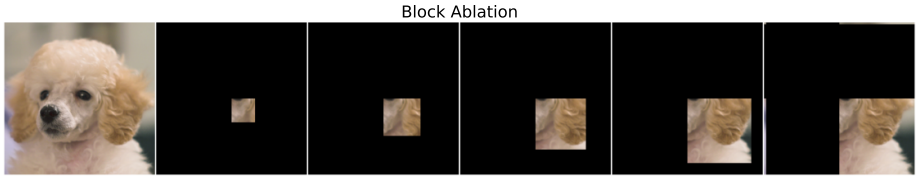
A.5. Example ablations

In Figure 7, we display examples of ablations of various types (column, block) and sizes.

[2]These are TorchVision's official checkpoints, and can be found here https://pytorch.org/vision/stable/models.html.
[3]We use the DeiT checkpoints of [47] which can be found here https://github.com/facebookresearch/deit.



(a) **Column ablations** with the following ablation size from left to right: original image, 13px, 19px, 25px, 31px, 37px.



(b) **Block ablations** with the following ablation size from left to right: original image, 35px, 55px, 75px, 95px, 115px.

Figure 7. Example ablations that we use in our paper.

A.6. Differences in setup from [25]

Our work builds on top of that of [25]. We use their robustness guarantee as is (see Section 2.1), but there are a few differences in the setup of our experiments. All experimental results (including the de-randomized smoothing baseline) are run using the same experimental setup in order to remain fair, which only improved the baseline over what was previously reported in the literature. For completeness, we describe the differences in setup here.

Encoding *null* inputs. The first difference is that [25] encode part of the input as being *null* or ablated by adding additional color channels, as described in [26], so that the *null* value is distinct from all real pixel colors. In practice, we found this to be unnecessary, and were able to replicate their results with ablations that simply replace masked pixels with 0.

Early stopping. We find that ResNets substantially benefit from early stopping when trained with ablations, and otherwise experience severe overfitting to the ablations with substantially reduced test accuracies. In our replication, we find that the ResNet-50 result reported by [25] can be substantially improved with an earlier checkpoint (improving certified accuracy by nearly 10%), and thus we use early-stopping in all of our ResNet baselines.

Starting from pretrained models. To reduce training time, for both ImageNet and CIFAR-10 experiments, we start from pre-trained ImageNet checkpoints (see Section A.3). This step is especially necessary for the CIFAR-10 experiments, as it is quite challenging to train a ViT from scratch on CIFAR-10 (these models tend to require a large amount of data).

Upsampled CIFAR-10. In order to use the pretrained ImageNet checkpoints when training our base classifiers for CIFAR-10, we (nearest neighbor) upsample the CIFAR-10 inputs to 224×224 as part of the model architecture. We verify robustness in the original 32×32 images.

Sweeping over ablation size. We note that [25] tested various ablations sizes only on CIFAR-10. Due to our speed-ups, we were able to sweep over ablations sizes for ImageNet.

B. Are standard ViTs more robust than CNNs against adversarial patches?

In this appendix, we demonstrate that neither standard ViTs nor CNNs are naturally robust to patch attacks. This is in contrast to what previous work have shown that ViTs might have non-trivial robustness levels against patch attacks [2, 36].

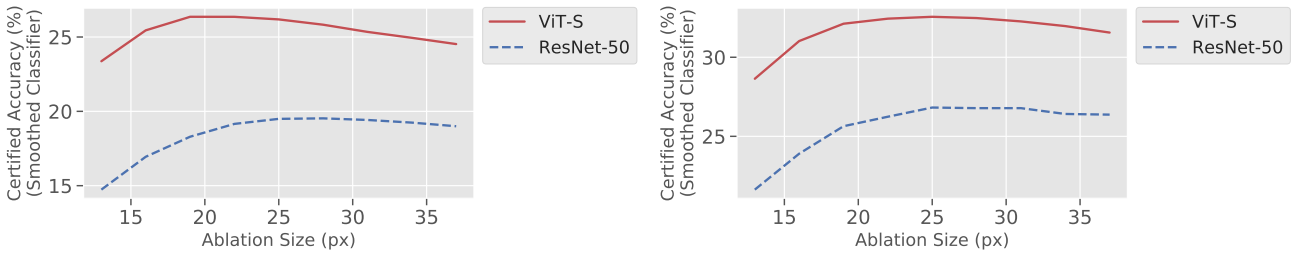
We attack all the pretrained ViTs and CNNs that we use in our paper and demonstrate that all of them are broken. We conduct this analysis on ImageNet. We simply optimize a 32×32 adversarial patch (2% of the size of ImageNet images) for 5000 randomly sampled datapoints of the ImageNet validation set. We did not need to optimize the location of the patch; we simply placed the patch at a fixed location (upper left corner of the patch placed at (100,100)). Our attack is able to break all models as shown in Table 5.

