# Are Vision Transformers Robust to Patch Perturbations?

Jindong Gu[1], Volker Tresp[1], and Yao Qin[2]

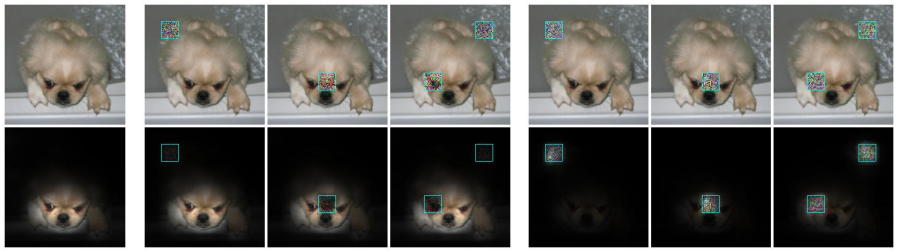
[1] University of Munich [2] Google Research

**Abstract.** Recent advances in Vision Transformer (ViT) have demonstrated its impressive performance in image classification, which makes it a promising alternative to Convolutional Neural Network (CNN). Unlike CNNs, ViT represents an input image as a sequence of image patches. The patch-based input image representation makes the following question interesting: How does ViT perform when individual input image patches are perturbed with natural corruptions or adversarial perturbations, compared to CNNs? In this work, we study the robustness of ViT to patch-wise perturbations. Surprisingly, we find that ViTs are more robust to naturally corrupted patches than CNNs, whereas they are more vulnerable to adversarial patches. Furthermore, we discover that the attention mechanism greatly affects the robustness of vision transformers. Specifically, the attention module can help improve the robustness of ViT by effectively ignoring natural corrupted patches. However, when ViTs are attacked by an adversary, the attention mechanism can be easily fooled to focus more on the adversarially perturbed patches and cause a mistake. Based on our analysis, we propose a simple temperature-scaling based method to improve the robustness of ViT against adversarial patches. Extensive qualitative and quantitative experiments are performed to support our findings, understanding, and improvement of ViT robustness to patch-wise perturbations across a set of transformer-based architectures.

**Keywords:** Understanding Vision Transformer, Adversarial Robustness

## 1 Introduction

Recently, Vision Transformer (ViT) has demonstrated impressive performance [10,47,49,50,14,8,15,7,25], which makes it become a potential alternative to convolutional neural networks (CNNs). Meanwhile, the robustness of ViT has also received great attention [5,20,38,39,41,42,45]. On the one hand, it is important to improve its robustness for safe deployment in the real world. On the other hand, diagnosing the vulnerability of ViT can also give us a deeper understanding of its underlying working mechanisms. Existing works have intensively studied the robustness of ViT and CNNs when the whole input image is perturbed with natural corruptions or adversarial perturbations [5,41,28,3,2]. Unlike CNNs, ViT processes the input image as a sequence of image patches. Then, a self-attention



(a) Clean Image (b) with Naturally Corrupted Patch (c) with Adversarial Patch

Fig. 1: Images with patch-wise perturbations (top) and their corresponding attention maps (bottom). The attention mechanism in ViT can effectively ignore the naturally corrupted patches to maintain a correct prediction in Fig. b, whereas it is forced to focus on the adversarial patches to make a mistake in Fig. c. The images with corrupted patches (Fig. b) are all correctly classified. The images with adversary patches (Fig. c) are misclassified as *dragonfly*, *axolotl*, and *lampshade*, respectively.

mechanism is applied to aggregate information from all patches. Based on the special patch-based architecture of ViT, we mainly focus on studying the robustness of ViT to patch-wise perturbations.

In this work, two typical types of perturbations are considered to compare the robustness between ViTs and CNN (e.g., ResNets [16]). One is natural corruptions [17], which is to test models' robustness under distributional shift. The other is adversarial perturbations [44,13], which are created by an adversary to specifically fool a model to make a wrong prediction. Surprisingly, we find ViT does *not always* perform more robustly than ResNet. When individual image patches are naturally corrupted, ViT is more robust compared to ResNet. However, when input image patch(s) are adversarially attacked, ViT shows a higher vulnerability than ResNet.

Digging down further, we revealed that ViT's stronger robustness to natural corrupted patches and higher vulnerability against adversarial patches are both caused by the attention mechanism. Specifically, the self-attention mechanism of ViT can effectively ignore the natural patch corruption, while it's also easy to manipulate the self-attention mechanism to focus on an adversarial patch. This is well supported by rollout attention visualization [1] on ViT. As shown in Fig. 1 (a), ViT successfully attends to the class-relevant features on the clean image, *i.e.*, the head of the dog. When one or more patches are perturbed with natural corruptions, shown in Fig. 1 (b), ViT can effectively ignore the corrupted patches and still focus on the main foreground to make a correct prediction. In Fig. 1 (b), the attention weights on the positions of naturally corrupted patches are much smaller even when the patches appear on the foreground. In contrast, when the patches are perturbed with adversarial perturbations by an adversary, ViT is successfully fooled to make a wrong prediction, as shown in Fig. 1 (c).

This is because the attention of ViT is misled to focus on the adversarial patch instead.

Based on this understanding that the attention mechanism leads to the vulnerability of ViT against adversarial patches, we propose a simple Smoothed Attention to discourage the attention mechanism to a single patch. Specifically, we use a temperature to smooth the attention weights computed by a *softmax* operation in the attention. In this way, a single patch can hardly dominate patch embeddings in the next layer, which can effectively improve the robustness of ViT against adversarial patch attacks.

