From Code to Reality: A Study on Certified Patch Robustness

Vanshika Jain
Department of CVMLA
Northeastern University
Boston, MA 02120
jain.van@northeastern.edu

Abstract

This study primarily explores the robustness within the realm of adversarial patches, encompassing diverse modifications confined to a small, contiguous area. While ensuring model robustness with formal guarantees can be intriguing, the more captivating aspect lies in verifying models throughout their entire lifecycle, from training to deployment. I certified CNN and ViT models on CIFAR-10, confirming improved guarantees. Subsequent adversarial patch tests aligned closely with reported results. Extending to real-world datasets, I explored patch robustness with a focus on chair images, yielding unexpected insights and making way for practical applications. In a physical-world test, printed images subjected to patches revealed less favorable outcomes, yet the experiment needs more evaluation for robustness verification.

1 Introduction

While neural networks have excelled in tasks like image classification and speech recognition, they display high susceptibility to minor, adversarially chosen alterations in their inputs. The field of machine learning has witnessed a growing fascination with adversarial attacks in recent years. It's noteworthy that in numerous cases, the threat models associated with these attacks implicitly assume that the attacker possesses the capability to directly manipulate the input fed into a neural network.

In critical situations, the priority is to create models that are consistently robust, guaranteeing resistance to diverse transformations. This need has practical implications in fields such as aviation, where ensuring the safety of aircraft controllers in the presence of nearby planes is essential, and in automotive systems, where stability must be maintained despite sensor noise.

The focus of this study is centered on examining the robustness within the domain of adversarial patches, which involves a varied set of alterations localized to a small, contiguous area. This category encompasses a wide range, incorporating intentionally designed physical objects like adversarial glasses, stickers, and clothing with malicious intent. Adversarial patches have been utilized to mislead image classifiers, manipulate object detectors, and interfere with optical flow estimation.

Evaluating defenses against adversarial patches poses challenges as recent findings revealed weaknesses in empirical defenses against more potent adaptive attacks. This difficulty prompted the creation of certified defenses, providing models with provable robustness without depending on empirical evaluations. Nevertheless, certified guarantees often come with trade-offs, such as a significant drop in standard accuracy and longer inference times.

The question at stake is this: Does certified robustness really need to come at such a high price? Recent research despite being top-performing, sacrifices 30% of standard accuracy and experiences a two-order-of-magnitude increase in inference time. Moreover, it achieves only 13.9% robust accuracy

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Table 1: Experiments

S.No.	Description	Results
1	Train Models with Certification for ResNet-18 and ViT-T	Improved Certified Accuracy
2	Attack and Certify Robustness on CIFAR-10 and ImageNet	Almost matched the reported results
3	Certified Patch Robustness in action on real-world dataset	Surprising results
4	Certify real-world Patch Attack on printed patch	Interesting experiment

on ImageNet against patches occupying 2% of the image. While these drawbacks are commonly acknowledged as the price of certification, they significantly constrain the practicality of certified defenses.

1.1 Main Contribution

This paper is a study on certified adversarial patch robustness. In general, the main objective of this study is to understand why a model makes a particular prediction given the inherent complexity of interpreting machine learning models. In the domain of Verifiable Machine Learning, people are always coming up with new adversarial attacks and defenses where we intentionally manipulate input data to mislead a model and develop robust defenses against such attacks. While ensuring model robustness through formal guarantees can be engaging, the more intriguing aspect lies in verifying models across their entire lifecycle, spanning from training to deployment.

To achieve my objectives, I conducted a series of experiments 1. Initially, I trained Convolutional Neural Network (CNN) and Vision Transformer (ViT) architectures with certification, observing improved guarantees on the CIFAR-10 dataset. Subsequently, I subjected these robust models to adversarial patch attacks to validate their performance against the promised robustness. The obtained results closely aligned with the reported numbers, with further details to be discussed later.

Moving forward, I sought to extend the certification of patch robustness to real-world datasets, specifically focusing on images of chairs, yielding some unexpected and noteworthy results. For the final experiment, I aimed to validate the robust model in a physical-world setting. I printed out images that had been subjected to patch attacks and tested the model's performance on them. The outcomes were not as favorable as expected, yet the entire experiment proved to be captivating and insightful.