Table 5. A collection of neural network architectures we use in our paper.

Architecture	ViT-T	ResNet-18	ViT-S	ResNet-50	ViT-B	WRN-101-2
Standard Accuracy (%)	72.2	69.8	79.9	76.1	81.8	78.9
Robust Accuracy (%)	0.04	0.06	0.06	1.51	1.04	4.02

C. Effect of data augmentation

Recent work has shown that ViTs and CNNs trained with the same data augmentation procedures have similar levels of empirical patch robustness [2]. In this section, we control for data augmentation and measure the effect of the ViT architecture change alone. Specifically, rather than starting from pretrained models, we train from scratch a ResNet-50 and ViT-S models using 1) ResNet standard data-augmentations (random resized crop + random horizontal flip), or 2) ViT data-augmentations (RandAug [8], MixUp[61], CutMix [59]). We then smooth these models as done in the rest of the paper. We report the certified accuracy curves in Figure 8. We find that while data augmentation does help both architectures, ViTs outperform CNNs even when using the same augmentation.



(a) ResNet Augmentations: Random resized crop + random horizontal flip.

(b) ViT Augmentations: RandAug + MixUp + CutMix.

Figure 8. Certified accuracies for smoothed ViTs and smoothed ResNets trained using unified data augmentations. ViTs outperform ResNets regardless of the data augmentation used.

D. Ablation sweeps

In this section, we further explore the impact of changing the ablation size on both standard and certified performance. In Section D.1, we explore the effect of modifying the ablation size at training time. In Section D.2, similar to the experiment on ImageNet from Section 3.2, we present additional results on adjusting the ablation size at test time for CIFAR10.

D.1. Train-time ablation

We first explore varying the ablation size used during training for ImageNet. Specifically, we train and certify a ResNet-50 and ViT-S over a range of column widths from 1 to 67 pixels (Figure 9).

For ViTs, we find that once the columns are wide enough, we see only marginal improvements in certified accuracy (i.e. only 1.3% higher certified accuracy over $b = 19$). This suggests that small ablations are sufficient at training time, allowing for fast training of ViTs when using cropped ablations.

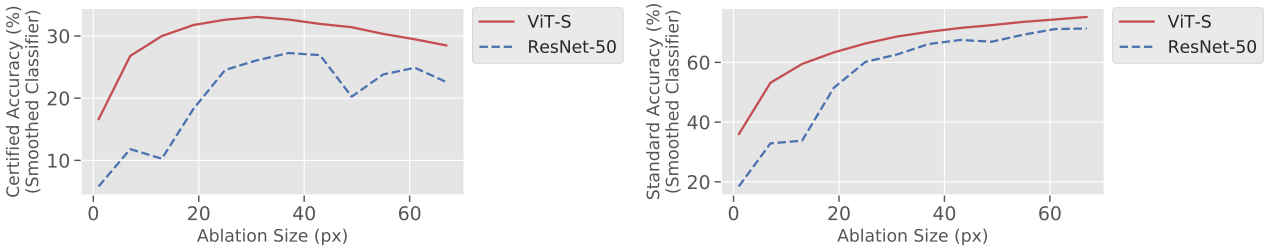


Figure 9. Certified and standard accuracy for a smoothed model trained and evaluated on ImageNet column ablations with varying widths. The ResNet-50 requires a substantially larger ablation size for certification, whereas the ViT-S is more flexible.

On the other hand, ResNets require a substantially larger column width than was previously explored. Specifically, the certified accuracy of the ResNet baseline can be greatly improved from 18% to 27% when the ablation size is increased to $b = 37$. This ablation size is optimal for the ResNet, but is still 6% lower certified accuracy when compared to the ViT.