Our main contributions can be summarized as follows:

- **Finding**: Based on a fair comparison, we discover that ViT is more robust to natural patch corruption than ResNet, whereas it is more vulnerable to adversarial patch perturbation.
- **Understanding**: We reveal that the self-attention mechanism can effectively ignore natural corrupted patches to maintain a correct prediction but be easily fooled to focus on adversarial patches to make a mistake.
- **Improvement**: Inspired by our understanding, we propose Smoothed Attention, which can effectively improve the robustness of ViT against adversarial patches by discouraging the attention to a single patch.

## 2 Related Work

**Robustness of Vision Transformer.** The robustness of ViT have achieved great attention due to its great success [5,33,41,4,28,3,34,2,39,51,18,30,29,33,38]. On the one hand, [5,36] show that vision transformers are more robust to natural corruptions [17] compared to CNNs. On the other hand, [5,41,36] demonstrate that ViT achieves higher adversarial robustness than CNNs under adversarial attacks. These existing works, however, mainly focus on investigating the robustness of ViT when a whole image is naturally corrupted or adversarially perturbed. Instead, our work focuses on patch perturbation, given the patch-based architecture trait of ViT. The patch-based attack [20,12] and defense [32,42] methods have also been proposed recently. Different from their work, we aim to understand the robustness of patch-based architectures under patch-based natural corruption and adversarial patch perturbation.

**Adversarial Patch Attack.** The seminal work [35] shows that adversarial examples can be created by perturbing only a small amount of input pixels. Further, [6,24] successfully creates universal, robust, and targeted adversarial patches. These adversarial patches therein are often placed on the main object in the images. The works [11,31] shows that effective adversarial patches can be created without access to the target model. However, both universal patch attacks and black-box attacks are weak to be used for our study. They can only achieve very low fooling rates when a single patch of ViT (only 0.5% of image) is attacked. In contrast, the white-box attack [21,23,48,37,26] can fool models by attacking only a very small patch. In this work, we apply the most popular adversarial patch attack in [21] to both ViT and CNNs for our study.

Table 1: Comparison of popular ResNet and ViT models. The difference in model robustness can not be blindly attributed to the model architectures. It can be caused by different training settings. WS, GN and WD correspond to Weight Standardization, Group Normalization and Weight Decay, respectively.

| Model | Pretraining | DataAug | Input Size | WS | GN | WD |
|---|---|---|---|---|---|---|
| ResNet [16] | N | N | 224 | N | N | Y |
| BiT [22] | Y | N | 480 | Y | Y | N |
| ViT [10] | Y | N | 224/384 | N | N | N |
| DeiT [47] | N | Y | 224/384 | N | N | N |

## 3 Experimental Settings to Compare ViT and ResNet

**Fair Base Models.** We list the state-of-the-art ResNet and ViT models and part of their training settings in Tab. 1. The techniques applied to boost different models are different, *e.g.*, pretraining. A recent work [3] points out the necessity of a fair setting. Our investigation finds weight standardization and group normalization also have a significant impact on model robustness (More in Appendix A). This indicates that the difference in model robustness can not be blindly attributed to the model architectures if models are trained with different settings. Hence, we build fair models to compare ViT and ResNet as follows.

First, we follow [47] to choose two pairs of fair model architectures, DeiT-small vs. ResNet50 and DeiT-tiny vs. ResNet18. The two models of each pair (*i.e.* DeiT and its counter-part ResNet) are of similar model sizes. Further, we train ResNet50 and ResNet18 using the **exactly same setting** as DeiT-small and Deit-tiny in [47]. In this way, we make sure the two compared models, *e.g.*, DeiT-samll and ResNet50, have similar model sizes, use the same training techniques, and achieve similar test accuracy (See Appendix A). The two fair base model pairs are used across this paper for a fair comparison.

**Adversarial Patch Attack.** We now introduce adversarial patch attack [21] used in our study. The first step is to specify a patch position and replace the original pixel values of the patch with random initialized noise $\delta$. The second step is to update the noise to minimize the probability of ground-truth class, *i.e.* maximize the cross-entropy loss via multi-step gradient ascent [27]. The adversary patches are specified to align with input patches of DeiT.

**Evaluation Metric.** We use the standard metric **Fooling Rate (FR)** to evaluate the model robustness. First, we collect a set of images that are correctly classified by both models that we compare. The number of these collected images is denoted as $P$. When these images are perturbed with natural patch corruption or adversarial patch attack, we use $Q$ to denoted the number of images that are misclassified by the model. The Fooling Rate is then defined as $\text{FR} = \frac{Q}{P}$. The lower the FR is, the more robust the model is.

Table 2: Fooling Rates (in %) are reported. DeiT is more robust to naturally corrupted patches than ResNet, while it is significantly more vulnerable than ResNet against adversarial patches. Bold font is used to mark the lower fooling rate, which indicates the higher robustness.

| Model | # Naturally Corrupted Patches | | | | # Adversarial Patches | | | |
|---|---|---|---|---|---|---|---|---|
| | 32 | 96 | 160 | 196 | 1 | 2 | 3 | 4 |
| ResNet50 | 3.7 | 18.2 | 43.4 | 49.8 | **30.6** | **59.3** | **77.1** | **87.2** |
| DeiT-small | **1.8** | **7.4** | **22.1** | **38.9** | 61.5 | 95.4 | 99.9 | 100 |
| ResNet18 | 6.8 | 31.6 | 56.4 | 61.3 | **39.4** | **73.8** | **90.0** | **96.1** |
| DeiT-tiny | **6.4** | **14.6** | **35.8** | **55.9** | 63.3 | 95.8 | 99.9 | 100 |

## 4 ViT Robustness to Patch-wise Perturbations

Following the setting in [47], we train the models DeiT-small, ResNet50, DeiT-tiny, and ResNet18 on ImageNet 1k training data respectively. Note that no distillation is applied. The input size for training is $H = W = 224$, and the patch size is set to 16. Namely, there are 196 image patches totally in each image. We report the clean accuracy in Appendix A where DeiT and its counter-part ResNet show similar accuracy on clean images.

