Improving Robustness of Vision Transformers by Reducing Sensitivity to Patch Corruptions

Yong Guo, David Stutz, Bernt Schiele
Max Planck Institute for Informatics, Saarland Informatics Campus
{yongguo,david.stutz,schiele}@mpi-inf.mpg.de

Abstract

*Despite their success, vision transformers still remain vulnerable to image corruptions, such as noise or blur. Indeed, we find that the vulnerability mainly stems from the unstable self-attention mechanism, which is inherently built upon patch-based inputs and often becomes overly sensitive to the corruptions across patches. For example, when we only occlude a small number of patches with random noise (e.g., 10%), these patch corruptions would lead to severe accuracy drops and greatly distract intermediate attention layers. To address this, we propose a new training method that improves the robustness of transformers from a new perspective – **reducing sensitivity to patch corruptions (RSPC)**. Specifically, we first identify and occlude/corrupt the most vulnerable patches and then explicitly reduce sensitivity to them by aligning the intermediate features between clean and corrupted examples. We highlight that the construction of patch corruptions is learned adversarially to the following feature alignment process, which is particularly effective and essentially different from existing methods. In experiments, our RSPC greatly improves the stability of attention layers and consistently yields better robustness on various benchmarks, including CIFAR-10/100-C, ImageNet-A, ImageNet-C, and ImageNet-P.*

1. Introduction

Despite the success of vision transformers [10] in recent years, they still lack robustness against common image corruptions [24, 52], such as noise or blur, and adversarial perturbations [13, 15, 42]. For example, even for the state-of-the-art robust architectures, e.g., RVT [34] and FAN [61], the accuracy drops by more than 15% on corrupted examples, e.g., with Gaussian noise, as shown in Figure 2 (blue star on the right). We suspect that this vulnerability is inherent to the used self-attention mechanism, which relies on patch-based inputs and may easily become overly sensitive to corruptions or perturbations upon them.

Figure 1. Sensitivity to patch perturbations/corruptions in terms of confidence score of the ground-truth class. We randomly select 10% patches to be perturbed/corrupted for RVT-Ti [34]. In practice, adversarial patch perturbations (often invisible) significantly reduce the confidence, indicating the high sensitivity of transformers to patches. However, directly adding random noise only yields marginal degradation even with the highest severity in ImageNet-C [24]. By contrast, occluding patches with noise greatly reduces the confidence and can be used as a good proxy of adversarial patch perturbations to reveal the patch sensitivity issue.

A piece of empirical evidence is that transformers can be easily misled by the adversarial perturbations only on very few patches (even a single patch [13]). As shown in Figure 1, given a clean image, we randomly sample a small number of patches, e.g., 10%, and introduce perturbations/corruptions into them. Considering RVT [34] as a strong baseline, when we generate adversarial perturbations using PGD-5, these perturbed patches greatly reduce the confidence score from 63.8% to 3.1% and result in a misclassification. Nevertheless, generating adversarial perturbations can be very computationally expensive (e.g., 5× longer training time for PGD-5), which makes adversarial training often infeasible on large-scale datasets [26, 39, 53], e.g., ImageNet. Instead, an efficient way is directly adding corruptions, e.g., random noise, on top of these patches. In practice, even with the highest severity in ImageNet-C [24], these corrupted patches only yield a marginal degradation in terms of confidence score. Thus, how to construct patch corruptions that greatly misleads the model and can be produced very efficiently becomes a critical problem.

Interestingly, if we totally discard these patches and occlude them with random noise, the model becomes very vulnerable again, e.g., with the confidence score dropping from 63.8% to 17.3% in Figure 1. More critically, these corrupted

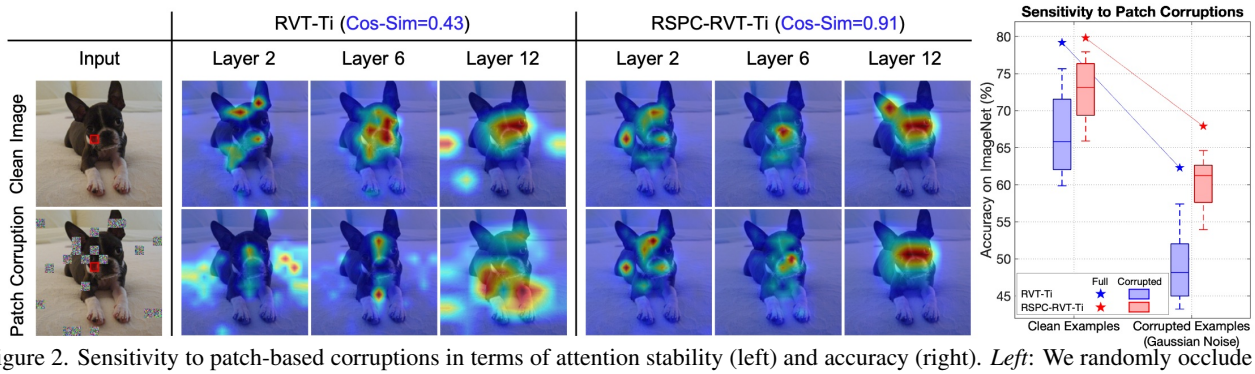

Figure 2. Sensitivity to patch-based corruptions in terms of attention stability (left) and accuracy (right). *Left*: We randomly occlude 10% patches with noise and show the attention maps of different layers in RVT-Ti [34] and our RSPC-RVT-Ti. Following [13], we choose the center patch (red square) as the query and average the attention scores across all the attention heads for visualization. Regarding this example, we also compute the average cosine similarity (Cos-Sim) between the clean and corrupted attentions across different layers. Clearly, our RSPC model yields more stable attention maps. *Right*: On ImageNet, we plot the distribution of accuracy on the occluded examples with different occlusion masks. Here, we randomly sample 100 different masks for each image. We show that RVT is very sensitive to the patch-based corruptions and has a much larger variance of accuracy than our RSPC model.

patches also have significant impact on the attention maps across layers, as shown in Figure 2 (left). We suspect this to be the case due to the global interactions across tokens in the attention mechanism – even when occluding only few patches. Quantitatively, this can be captured by computing the average cosine similarity between the attentions on clean and corrupted images across layers, denoted by Cos-Sim. Regarding the considered example in Figure 2, the Cos-Sim of only 0.43 for RVT indicates a significant shift in attention – a phenomenon that we can observe across the entire ImageNet dataset (see Figure 5). In fact, these attention shifts also have direct and severe impact on accuracy: In Figure 2 (right), we randomly sample 100 occlusion masks for each image and show the distribution of accuracy (blue box). Unsurprisingly, the accuracy decreases significantly when facing patch-based corruptions, compared to the original examples (blue star). These experiments highlight the need for an inherently more robust attention mechanism in order to improve the overall robustness of transformers.

