Towards Robust Vision Transformer

Xiaofeng Mao[1] Gege Qi[1] Yuefeng Chen[1] Xiaodan Li[1] Ranjie Duan[2] Shaokai Ye[3]
Yuan He[1] Hui Xue[1]
[1]Alibaba Group [2]Swinburne University of Technology [3]EPFL
{mxf164419,qigege.qgg,yuefeng.chenyf,fiona.lxd}@alibaba-inc.com

Abstract

*Recent advances on Vision Transformer (ViT) and its improved variants have shown that self-attention-based networks surpass traditional Convolutional Neural Networks (CNNs) in most vision tasks. However, existing ViTs focus on the standard accuracy and computation cost, lacking the investigation of the intrinsic influence on model robustness and generalization. In this work, we conduct systematic evaluation on components of ViTs in terms of their impact on robustness to adversarial examples, common corruptions and distribution shifts. We find some components can be harmful to robustness. By leveraging robust components as building blocks of ViTs, we propose **Robust Vision Transformer (RVT)**, which is a new vision transformer and has superior performance with strong robustness. Inspired by the findings during the evaluation, we further propose two new plug-and-play techniques called position-aware attention scaling and patch-wise augmentation to augment our RVT, which we abbreviate as RVT*. The experimental results of RVT on ImageNet and six robustness benchmarks demonstrate its advanced robustness and generalization ability compared with previous ViTs and state-of-the-art CNNs. Furthermore, RVT-S* achieves Top-1 rank on multiple robustness leaderboards including ImageNet-C, ImageNet-Sketch and ImageNet-R.*

1. Introduction

Following the popularity of transformers in Natural Language Processing (NLP) applications, e.g., BERT [8] and GPT [30], there has sparked particular interest in investigating whether transformer can be a primary backbone for computer vision applications previously dominated by Convolutional Neural Networks (CNNs). Recently, Vision Transformer (ViT) [10] successfully applies a pure transformer for classification which achieves an impressive speed-accuracy trade-off by capturing long-range dependencies via self-attention. Base on this seminal work, numerous variants have been proposed to improve ViTs from different perspectives containing training data efficiency [40], self-attention mechanism [25], introducing convolution [23,45,50] or pooling layers [20,43], etc. However, these works only focus on the standard accuracy and computation cost, lacking the investigation of the intrinsic influence on model robustness and generalization.

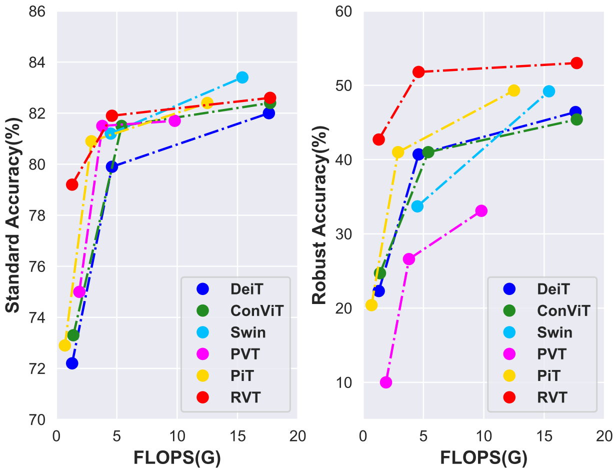

Figure 1. Comparison between RVT and the baseline transformers. The robust accuracy in figure is recorded under FGSM [11] adversary.

In this work, we take initiatives to explore a ViT model with strong robustness. To this end, we first give an empirical assessment of existing ViT models in Figure 1. Surprisingly, although all ViT variants reproduce the standard accuracy claimed in the paper, some of their modifications may bring devastating damages on the model robustness. A vivid example is PVT [43], which achieves a high standard accuracy but suffered with large drop of robust accuracy. We show that PVT-Small obtains only 26.6% robust accuracy, which is 14.1% lower than original DeiT-S in Figure 1.

To demystify the trade-offs between accuracy and robustness, we analyze ViT models with different patch embedding, position embedding, transformer blocks and classification head whose impact on the robustness that has never been thoroughly studied. Based on the valuable findings revealed by exploratory experiments, we propose a Robust Vision Transformer (RVT), which has significant improvement on robustness, but also exceeds most other transformers in accuracy. In addition, we propose two new plug-and-play techniques to further boost the RVT. The first is Position-Aware Attention Scaling (PAAS), which plays the role of position encoding in RVT. PAAS improves the self-attention mechanism by filtering out redundant and noisy position correlation and activating only major attention with strong correlation, which leads to the enhancement of model robustness. The second is a simple and general patch-wise augmentation method for patch sequences which adds rich affinity and diversity to training data. Patch-wise augmentation also contributes to the model generalization by reducing the risk of over-fitting. With the above proposed methods, we can build an augmented Robust Vision Transformer* (RVT*). Contributions of this paper are three-fold:

- We give a systematic robustness analysis of ViTs and reveal harmful components. Inspired by it, we reform robust components as building blocks as a new transformer, named Robust Vision Transformer (RVT).
- To further improve the RVT, we propose two new plug-and-play techniques called position-aware attention scaling and patch-wise augmentation. Both of them can be applied to other ViT models and yield significant enhancement on robustness and standard accuracy.
- Experimental results on ImageNet and six robustness benchmarks show that RVT exhibits best trade-offs between standard accuracy and robustness compared with previous ViTs and CNNs. Specifically, RVT-S* achieves Top-1 rank on ImageNet-C, ImageNet-Sketch and ImageNet-R.

2. Related Work

Robustness Benchmarks. The rigorous benchmarks are important for evaluating and understanding the robustness of deep models. Early works focus on the model safety under the adversarial examples with constrained perturbations [11, 38]. In real-world applications, the phenomenon of image corruption or out-of-distribution is more commonly appeared. Driven by this, ImageNet-C [17] benchmarks the model against image corruption which simulates distortions from real-world sources. ImageNet-R [16] and ImageNet-Sketch [42] collect the online images consisting of naturally occurring distribution changes such as image style, to measure the generalization ability to new distributions at test time. In this paper, we adopt all the above benchmarks as the fair-minded evaluation metrics.