Overall, we find that certified performance of ViTs on ImageNet remains largely stable with respect to the column ablation size used for training. We can thus use smaller ablation sizes during training (e.g $b = 19$) to improve training speed while certifying using larger ablation sizes.

D.2. Test-time ablations

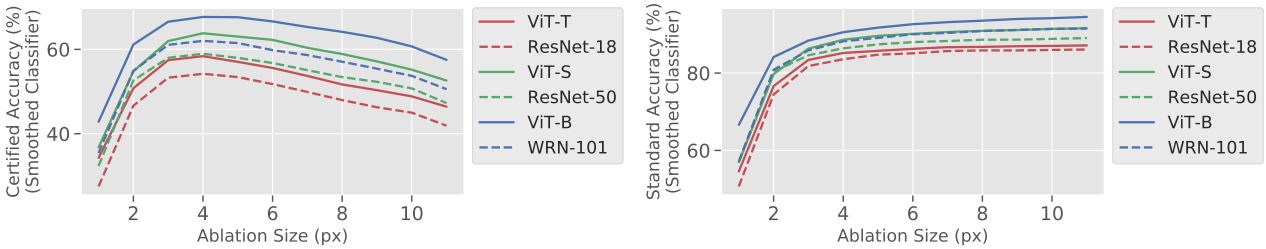


Figure 10. Certified and standard accuracy for a smoothed model on CIFAR-10 trained with a fixed ablation size ($b = 5$), and evaluated with varying ablation sizes.

Similar to the experiment on ImageNet from Section 3.2, we present analogous results for varying the ablation size used at test time for CIFAR-10. These results largely reflect what was previously observed by [25]. Specifically, the optimal ablation size for CIFAR10 is a column width of $b = 4$, with a steep drop-off in performance for larger ablation sizes. This is in contrast to what we observed in ImageNet, which did not see such a steep drop in performance.

E. Dropping tokens for ViTs

E.1. Computational complexity of ViTs with dropped tokens

We can now derive the computational complexity of the smoothed ViT when dropping tokens. Specifically, consider a ViT that divides an $h \times w$ pixel image into $p \times p$ patches, and positionally encodes them tokens with d hidden dimensions.

Recall that a ViT has two operation types: *attention operators* which scale quadratically with the number of tokens but linearly with hidden dimension d and *fully-connected operators* which scale linearly with the number of tokens but quadratically in d. Without dropping tokens, we have hw/p^2 tokens. A forward pass of processing an image ablation without dropping tokens thus has an overall complexity of

$$O\left(\left(\frac{hw}{p^2}\right)^2 d + \left(\frac{hw}{p^2}\right) d^2\right)$$

where the first term corresponds to the attention operations, and the second term corresponds to the fully-connected operations.

For column ablations with width b, dropping masked tokens reduces the number of tokens to hb/p^2. The complexity of the forward pass to process an image ablation when dropping masked tokens (i.e `ProcessAblation`) then drops to

$$O\left(\left(\frac{hb}{p^2}\right)^2 d + \left(\frac{hb}{p^2}\right) d^2\right)$$

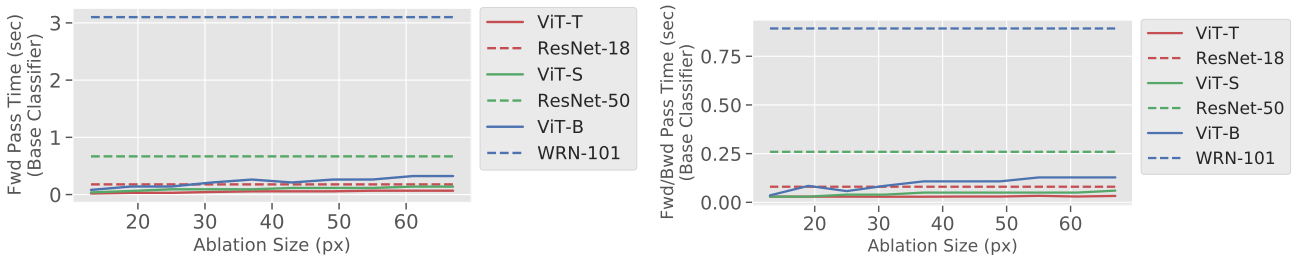
thus reducing the attention cost by a factor of $O(w^2/b^2)$ and the fully-connected cost by a factor of $O(w/b)$. In practice, the computation of fully-connected operations tends to dominate since $d > \frac{hw}{p^2}$.