We address this problem by finding particularly vulnerable patches to construct patch-based corruptions and stabilizing the intermediate attention layers against them. Since we use random noise to occlude patches, we move the focus from how to perturb patch content to finding which patch should be occluded. As shown in Figure 2 (right), with a fixed occlusion ratio, the accuracy varies a lot when occluding different patches (e.g., ranging from 60% to 75% in the blue box). Since we seek to reduce the sensitivity to patch corruptions, occluding the most vulnerable (often very important) patches and explicitly reducing the impact of them should bring the largest robustness improvement. Inspired by this, we seek to identify the most vulnerable patches to construct patch-based corruptions and then align the intermediate features to make the attention less sensitive to the corruptions in individual patches. In practice, we are able to reduce the impact of patch-based corruptions significantly, improving the Cos-Sim from 0.43 (for RVT-Ti) to 0.91 in Figure 2 (left). This is also directly observed in the visual results where these corruptions have little impact on the intermediate attention maps of our robust model. The stable attention mechanism also greatly improves the robustness of transformers. As shown in Figure 2 (right), compared with RVT, we obtain significantly higher accuracy when facing examples with different occlusion masks (red box), alongside the improved overall accuracy and robustness on full images (red star).

Contributions: In this paper, we study the sensitivity of transformers to patch corruptions and explicitly stabilize models against them to improve the robustness. Here, we make three key contributions: *1)* We propose a new training method that improves robustness by **reducing sensitivity to patch corruptions (RSPC)**. To this end, we first construct effective patch-based corruptions and then reduce the sensitivity to them by aligning the intermediate features. *2)* When constructing patch corruptions, we develop a patch corruption model to find particularly vulnerable patches that severely distract intermediate attention layers. In practice, the corruption model is trained adversarially to the classification model, which, however, is essentially different from adversarial training methods. To be specific, we only learn *which patch should be corrupted* instead of the pixel-level perturbations. *3)* In experiments, we demonstrate that the robustness improvement against patch corruptions (shown in Figure 2 (right)) can generalize well to diverse architectures on various robustness benchmarks, including ImageNet-A/C/P [24, 60]. More critically, we can show, both qualitatively and quantitatively, that these improvements stem from the more stable attention mechanism across layers. It is worth noting that, when compared with adversarial training methods, RSPC obtains a better tradeoff between accuracy and corruption robustness while keeping significantly lower training cost [57] (see Figure 7).

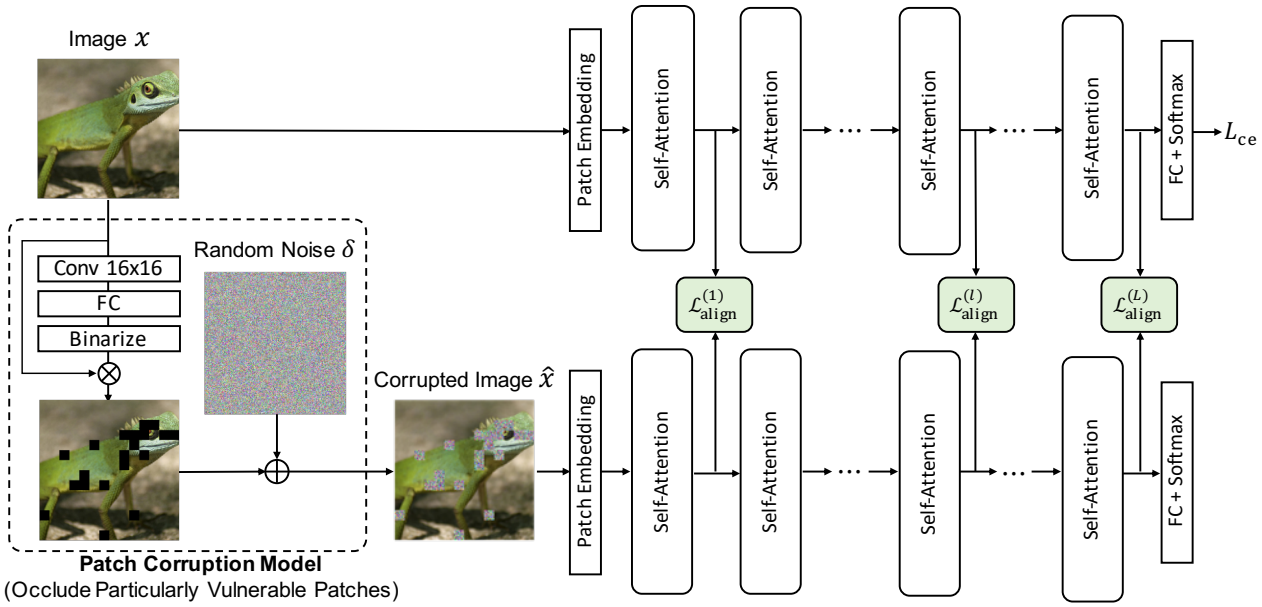

Figure 3. Overview of the proposed **reducing sensitivity to patch corruptions (RSPC)** training procedure. We present a patch corruption model to produce patch-based corruptions and align the features of each self-attention block between the clean and corrupted examples (the alignment loss is highlighted by green box). Unlike existing methods, we select the patches to be occluded/corrupted in an adversarial way, i.e., corrupting the most vulnerable patches that would greatly distract intermediate attention layers.

2. Related Work

Vision Transformers (ViTs) [10, 25, 49] have achieved remarkable performance in various learning tasks. Besides improving clean accuracy, many works seek to study and improve the robustness of ViTs [2, 4, 5, 18, 20, 37, 43]. Interestingly, ViTs are often more robust than convolutional networks against corruptions [38, 45, 47] and adversarial attacks [3, 14, 29, 31, 33, 36, 41, 46]. To further improve robustness, RVT [34] develops a robust transformer by comparing different designs for each component, and presents a patch-wise augmentation. FAN [61] combines token attention and channel attention [1] and yields new state-of-the-arts. However, even for these robust transformers, there is still a large gap between clean and robust accuracy. More critically, the impact of corruptions/perturbations on the key component of ViTs, i.e., self-attention, still remains poorly understood.