### 4.1 Patch-wise Natural Corruption

First, we investigate the robustness of DeiT and ResNet to patch-based natural corruptions. Specifically, we randomly select 10k test images from ImageNet-1k validation dataset [9] that are correctly classified by both DeiT and ResNet. Then for each image, we randomly sample $n$ input image patches $\boldsymbol{x}_i$ from 196 patches and perturb them with natural corruptions. As in [17], 15 types of natural corruptions with the highest level are applied to the selected patches, respectively. The fooling rate of the patch-based natural corruption is computed over all the test images and all corruption types. We test DeiT and ResNet with the same naturally corrupted images for a fair comparison.

We find that both DeiT and ResNet hardly degrade their performance when a small number of patches are corrupted (*e.g.*, 4). When we increase the number of patches, the difference between two architectures emerges: DeiT achieves a lower FR compared to its counter-part ResNet (See Tab. 2). This indicates that DeiT is more robust against naturally corrupted patches than ResNet. The same conclusion holds under the extreme case when the number of patches $n = 196$. That is: the whole image is perturbed with natural corruptions. This is aligned with the observation in the existing work [5] that vision transformers are more robust to ResNet under distributional shifts. More details on different corruption types are in Appendix B.

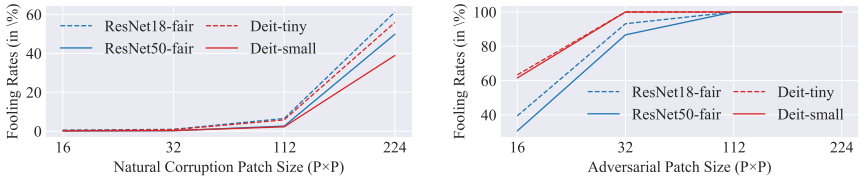



Fig. 2: DeiT with red lines shows a smaller FR to natural patch corruption and a larger FR to adversarial patch of different sizes than counter-part ResNet.

In addition, we also increase the patch size of the perturbed patches, *e.g.*, if the patch size of the corrupted patch is $32 \times 32$, it means that it covers 4 continuous and independent input patches as the input patch size is $16 \times 16$. As shown in Fig. 2 (Left), even when the patch size of the perturbed patches becomes larger, DeiT (marked with red lines) is still more robust than its counter-part ResNet (marked with blue lines) to natural patch corruption.

### 4.2 Patch-wise Adversarial Attack

In this section, we follow [21] to generate adversarial patch attack and then compare the robustness of DeiT and ResNet against adversarial patch attack. We first randomly select the images that are correctly classified by both models from imagenet-1k validation daset. Following [21], the $\ell_\infty$-norm bound, the step size, and the attack iterations are set to 255/255, 2/255, and 10K respectively. Each reported FR score is averaged over 19.6k images.

As shown in Tab. 2, DeiT achieves much higher fooling rate than ResNet when one of the input image patches is perturbed with adversarial perturbation. This consistently holds even when we increase the number of adversarial patches, sufficiently supports that DeiT is more vunerable than ResNet against patch-wise adversarial perturbation. When more than 4 patches ($\sim$2% area of the input image) are attacked, both DeiT and ResNet can be successfully fooled with almost 100% FR.

When we attack a large continuous area of the input image by increasing the patch size of adversarial patches, the FR on DeiT is still much larger than counter-part ResNet until both models are fully fooled with 100% fooling rate. As shown in Fig. 2 (Right), DeiT (marked with red lines) consistently has higher FR than ResNet under different adversarial patch sizes.

Taking above results together, we discover that DeiT is more robust to natural patch corruption than ResNet, whereas it is significantly more vulnerable to adversarial patch perturbation.

## 5 Understanding ViT Robustness to Patch Perturbation

In this section, we design and conduct experiments to analyze the robustness of ViT. Especially, we aim to obtain deep understanding of how ViT performs when its input patches are perturbed with natural corruption or adversary patches.


(a) on ResNet50 under Adversary Patch Attack


(b) on DeiT-small under Adversary Patch Attack

Fig. 3: Gradient Visualization. the clean image, the images with adversarial patches, and their corresponding gradient maps are visualized. We use a blue box on the gradient map to mark the location of the adversarial patch. The adversary patch on DeiT attracts attention, while the one on ResNet hardly do.

### 5.1 How ViT Attention Changes under Patch Perturbation?

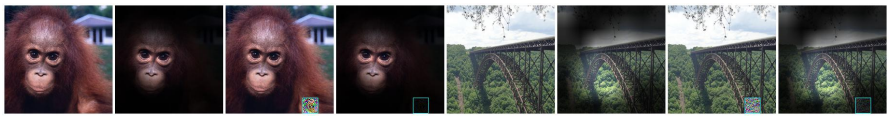
We visualize and analyze models' attention to understand the different robustness performance of DeiT and ResNet against patch-wise perturbations. Although there are many existing methods, *e.g.*, [40,43,53], designed for CNNs to generate saliency maps, it is not clear yet how suitable to generalize them to vision transformers. Therefore, we follow [21] to choose the **model-agnostic** vanilla gradient visualization method to compare the gradient (saliency) map [52] of DeiT and ResNet. Specifically, we consider the case where DeiT and ResNet are attacked by adversarial patches. The gradient map is created as follow: we obtain the gradients of input examples towards the predicted classes, sum the absolute values of the gradients over three input channels, and visualize them by mapping the values into gray-scale saliency maps.