2 Certified Patch Robustness via Derandomized Smoothing

Certified defenses often fall under the category of smoothing methods, encompassing techniques that combine a classifier's predictions across diverse input variations to generate predictions that can be reliably certified as robust. An exemplar within this category is derandomized smoothing which excels in achieving robustness against adversarial patches. This method involves aggregating a classifier's predictions based on various image ablations that effectively mask a substantial portion of the image. These approaches typically use CNNs, a common default model for computer vision tasks, to evaluate the image ablations. We initiate our approach by questioning the suitability of convolutional architectures for the task at hand. The core of our methodology involves harnessing the capabilities of vision transformers, as we illustrate their superior ability to adeptly manage the image ablations inherent 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]. In this study, the focus is primarily on column ablations which mask out the entire image except for a column of a fixed width. For an input $h\ x\ w$ sized image x, we denote by $S_b(x)$ 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 $S_b(x)$.

Derandomized smoothing Derandomized smoothing [25] is a popular approach for certified patch defenses that constructs a smoothed classifier comprising two main components: (1) a base

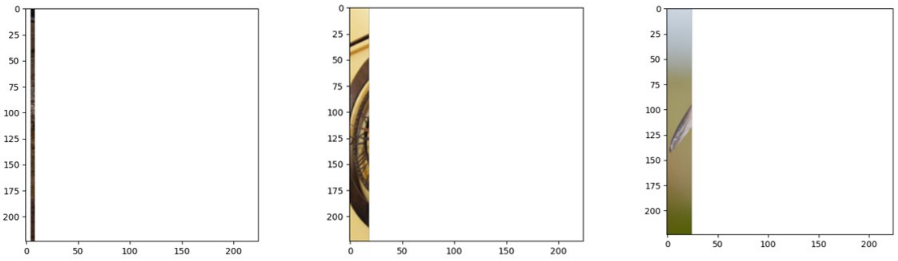


Figure 1: Different image ablation size for column type ablation from ablation size 4, 19, and 25.

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 $S_b(x)$. Specifically, for an input image x, ablation set $S_b(x)$, and a base classifier *f*, a smoothed classifier *g* is defined as:

$$g(x) = \arg\max_{c} n_c(x)$$

where

$$n_c(x) = \sum_{x' \in S_b(x)} \|\{f(x') = c\}$$

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 $S_b(x)$ that an adversarial patch can simultaneously intersect (e.g., for column ablations of size b, an *m x m* patch can intersect with at most $\Delta = m + b - 1$ ablations). Then, a smoothed classifier is certifiably robust on an input x if it is the case that for the predicted class *c*:

$$n_c(x) > \max_{c' \neq c} n'_c(x) + 2\Delta$$

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 as certified accuracy.

Vision transformers A key component of this approach is the vision transformer (ViT) architecture. In contrast to convolutional architectures, ViTs use self-attention layers instead of convolutional layers as their primary building block. ViTs process images in three main stages:
1. *Tokenization*: The ViTs split the image into p x p patches. Each patch is then embedded into a positionally encoded token.
2. *Self-Attention*: The set of tokens is then passed through a series of multi-headed self-attention layers.
3. *Classification head*: The resulting representation is fed into a fully connected layer to make predictions for classification.

Smoothed Vision Transformer 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.

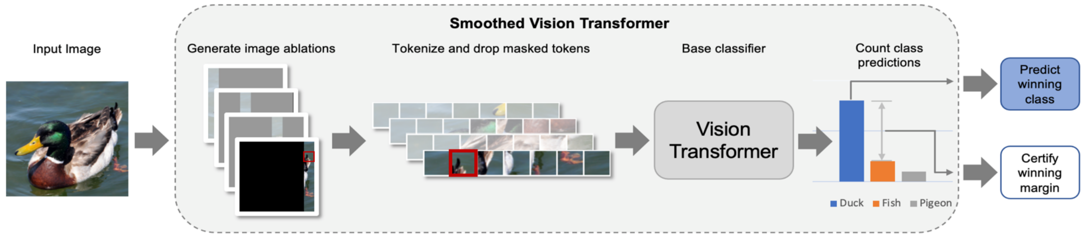


Figure 2: Illustration of the smoothed vision transformer. For a given image, first, generate a set of ablations, encode each ablation into tokens, and drop fully masked tokens. The remaining tokens for each ablation are then fed into a vision transformer, which predicts a class label for each ablation.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

Moreover, unlike convolutions which operate locally, the self-attention mechanism in ViTs shares information globally at every layer. 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, this methodology leverages the ViT architecture as the base classifier for processing the image ablations used in derandomized smoothing. The ViT architecture is modified for the smoothing procedure to drastically speed up the cost of inference of a smoothed ViT. The overview of this approach is presented in figure 2.1.