Overall, a smoothed ViT with stride s processes w/s ablations. Thus, the overall complexity of the smoothed ViT is:

$$O\left(\frac{w}{s}\left(\left(\frac{hb}{p^2}\right)^2 d + \left(\frac{hb}{p^2}\right) d^2\right)\right)$$

E.2. Effect of dropping tokens on speed

We extend the timing experiments comparing ViTs and ResNets to a range of ablation sizes (previously presented in Table 3 from Section 4 for a single column ablation size of $b = 19$). Empirically, even for substantially larger ablations, we find significantly faster training and inference times for ViTs over ResNets. In Figure 11, we compare the evaluation and training speeds for processing image ablations with ResNets and ViTs with dropped tokens.



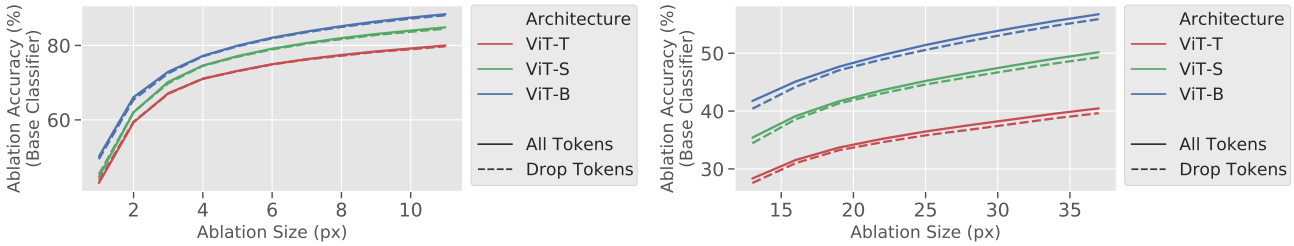
(a) Forward Pass Time

(b) Forward/Backward Pass Time

Figure 11. (a) Average time for computing a forward pass on a batch of 1024 image ablations on ImageNet (b) Average time for computing a full training step (forward and backward pass) on a batch of 128 image ablations on ImageNet

E.3. Effect of dropping tokens on performance

Since the tokens are individually positionally encoded, dropping tokens that are fully masked does not remove any information from the input. In Figure 12, we confirm that dropping masked tokens does not significantly change the accuracy of the ViT base classifier on ablations.



(a) CIFAR-10 models.

(b) ImageNet models.

Figure 12. We compare the ablation accuracies of dropping masked tokens versus processing all tokens for a collection of vision transformers on CIFAR-10 and ImageNet. Dropping masked tokens does not substantially degrade accuracy.

F. Strided ablations

In this section, we explore strided ablations for certification in more depth. In Section F.1 we present the threshold for certification when using strided ablations. In Section F.2 we show how striding affects performance.

F.1. Certification thresholds for strided ablation sets

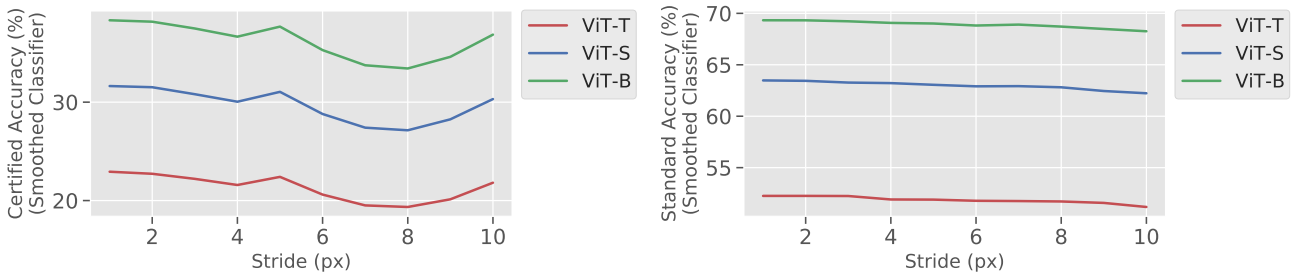
We briefly describe the new thresholds for certification when using strided ablations. Recall from equation 2 that a prediction is certified robust if

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

Thus Δ, the number of ablations that a patch can intersect, fully describes the certification threshold.