**Qualitative Evaluation.** As shown in Fig. 3 (a), when we use adversarial patch to attack a ResNet model, the gradient maps of the original images and the images with adversarial patch are similar. The observation is consistent with the one made in the previous work [21]. In contrast to the observation on ResNet, the adversarial patch can change the gradient map of DeiT by attracting more attention. As shown in Figure 3 (b), even though the main attention of DeiT is still on the object, part of the attention is misled to the adversarial patch. More visualizations are in Appendix C.

**Quantitative Evaluation.** We also measure our observation on the attention changes with the metrics in [21]. In each gradient map, we score each patch according to (1) the maximum absolute value within the patch (MAX); and (2) the sum of the absolute values within the patch (SUM). We first report the percentage of patches where the MAX is also the maximum of the whole gradient map. Then, we divide the SUM of the patch by the SUM of the all gradient values and report the percentage.

Table 3: Quantitative Evaluation. Each cell lists the percent of patches in which the maximum gradient value inside the patches is also the maximum of whole gradient map. SUM corresponds to the sum of element values inside patch divided by the sum of values in the whole gradient map. The average over all patches is reported.

| | Towards ground-truth Class | | | | Towards misclassified Class | | | |
|---|---|---|---|---|---|---|---|---|
| | SUM | | MAX | | SUM | | MAX | |
| Patch Size | 16 | 32 | 16 | 32 | 16 | 32 | 16 | 32 |
| ResNet50 | 0.42 | 1.40 | 0.17 | 0.26 | 0.55 | 2.08 | 0.25 | 0.61 |
| DeiT-small | **1.98** | **5.33** | **8.3** | **8.39** | **2.21** | **6.31** | **9.63** | **12.53** |
| ResNet18 | 0.24 | 0.74 | 0.01 | 0.02 | 0.38 | 1.31 | 0.05 | 0.13 |
| DeiT-tiny | **1.04** | **3.97** | **3.67** | **5.90** | **1.33** | **4.97** | **6.49** | **10.16** |



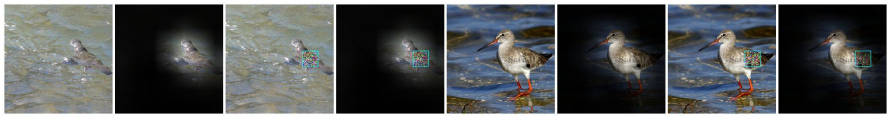
(a) Attention on ResNet18 under Adversary Patch Attack

(b) Attention on DeiT-tiny under Adversary Patch Attack

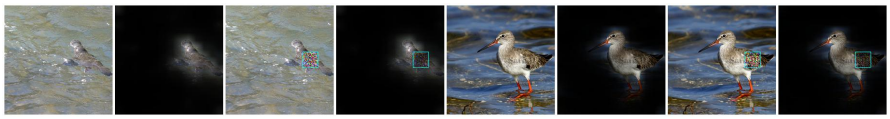
Fig. 4: Attention Comparison between ResNet and DeiT under Patch Attack. The clean image, the adversarial images, and their corresponding attention are visualized. The adversary patch on DeiT attract attention, while the ones on ResNet hardly do.

As reported in Tab. 3, the pixel with the maximum gradient value is more likely to fall inside the adversarial patch on DeiT, compared to that on ResNet. Similar behaviors can be observed in the metric of SUM. The quantitative experiment also supports our claims above that adversarial patches mislead DeiT by attracting more attention.

Besides the gradient analysis, another popular tool used to visualize ViT is Attention Rollout [1]. To further confirm our claims above, we also visualize DeiT with Attention Rollout in Fig. 4. The rollout attention also shows that the attention of DeiT is attracted by adversarial patches. The attention rollout is not applicable to ResNet. As an extra check, we visualize and compare the feature maps of classifications on ResNet. The average of feature maps along the channel dimension is visualized as a mask on the original image. The visualization also



(a) Attention on ResNet18 under Natural Patch Corruption



(b) Attention on DeiT-tiny under Natural Patch Corruption

Fig. 5: Attention Comparison between ResNet and DeiT under Natural Patch Corruption. The clean image, the naturally corrupted images, and their corresponding attention are visualized. The patch corruptions on DeiT are ignored by attending less to the corrupted patches, while the ones on ResNet are treated as normal patches.

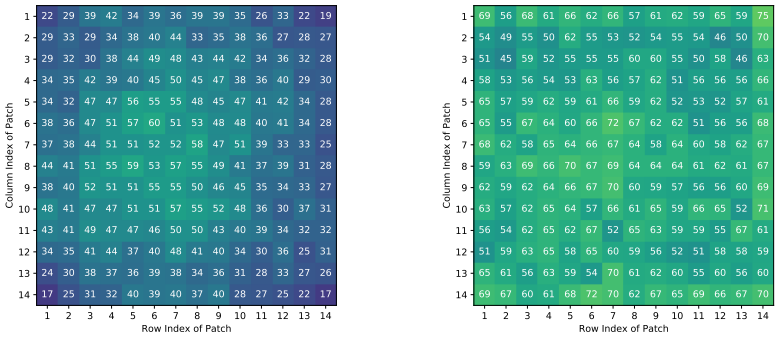
supports the claims above. More visualizations are in Appendix D. Both qualitative and quantitative analysis verifies our claims that the adversarial patch can mislead the attention of DeiT by attactting it.

However, the gradient analysis is not available to compare ViT and ResNet on images with natural corrupted patches. When a small number of patch of input images are corrupted, both Deit and ResNet are still able to classify them correctly. The slight changes are not reflected in vanilla gradients since they are noisy. When a large area of the input image is corrupted, the gradient is very noisy and semantically not meaningful. Due to the lack of a fair visualization tool to compare DeiT and ResNet on naturally corrupted images, we apply Attention Rollout to DeiT and Feature Map Attention visualization to ResNet for comparing the their attention.