3 Experiments

3.1 Train models with certification for ResNet-18 and ViT-T

2 is the result table for the first experiment on two different architectures, CNN and ViT, on the CIFAR-10 dataset with ablation size 4 and patch size 4. From the table, the training time for ViT is half of that of CNN because of the fact that ViT works on tokenization, and its easy to drop unnecessary tokens for speedup. The certification accuracy improved by a small margin of 6% on CNN and 4% on ViT.

Certified Defense The mechanism behind adversarial training involves identifying an input that results in maximum loss to achieve high accuracy, which is easier to optimize but poses challenges for robustness verification. In contrast, certified defense aims to find an output that attains maximum loss, excelling in certification but occasionally leading to lower accuracy and potentially presenting challenges for optimization.

Keeping this in consideration, the robustness-focused models that I developed were pre-trained using the ImageNet dataset. They were then directly trained on image ablations, specifically of the column type, with a fixed ablation size of 4 for CIFAR-10 and 19 for ImageNet. As CIFAR-10 has a resolution of 32x32, it is upsampled to an image size of 224. Importantly, it's crucial to note that this training approach does not involve adversarial training, meaning there are no adversarial attacks incorporated during the training process.

Results As outlined in the original paper, the use of CNNs for robust models and certification is discouraged due to a significant reported difference in certified accuracy between these architectures. Despite this, I had a suspicion that CNNs might have untapped potential. Through extended training epochs, the robust accuracy improved to 61% for CNNs and 62% for Vision Transformers (ViT). Thus, almost matching their robust accuracy.

Outcome The final outcome of this experiment is that models are data-hungry. As said by Schmidt et. al. more data helps provide a lower bound for adversarial robustness. (Provides lower bound on number of samples needed to achieve adversarial robustness, Schmidt et. al.)

Table 2: Exp 1 – Train models with certification for CIFAR-10 dataset with ablation size and patch size 4. The results for both architectures are in terms of accuracy (standard, smoothed and certified) with improved certified robustness by 6% on ResNet-18 and 4% on ViT-T

Arch. (params)	Time	Std.	Smoothed	Certified
ResNet-18 (12M)	1h	82.13	84.22	**61.32** (+6)
ViT-T (5M)	0.5h	85.62	85.07	**62.70** (+4)

Table 3: Patch Attack parameters and values as taken from [1]. Here c is CIFAR-10 and i is referred to ImageNet Dataset.

Attack Parameter	Value
Epsilon	1.0
Random start	True
Attack steps	150
Step size	0.05
Column size	4 (c) / 19 (i)
Patch size	4 (c) / 16 (i)

3.2 Attack and certify robustness on CIFAR-10 and ImageNet

Patch Attack The patch attack that is incorporated for this study is taken from [1] by Levine and Feizi where they introduce a certifiable defense against patch attacks that guarantees robustness for a given image and patch attack size such that no patch adversarial examples exist. Their method is related to the broad class of randomized smoothing robustness schemes which provide high confidence probabilistic robustness certificates. By exploiting the fact that patch attacks are more constrained than general sparse attacks, they can derive meaningfully large robustness certificates against them.

The development of physical adversarial attacks, in which small visible changes are made to real-world objects in order to disrupt the classification of images of these objects, represents a more concerning security threat. Patch attack is a type of physical adversarial attack which are more constrained than general sparse attacks. This attack is a special case of L0 (sparse) adversarial attacks where the adversary can choose a limited number of pixels and apply unbounded distortions to them. This type of attack is constrained to selecting only a block of adjacent pixels to attack, rather than any arbitrary pixels.

There are two types of such attacks: the universal patch attack is an effective physical sticker attack. The attack method is universal in the sense that pixels of the adversarial patch do not depend on the attacked image. Image-specific patch attacks have also been proposed, such as LaVAN, which reduces ImageNet classification accuracy to 0% using only a 42 x 42 pixel square patch (on images of size 299 x 299). In [1], they considered all attacks (image-specific or universal) on square patches of size m x m.