Besides the above, an intuitive way to improve robustness is to reduce the gap of intermediate features between clean and corrupted examples, e.g., using feature alignment [6, 8, 19, 44, 50, 55, 59]. Typically, feature alignment aligns the features of examples from two different domains. However, when considering corruption robustness, the corruption type of test data is often unknown and the corrupted training examples are also unavailable. To tackle this, we focus on ViTs and develop a patch corruption model to produce effective corrupted examples. Based on these patch corruptions, we investigate the sensitivity of ViTs to them and develop a training method to improve robustness.

3. Reducing Sensitivity to Patch Corruptions

We suspect that the vulnerability of state-of-the-art transformers stems from the inherent sensitivity of self-attention mechanism to the input patches. To verify this, we first investigate the sensitivity of transformers to diverse patch corruptions/perturbations in Section 3.1. To alleviate this issue, we explicitly **reduce sensitivity to patch corruptions (RSPC)** to improve the robustness. Specifically, in Section 3.2, we first develop an effective occlusion-based patch corruption scheme that identifies particularly vulnerable patches to construct patch corruptions. Then, in Section 3.3, we propose to stabilize attention layers against patch corruptions by aligning the intermediate features between the corrupted and clean images, as shown in Figure 3.

3.1. Sensitivity of Transformers to Patches

In this part, we comprehensively compare different patch perturbations/corruptions to investigate the sensitivity of transformers to their input patches. As shown in Figure 1, RVT yields extremely low confidence against adversarial patch perturbations, showing that transformers are very sensitive to individual patches. Nevertheless, generating adversarial perturbations is very computationally expensive and often becomes infeasible to perform adversarial training on large-scale datasets [32, 54, 57]. This encourages us to explore how to reveal and alleviate the patch sensitivity issue of transformers in a more efficient way. Actually, a simple way is directly introducing corruptions, e.g., random noise, into patches. In Figure 1, based on the same patch mask,

adding noise (even with the highest severity in ImageNet-C [24]) only slightly reduces the confidence by 2.5%, sharing a similar observation with concurrent work [16]. The main reason is that the corrupted patches still contain abundant information that can be easily used to build strong correlations with their neighboring patches.

Interestingly, we find that the model can be easily misled again by only incorporating small modifications into the standard patch corruption method. Specifically, we propose an occlusion-based patch corruption scheme that occludes patches with random noise, as shown in the last column in Figure 1. Formally, given an example x and a binary mask $M(x)$ on patches, we occlude the selected ones with a random noise δ sampled from the uniform distribution. Thus, the corrupted example can be represented by

$$\hat{x} = M(x) \cdot x + (1 - M(x)) \cdot \delta. \tag{1}$$

As for the example in Figure 1, the proposed patch corruption scheme greatly reduces the confidence score by 56.5%. More critically, as shown in Figure 2, these corrupted patches also significantly distract the intermediate attention layers. In this way, the occlusion-based corruptions can be regarded as a good proxy of adversarial patch perturbations to reveal the patch sensitivity issue. Moreover, unlike directly dropping patches, occluding patches with noise is more challenging for the model and particularly effective in practice (see results in supplementary). We highlight that we are not trying to improve the accuracy against patch corruptions but to improve the overall robustness from a new perspective, i.e., reducing sensitivity to patch corruptions.

3.2. Finding Vulnerable Patches to be Corrupted

As shown in Figure 2 (right, blue box), randomly occluding/corrupting patches often incurs significantly different impacts on accuracy between the best and worst case. This leads us to consider what patches should be occluded to perform effective feature alignment. Since we seek to reduce the patch sensitivity, we propose a patch corruption model to construct the worst case. As shown in Figure 3 (bottom left), given an example x and an occlusion ratio ρ, the corruption model predicts a binary mask $M(x) := \mathcal{C}(x; \rho)$ to determine which patches should be corrupted. Due to page limit, we put the ablation study on ρ in supplementary.

Recall that we seek to reduce the patch sensitivity in all the attention layers, we propose to find those vulnerable patches that, once occluded, can greatly distract the intermediate attention layers. To this end, we train the patch corruption model by maximizing the distance of intermediate features between clean and corrupted examples. Let $\mathcal{F}_l(x)$ be the features obtained at the l-th layer for x. Given a model with L layers, the training objective of the patch corruption model becomes

$$\max_{\mathcal{C}} \ \mathbb{E}_{x \sim \mathcal{D}} \ \mathcal{L}_{\text{align}}(x, \hat{x}),$$
$$\text{where } \ \mathcal{L}_{\text{align}}(x, \hat{x}) = \frac{1}{L} \sum\nolimits_{l=1}^{L} \|\mathcal{F}_l(x) - \mathcal{F}_l(\hat{x})\|^2. \tag{2}$$

Here, $\mathcal{D}$ denotes the distribution of data and $\mathcal{L}_{\text{align}}(x, \hat{x})$ denotes the feature alignment loss that measures the average feature distance over all the attention layers. Compared with directly maximizing the cross-entropy loss, maximizing $\mathcal{L}_{\text{align}}$ explicitly distracts the attention layers and in practice performing alignment against it brings larger robustness improvement (see results in supplementary).

Architecture of patch corruption model. As shown in Figure 3 (bottom left), the corruption model is a lightweight network that contains a convolution followed by a fully connected layer and a binarization layer. The binarization layer is essentially a (hard) threshold function that selects top ρ of the patches to be occluded and keeps the rest unchanged. Following [27], we use the Straight Through Estimator (STE) to make binarization operation differentiable.

Algorithm 1 Training transformer models by **reducing sensitivity to patch corruptions (RSPC)**. We train the classification model $\mathcal{F}$ and the corruption model $\mathcal{C}$ in an end-to-end manner. In each iteration, we descend the gradient for $\mathcal{F}$ and ascend the gradient for $\mathcal{C}$, respectively.

Require: Training data $\mathcal{D}$, model parameters $\theta_{\mathcal{C}}$ and $\theta_{\mathcal{F}}$, occlusion ratio ρ, step size η, hyper-parameter λ.

for each training iteration **do**
  Sample a data batch $\{x_i\}_{i=1}^N$ from $\mathcal{D}$
  // Construct patch-based corruptions $\hat{x}$
  Sample the random noise δ from a uniform distribution
  Construct $\hat{x}$ using the patch corruption model $\mathcal{C}$:
  $\hat{x} = \mathcal{C}(x; \rho) \cdot x + (1 - \mathcal{C}(x; \rho)) \cdot \delta$
  // Update the classification model $\mathcal{F}$
  Update $\theta_{\mathcal{F}}$ by descending the gradient:
  $\theta_{\mathcal{F}} = \theta_{\mathcal{F}} - \eta \frac{1}{N} \sum_{i=1}^N \nabla_{\theta_{\mathcal{F}}} [\mathcal{L}_{\text{ce}}(x_i) + \lambda \mathcal{L}_{\text{align}}(x_i, \hat{x}_i)]$
  // Update the patch corruption model $\mathcal{C}$
  Update $\theta_{\mathcal{C}}$ by ascending the gradient:
  $\theta_{\mathcal{C}} = \theta_{\mathcal{C}} + \eta \frac{1}{N} \sum_{i=1}^N \nabla_{\theta_{\mathcal{C}}} \lambda \mathcal{L}_{\text{align}}(x_i, \hat{x}_i)$
end for