Robustness Study for CNNs. The robustness research of CNNs has experienced explosive development in recent years. Numerous works conduct thorough study on the robustness of CNNs and aim to strengthen it in different ways, e.g., stronger data augmentation [16, 18, 33], carefully designed [36, 44] or searched [9, 13] network architecture, improved training strategy [22, 26, 47], quantization [24] and pruning [49] of the weights, better pooling [41, 53] or activation functions [46], etc. Although the methods mentioned above perform well on CNNs, there is no evidence that they also keep the effectiveness on ViTs. A targeted research for improving the robustness of ViTs is still blank.

Robustness Study for ViTs. Until now, there are several works attempting at studying the robustness of ViTs. Early works focus on the adversarial robustness of ViTs. They find that ViTs are more adversarially robust than CNNs [34] and the transferability of adversarial examples between CNNs and ViTs is remarkably low [27]. Follow up works [2, 29] extend the robustness study on ViTs to much common image corruption and distribution shift, and indicate ViTs are more robust learners. Although some findings are consistent with above works, in this paper, we do not make simple comparison of robustness between ViTs and CNNs, but take a step further by analyzing the detailed robust components in ViT and its variants. Based on the analysis, we design a robust vision transformer and introduce two novel techniques to further reduce the fragility of ViT models.

3. Robustness Analysis of Designed Components

We give the robustness analysis of four main components in ViTs: patch embedding, position embedding, transformer blocks and classification head. DeiT-Ti [40] is used as the base model. All the robustness benchmarks mentioned in section 2 are considered comprehensively. There is a positive correlation between these benchmarks in most cases. Due to the limitation of space, we show the robust accuracy under FGSM [11] adversary in the main body and other results in Appendix A.

3.1. Patch Embedding

F1: Low-level feature of patches helps for the robustness. ViTs [10] tokenize an image by splitting it into patches with size of 16×16 or 32×32. Such simple tokenization makes the models hard to capture low-level structures such as edges and corners. To extract low-level features of patches, CeiT [50], LeViT [12] and TNT [14] use a convolutional stem instead of the original linear layer, T2T-ViT [51] leverages self-attention to model dependencies among neighboring pixels. However, these methods merely focus on the standard accuracy. To answer how is the robustness affected by leveraging low-level features

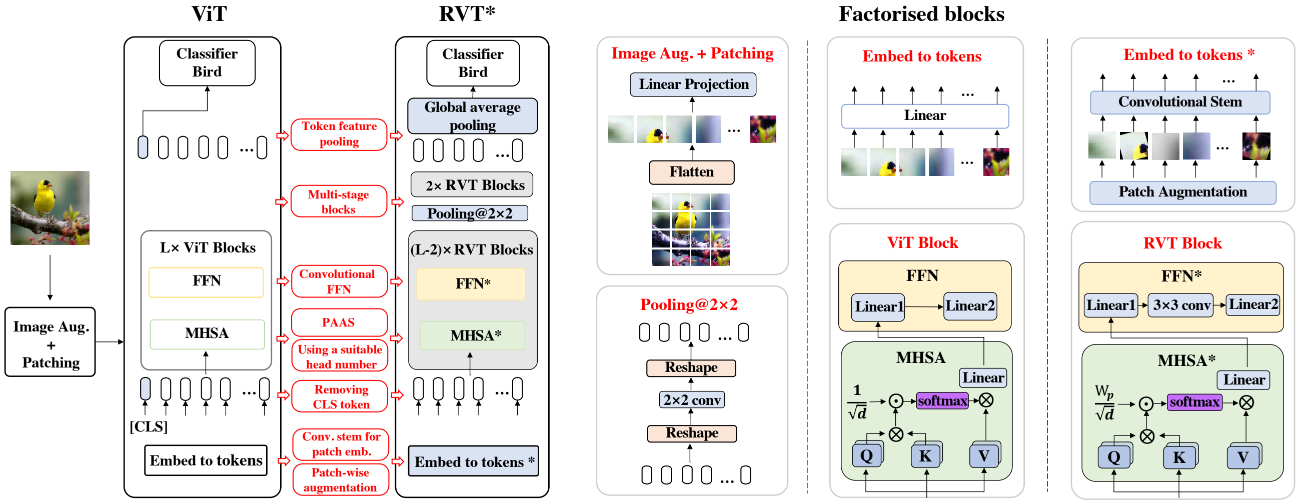

Figure 2. **Overall architecture of the proposed Robust Vision Transformer (RVT).**

of patches, we compare the original linear projection with two new convolution and tokens-to-tokens embedders, proposed by CeiT and T2T-ViT respectively. As shown in Table 2, low-level patch embedding has a positive effect on the model robustness and standard accuracy as more detailed visual features are exploited. Among them tokens-to-tokens embedder is the best, but it has quadratic complexity with the expansion of image size. We adopt the convolutional embedder with less computation cost.

	positional embedding	Acc	Robust Acc
(i)	none	68.3	15.8
(ii)	learned absolute position	72.2	**22.3**
(iii)	sin-cos absolute position	72.0	21.9
(iv)	learned relative position [35]	71.8	22.3
(v)	input-conditioned position [3]	**72.4**	21.5

Table 1. **Effect of different positional embeddings.** We use Deit-Ti as the base model.