Certification The certifiably robust classification scheme is based on randomized smoothing, a class of certifiably robust classifiers that have been proposed for various threat models, including L2, L1, and L0, and Wasserstein metrics. All of these methods rely on a similar mechanism where noisy versions of an input image x are used in the classification. Such noisy inputs are created either by adding random noise to all pixels or by removing (ablating) some of the pixels.

A large number of noisy images are then classified by a base classifier and then the consensus of these classifications is reported as the final classification result. For an adversarial image x0 at a bounded distance from x, the probability distributions of possible noisy images which can be produced from x and x0 will substantially overlap. This implies that, if a sufficiently large fraction of noisy images derived from x are classified to some class c, then with high confidence, a plurality of noisy images derived from x0 will also be assigned to this class.

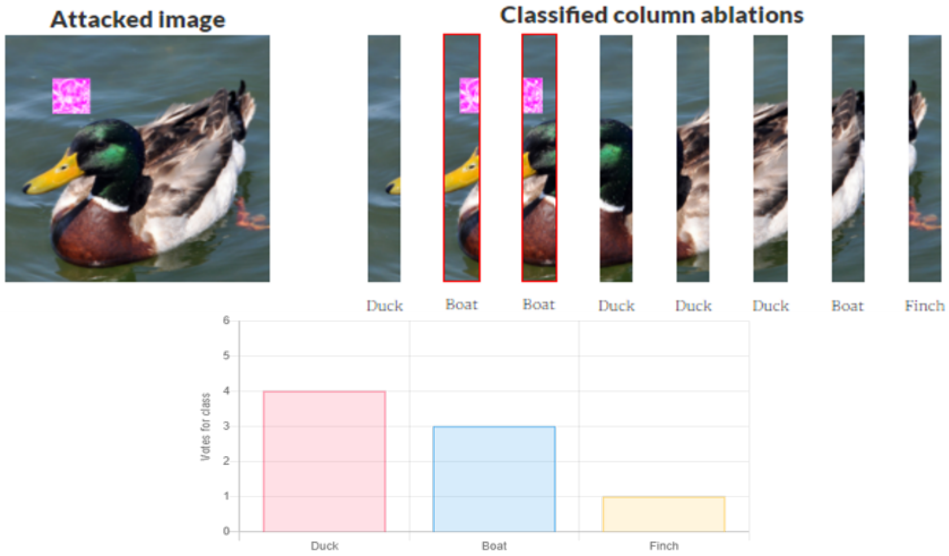


Figure 3: Exp 2 - A simplified illustration of the derandomized smoothing defense for patch robustness is presented here. In this example, an image of a duck is divided into column ablations, each classified individually. As 6 out of 8 column ablations are identified as a duck, the smoothed model predicts the overall image as a duck. Since a 32x32 patch can only impact a maximum of 2 column ablations, it is insufficient to alter the prediction. Therefore, this prediction is certifiably robust against 32x32 patches. Conversely, a 64x64 patch can influence up to 3 column ablations, potentially causing the most frequent class to change to a boat. Consequently, the model is not certifiably robust against 64x64 patches.

The defense method is based on structured ablation where instead of randomized ablation which would randomly select pixels to use for classification, these column pixels are selected in a correlated way which intuitively reduces the probability that the adversarial patch is sampled.

By reducing the total number of possible ablations of an image, structured ablation allows us to de-randomize the algorithm, yielding improved, deterministic certificates. It is not feasible to evaluate precisely the probability that f(x) returns any particular class c: one must estimate this based on random samples.

Using the derandomized method, the number of possible ablations is small enough so that it is tractable to classify using all possible ablations: we can exactly evaluate the probability that f(x) returns each class.