3.3. Reducing Sensitivity via Feature Alignment

Based on the constructed patch corruptions, we seek to stabilize the self-attention layers by aligning the intermediate features between clean and patch-based corrupted examples, as shown in Figure 3. To make sure that we can always construct the most challenging corrupted examples w.r.t. the latest classification model, we simultaneously train the corruption model $\mathcal{C}$ and the classification model $\mathcal{F}$ using an adversarial objective. Specifically, we minimize both the cross-entropy loss $\mathcal{L}_{\text{ce}}(x)$ and the alignment loss

	Model	#Params (M)	CIFAR-10 (%)	CIFAR-10-C (%)	CIFAR-100 (%)	CIFAR-100-C (%)
CNN	ResNet50 [22]	23.5	94.77	84.81	76.43	66.75
	ResNet50†	23.5	96.01	87.53	81.16	68.08
	PRIME [35]	11.7	93.06	89.05	77.60	68.28
	NoisyMix [12]	6.1	96.73	92.78	81.16	72.06
	AdA [7]	23.5	94.93	92.17	-	-
	CARD-Decks [9]	44.6	96.80	92.75	80.60	71.30
ViT	Swin-T [30]	27.6	95.84	90.25	81.83	70.87
	DeiT-S [48]	21.7	95.30	89.01	79.84	68.79
	ConViT-S [11]	27.4	96.90	91.73	82.43	71.39
	RVT-S [34]	23.0	97.21	92.35 (+0.00)	84.13	73.43 (+0.00)
	+ RSPC (Ours)	23.0	97.73	94.14 (+1.79)	84.81	74.94 (+1.51)
	FAN-S-Hybrid [61]	25.7	97.69	93.14 (+0.00)	84.92	74.19 (+0.00)
	+ RSPC (Ours)	25.7	**98.06**	**94.59 (+1.45)**	**85.30**	**75.72 (+1.53)**

Table 1. Comparisons with the state-of-the-art on CIFAR-10 and CIFAR-100. We evaluate clean accuracy on the original test set and robust accuracy on the corresponding corrupted datasets, i.e., CIFAR-10-C and CIFAR-100-C. We show that our RSPC significantly improves the robustness on both datasets. † denotes models with the same training recipe as used for our RSPC.

$\mathcal{L}_{\text{align}}(x, \hat{x})$ to train $\mathcal{F}$, while maximizing the alignment loss for $\mathcal{C}$. Since $\mathcal{L}_{\text{ce}}(x)$ only relies on the clean example, our training objective can be equivalently formulated as

$$\min_{\mathcal{F}} \max_{\mathcal{C}} \ \mathbb{E}_{x\sim\mathcal{D}}[\mathcal{L}_{\text{ce}}(x) + \lambda\mathcal{L}_{\text{align}}(x, \hat{x})], \qquad (3)$$

where λ determines the importance of $\mathcal{L}_{\text{align}}$. To solve the minimax problem (3), we update the models $\mathcal{F}$ and $\mathcal{C}$ by descending and ascending the gradients, respectively. As shown in Algorithm 1, we first produce the corrupted examples $\hat{x}$ using $\mathcal{C}$ and descend the gradient w.r.t. Eqn. (3) to update the parameters $\theta_{\mathcal{F}}$ of the classification model $\mathcal{F}$. Then, we ascend the gradients to update the parameters $\theta_{\mathcal{C}}$ of the corruption model $\mathcal{C}$, enforcing it to produce the worst-case corrupted examples. When ascending the gradient, we can directly change the sign of gradients, making it possible for end-to-end training.

4. Experiments

We conduct extensive experiments to evaluate our RSPC based on two state-of-the-art robust architectures, including RVT [34] and FAN [61]. In Section 4.1, we first justify our method on CIFAR datasets and show that RSPC achieves new state-of-the-arts on two corruption benchmarks, namely CIFAR-10-C and CIFAR-100-C. Then, in Section 4.2, we conduct comparisons on ImageNet and demonstrate that RSPC greatly improves the robustness on various robustness benchmarks, including ImageNet-A, ImageNet-C, and ImageNet-P. Our code is available at https://github.com/guoyongcs/RSPC.

4.1. Comparisons on CIFAR-10 and CIFAR-100

In this experiment, we train the models from scratch on CIFAR-10/100 and compare both the accuracy and corruption robustness. Following [7], we use DeepAugment [23] and train the models for 200 epochs. We adopt the batch size of 128 and use cosine decay to adjust the learning rate. For fair comparisons, we consider RVT-S [34] and FAN-S-Hybrid [61] as the baselines since they contain approximately the same number of parameters with popular CNNs and transformers. In all the experiments, by default, we set $\lambda = 5\times10^{-3}$ and $\rho = 10\%$ for our RSPC models. Due to the page limit, we put the ablations on these hyper-parameters in our supplementary.

In Table 1, we compare our RSPC models with both state-of-the-art CNNs [7,9,12,17,28,35] and popular transformer models [11,30,34,48,61]. To make fair comparisons with CNNs, we also apply the training recipe of transformers to train a ResNet50 model, denoted by ResNet50† in Table 1. Specifically, we do not exploit our patch corruption model or feature alignment, but directly apply the same augmentation for training. Compared with CNNs, transformers tend to obtain higher accuracy but do not necessarily exhibit better robustness, such as Swin [30], DeiT [48], and ConViT [11]. As for the carefully designed robust architecture RVT [34] and FAN [61], they both greatly improve the robustness and outperform existing methods in most cases. Compared with the RVT and FAN baselines, our RSPC models further improve the corruption robustness by a large margin, i.e., with the improvement larger than 1.4% on both CIFAR-10-C and CIFAR-100-C. More critically, our RSPC-FAN-S-Hybrid modes achieve new state-of-the-art results on both benchmarks.