The attention visualization of these images is shown in Fig. 5. We can observe that ResNet treats the naturally corrupted patches as normal ones. The attention of ResNet on natually patch-corrupted images is almost the same as that on the clean ones. Unlike CNNs, DeiT attends less to the corrupted patches when they cover the main object. When the corrupted patches are placed in the background, the main attention of DeiT is still kept on the main object. More figures are in Appendix E.

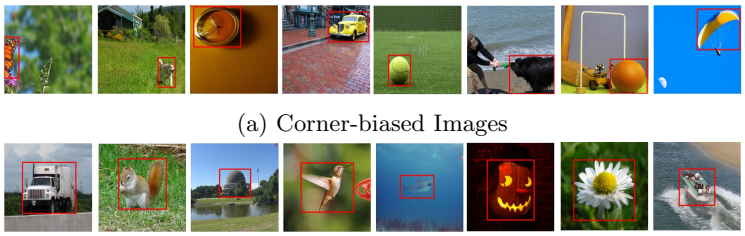
### 5.2 How Sensitive Is ViT Vulnerability to Attack Patch Positions?

To investigate the sensitivity against the location of adversarial patch, we visualize the FR on each patch position in Fig. 6. We can clearly see that adversarial patch achieves higher FR when attacking DeiT-tiny than ResNet18 in different patch positions. Interestingly, we find that the FRs in different patch positions of DeiT-tiny are similar, while the ones in ResNet18 are center-clustered. A similar pattern is also found on DeiT-small and ResNet50 in Appendix F.




(a) Patch Attack FRs on ResNet18 (b) Patch Attack FRs on DeiT-tiny

Fig. 6: Patch Attack FR (in %) in each patch position is visualized. FRs in different patch positions of DeiT-tiny are similar, while the ones in ResNet18 are center-clustered.



(a) Corner-biased Images

(b) Center-biased Images

Fig. 7: Collection of two sets of biased data. The fist set contains only images with corner-biased object(s), and the other set contains center-biased images.

Considering that ImageNet are center-biased where the main objects are often in the center of the images, we cannot attribute the different patterns to the model architecture difference without further investigation.

Hence, we design the following experiments to disentangle the two factors, *i.e.*, model architecture and data bias. Specifically, we select two sets of correctly classified images from ImageNet 1K validation dataset. As shown in Fig. 7a, the first set contains images with corner bias where the main object(s) is in the image corners. In contrast, the second set is more center-biased where the main object(s) is exactly in the central areas, as shown in Fig. 7b.

We apply patch attack to corner-biased images (*i.e.*, the first set) on ResNet. The FRs of patches in the center area are still significantly higher than the ones in the corner (See Appendix G). Based on this, we can conclude that such a relation of FRs to patch position on ResNet is caused by ResNet architectures instead of data bias. The reason behind this might be that pixels in the center can affect more neurons of ResNet than the ones in corners.

Table 4: Transferability of adversarial patch across different patch positions of the the image. Translation X/Y stands for the number of pixels shifted in rows or columns. When they are shifted to cover other patches exactly, adversarial patches transfer well, otherwise not.

| Trans-(X,Y) | (0, 1) | **(0, 16)** | **(0, 32)** | (1, 0) | **(16, 0)** | **(32, 0)** | (1, 1) | **(16, 16)** |
|---|---|---|---|---|---|---|---|---|
| ResNet50 | 0.06 | 0.31 | 0.48 | 0.06 | 0.18 | 0.40 | 0.08 | 0.35 |
| DeiT-small | 0.27 | **8.43** | **4.26** | 0.28 | **8.13** | **3.88** | 0.21 | **4.97** |
| ResNet18 | 0.22 | 0.46 | 0.56 | 0.19 | 0.49 | 0.68 | 0.15 | 0.49 |
| DeiT-tiny | 2.54 | **29.15** | **18.19** | 2.30 | **28.37** | **17.32** | 2.11 | **21.23** |

Similarly, we also apply patch attack to center-biased images (the second set) on DeiT. We observe that the FRs of all patch positions are still similar even the input data are highly center-biased (See Appendix H). Hence, we draw the conclusion that DeiT shows similar sensitivity to different input patches regardless of the content of the image. We conjecture it can be explained by the architecture trait of ViT, in which each patch equally interact with other patches regardless of its position.

### 5.3 Are Adversarial Patches on ViT Still Effective When Shifted?

The work [21] shows that the adversarial patch created on an image on ResNet is not effective anymore even if a single pixel is shifted away. Similarly, we also find that the adversarial patch perturbation on DeiT does not transfer as well when shifting a single pixel away. However, when an adversarial patch is shifted to exactly match another input patch, it remains highly effective, as shown in Tab. 4. This mainly because the attention can still be misled to focus on the adversarial patch as long as it is perfectly aligned with the input patch. In contrast, if a single pixel is shifted away, the structure of the adversarial perturbation is destroyed due to the misalignment between the input patch of DeiT and the constructed adversarial patch. Additionally, We find that the adversarial patch perturbation can hardly transfer across images or models regardless of the alignment. Details can be found in Appendix I.

## 6 Improving ViT Robustness to Adversarial Patch

Given an input image $\boldsymbol{x} \in \mathbb{R}^{H \times W \times C}$, ViT [10] first reshapes the input $\boldsymbol{x}$ into a sequence of image patches $\{\boldsymbol{x}_i \in \mathbb{R}^{(\frac{H}{P} \cdot \frac{W}{P}) \times (P^2 \cdot C)}\}_{i=1}^{N}$ where $P$ is the patch size and $N$ is the number of patches. A class-token patch $\boldsymbol{x}_0$ is concatenated to the patch sequence. A set of self-attention blocks is applied to obtain patch embeddings of the $l$-th block $\{\boldsymbol{x}_i^l\}_{i=0}^{N}$. The class-token patch embedding of the last block is mapped to the output.