3.2. Position Embedding

F2: Position encoding is critical for learning shape-bias based semantic features which are robust to texture changes. Besides, existing position encoding methods have no big impact on the robustness. We first explore the necessity of position embeddings. Previous work [3] shows ViT trained without position embeddings has 4% drop of standard accuracy. In this work, we find this gap even can be larger on robustness. In Appendix A, we find with no position encoding, ViT fails to recognize shape-bias objects, which leads to 8% accuracy drop on ImageNet-Sketch. Concerning the ways of positional encoding, learned absolute, sin-cos absolute, learned relative [35], input-conditioned [3] position representations are compared. In Table 1, the result suggests that most position encoding methods have no big impact on the robustness, and a minority even have a negative effect. Especially, CPE [3] encodes position embeddings conditioned on inputs. Such a conditional position representation makes it changed easily with the input, and causes the poor robustness. The fragility of position embeddings also motivates us to design a more robust position encoding method.

Table 2. Ablations on other ViT components, where ✓indicates the use of the corresponding component.

Patch Emb.			Local	Conv.	CLS	Acc	Rob.
Linear	Conv.	T2T	SA	FFN			Acc
✓					✓	72.2	22.3
	✓				✓	73.6	23.2
		✓			✓	74.9	25.4
✓			✓		✓	69.1	21.0
✓				✓		73.9	31.9
✓					✓	72.4	28.4

3.3. Transformer Blocks

F3: An elaborate multi-stage design is required for constructing robust vision transformers. Modern CNNs always start with a feature of large spatial sizes and a small channel size and gradually increase the channel size while decreasing the spatial size. The different sizes of feature maps constitute the multi-stage convolution blocks. As shown by previous works [4], such a design contributes to the expressiveness and generalization performance of the network. PVT [43], PiT [20] and Swin [25] employ this design principle into ViTs. To measure the robustness variance with changing of stage distribution, we slightly modify the DeiT-Ti architecture to get five variants (V2-V6) in Table 3. We keep the overall number of transformer blocks consistent to 12 and replace some of them with smaller or larger spatial resolution. Detailed architecture is shown in Appendix A. By comparing with DeiT-Ti, we find all five variants improve the standard accuracy, benefit from the extraction of hierarchical image features. In terms of robustness, transformer blocks with different spatial sizes show different effects. An experimental conclusion is that the model will get worse on robustness when it contains more transformer blocks with large spatial resolution. On the contrary, reducing the spatial resolution gradually at later transformer blocks contributes to the modest enhancement of robustness. Besides, we also observe that having more blocks with larger input spatial size will increase the number of FLOPs and memory consumption. To achieve the best trade-off on speed and performance, we think V2 is the most compromising choice in this paper.

F4: Robustness can be benefited from the completeness and compactness among attention heads, by choosing an appropriate head number. ConViT [6], Swin [25] and LeViT [12] both use more self-attention heads and smaller dimensions of keys and queries to achieve better performance at a controllable FLOPs. To study how does the number of heads affect the robustness, we train DeiT-Ti with different head numbers. Once the number of heads increases, we meanwhile reduce the head dimensions to ensure the overall feature dimensions are unchanged. Similar with generally understanding in NLP [28], we find the completeness and compactness among attention heads are important for ViTs. As shown in the Table 4, the robustness and standard accuracy still gain great improvement with the head increasing till to 8. We think that an appropriate number of heads supplies various aspects of attentive information on the input. Such complete and non-redundant attentive information also introduces more fine-grained representations which are prone to be neglect by model with less heads, thus increases the robustness.

variants	$[S_1, S_2, S_3, S_4]$	FLOPs	Mem	Acc	Robust Acc
V1	[0, 0, 12, 0]	1.3	1.1	72.2	22.3
V2	[0, 0, 10, 2]	**1.2**	**1.1**	74.8	**24.3**
V3	[0, 2, 10, 0]	1.5	1.7	73.8	22.0
V4	[0, 2, 8, 2]	1.4	1.7	76.4	22.3
V5	[2, 2, 8, 0]	3.4	6.0	73.4	17.0
V6	[2, 2, 6, 2]	3.4	6.0	**76.4**	17.5

Table 3. **Effect of stage distribution.** We ablate the number of blocks in stages S_1, S_2, S_3, S_4 of **DeiT-Ti**, where S_1 is the stage with the largest 56 × 56 input spatial dimension, and gradually reduced to half of the original in later stages. The GPU memory consumption is tested on input with batch size of 64.

F5: The locality constraints of self-attention layer may do harm for the robustness. Vanilla self-attention calculates the pair-wise attention of all sequence elements. But for image classification, local region needs to be paid more attention than remoter regions. Swin [25] limits the self-attention computation to non-overlapping local windows on the input. This hard coded locality of self-attention enjoys great computational efficiency and has linear complexity with respect to image size. Although Swin can also get competitive accuracy, in this work we find such local window self-attention is harmful to the model robustness. The result in Table 2 shows after modifying self-attention to the local version, the robust accuracy is getting worse. We think this phenomenon may be partly caused by the destruction of long-range dependencies modeling in ViTs.

Heads	1	2	4	6	8	12
Acc	69.0	71.7	73.1	73.4	**73.9**	73.5
Rob. Acc	17.6	21.4	22.8	24.6	**25.2**	24.7

Table 4. **The performance variance with the number of heads. DeiT-Ti** with head number of 1, 2, 4, 6, 8 and 12 are trained for comparison.

F6: Feed-forward networks (FFN) can be extended to convolutional FFN by encoding multiple tokens in local regions. Such information exchange of local tokens in FFN makes ViTs more robust. LocalViT [23] and CeiT [50] introduce connectivity of local regions into ViTs by adding a depth-wise convolution in feed-forward networks (FFN). Our experiment in Table 2 verifies that the convolutional FFN greatly improves both the standard accuracy and robustness. We think the reason lies in two aspects. First, compared with locally self-attention, convolutional FFN will not damage the long-term dependencies modeling ability of ViTs. The merit of ViTs can be inherited. Second, original FFN only encodes single token representation, while convolutional FFN encodes both the current token and its neighbors. Such information exchange within a local region makes ViTs more robust.