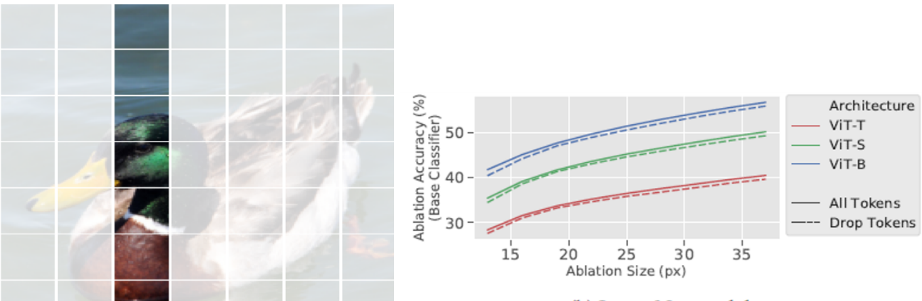
This certificate is therefore exact, rather than probabilistic, so the classifications are provably robust in an absolute sense. Determinism provides another benefit: the absence of estimation error increases the certified accuracies that can be reported. Additionally, because estimation error is no longer a concern, derandomization allows us to use more rich information from the base classifier without incurring an additional cost in increased estimation error.

We take advantage of this to allow the base classifier to abstain in cases where it cannot make a high-confidence prediction towards any class. Delta is the maximum number of ablations in the ablation set that an adversarial patch can simultaneously intersect. For instance, for a patch size of 4 and ablation size 4, delta is 7. So, there is a possibility that a 4 x 4 patch can intersect 7 ablations of an image.

Drop Tokens in ViT Randomized smoothed classifier is slow as it has to classify for each input variation but deterministic smoothing reduced the number of ablations and the inference time. Vits have an inherent advantage as they process images as tokens (specifically 16 x 16 patch) and less the number of tokens, less run time. So, the idea is to drop the masked-out tokens.

Table 4: Exp 2 – Attack and certification results for ViT-T model with CIFAR-10 and ImageNet dataset for patch size of 4 and 16 respectively.

Model	Time	Images	Drop Tokens	Std.	Smoothed	Certified	Ablation	Delta
ViT-T (c4)	9	1400	True	85.1	38.6	33.6	70.0	-19
ViT-T (c4)	5	900	False	86.1	38.6	36.4	76.3	-19
ViT-T (i16)	1.7	500	True	53.3	52.3	**26.5** (-2.7)	63.2 (-4.3)	20
ViT-T (i16)	1.7	500	False	57.8	52.8	**30.6** (-1.4)	70.5 (-3.7)	23



(b) ImageNet models.

Figure 4: Different image ablation size for column type ablation from ablation size 4, 19, and 25.

In the paper, they point out that there is not much difference in the accuracy when tokens are dropped but I found a gap of 3-4% on certification accuracy and 6-7% on ablation accuracy. So, I think the concept of drop tokens has not been addressed properly.

Outcome The final outcome of this experiment is unclear. The issue lies in either the potency of the attack being excessively strong or a flaw in the reported defense mechanism [2]. Specifically, the failure to properly attend to drop tokens is a critical concern.

3.3 Certified patch robustness in action on real-world dataset

To verify my model's patch robustness on real-world dataset for my next experiment, I tested my model against a dataset that I collected in the robotics lab as part of another project. At first, I tested a handful of cropped images of chairs and got obviously poor results. So, I fine-tuned my model and tested it again to get 80% certified robustness on 16 patch size.

Evaluation on CIFAR-100 In the initial phase of this experiment, Vit-T was trained on the CIFAR-100 dataset, adhering to standard procedures. This choice was made because ImageNet lacks an explicit "chair" class; instead, it includes classes like "barber chair" and "rocking chair." The obtained low accuracy for clean data initially led to a sense of discouragement and a belief that the practicality of the model might be limited after all. Clean accuracy of 2% on the cropped chair dataset was not good enough to further test robustness.

Fine Tuning on chair dataset However, given the abundance of video data collected, I decided to create a training and test set to fine-tune my model. To preprocess these camera frames, I employed a FastRCNN object detector and cropped all images containing chairs using the corresponding bounding box coordinates.

To avoid overfitting solely to chairs, I implemented a 0.6 split and trained the model for 10 epochs, achieving a commendable 79% accuracy on clean data in the test set. Using this model, I assessed certified accuracy on both 16-patch and 32-patch scenarios, both of which demonstrated satisfactory performance.

While this experiment may seem biased by focusing on a single class, my main objective was to establish that careful training can ensure robustness on real-world data. This highlights the potential deployment of the model for practical image classification applications.