The patch embedding of the $i$-th patch in the $l$-th layer is the weighted sum of all patch embedding $\{\boldsymbol{x}_j^{l-1}\}_{i=0}^N$ of the previous layer. The weights are the attention weights obtained from the attention module. Formally, the patch embedding $\boldsymbol{x}_i^l$ is computed with following equation

$$\boldsymbol{x}_i^l = \sum_{j=0}^{N} \alpha_{ij} \cdot \boldsymbol{x}_j^{l-1}, \qquad \alpha_{ij} = \frac{\exp(Z_{ij})}{\sum_{j=0}^{N} \exp(Z_{ij})} \tag{1}$$

where $\alpha_{ij}$ is the attention weight that stands for the attention of the $i$-th patch of the $l$-th layer to the $j$-th patch of the ($l$-1)-th layer. $Z_{ij}$ is the scaled dot-product between the key of the $j$-th patch and the query of of the $i$-th patch in the ($l$-1)-th layer, i.e., the logits before *softmax* attention.

Given a classification task, we denote the patch embedding of the clean image as $\boldsymbol{x}_i^{*l}$. When the $k$-th patch is attacked, the patch embedding of the $i$-th patch in the $l$-th layer deviates from $\boldsymbol{x}_i^{*l}$. The deviation distance is described as

$$d(\boldsymbol{x}_i^l, \boldsymbol{x}_i^{*l}) = \sum_{j=0}^{N} \alpha_{ij} \cdot \boldsymbol{x}_j^{l-1} - \sum_{j=0}^{N} \alpha_{ij}^* \cdot \boldsymbol{x}_j^{l-1}, \tag{2}$$

where $\alpha_{ij}^*$ is the attention weight corresponding to the clean image. Our analysis shows that the attention is misled to focus on the attacked patch. In other words, $\alpha_{ik}$ is close to 1, and other attention weights are close to zero.

To address this, we replace the original attention with smoothed attention using temperature scaling in the *softmax* operation. Formally, the smoothed attention is defined as

$$\alpha_{ij}^{\diamondsuit} = \frac{\exp(Z_{ij}/T)}{\sum_{j=0}^{N} \exp(Z_{ij}/T)}, \tag{3}$$

where $T(>1)$ is the hyper-parameter that determines the smoothness of the proposed attention. With the smoothed attention, the deviation of the patch embedding from the clean patch embedding is smaller.

$$d(\boldsymbol{x}_i^{\diamondsuit l}, \boldsymbol{x}_i^{*l}) = \sum_{j=0}^{N} \alpha_{ij}^{\diamondsuit l} \cdot \boldsymbol{x}_j^{l-1} - \sum_{j=0}^{N} \alpha_{ij}^* \cdot \boldsymbol{x}_j^{l-1} < d(\boldsymbol{x}_i^l, \boldsymbol{x}_i^{*l}) \tag{4}$$

We can see that the smoothed attention naturally encourages self-attention not to focus on a single patch. To validate if ViT becomes more robust to adversarial patches, we apply the method to ViT and report the results in Fig. 8. Under different temperatures, the smoothed attention can improve the adversarial robustness of ViT to adversarial patches and rarely reduce the clean accuracy. In addition, the effectiveness of smoothed attention also verifies our understanding of the robustness of ViT in Sec. 5: it is the attention mechanism that causes the vulnerability of ViT against adversarial patch attacks.

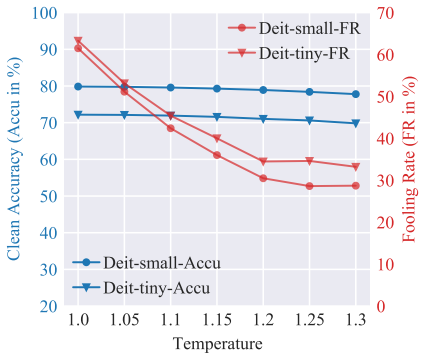



Fig. 8: The robustness of ViT can be improved with Smoothed Attention.

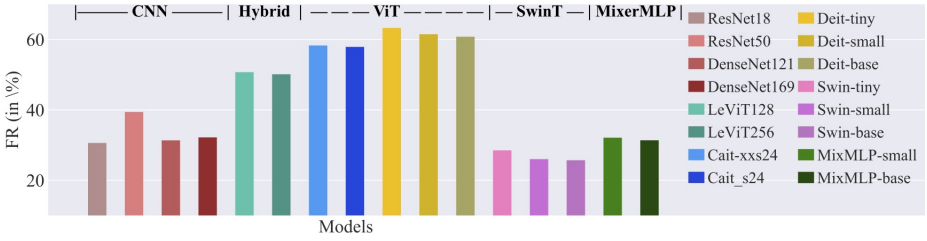



Fig. 9: We report Fooling Rates on different versions of ViT, CNN as well as Hybrid architectures under adversarial patch attacks.

## 7 Discussion

In previous sections, we mainly focus on studying the state-of-the-art patch attack methods on the most primary ViT architecture and ResNet. In this section, we further investigate different variants of model architectures as well as adversarial patch attacks.

***Different Model Architectures*** In addition to DeiT and ResNet, we also investigate the robustness of different versions of ViT [10,47,25], CNN [16,19] as well as Hybrid architectures [14] under adversarial patch attacks. Following the experimental setting in section 3, we train all the models and report fooling rate on each model in Fig. 9. Four main conclusions can be drawn from the figure.