Column smoothing. For column smoothing with width b and stride s, the maximum number of ablations that an $m \times m$ patch can intersect with is at most $\Delta_{column+stride} = \lceil (m + s - 1)/s \rceil$.

F.2. Performance under strided ablations



(a) Certified Accuracy

(b) Standard Accuracy

Figure 13. Certified and standard accuracy of various ViTs for ImageNet when using strided column ablations with varying stride lengths.

In this section, we explore how striding affects standard and certified performance. We find that striding does not result in a monotonic change in certified accuracy—certification accuracy can both decrease and increase as the stride increases.

For a few choices in striding, it is possible to not substantially change the accuracy of the ViT at classifying ablations, as shown in Figure 13. For example, a ViT-B which normally obtains 38.3% certified accuracy without striding, maintains certified accuracies of 37.6% at stride $s = 5$ and 36.8% at stride $s = 10$. For these small drops in certified accuracy, striding directly enables 5x or 10x faster inference times.

G. Block smoothing

In this section, we investigate an alternative type of smoothing known as *Block Smoothing*, previously investigated in the CIFAR-10 setting [25]. In block smoothing, we ablate (square) blocks of pixels instead of columns of pixels. This procedure is prohibitively expensive for ImageNet due to its quadratic complexity. For example, smoothing a 224×224 image with block ablations takes a majority vote over $224 \times 224 = 50,176$ ablations, which is four orders of magnitude slower than a standard forward pass. We alleviate this obstacle for larger image settings such as ImageNet with the token-based speedups for ViTs from Section 4.1 and the striding from Section 4.2. In combination, these improvements in speed allow us to perform a practical investigation into block smoothing on ImageNet.

Certification. Certification of derandomized smoothing models with block ablations is similar to that of column ablations, and depends on the maximum number of ablations in the ablation set that an adversarial patch can simultaneously intersect. Recall that for column ablations of size b, the certification threshold is $\Delta = m + b - 1$ ablations. For block ablations of size b (where b here is the side of the retained block/square of pixels), $\Delta = (m + b - 1)^2$. The threshold can then be plugged as before into Equation equation 2 to check whether the model is certifiably robust.

G.1. Practical inference speeds for block smoothing

We first demonstrate how dropping masked tokens significantly increases the speed of evaluating block ablations for the base classifier. In Figure 14, we show that dropping masked tokens substantially reduces the time needed to process 1024 block ablations for various sizes of ViTs. This directly leads to a 4.85x speedup for ViT-S with ablation size 75.