Table 5: Exp 3 - Certification on real-world dataset taken from Intel RealSense Camera, fine-tuned on chairs produces certified accuracy to be 80% on 16 patch size.

Model	Images	Time	Patch size	Std.	Smoothed	Certified
ViT-T + C100	70	-	-	2	-	-
ViT-T + C100 + chair	500	1.7	16	78.0	97.0	**80.0**
ViT-T + C100 + chair	500	1.7	32	79.0	97.0	**46.0**

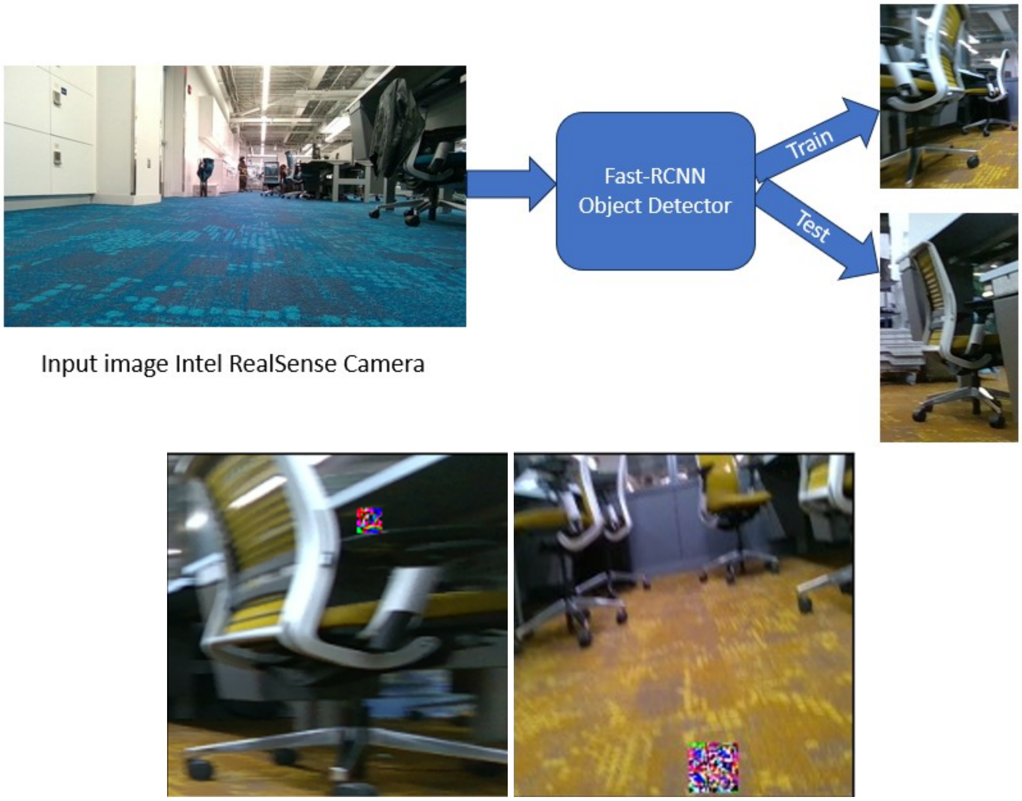


Figure 5: Setup for Exp 3 to create a train and test set of chairs dataset. Some of the examples with patch attack of 16 and 32 patch size.

Outcome The end goal of this experiment is that this method ensures robustness on real-world datasets, making it suitable for practical applications.

3.4 Certify real-world patch attack on printed patch

In my conclusive experiment, I aimed to validate robustness in real-world settings. I randomly selected 50 images with different labels from ImageNet and assessed the robustness of the Vit-T model against a 16-patch attack, achieving a 50% accuracy on my machine. Subsequently, I scanned printed images, both raw and attacked, using my laptop camera, resulting in an accuracy of 20%.

It became evident that printing out an adversarial attack is not a straightforward process, as factors such as color differences from the printer and lighting effects during scanning were not initially considered. These discrepancies in numbers are reasonable, and the primary goal of this verification was to observe how an attack would perform in real-world scenarios. However, additional experiments are necessary to thoroughly confirm proper robustness.

Outcome The strength of both the attack and certification in real-world settings requires further experimentation to validate robustness thoroughly.