1. CNN variants are consistently more robust than ViT models.
2. The robustness of LeViT model [14] with hybrid architecture (*i.e.*, Conv Layers + Self-Attention Blocks) lives somewhere between ViT and CNNs.
3. Swin Transformers [25] are as robust as CNNs. We conjecture this is because attention cannot be manipulated by a single patch due to hierarchical attention and the shifted windows therein. Specifically, the self-attention in Swin Transformers is conducted on patches within a local region rather than the whole image. In addition, a single patch will interact with patches from different groups in different layers with shifted windows. This makes effective adversarial patches challenging.
4. Mixer-MLP [46] uses the same patch-based architecture as ViTs and has no attention module. Mixer-base with FR (31.36) is comparable to ResNet and more robust than ViTs. The results further confirm that the vulnerability of ViT can be attributed to self-attention mechanism.

Our proposed attention smoothing by temperature scaling can effectively improve the robustness of DeiT and Levit. However, the improvement on Swin Transformers is tiny due to its architecture design.

***Different Patch Attacks*** Other than adversarial patch attacks studied previously, we also investigate the robustness of ViT and ResNet against the following variants of adversarial patch attacks.

*Imperceptible patch attack* In previous sections, we use unbounded local patch attacks where the pixel intensity can be set to any value in the image range $[0, 1]$. The adversarial patches are often visible, as shown in Fig. 1. In this section, we compare DeiT and ResNet under a popular setting where the adversarial perturbation is imperceptible to humans, bounded by $8/225$. In the case of a single patch attack, the attacker achieves FR of 2.9% on ResNet18 and 11.2% on DeiT-tiny (see Appendix J for more results). That is: DeiT is still more vulnerable than ResNet when attacked with imperceptible patch perturbation.

*Targeted patch attack* We also compare DeiT and ResNet under targeted patch attacks, which can be achieved by maximizing the probability of the target class. Specifically, we randomly select a target class other than the ground-truth class for each image. Under a single targeted patch attack, the FR is 15.4% for ResNet18 vs. 32.3% for DeiT-tiny, 7.4% for ResNet50 vs. 24.9% for DeiT-small. The same conclusion holds: DeiT is more vulnerable than ResNet. Visualization of adversarial patches is in Appendix K.

*Patch attack generated with different iterations* Following [21], we generate adversarial patch attacks with 10k iterations. In this section, we further study the minimum iterations required to successfully attack the classifier, which is averaged over all patch positions of the misclassified images. We find that the minimum attack iterations on DeiT-tiny is much smaller than that on ResNet18 (65 vs. 342). Similar results on DeiT-small and ResNet50 (294 vs. 455). This further validates DeiT is more vulnerable than ResNet.

*ViT-agnostic patch attack* In this section, we study ViT-agnostic patch attack where the adversarial patch of the same size as an input patch is placed to a random area of the image. The covered area can involve pixels from multiple input patches. We find that DeiT becomes less vulnerable to adversarial patch attack, *e.g.*, the FR on DeiT-small decreases from 61.5% to 47.9%. When the adversarial patch is not aligned with the input patch, *i.e.*, only part of patch pixels can be manipulated, the attention of DeiT is less likely to be misled. Under such ViT-agnostic patch attack, ViT is still more vulnerable than ResNet.

## 8 Conclusion

This work starts with an interesting observation on the robustness of ViT to patch perturbations. Namely, vision transformer (e.g., DeiT) is more robust to natural patch corruption than ResNet, whereas it is significantly more vulnerable against adversarial patches. Further, we discover the self-attention mechanism of ViT can effectively ignore natural corrupted patches but be easily misled to adversarial patches to make mistakes. Based on our analysis, we propose attention smoothing to improve the robustness of ViT to adversarial patches, which further validates our developed understanding. We believe this study can help the community better understand the robustness of ViT to patch perturbations.

## References


1. Abnar, S., Zuidema, W.: Quantifying attention flow in transformers. In: Annual Meeting of the Association for Computational Linguistics (ACL) (2020)
2. Aldahdooh, A., Hamidouche, W., Deforges, O.: Reveal of vision transformers robustness against adversarial attacks. arXiv:2106.03734 (2021)
3. Bai, Y., Mei, J., Yuille, A., Xie, C.: Are transformers more robust than cnns? arXiv:2111.05464 (2021)
4. Benz, P., Ham, S., Zhang, C., Karjauv, A., Kweon, I.S.: Adversarial robustness comparison of vision transformer and mlp-mixer to cnns. arXiv preprint arXiv:2110.02797 (2021)
5. Bhojanapalli, S., Chakrabarti, A., Glasner, D., Li, D., Unterthiner, T., Veit, A.: Understanding robustness of transformers for image classification. arXiv:2103.14586 (2021)
6. Brown, T.B., Mané, D., Roy, A., Abadi, M., Gilmer, J.: Adversarial patch. arXiv:1712.09665v1 (2017)
7. Chen, C.F., Fan, Q., Panda, R.: Crossvit: Cross-attention multi-scale vision transformer for image classification. arXiv:2103.14899 (2021)
8. Chen, Z., Xie, L., Niu, J., Liu, X., Wei, L., Tian, Q.: Visformer: The vision-friendly transformer. arXiv:2104.12533 (2021)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2009)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929 (2020)
11. Fawzi, A., Frossard, P.: Measuring the effect of nuisance variables on classifiers. In: Proceedings of the British Machine Vision Conference (BMVC) (2016)
12. Fu, Y., Zhang, S., Wu, S., Wan, C., Lin, Y.: Patch-fool: Are vision transformers always robust against adversarial perturbations? In: International Conference on Learning Representations (2021)
13. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv:1412.6572 (2014)
14. Graham, B., El-Nouby, A., Touvron, H., Stock, P., Joulin, A., Jégou, H., Douze, M.: Levit: a vision transformer in convnet's clothing for faster inference. arXiv:2104.01136 (2021)
15. Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., Wang, Y.: Transformer in transformer. arXiv:2103.00112 (2021)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
17. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. In: International Conference on Learning Representations (ICLR) (2019)
18. Hu, H., Lu, X., Zhang, X., Zhang, T., Sun, G.: Inheritance attention matrix-based universal adversarial perturbations on vision transformers. IEEE Signal Processing Letters **28**, 1923–1927 (2021)
19. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)