3.4. Classification Head

F7: Is the classification token (`CLS`) important for ViTs? The answer is not, and replacing `CLS` with global average pooling on output tokens even improves the robustness. CNNs adopt a global average pooling layer before the classifier to integrate visual features at different spatial locations. This practice also inherently takes advantage of the translation invariance of the image. However, ViTs use an additional classification token (`CLS`) to perform classification, are not translation-invariant. To get over this shortcoming, CPVT [3] and LeViT [12] remove the `CLS` token and replace it by average pooling along with the last layer sequential output of the Transformer. We compare models trained with and without `CLS` token in Table 2. The result shows the adversarial robustness can be greatly improved by removing `CLS` token. Also we find removing `CLS` token has slight help for the standard accuracy, which can be benefited from the desired translation-invariance.

3.5. Combination of Robust Components

In the above, we separately analyze the effect of each designed component in the ViTs. To make use of these findings, we combine the selected useful components, listed in follows: 1) Extract low-level feature of patches using a convolutional stem; 2) Adopt the multi-stage design of ViTs and avoid blocks with larger spatial resolution; 3) Choose a suitable number of heads; 4) Use convolution in FFN; 5) Replace `CLS` token with token feature pooling. As we find the effects of the above modifications are superimposed, we adopt all of these robust components into ViTs, the resultant model is called Robust Vision Transformer (RVT). RVT has achieved the new state-of-the-art robustness compared to other ViT variants. To further improve the performance, we propose two novel techniques, position-aware attention scaling and patch-wise data augmentation, to train our RVT. Both of them are also applicable to other ViT models.

4. Position-Aware Attention Scaling

In this section, we introduce our proposed position encoding mechanism called Position-Aware Attention Scaling (PAAS), which modifies the rescaling operation in the dot product attention to a more generalized version. To start with, we illustrate the scaled dot-product attention in transformer firstly. And then the modification of PAAS will be explained.

Scaled Dot-product Attention. Scaled dot-product attention is a key component in Multi-Head Self Attention layer (MHSA) of Transformer. MHSA first generates set of queries $Q \in \mathbb{R}^{N \times d}$, keys $K \in \mathbb{R}^{N \times d}$, values $V \in \mathbb{R}^{N \times d}$ with the corresponding projection. Then the query vector $q \in \mathbb{R}^d$ is matched against the each key vector in K. The output is the weighted sum of a set of N value vectors v based on the matching score. This process is called scaled dot-product attention:

$$\text{Attention}(Q, K, V) = \text{Softmax}(QK^T/\sqrt{d})V \quad (1)$$

For preventing extremly small gradients and stabilizing the training process, each element in QK^T multiplies by a constant $\frac{1}{\sqrt{d}}$ to be rescaled into a standard range.

Position-Aware Attention Scaling. In this work, a more effective position-aware attention scaling method is proposed. To make the original rescaling process of dot-product attention position-aware, we define a learnable position importance matrix $W_p \in \mathbb{R}^{N \times N}$, which presents the importance of each pair of q-k. The oringinal scaled dot-product attention is modified as follows:

$$\text{Attention}(Q, K, V) = \text{Softmax}(QK^T \odot (W_p/\sqrt{d}))V \quad (2)$$

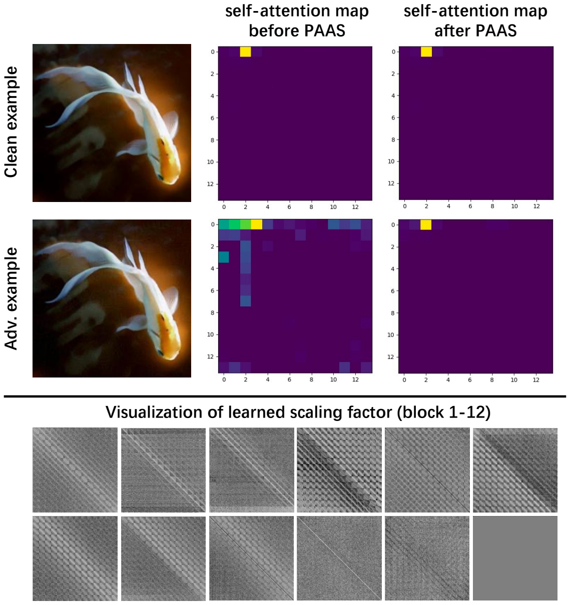

Figure 3. **Top:** visualization of self-attention before and after the position-aware attention scaling. **Bottom:** visualization of learned scaling factor by our PAAS.

where $\odot$ is the element-wise product. As W_p is input independent and only determined by the position of each q, k in the sequence, our position-aware attention scaling can also serve as a position representation. Thus, we replace the traditional position embedding with our PAAS in RVT. After that the overall self-attention can be decoupled into two parts: the QK^T term presents the content-based attention, and $W_p/\sqrt{d}$ term acts as the position-based attention. This untied design offers more expressiveness by removing the mixed and noisy correlations [21].

Robustness of PAAS. As mentioned in section 3.2, most existing position embeddings have no contribution to the model robustness, and some of them even do a negative effect. Differently, our proposed PAAS can improve the model robustness effectively. This superior property relies on the position importance matrix W_p, which acts as a soft attention mask on each position pair of q-k. As shown in Figure 3, we visualize the attention map of 3*th* query patch in 3*th* transformer block. Without PAAS, an adversarial input can make some unrelated regions activated and produce a noisy self-attention map. To filter out these noises, PAAS suppresses the redundant positions irrelevant for classification in self-attention map, by a learned small multiplier in W_p. Finally only the regions important for classification are activated. We experimentally validate that PAAS can provide certain defense power against some white-box adversaries, e.g., FGSM [11]. Not limited to adversarial attack, it also helps to the corruption and out-of-distribution generalization. Details can be referred to section 6.3.