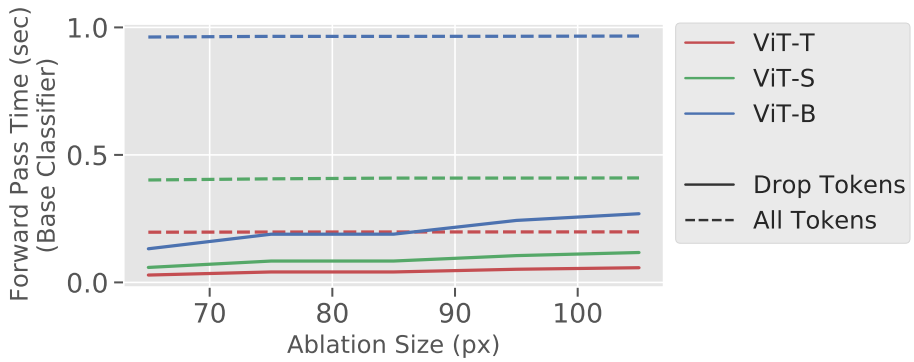


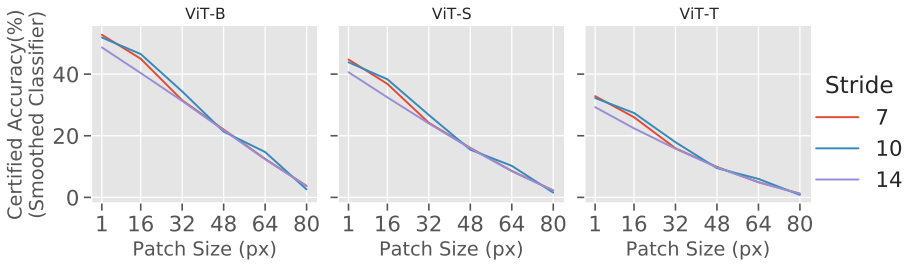
Figure 14. Average time to compute a forward pass for ViTs on 1024 block ablated images with varying ablation sizes with and without dropping masked tokens.

Even with this optimization, however, block smoothing is quite expensive. A forward pass through the smoothed model still requires around 50k passes through the base classifier. We thus leverage our second speedup from strided ablations and use *strided* block smoothing. Similar to strided column ablations, for a stride length of s, we only consider block ablations that are s pixels apart, vertically and horizontally. This changes the certification threshold Δ to be, $\Delta_{block+stride} = \lceil (m + s - 1)/s \rceil^2$. With dropping fully masked tokens and using a stride of 10, a smoothed ViT-S using an ablation size of 75 is only 2.8x slower than a standard (non-robust) ResNet-50.

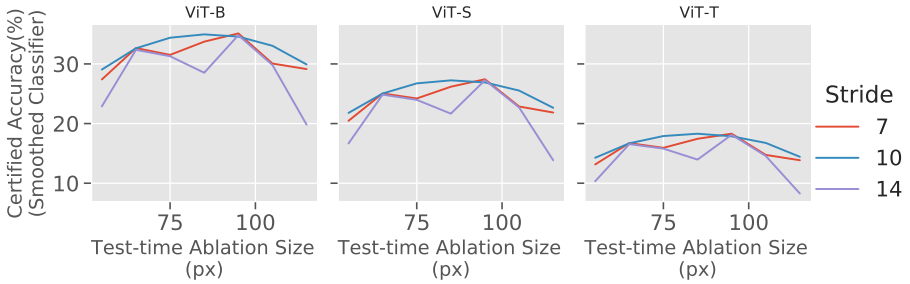
Certified accuracy. We find that, despite an systematic search over stride length and block size (both at training and evaluation), block smoothing on ImageNet remains significantly worse than column smoothing. For example, with optimal stride and ablation size, we see up to 5% lower certified accuracy than column smoothing on the largest model, ViT-B. We checked a range of ablation sizes from 55 to 115 as well as three stride lengths $\{7, 10, 14\}$ (Figure 15).

Similar to striding with column ablations, there is a significant amount of variability based on the stride length. To pinpoint the effect of striding, we certify one of the best-performing block sizes ($b = 75$) over a full range of strides from $s = 1$ to $s = 20$ (Figure 16). This is a fairly expensive calculation, as using stride $s = 1$ corresponds to the full block ablation with 50k ablations.

Even when using all possible block ablations ($s = 1$), block smoothing does not improve over column smoothing. However, we do find that certain stride lengths ($s = 18$) can achieve similar performance to non-strided block ablations ($s = 1$),



(a) We fix the test-time ablation size at $b = 75$ and plot the certified accuracy as a function of the adversarial patch size, for various stride length.



(b) We fix the adversarial patch size $m = 32$ and plot the certified accuracy as a function of the test-time ablation size, for various stride length.

Figure 15. Strided block smoothing on ImageNet for a collection of ViT models trained with block ablations of size $b = 75$.

which means that we can speed up certification (by 18x) without sacrificing certified accuracy. Thus, while our methods can make block smoothing computationally feasible, further investigation is needed to make block smoothing match column smoothing in terms of certified and standard accuracies.

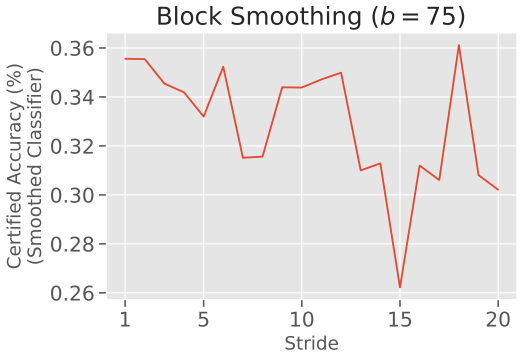


Figure 16. Strided block smoothing on ImageNet for ViT-B with a fixed ablation size $b = 75$. The reported certified accuracy are against adversarial patches of size 32×32. Note how some stride lengths ($s = 18$ for example) can achieve similar performance to non-strided block ablations ($s = 1$).