4.2. Comparisons on ImageNet

On ImageNet, we apply our method on top of both RVT [34] and FAN [61]. Again, we closely follow the settings of them for training. To evaluate the robustness, we consider several robustness benchmarks, including ImageNet-A (IN-A) [60], ImageNet-C [24], and ImageNet-P (IN-P) [24]. Since we also introduce noise to construct the patch-based corruptions, we also report the results on IN-C without the corruption types related to noise (i.e., excluding Gaussian Noise, Shot Noise, and Impulse Noise from the

	Model	#FLOPs (G)	#Params (M)	ImageNet	Robustness Benchmarks			
					IN-A	IN-C ↓	IN-C w/o Noise ↓	IN-P ↓
CNN	ResNet50 [22]	4.1	25.6	76.1	0.0	76.7	76.0	58.0
	ANT [40]	4.1	25.6	76.1	1.1	63.0	64.3	53.2
	EWS [17]	4.1	25.6	77.3	5.9	58.7	60.2	30.9
	DeepAugment [23]	4.1	25.6	75.8	3.9	60.6	52.2	32.1
ViT-Tiny	DeiT-Ti [48]	1.3	5.7	72.2	7.3	71.1	72.9	56.7
	ConViT-Ti [11]	1.4	5.7	73.3	8.9	68.4	70.4	53.7
	PVT-Tiny [51]	1.9	13.2	75.0	7.9	69.1	70.0	60.1
	RVT-Ti [34]	1.3	10.9	79.2	14.6 (+0.0)	57.0 (-0.0)	58.9 (-0.0)	39.1 (-0.0)
	+ RSPC (Ours)	1.3	10.9	79.5	16.5 (+1.9)	**55.7 (-1.3)**	**57.5 (-1.4)**	38.0 (-1.1)
	FAN-T-Hybrid [61]	3.5	7.5	80.1	21.9 (+0.0)	58.3 (-0.0)	59.8 (-0.0)	38.3 (-0.0)
	+ RSPC (Ours)	3.5	7.5	**80.3**	**23.6 (+1.7)**	57.2 (-1.1)	58.4 (-1.4)	**37.3 (-1.0)**
ViT-Small	DeiT-S [48]	4.6	22.1	79.9	6.3	54.6	56.6	36.9
	ConViT-S [11]	5.4	27.8	81.5	18.9	49.8	52.1	35.8
	Swin-T [30]	4.5	28.3	81.2	21.6	62.0	64.2	38.3
	PVT-Small [51]	3.8	24.5	79.9	18.0	66.9	70.0	45.1
	T2T-ViT_t-14 [56]	6.1	21.5	81.7	23.9	53.2	54.4	36.2
	RVT-S [34]	4.7	23.3	81.9	25.7 (+0.0)	49.4 (-0.0)	51.6 (-0.0)	35.2 (-0.0)
	+ RSPC (Ours)	4.7	23.3	82.2	27.9 (+2.2)	48.4 (-1.0)	50.4 (-1.2)	34.3 (-0.9)
	FAN-S-Hybrid [61]	6.7	25.7	83.5	33.9 (+0.0)	48.5 (-0.0)	50.7 (-0.0)	34.5 (-0.0)
	+ RSPC (Ours)	6.7	25.7	**83.6**	**36.8 (+2.9)**	**47.5 (-1.0)**	**49.4 (-1.3)**	**33.5 (-1.0)**
ViT-Base	MAE (ViT-B) [21]	17.6	86.6	83.6	35.9	51.7	-	-
	DeiT-B [48]	17.6	86.6	82.0	27.4	48.5	50.9	32.1
	ConViT-B [11]	17.7	86.5	82.4	29.0	46.9	49.3	32.2
	Swin-B [30]	15.4	87.8	83.4	35.8	54.4	57.0	32.7
	PVT-Large [51]	9.8	61.4	81.7	26.6	59.8	63.0	39.3
	T2T-ViT_t-24 [56]	15.0	64.1	82.6	28.9	48.0	49.3	31.8
	RVT-B [34]	17.7	91.8	82.6	28.5 (+0.0)	46.8 (-0.0)	49.8 (-0.0)	31.9 (-0.0)
	+ RSPC (Ours)	17.7	91.8	82.8	32.1 (+3.6)	45.7 (-1.1)	48.5 (-1.3)	31.0 (-0.8)
	FAN-B-Hybrid [61]	11.3	50.5	83.9	39.6 (+0.0)	46.1 (-0.0)	48.1 (-0.0)	31.3 (-0.0)
	+ RSPC (Ours)	11.3	50.5	**84.2**	**41.1 (+1.5)**	**44.5 (-1.6)**	**46.8 (-1.3)**	**30.0 (-1.2)**

Table 2. Comparisons of robustness on ImageNet. We report the mean corruption error (mCE) on ImageNet-C and mean flip rate (mFR) on ImageNet-P. The lower mCE or mFR is, the more robust the model is. Across different model sizes, our RSPC models consistently improve the robustness compared with the considered baseline.

15 corruption types). Following [24], we report the mean corruption error (mCE) on IN-C (also IN-C w/o Noise) and mean flip rate (mFR) on IN-P. For both metrics, *lower is better*. As shown in Table 2, compared with the baselines, our RSPC models consistently improve the robustness on IN-A by $>1.5\%$ across different model sizes while keeping comparable clean accuracy. Moreover, we reduce the corruption error by $>1.0\%$ on IN-C and by $>1.2\%$ on IN-C without noise corruption types. This indicates that our method not only improves the robustness on noise corruptions but also generalizes well to the other corruptions (see results on individual corruption type in supplementary). When evaluating the stability against perturbations on IN-P, our RSPC models also show clear superiority over the RVT and FAN baselines. Moreover, as will be shown in supplementary, our RSPC consistently obtains better robustness against patch perturbations (e.g., Patch-Fool [13]), patch corruptions, and adversarial attacks (e.g., PGD [32]). Overall, these experiments indicate that explicitly reducing sensitivity to patches is particularly effective in improving robustness.