5. Patch-Wise Augmentation

Image augmentation is a strategy especially important for ViTs since a biggest shortcoming of ViTs is the worse generalization ability when trained on relatively small-size datasets, while this shortcoming can be remedied by sufficient data augmentation [40]. On the other hand, a rich data augmentation also helps with robustness and generalization, which has been verified in previous works [18]. For improving the diversity of the augmented training data, we propose the patch-wise data augmentation strategy for ViTs, which imposes diverse augmentation on each input image patches at training time. Our motivation comes from the difference of ViTs and CNNs that ViTs not only extract intra-patch features but also concern the inter-patch relations. We think the traditional augmentation which randomly transforms the whole image could provide enough intra-patch augmentation. However, it lacks the diversity on inter-patch augmentation, as all of patches have the same transformation at one time. To impose more inter-patch diversity, we retain the original image-level augmentation, and then add the following patch-level augmentation on each image patch. For simplicity, only three basic image transformations are considered for patch-level augmentation: *random resized crop*, *random horizontal flip* and *random gaussian noise*.

Robustness of Patch-Wise Augmentation. Same with the augmentations like MixUp [52], AugMix [18], RandAugment [5], patch-wise augmentation also benefit the model robustness. It effects on the phases after conventional image-level augmentations, and provides the meaningful augmentation on patch sequence input. Different from RandAugment, which adopts augmentations conflicting with ImageNet-C, we only use simple image transform for patch-wise augmentation. It confirms that the most part of robustness improvement is derived from the strategy itself but not the used augmentation. A significant advantage of patch-wise augmentation is that it can be in common use across different ViT models and bring more than 1% and 5% improvement on standard and robust accuracy. Details can be referred to section 6.3.

6. Experiments

6.1. Experimental Settings

Implementation Details. All of our experiments are performed on the NVIDIA 2080Ti GPUs. We implement RVT in three sizes named by RVT-Ti, RVT-S, RVT-B respectively. All of them adopt the best settings investigated in section 2. For RVT*, we add PAAS on multiple transformer blocks. The patch-wise augmentation uses the combination of base augmentation introduced in section 6.4. Other training hyperparameters are same with DeiT [40].

Evaluation Benchmarks. We adopt the ImageNet-1K [7] dataset for training and standard performance evaluation. No other large-scale dataset is needed for pre-training. For robustness evaluation, we test our RVT in three aspect: 1) for adversarial robustness, we test the adversarial examples generated by white-box attack algorithms FGSM [11] and PGD [26] on ImageNet-1K validation set. ImageNet-A [19] is used for evaluating the model under natural adversarial example. 2) for common corruption robustness, we adopt ImageNet-C [17] which consists of 15 types of algorithmically generated corruptions with five levels of severity. 3) for out-of-distribution robustness, we evaluate on ImageNet-R [16] and ImageNet-Sketch [42]. They contain images with naturally occurring distribution changes. The difference is that ImageNet-Sketch only contains sketch images, which can be used for testing the classification ability when texture or color information is missing.

6.2. Standard Performance Evaluation

For standard performance evaluation, we compare our method with state-of-the-art classification methods including Transformer-based models and representative CNN-based models in Table 5. Compared to CNNs-based models, RVT has surpassed most of CNN architectures with fewer parameters and FLOPs. RVT-Ti* achieves 79.2% Top-1 accuracy on ImageNet-1K validation set, which is competitive with currently popular ResNet and RegNet series, but only has 1.3G FLOPs and 10.9M parameters (around 60% smaller than CNNs). With the same computation cost, RVT-S* obtains 81.9% test accuracy, 2.9% higher than ResNet-50. This result is closed to EfficientNet-B4, however EfficientNet-B4 requires larger 380×380 input size and has much lower throughput.

Compared to Transformer-based models, our RVT also achieves the comparable standard accuracy. We find just combining the robust components can make RVT-Ti get 78.4% Top-1 accuracy and surpass the existing state-of-the-art on ViTs with tiny version. By adopting our newly proposed position-aware attention scaling and patch-wise data augmentation, RVT-Ti* can further improve 0.8% on RVT-Ti with little additional computation cost. For other scales of the model, RVT-S* and RVT-B* also achieve a good promotion compared with DeiT-S and DeiT-B. Although the improvement becomes smaller with the increase of model capacity, we think the advance of our model is still obvious as it strengthen the model ability in various views such as robustness and out-of-domain generalization.

6.3. Robustness Evaluation

We employ a series of benchmarks to evaluate the model robustness on different aspects. Among them, ImageNet-C (IN-C) calculates the mean corruption error (mCE) as metric. The smaller mCE means the more robust of the model

Table 5. The performance of RVT and several SOTA CNNs and Transformers on ImageNet and six robustness benchmarks. RVT* represents the RVT model but trained with our proposed PAAS and patch-wise augmentation. Except for different architectures, we also compare some methods such as AugMix, which aims at improving the model robustness based on ResNet-50.