20. Joshi, A., Jagatap, G., Hegde, C.: Adversarial token attacks on vision transformers. arXiv:2110.04337 (2021)
21. Karmon, D., Zoran, D., Goldberg, Y.: Lavan: Localized and visible adversarial noise. In: International Conference on Machine Learning (ICML) (2018)
22. Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., Houlsby, N.: Big transfer (bit): General visual representation learning. In: European Conference on Computer Vision (ECCV) (2020)
23. Liu, A., Liu, X., Fan, J., Ma, Y., Zhang, A., Xie, H., Tao, D.: Perceptual-sensitive gan for generating adversarial patches. In: AAAI (2019)
24. Liu, A., Wang, J., Liu, X., Cao, B., Zhang, C., Yu, H.: Bias-based universal adversarial patch attack for automatic check-out. In: European conference on computer vision. pp. 395–410. Springer (2020)
25. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. arXiv:2103.14030 (2021)
26. Luo, J., Bai, T., Zhao, J.: Generating adversarial yet inconspicuous patches with a single image (student abstract). In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 15837–15838 (2021)
27. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: arXiv:1706.06083 (2017)
28. Mahmood, K., Mahmood, R., Van Dijk, M.: On the robustness of vision transformers to adversarial examples. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7838–7847 (2021)
29. Mao, X., Qi, G., Chen, Y., Li, X., Duan, R., Ye, S., He, Y., Xue, H.: Towards robust vision transformer. arXiv:2105.07926 (2021)
30. Mao, X., Qi, G., Chen, Y., Li, X., Ye, S., He, Y., Xue, H.: Rethinking the design principles of robust vision transformer. arXiv:2105.07926 (2021)
31. Metzen, J.H., Finnie, N., Hutmacher, R.: Meta adversarial training against universal patches. arXiv preprint arXiv:2101.11453 (2021)
32. Mu, N., Wagner, D.: Defending against adversarial patches with robust self-attention. In: ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning (2021)
33. Naseer, M., Ranasinghe, K., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Intriguing properties of vision transformers. arXiv:2105.10497 (2021)
34. Naseer, M., Ranasinghe, K., Khan, S., Khan, F.S., Porikli, F.: On improving adversarial transferability of vision transformers. arXiv:2106.04169 (2021)
35. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A.: The limitations of deep learning in adversarial settings. In: 2016 IEEE European symposium on security and privacy (EuroS&P) (2016)
36. Paul, S., Chen, P.Y.: Vision transformers are robust learners. arXiv:2105.07581 (2021)
37. Qian, Y., Wang, J., Wang, B., Zeng, S., Gu, Z., Ji, S., Swaileh, W.: Visually imperceptible adversarial patch attacks on digital images. arXiv preprint arXiv:2012.00909 (2020)
38. Qin, Y., Zhang, C., Chen, T., Lakshminarayanan, B., Beutel, A., Wang, X.: Understanding and improving robustness of vision transformers through patch-based negative augmentation. arXiv preprint arXiv:2110.07858 (2021)
39. Salman, H., Jain, S., Wong, E., Madry, A.: Certified patch robustness via smoothed vision transformers. arXiv:2110.07719 (2021)

40. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: ICCV (2017)
41. Shao, R., Shi, Z., Yi, J., Chen, P.Y., Hsieh, C.J.: On the adversarial robustness of visual transformers. arXiv:2103.15670 (2021)
42. Shi, Y., Han, Y.: Decision-based black-box attack against vision transformers via patch-wise adversarial removal. arXiv preprint arXiv:2112.03492 (2021)
43. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: International Conference on Machine Learning (ICML) (2017)
44. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. International Conference on Learning Representations (ICLR) (2014)
45. Tang, S., Gong, R., Wang, Y., Liu, A., Wang, J., Chen, X., Yu, F., Liu, X., Song, D., Yuille, A., et al.: Robustart: Benchmarking robustness on architecture design and training techniques. arXiv preprint arXiv:2109.05211 (2021)
46. Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Keysers, D., Uszkoreit, J., Lucic, M., et al.: Mlp-mixer: An all-mlp architecture for vision. In: arXiv:2105.01601 (2021)
47. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning (ICML) (2021)
48. Wang, J., Liu, A., Bai, X., Liu, X.: Universal adversarial patch attack for automatic checkout using perceptual and attentional bias. IEEE Transactions on Image Processing **31**, 598–611 (2021)
49. Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J., Keutzer, K., Vajda, P.: Visual transformers: Token-based image representation and processing for computer vision. arXiv:2006.03677 (2020)
50. Xiao, T., Singh, M., Mintun, E., Darrell, T., Dollár, P., Girshick, R.: Early convolutions help transformers see better. arXiv:2106.14881 (2021)
51. Yu, Z., Fu, Y., Li, S., Li, C., Lin, Y.: Mia-former: Efficient and robust vision transformers via multi-grained input-adaptation. arXiv preprint arXiv:2112.11542 (2021)
52. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European conference on computer vision (2014)
53. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)