5. Analysis and Discussions

5.1. Visualization Results and More Analysis

Stability of intermediate attention maps. We directly visualize how much the intermediate attentions would be changed when facing patch-based corruptions. We take RVT-Ti as the baseline and compare the attention maps between RVT-Ti and our RSPC-RVT-Ti. Following [13], we average the attention maps across all the attention heads in each layer and visualize the attention map for a query token, e.g., the center token highlighted by the red box. In Figure 4, we show that RVT often incurs significant changes in the attention maps. By contrast, our RSPC effectively preserves most of the regions with relatively high attention scores across layers. We also quantitatively evaluate the attention stability by computing the *cosine similarity* between the attention maps extracted from the clean and patch-based corrupted examples. Here, we compute the cosine similarity for each head in all the layers and then report the average score over them. Figure 5 plots histograms of attention sim-

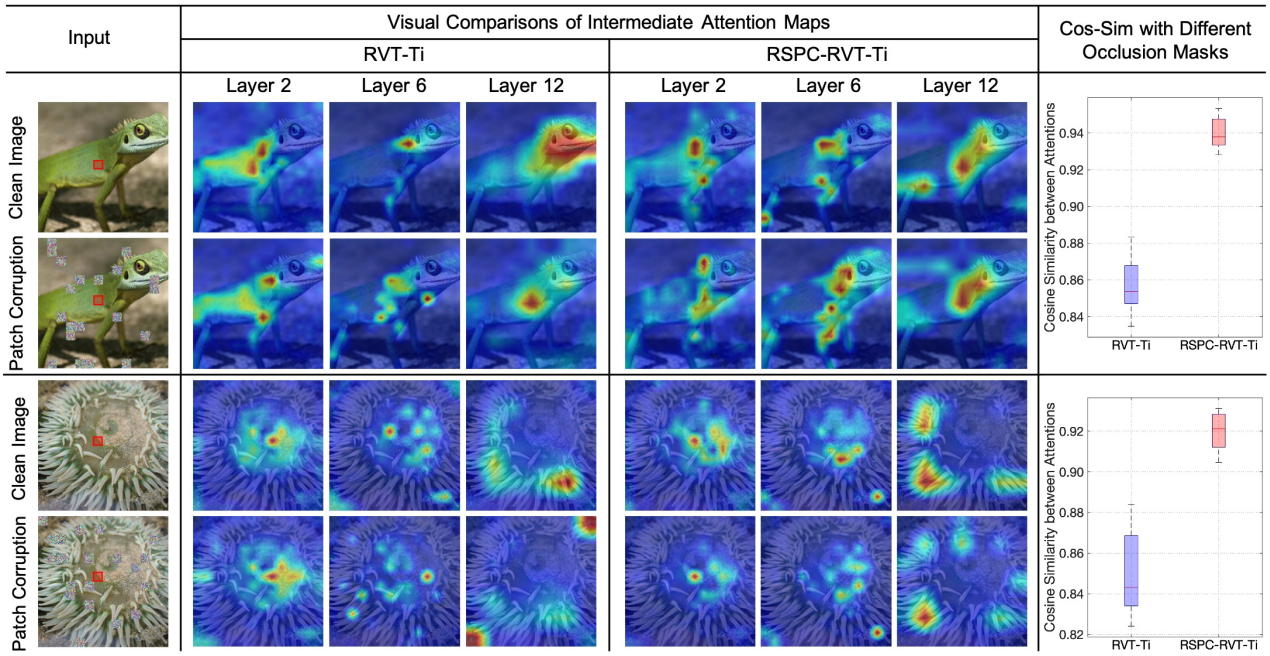

Figure 4. Comparisons of attention stability between RVT-Ti and our RSPC-RVT-Ti. We adopt the same method as that in Figure 2 to obtain the attention maps for visualization. In the last column, we also investigate the impact of different occlusion masks (1000 random masks) on each example and quantitatively evaluate the stability using cosine similarity (Cos-Sim). Clearly, our RSPC model yields much more stable attention maps both qualitatively and quantitatively.

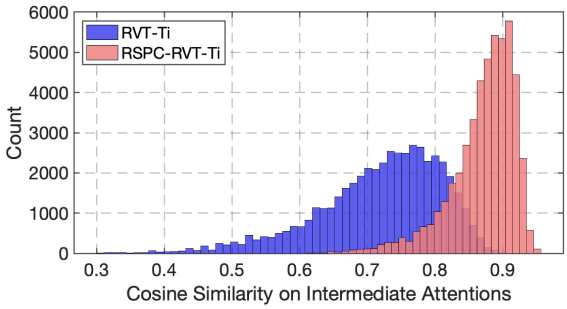

Figure 5. Histogram of cosine similarity on intermediate attention maps on ImageNet. For each image, we construct the corrupted example using a random occlusion mask and compute the average cosine similarity across layers. Clearly, our RSPC model yields much more stable attentions than RVT.

ilarity scores across the whole validation set of ImageNet. Clearly, RSPC increases the similarity both on average and in the worst-case across the whole dataset. In addition, Figure 4 (last column) also studies impact of different occlusion masks and shows the distribution of this score for two example images, each with 1000 random occlusion masks.

Patch-based corruptions generated by $\mathcal{C}$. We also visualize the patch-based corruptions generated by our patch corruption model $\mathcal{C}$ in Figure 6. By maximizing the feature alignment loss, the corruption model often identifies those patches that have major contributions in the attention mechanism but on the other hand would make the attention very unstable if they are corrupted. In practice, the patch corruption model tends to occlude the patches that are mainly located in the key part of the object, e.g., the eyes of the dog in the first example. Moreover, as detailed in supplementary, we also observe that the generated patch corruptions often greatly reduce the confidence score. By contrast, as we explicitly perform feature alignment against these patch corruptions, our RSPC model still yields promising confidence score and thus comes with better robustness.

5.2. More Results and Ablations

Comparisons with adversarial training methods. Both RSPC and adversarial training (AT) exploit an adversarial objective, but our they still have several essential differences. **1)** They have different goals. AT learns pixel perturbations and aims to improve adversarial robustness, while putting less importance on accuracy (often with large drop) and corruption robustness (marginal improvement). To be specific, we compare a popular adversarial training method TRADES [58] with various attack steps $K = \{1, 2, 3\}$. From Figure 7, adversarial training is not very attractive due to its large accuracy drop, marginal robustness improvement against patch perturbations (e.g., Patch-Fool [13]) and corruptions on IN-C, and high training cost. By contrast, RSPC learns which patches should be corrupted and finds a better tradeoff between accuracy and robustness. **2)** They produce examples in different ways. AT iteratively optimizes pixels but RSPC learns to *generate* examples. Unlike AT with multiple forward-backward propagations on the full model $\mathcal{F}$, RSPC produces examples more efficiently with single