Group	Model	FLOPs	Params	ImageNet		Robustness Benchmarks					
		(G)	(M)	Top-1	Top-5	FGSM	PGD	IN-C (↓)	IN-A	IN-R	IN-SK
CNNs	ResNet-50 [15]	4.1	25.6	76.1	86.0	12.2	0.9	76.7	0.0	36.1	24.1
	ResNet-50* [15]	4.1	25.6	79.0	94.4	36.3	12.5	65.5	5.9	42.5	31.5
	Inception v3 [37]	5.7	27.2	77.4	93.4	22.5	3.1	80.6	10.0	38.9	27.6
	RegNetY-4GF [31]	4.0	20.6	79.2	94.7	15.4	2.4	68.7	8.9	38.8	25.9
	EfficientNet-B4 [39]	4.4	19.3	83.0	96.3	44.6	18.5	71.1	26.3	47.1	34.1
	ResNeXt50-32x4d [48]	4.3	25.0	79.8	94.6	34.7	13.5	64.7	10.7	41.5	29.3
	DeepAugment [16]	4.1	25.6	75.8	92.7	27.1	9.5	53.6	3.9	46.7	32.6
	ANT [33]	4.1	25.6	76.1	93.0	17.8	3.1	63.0	1.1	39.0	26.3
	AugMix [18]	4.1	25.6	77.5	93.7	20.2	3.8	65.3	3.8	41.0	28.5
	Anti-Aliased CNN [53]	4.2	25.6	79.3	94.6	32.9	13.5	68.1	8.2	41.1	29.6
	Debiased CNN [22]	4.1	25.6	76.9	93.4	20.4	5.5	67.5	3.5	40.8	28.4
Transformers	DeiT-Ti [40]	1.3	5.7	72.2	91.1	22.3	6.2	71.1	7.3	32.6	20.2
	ConViT-Ti [6]	1.4	5.7	73.3	91.8	24.7	7.5	68.4	8.9	35.2	22.4
	PiT-Ti [20]	0.7	4.9	72.9	91.3	20.4	5.1	69.1	6.2	34.6	21.6
	PVT-Tiny [43]	1.9	13.2	75.0	92.5	10.0	0.5	79.6	7.9	33.9	21.5
	RVT-Ti	1.3	8.6	78.4	94.2	34.8	11.7	58.2	13.3	43.7	30.0
	RVT-Ti*	1.3	10.9	**79.2**	**94.7**	**42.7**	**18.9**	**57.0**	**14.4**	**43.9**	**30.4**
	DeiT-S [40]	4.6	22.1	79.9	95.0	40.7	16.7	54.6	18.9	42.2	29.4
	ConViT-S [6]	5.4	27.8	81.5	95.8	41.0	17.2	49.8	24.5	45.4	33.1
	Swin-T [25]	4.5	28.3	81.2	95.5	33.7	7.3	62.0	21.6	41.3	29.1
	PVT-Small [43]	3.8	24.5	79.9	95.0	26.6	3.1	66.9	18.0	40.1	27.2
	PiT-S [20]	2.9	23.5	80.9	95.3	41.0	16.5	52.5	21.7	43.6	30.8
	TNT-S [14]	5.2	23.8	81.5	95.7	33.2	4.2	53.1	24.7	43.8	31.6
	T2T-ViT_t-14 [51]	6.1	21.5	81.7	**95.9**	40.9	11.4	53.2	23.9	45.0	32.5
	RVT-S	4.7	22.1	81.7	95.7	51.3	26.2	50.1	24.1	46.9	**35.0**
	RVT-S*	4.7	23.3	**81.9**	95.8	**51.8**	**28.2**	**49.4**	**25.7**	**47.7**	34.7
	DeiT-B [40]	17.6	86.6	82.0	95.7	46.4	21.3	48.5	27.4	44.9	32.4
	ConViT-B [6]	17.7	86.5	82.4	96.0	45.4	20.8	46.9	29.0	48.4	35.7
	Swin-B [25]	15.4	87.8	**83.4**	96.4	49.2	21.3	54.4	**35.8**	46.6	32.4
	PVT-Large [43]	9.8	61.4	81.7	95.9	33.1	7.3	59.8	26.6	42.7	30.2
	PiT-B [20]	12.5	73.8	82.4	95.7	49.3	23.7	48.2	33.9	43.7	32.3
	T2T-ViT_t-24 [51]	15.0	64.1	82.6	96.1	46.7	17.5	48.0	28.9	47.9	35.4
	RVT-B	17.7	86.2	82.5	96.0	52.3	27.4	47.3	27.7	48.2	35.8
	RVT-B*	17.7	91.8	82.7	**96.5**	**53.0**	**29.9**	**46.8**	28.5	**48.7**	**36.0**

under corruptions. All other benchmarks use Top-1 accuracy on test data if no special illustration. The results are reported in Table 5.

Adversarial Robustness. For evaluating the adversarial robustness, we adopt single-step attack algorithm FGSM [11] and multi-step attack algorithm PGD [26] with steps $t = 5$, step size $\alpha = 0.5$. Both attackers perturb the input image with max magnitude $\epsilon = 1$. Table 5 suggests that the adversarial robustness has a strong correlation with the design of model architecture. With similar model scale and FLOPs, most Transformer-based models have higher robust accuracy than CNNs under adversarial attacks. This conclusion is also consistent with [34]. Some modifications on ViTs or CNNs will also weaken or strengthen the adversarial robustness. For example, Swin-T [25] introduces window self-attention for reducing the computation cost but damaging the adversarial robustness, and EfficientNet-B4 [39] uses smooth activation functions which is helpful with adversarial robustness.

We summarize the robust design experiences of ViTs in this work. The resultant RVT model achieves superior performance on both FGSM and PGD attackers. In detail, RVT-Ti and RVT-S get over 10% improvement on FGSM, compared with the previous ViT variants. This advance is further expanded by our PAAS and patch-wise augmentation. Adversarial robustness seems unrelated with standard performance. Although models like Swin-T, TNT-S get higher standard accuracy than DeiT-S, their adversarially robust accuracy is well below the baseline. However, our RVT model can achieve the best trade-off between standard performance and adversarial robustness.

Common Corruption Robustness. To metric the model degradation on common image corruptions, we present the mCE on ImageNet-C (IN-C) in Table 5. We also list some methods from ImageNet-C Leaderboard, which are built based on ResNet-50. Our RVT-S* gets 49.4 mCE, which has 4.2 improvement on top-1 method DeepAugment [16] in the leaderboard, and bulids the new state-of-the-art. The result also indicates that Transformer-based models have a natural advantage in dealing with image corruptions. Attributed to its ability of long-range dependencies modeling, ViTs are easier to learn the shape-bias features. Note that in this work we are not considering RandAugment. As a training augmentation of ViTs, RandAugment adopts conflicted augmentation with ImageNet-C and may cause the unfairness of the comparison proposed by [1].