H. Extended experimental results

Table 6. **An extended version of Table 1.** Summary of our ImageNet results and comparisons to certified patch defenses from the literature: Clipped Bagnet (CBN), 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.

Standard and Certified Accuracy on ImageNet (%)					
Patch Size	Clean	1% pixels	2% pixels	3% pixels	Time (sec)
CBN [63]	49.5	13.4	7.1	3.1	3.05 ± 0.02
DS [25]	44.4	17.7	14.0	11.2	149.52 ± 0.33
PG [56] (1% pixels)	55.1	32.3	0.0	0.0	3.05 ± 0.02
PG [56] (2% pixels)	54.6	26.0	26.0	0.0	3.05 ± 0.02
PG [56] (3% pixels)	54.1	19.7	19.7	19.7	3.05 ± 0.02
Vary Ablation Size (Stride = 1)					
ResNet-18 (b = 19)	50.6	24.1	19.8	16.9	39.84 ± 0.97
ResNet-18 (b = 25)	52.7	24.2	20.0	17.1	39.84 ± 0.97
ResNet-18 (b = 37)	54.3	22.4	18.6	15.7	39.84 ± 0.97
ViT-T (b = 19)	52.3	27.3	22.9	19.9	6.81 ± 0.05
ViT-T (b = 25)	53.7	26.9	22.8	19.7	6.82 ± 0.05
ViT-T (b = 37)	55.6	25.5	21.7	18.8	12.64 ± 0.10
ResNet-50 (b = 19)	51.5	22.8	18.3	15.3	149.52 ± 0.33
ResNet-50 (b = 25)	54.7	23.8	19.5	16.4	149.52 ± 0.33
ResNet-50 (b = 37)	57.8	23.1	19.0	16.1	149.52 ± 0.33
ViT-S (b = 19)	63.5	36.8	31.6	27.9	14.00 ± 0.16
ViT-S (b = 25)	65.1	36.8	31.9	28.2	20.58 ± 0.18
ViT-S (b = 37)	67.1	35.3	30.7	27.1	20.61 ± 0.16
WRN-101-2 (b = 19)	61.4	33.3	28.1	24.1	694.50 ± 0.58
WRN-101-2 (b = 25)	64.2	34.3	29.1	25.3	694.50 ± 0.58
WRN-101-2 (b = 37)	67.2	33.7	28.8	25.2	694.50 ± 0.58
ViT-B (b = 19)	69.3	43.8	38.3	34.3	31.51 ± 0.17
ViT-B (b = 25)	71.1	44.0	38.8	34.8	31.52 ± 0.21
ViT-B (b = 37)	73.2	43.0	38.2	34.1	58.74 ± 0.17
Vary Ablation Stride					
WRN-101-2 (b = 19, s = 5)	61.1	30.1	27.3	21.9	138.90 ± 0.12
WRN-101-2 (b = 19, s = 10)	59.7	25.8	25.8	20.9	69.45 ± 0.06
ViT-B (b = 19, s = 5)	69.0	40.6	37.7	32.0	6.30 ± 0.03
ViT-B (b = 19, s = 10)	68.3	36.9	36.9	31.4	3.15 ± 0.02
WRN-101-2 (b = 37, s = 5)	66.9	32.6	27.2	24.7	138.90 ± 0.12
WRN-101-2 (b = 37, s = 10)	66.1	31.7	26.7	21.7	69.45 ± 0.06
ViT-B (b = 37, s = 5)	73.1	41.9	36.4	33.5	11.75 ± 0.03
ViT-B (b = 37, s = 10)	72.6	41.3	36.1	30.8	5.87 ± 0.02

Table 7. **An extended version of 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). b is the column ablation size out of 32 pixels.