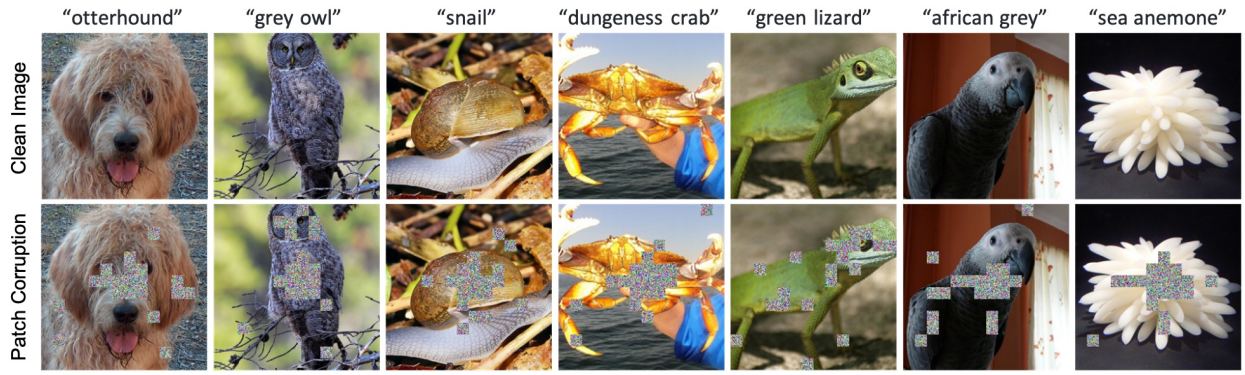

Figure 6. Visualization of the patch-based corrupted examples produced by the proposed patch corruption model. The corruption model often identifies those patches that are often located at the key part of the object, e.g., face or body of an animal.

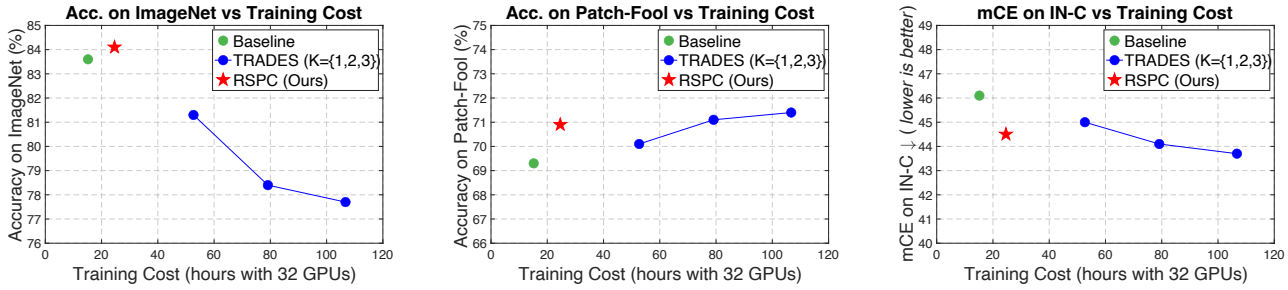

Figure 7. Comparisons with adversarial training (TRADES with $\epsilon=1/255$) on FAN-B-Hybrid. For mCE (last plot), lower is better. Clearly, our RSPC model obtains a better tradeoff between accuracy and robustness than TRADES, along with significantly lower training cost.

forward propagation on a *lightweight* corruption model $\mathcal{C}$ ($\sim$80x faster than AT even with single iteration). **3)** They train models in different manners. AT trains alternatingly by computing adversarial examples (via iterative optimization) and then updating models in each iteration. By contrast, RSPC trains the model end-to-end and is much more efficient than AT, as shown in Figure 7.

Patch selection strategy. As mentioned in Section 3.2, we find the vulnerable patches to be occluded in an adversarial way. To justify this, we compare our method with the random patch selection strategy. Besides the baseline model, we additionally compare the model trained on both clean samples and the ones with patch corruptions. As shown in Table 3, training with the adversarial patch selection strategy greatly outperforms the random strategy on IN-C. This experiment indicates that adversarially selecting patches to introduce corruptions is particularly effective.

Effect of RSPC on diverse architectures. Besides RVT and FAN, we apply our RSPC on top of more architectures, including DeiT [48] and Swin [30]. Based on DeiT-Ti, we greatly improve the robustness on IN-C and reduce the mCE by 1.4% while yielding a promising improvement of 0.4% on clean data. As for Swin-T, we obtain a similar observation that our RSPC is particularly effective in improving corruption robustness, reducing mCE from 62.0% to 61.0% (see details in supplementary). These results indicate that our RSPC can generalize well across diverse architectures.

Training Method	FAN-B-Hybrid	
	Imagenet	IN-C ↓
Training on Clean Data (Baseline)	83.9	46.1 (-0.0)
Training on Both Clean and Corrupted Data	84.1	45.8 (-0.3)
RSPC on Randomly Selected Patches	84.1	45.6 (-0.5)
RSPC on Patches Selected by $\mathcal{C}$	**84.2**	**44.5 (-1.6)**

Table 3. Comparisons of different strategies for feature alignment on ImageNet and ImageNet-C (IN-C). We take RVT-B and FAN-B-Hybrid as the baselines in this experiment. We show that finding and occluding particularly vulnerable patches yields significantly better robustness the random strategy while keeping comparable accuracy with the baseline without the alignment loss.

6. Conclusion

In this paper, we study the robustness of transformer models by investigating the sensitivity to the input patches. For most self-attention modules, the features and the corresponding attentions over them are very vulnerable, which, however, contributes to the lack of overall robustness. To alleviate this, we propose a new training method by explicitly reducing sensitivity to patch corruptions (RSPC). Specifically, we develop a patch corruption model to identify the particularly vulnerable patches to be corrupted and stabilize intermediate attention layers using feature alignment. In practice, RSPC greatly improves the stability of self-attention as well as the overall robustness.

References

[1] Alaaeldin Ali, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, et al. Xcit: Cross-covariance image transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021. 3

[2] Yutong Bai, Jieru Mei, Alan L Yuille, and Cihang Xie. Are transformers more robust than cnns? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021. 3

[3] Philipp Benz, Soomin Ham, Chaoning Zhang, Adil Karjauv, and In So Kweon. Adversarial robustness comparison of vision transformer and mlp-mixer to cnns. In *Proc. of the British Machine Vision Conference (BMVC)*, 2021. 3

[4] Philipp Benz, Chaoning Zhang, Soomin Ham, Adil Karjauv, and I Kweon. Robustness comparison of vision transformer and mlp-mixer to cnns. In *Proceedings of the CVPR 2021 Workshop on Adversarial Machine Learning in Real-World Computer Vision Systems and Online Challenges (AML-CV)*, pages 21–24, 2021. 3

[5] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. In *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, pages 10231–10241, 2021. 3

[6] Collin Burns and Jacob Steinhardt. Limitations of post-hoc feature alignment for robustness. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[7] Dan A Calian, Florian Stimberg, Olivia Wiles, Sylvestre-Alvise Rebuffi, Andras Gyorgy, Timothy Mann, and Sven Gowal. Defending against image corruptions through adversarial augmentations. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2022. 5