Out-of-distribution Robustness. We test the generalization ability of RVT on out-of-distribution data by reporting the top@1 accuracy on ImageNet-R (IN-R) and ImageNet-Sketch (IN-SK) in Table 5. Our RVT and RVT* also beat other ViT models on out-of-distribution generalization. As the superiority of Transformer-based models on capturing shape-bias features mentioned above, our RVT-S also surpasses most CNN and ViT models and get 35.0% and 46.9% test accuracy on ImageNet-Sketch and ImageNet-R, buliding the new state-of-the-art.

Layers	Pos. Emb.	Acc	Rob. Acc
0-1	Ori.	78.2	34.1
	Ours	78.4	34.3
0-5	Ori.	78.4	34.6
	Ours	78.6	35.2
0-10	Ori.	78.4	34.8
	Ours	**78.6**	**35.3**

Table 6. Comparison of single and multiple block PAAS. Ori. stands for the learned absolute position embedding in original ViTs.

Augmentations			Acc	Rob. Acc
RC	GN	HF		
✓			78.9	41.5
	✓		79.0	**42.0**
		✓	79.1	41.3
✓		✓	78.8	41.3
	✓	✓	79.0	41.9
✓	✓	✓	**79.2**	41.7

Table 7. Ablation experiments on patch-wise augmentation. RC, GN, HF represent *random resized crop*, *random gaussian noise* and *random horizontal flip* respectively.

6.4. Ablation Studies

we conduct ablation studies on the proposed components of PAAS and patch-wise augmentation in this section. Other modifications of RVT are not involved since they have been analyzed in section 2. All of our ablation experiments are based on the RVT-Ti model on ImageNet.

Single layer PAAS vs. Multiple layer PAAS. We evaluate whether using PAAS on multiple transformer blocks can benefit the performance or robustness. The result is suggested in Table 6. Learned absolute position embedding in original ViT model is adopted for comparison. With more transformer blocks using PAAS, the standard and robust accuracy gain greater enhancement. After applying PAAS on 5 blocks, the benefit of PAAS gets saturated. There will be the same trend if we replace PAAS with the original position embedding. But the original position embedding is not performed as good as our PAAS on both standard and robust accuracy.

Different types of basic augmentation. Due to the limited training resources, we only test three basic image augmentations: *random resized crop*, *random horizontal flip* and *random gaussian noise*. For random resized crop, we crop the patch according to the scale sampled from [0.85, 1.0], then resize it to original size with aspect ratio unchanged. We set the mean and standard deviation as 0 and 0.01 for random gaussian noise. For each transformation, we set the applying probability $p = 0.1$. Other hyper-parameters are consistent with the implementation in Kornia [32]. As shown in Table 7, we can see both three augmentations are beneficial of standard and robust accuracy. Among them, random gaussian noise is the better choice as it helps for more robustness improvement.

Combination of basic augmentations. We further evaluate the combination of basic patch-wise augmentations. For traditional image augmentation, combining multiple basic transformation [5] can largely improve the standard accuracy. Differently, as shown in Table 7, the benefit is marginal for combining basic patch-wise augmentations, but combination of three is still better than using only single augmentation. In this paper, we adopt the combination of all basic augmentations.

Effect on other ViT architectures. For showing the effectiveness of our proposed position-aware attention scaling and patch-wise augmentation, we apply them to train other ViT models. DeiT-Ti, ConViT-Ti and PiT-Ti are adopted as the base model. The experimental results are shown in Table 8, with combining the proposed techniques into these base models, all the augmented models achieve significant improvement. Specifically, all the improved models yield more than 1% and 5% promotion on standard and robust accuracy on average.

Vanilla models	Acc	Rob. Acc	Improved models	Acc	Rob. Acc
DeiT-Ti	72.2	22.3	DeiT-Ti*	**74.4**	**29.9**
ConViT-Ti	73.3	24.7	ConViT-Ti*	**74.4**	**30.7**
PiT-Ti	72.9	20.4	PiT-Ti*	**74.3**	**27.7**

Table 8. Effect of our proposed PAAS and patch-wise augmentation on other ViT architectures.

7. Conclusion

We systematically study the robustness of key components in ViTs, and propose Robust Vision Transformer (RVT) by alternating the modifications which would damage the robustness. Furthermore, we have devised a novel patch-wise augmentation which adds rich affinity and diversity to training data. Considering the lack of spatial information correlation in scaled dot-product attention, we present position-aware attention scaling (PAAS) method to further boost the RVT. Experiments show that our RVT achieves outstanding performance consistently on ImageNet and six robustness benchmarks. Under the exhaustive trade-offs between FLOPs, standard and robust accuracy, extensive experiment results validate the significance of our RVT-Ti and RVT-S.

References

[1] Yutong Bai, Jieru Mei, Alan L Yuille, and Cihang Xie. Are transformers more robust than cnns? *Advances in Neural Information Processing Systems*, 34, 2021. 8

[2] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. *arXiv preprint arXiv:2103.14586*, 2021. 2

[3] Xiangxiang Chu, Bo Zhang, Zhi Tian, Xiaolin Wei, and Huaxia Xia. Do we really need explicit position encodings for vision transformers? *arXiv preprint arXiv:2102.10882*, 2021. 3, 4

[4] Nadav Cohen and Amnon Shashua. Inductive bias of deep convolutional networks through pooling geometry. In *Proceedings of the International Conference on Learning Representations*, 2017. 3

[5] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020. 6, 8

[6] Stéphane d'Ascoli, Hugo Touvron, Matthew Leavitt, Ari Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. *arXiv preprint arXiv:2103.10697*, 2021. 4, 7

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009. 6

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019. 1

[9] Minjing Dong, Yanxi Li, Yunhe Wang, and Chang Xu. Adversarially robust neural architectures. *arXiv preprint arXiv:2009.00902*, 2020. 2

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

[11] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proceedings of the International Conference on Learning Representations*, 2015. 1, 2, 5, 6, 7