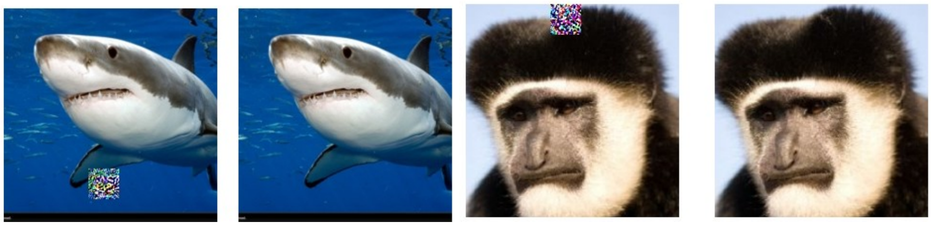


Figure 6: Exp 4 - Physical world settings. Some example images of patch attacked image with patch size 16

4 Related Work

Certified defenses Extensive research has explored the development of certified or provable defenses against adversarial perturbations, categorizing this line of study into tighter or exact verifiers [1] and smoothing-based defenses [2]. In the context of patches, the earliest certified defense employed convex relaxation (interval bounds) to provide provable guarantees against adversarial patches. Subsequent research [1] has concentrated on randomized smoothing, a technique that smoothens classifiers over random noise but is typically significantly more computationally expensive (4-5 orders of magnitude slower) than standard, non-robust models.

Deterministic smoothing To address the computational cost associated with the lengthy inference times of randomized smoothing, [1] introduced derandomized smoothing. This method employs a finite set of ablations to smooth a base classifier, significantly reducing computational requirements. However, it remains two orders of magnitude slower than standard models. Other defenses, such as Clipped BagNet and PatchGuard, adopt an approach of restricting the model's receptive field to enhance efficiency. While these methods are faster than derandomized smoothing, they come with certain limitations. Clipped BagNet (CBN) offers substantially weaker robustness guarantees compared to derandomized smoothing. On the other hand, PatchGuard provides higher but brittle guarantees, defended models are optimally protected against a specific patch size and exhibit no robustness against patches even slightly larger than the considered size.

Vision transformers This study utilizes the vision transformer (ViT) architecture, which transforms the widely used attention-based model from the language domain to the vision domain. Recent advancements include more efficient training methods and the availability of pre-trained ViTs, making these architectures more accessible to the broader research community.

5 Conclusion

This study demonstrates the substantial enhancement of certified robustness against adversarial patches by incorporating visual transformers (ViTs) into the smoothing framework. This improvement is achieved while maintaining standard accuracies comparable to those of regular (non-robust) models. Additionally, the introduction of modifications to the ViT architecture and the associated smoothing procedure results in significantly faster inference times. These advancements position models that are certifiably robust to adversarial patches as a practical alternative to standard (non-robust) models. This study also exemplifies the application of this technology in real-world scenarios, specifically on datasets involving chairs and by assessing performance in physical-world settings through the printing of posters.

Limitations Similar to other certified defenses, this method specifically addresses patch attacks and doesn't assure robustness against attacks beyond this threat model. Although faster than other smoothed models, smoothed ViTs are still slightly slower than standard (non-robust) models. The standard accuracy may suffer if the predictive signal relies on a small image region, potentially absent in many ablations. A potential drawback is the risk of instilling overconfidence in the model. While robustness guarantees ensure stable predictions at test time, correctness is not guaranteed, urging users to acknowledge these nuances before employing the technique.

References

[1] Alexander Levine & Soheil Feizi (2020) (De)Randomized Smoothing for Certifiable Defense against Patch Attacks, *NeurIPS*.

[2] Hadi Salman, Saachi Jain, Eric Wong, & Aleksander Madry (2021) Certified Patch Robustness via Smoothed Vision Transformers. ArXiv preprint arXiv:2110.07719.

[3] Hadi Salman, Greg Yang, Jerry Li, Pengchuan Zhang, Huan Zhang, Ilya Razenshteyn, & Sebastien Bubeck (2020) Provably robust deep learning via adversarially trained smoothed classifiers.

[4] PingYeh Chiang, Renkun Ni, Ahmed Abdelkader, Chen Zhu, Christoph Studor, & Tom Goldstein (2020) Certified defenses for adversarial patches. *In International Conference on Learning Representations.*