Standard and Certified Accuracy on CIFAR-10 (%)			
Patch Size	Clean	2×2	4×4
Baselines			
CBN [63]	84.2	44.2	9.3
DS [25]	83.9	68.9	56.2
PG [56] (2×2)	84.7	69.2	0.0
PG [56] (4×4)	84.6	57.7	57.7
Smoothed models			
ResNet-18 (b = 4)	83.6	67.0	54.2
ViT-T (b = 4)	**85.5**	**70.0**	**58.5**
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**

Table 8. Standard accuracies of regularly trained architectures vs. smoothed architectures with column ablations of size $b = 4$ for CIFAR-10 and $b = 19$ for ImageNet.

		Standard accuracy of architecture (%)					
		ViT-T	ResNet-18	ViT-S	ResNet-50	ViT-B	WRN-101-2
ImageNet	Standard	72.03	69.76	79.72	76.13	81.74	78.85
	Smoothed	52.25	50.62	63.48	51.47	69.33	61.38
	Difference	19.77	19.14	16.24	24.66	12.41	17.47
CIFAR-10	Standard	93.13	95.72	93.33	96.16	97.07	97.85
	Smoothed	85.53	88.41	86.39	83.57	90.75	88.20
	Difference	7.60	7.31	6.94	12.59	6.32	9.65

Table 9. ImageNet certified models trained on ablations of size 19, with a variety of test-time ablations sizes b and stride lengths s.

Architecture	s	b	Standard accuracy(%)	Certified Accuracy(%) 1% pixels	2% pixels	3% pixels
ResNet-18	1	19	50.6	24.1	19.8	16.9
		25	52.7	24.2	20.0	17.1
		37	54.3	22.4	18.6	15.7
	5	19	50.3	21.5	19.3	15.3
		25	52.4	22.1	17.9	15.8
		37	54.2	21.5	17.4	15.4
	10	19	49.3	18.7	18.7	14.8
		25	51.5	21.5	17.3	13.6
		37	53.7	21.0	17.1	13.5
ViT-T	1	19	52.3	**27.3**	**22.9**	**19.9**
		25	53.7	26.9	22.8	19.7
		37	**55.6**	25.5	21.7	18.8
	5	19	51.9	24.6	22.4	18.2
		25	53.5	25.1	20.6	18.5
		37	55.4	24.7	20.5	18.5
	10	19	51.2	21.8	21.8	17.8
		25	53.1	24.6	20.4	16.4
		37	55.1	24.4	20.3	16.5
ResNet-50	1	19	51.5	22.8	18.3	15.3
		25	54.7	23.8	19.5	16.4
		37	57.8	23.1	19.0	16.1
	5	19	51.0	20.1	17.9	13.6
		25	54.5	21.7	17.2	15.1
		37	57.7	22.1	17.7	15.8
	10	19	49.9	17.2	17.2	13.2
		25	53.7	21.0	16.7	12.8
		37	57.1	21.7	17.6	13.7
ViT-S	1	19	63.5	**36.8**	31.6	27.9
		25	65.1	**36.8**	**31.9**	**28.2**
		37	**67.1**	35.3	30.7	27.1
	5	19	63.1	33.8	31.1	25.7
		25	64.9	34.4	29.3	26.7
		37	67.0	34.3	29.1	26.7
	10	19	62.2	30.3	30.3	25.2
		25	64.3	33.9	28.7	23.7
		37	66.5	33.8	29.0	24.2
WRN-101	1	19	61.4	33.3	28.1	24.1
		25	64.2	34.3	29.1	25.3
		37	67.2	33.7	28.8	25.2
	5	19	61.1	30.1	27.3	21.9
		25	63.8	31.8	26.3	23.7
		37	66.9	32.6	27.2	24.7
	10	19	59.7	25.8	25.8	20.9
		25	62.7	30.5	25.3	20.5
		37	66.1	31.7	26.7	21.7
ViT-B	1	19	69.3	43.8	38.3	34.3
		25	71.1	**44.0**	**38.8**	**34.8**
		37	**73.2**	43.0	38.2	34.1
	5	19	69.0	40.6	37.7	32.0
		25	70.8	41.6	36.0	33.0
		37	73.1	41.9	36.4	33.5
	10	19	68.3	36.9	36.9	31.4
		25	70.3	40.9	35.2	29.8
		37	72.6	41.3	36.1	30.8