[12] Ben Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. Levit: a vision transformer in convnet's clothing for faster inference. *arXiv preprint arXiv:2104.01136*, 2021. 2, 4

[13] Minghao Guo, Yuzhe Yang, Rui Xu, Ziwei Liu, and Dahua Lin. When nas meets robustness: In search of robust architectures against adversarial attacks. In *Proceedings of the Computer Vision and Pattern Recognition*, pages 631–640, 2020. 2

[14] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *arXiv preprint arXiv:2103.00112*, 2021. 2, 7

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the Computer Vision and Pattern Recognition*, pages 770–778, 2016. 7

[16] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv preprint arXiv:2006.16241*, 2020. 2, 6, 7

[17] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *Proceedings of the International Conference on Learning Representations*, 2019. 2, 6

[18] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. In *Proceedings of the International Conference on Learning Representation*, 2020. 2, 6, 7

[19] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *arXiv preprint arXiv:1907.07174*, 2019. 6

[20] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. *arXiv preprint arXiv:2103.16302*, 2021. 1, 3, 7

[21] Guolin Ke, Di He, and Tie-Yan Liu. Rethinking the positional encoding in language pre-training. *arXiv preprint arXiv:2006.15595*, 2020. 5

[22] Yingwei Li, Qihang Yu, Mingxing Tan, Jieru Mei, Peng Tang, Wei Shen, Alan Yuille, and Cihang Xie. Shape-texture debiased neural network training. In *Proceedings of the International Conference on Learning Representations*, 2021. 2, 7

[23] Yawei Li, Kai Zhang, Jiezhang Cao, Radu Timofte, and Luc Van Gool. Localvit: Bringing locality to vision transformers. *arXiv preprint arXiv:2104.05707*, 2021. 1, 4

[24] Ji Lin, Chuang Gan, and Song Han. Defensive quantization: When efficiency meets robustness. In *Proceedings of the International Conference on Learning Representations*, 2019. 2

[25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021. 1, 3, 4, 7

[26] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Rtowards deep learning models resistant to adversarial attacks. In *Proceedings of the International Conference on Learning Representations*, 2018. 2, 6, 7

[27] Kaleel Mahmood, Rigel Mahmood, and Marten Van Dijk. On the robustness of vision transformers to adversarial examples. *arXiv preprint arXiv:2104.02610*, 2021. 2

[28] Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? *Advances in Neural Information Processing Systems*, 2019. 4

[29] Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. *arXiv preprint arXiv:2105.07581*, 2021. 2

[30] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018. 1

[31] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the Computer Vision and Pattern Recognition*, pages 10428–10436, 2020. 7

[32] Edgar Riba, Dmytro Mishkin, Daniel Ponsa, Ethan Rublee, and Gary Bradski. Kornia: an open source differentiable computer vision library for pytorch. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 3674–3683, 2020. 8

[33] Evgenia Rusak, Lukas Schott, Roland S Zimmermann, Julian Bitterwolf, Oliver Bringmann, Matthias Bethge, and Wieland Brendel. A simple way to make neural networks robust against diverse image corruptions. In *Proceedings of the European Conference on Computer Vision*, pages 53–69. Springer, 2020. 2, 7

[34] Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. On the adversarial robustness of visual transformers. *arXiv preprint arXiv:2103.15670*, 2021. 2, 7

[35] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018. 3

[36] Dong Su, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao. Is robustness the cost of accuracy?–a comprehensive study on the robustness of 18 deep image classification models. In *Proceedings of the European Conference on Computer Vision*, pages 631–648, 2018. 2

[37] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the Computer Vision and Pattern Recognition*, pages 2818–2826, 2016. 7

[38] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *Proceedings of the International Conference on Learning Representations*, 2014. 2

[39] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. 7

[40] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020. 1, 2, 6, 7

[41] Cristina Vasconcelos, Hugo Larochelle, Vincent Dumoulin, Nicolas Le Roux, and Ross Goroshin. An effective anti-aliasing approach for residual networks. *arXiv preprint arXiv:2011.10675*, 2020. 2

[42] Haohan Wang, Songwei Ge, Eric P Xing, and Zachary C Lipton. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 2019. 2, 6

[43] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *arXiv preprint arXiv:2102.12122*, 2021. 1, 3, 7

[44] Boxi Wu, Jinghui Chen, Deng Cai, Xiaofei He, and Quanquan Gu. Do wider neural networks really help adversarial robustness? *arXiv preprint arXiv:2010.01279*, 2020. 2

[45] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. *arXiv preprint arXiv:2103.15808*, 2021. 1

[46] Cihang Xie, Mingxing Tan, Boqing Gong, Alan Yuille, and Quoc V Le. Smooth adversarial training. *arXiv preprint arXiv:2006.14536*, 2020. 2

[47] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the Computer Vision and Pattern Recognition*, pages 10687–10698, 2020. 2

[48] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the Computer Vision and Pattern Recognition*, pages 1492–1500, 2017. 7

[49] Shaokai Ye, Kaidi Xu, Sijia Liu, Hao Cheng, Jan-Henrik Lambrechts, Huan Zhang, Aojun Zhou, Kaisheng Ma, Yanzhi Wang, and Xue Lin. Adversarial robustness vs. model compression, or both? In *Proceedings of the International Conference on Computer Vision*, pages 111–120, 2019. 2

[50] Kun Yuan, Shaopeng Guo, Ziwei Liu, Aojun Zhou, Fengwei Yu, and Wei Wu. Incorporating convolution designs into visual transformers. *arXiv preprint arXiv:2103.11816*, 2021. 1, 2, 4

[51] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv preprint arXiv:2101.11986*, 2021. 2, 7

[52] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *Proceedings of the International Conference on Learning Representations*, 2018. 6

[53] Richard Zhang. Making convolutional networks shift-invariant again. In *Proceedings of International Conference on Machine Learning*, pages 7324–7334. PMLR, 2019. 2, 7