[8] Chaoqi Chen, Weiping Xie, Wenbing Huang, Yu Rong, Xinghao Ding, Yue Huang, Tingyang Xu, and Junzhou Huang. Progressive feature alignment for unsupervised domain adaptation. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 627–636, 2019. 3

[9] James Diffenderfer, Brian Bartoldson, Shreya Chaganti, Jize Zhang, and Bhavya Kailkhura. A winning hand: Compressing deep networks can improve out-of-distribution robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021. 5

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2021. 1, 3

[11] Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *Proc. of the International Conference on Machine Learning (ICML)*, pages 2286–2296. PMLR, 2021. 5, 6

[12] N Benjamin Erichson, Soon Hoe Lim, Francisco Utrera, Winnie Xu, Ziang Cao, and Michael W Mahoney. Noisymix: Boosting robustness by combining data augmentations, stability training, and noise injections. *arXiv.org*, 2202.01263, 2022. 5

[13] Yonggan Fu, Shunyao Zhang, Shang Wu, Cheng Wan, and Yingyan Lin. Patch-fool: Are vision transformers always robust against adversarial perturbations? In *Proc. of the International Conference on Learning Representations (ICLR)*, 2022. 1, 2, 6, 7

[14] Jindong Gu, Volker Tresp, and Yao Qin. Evaluating model robustness to patch perturbations. In *ICML 2022 Shift Happens Workshop*. 3

[15] Jindong Gu, Volker Tresp, and Yao Qin. Are vision transformers robust to patch perturbations? In *Proc. of the European Conference on Computer Vision (ECCV)*, pages 404–421. Springer, 2022. 1

[16] Jindong Gu, Volker Tresp, and Yao Qin. Are vision transformers robust to patch perturbations? In *Proc. of the European Conference on Computer Vision (ECCV)*. Springer, 2022. 4

[17] Yong Guo, David Stutz, and Bernt Schiele. Improving robustness by enhancing weak subnets. In *Proc. of the European Conference on Computer Vision (ECCV)*. Springer, 2022. 5, 6

[18] Yong Guo, David Stutz, and Bernt Schiele. Robustifying token attention for vision transformers. *arXiv preprint arXiv:2303.11126*, 2023. 3

[19] Yong Guo, Jingdong Wang, Qi Chen, Jiezhang Cao, Zeshuai Deng, Yanwu Xu, Jian Chen, and Mingkui Tan. Towards lightweight super-resolution with dual regression learning. *arXiv preprint arXiv:2207.07929*, 2022. 3

[20] Xing Han, Tongzheng Ren, Tan Minh Nguyen, Khai Nguyen, Joydeep Ghosh, and Nhat Ho. Robustify transformers with robust kernel density estimation. *arXiv.org*, 2210.05794, 2022. 3

[21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 6

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5, 6

[23] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Lixuan Zhu, Samyak Parajuli, Mike Guo, Dawn Xiaodong Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv.org*, abs/2006.16241, 2020. 5, 6

[24] Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2019. 1, 2, 4, 5, 6

[25] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, pages 11936–11945, 2021. 3

[26] Charles Herrmann, Kyle Sargent, Lu Jiang, Ramin Zabih, Huiwen Chang, Ce Liu, Dilip Krishnan, and Deqing Sun. Pyramid adversarial training improves vit performance. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13419–13429, 2022. 1

[27] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 2016. 4

[28] Klim Kireev, Maksym Andriushchenko, and Nicolas Flammarion. On the effectiveness of adversarial training against common corruptions. In *Proc. of the Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 1012–1021. PMLR, 2022. 5

[29] Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. *arXiv.org*, 2202.07800, 2022. 3

[30] 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. In *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021. 5, 6, 8

[31] Giulio Lovisotto, Nicole Finnie, Mauricio Munoz, Chaithanya Kumar Mummadi, and Jan Hendrik Metzen. Give me your attention: Dot-product attention considered harmful for adversarial patch robustness. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15234–15243, 2022. 3

[32] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2018. 3, 6

[33] Kaleel Mahmood, Rigel Mahmood, and Marten van Dijk. On the robustness of vision transformers to adversarial examples. In *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 3

[34] Xiaofeng Mao, Gege Qi, Yuefeng Chen, Xiaodan Li, Ranjie Duan, Shaokai Ye, Yuan He, and Hui Xue. Towards robust vision transformer. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 3, 5, 6

[35] Apostolos Modas, Rahul Rade, Guillermo Ortiz-Jiménez, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Prime: A few primitives can boost robustness to common corruptions. *arXiv.org*, 2112.13547, 2021. 5

[36] Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Fahad Shahbaz Khan, and Fatih Porikli. On improving adversarial transferability of vision transformers. *arXiv.org*, 2106.04169, 2021. 3

[37] Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. In *Proc. of the Conference on Artificial Intelligence (AAAI)*, volume 36, pages 2071–2081, 2022. 3

[38] Yao Qin, Chiyuan Zhang, Ting Chen, Balaji Lakshminarayanan, Alex Beutel, and Xuezhi Wang. Understanding and improving robustness of vision transformers through patch-based negative augmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3

[39] Leslie Rice, Eric Wong, and J. Zico Kolter. Overfitting in adversarially robust deep learning. In *Proc. of the International Conference on Machine Learning (ICML)*, 2020. 1

[40] E. Rusak, Lukas Schott, R. S. Zimmermann, Julian Bitterwolf, O. Bringmann, M. Bethge, and W. Brendel. A simple way to make neural networks robust against diverse image corruptions. In *Proc. of the European Conference on Computer Vision (ECCV)*, 2020. 6

[41] Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. On the adversarial robustness of vision transformers. *arXiv.org*, 2103.15670, 2021. 3

[42] Yucheng Shi and Yahong Han. Decision-based black-box attack against vision transformers via patch-wise adversarial removal. *arXiv.org*, 2112.03492, 2021. 1

[43] Zhouxing Shi, Huan Zhang, Kai-Wei Chang, Minlie Huang, and Cho-Jui Hsieh. Robustness verification for transformers. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2020. 3

[44] Chuanbiao Song, Kun He, Jiadong Lin, Liwei Wang, and John E Hopcroft. Robust local features for improving the generalization of adversarial training. *Proc. of the International Conference on Learning Representations (ICLR)*, 2019. 3

[45] Shiyu Tang, Ruihao Gong, Yan Wang, Aishan Liu, Jiakai Wang, Xinyun Chen, Fengwei Yu, Xianglong Liu, Dawn Song, Alan L. Yuille, Philip H. S. Torr, and Dacheng Tao. Robustart: Benchmarking robustness on architecture design and training techniques. *arXiv.org*, abs/2109.05211, 2021. 3

[46] Shiyu Tang, Siyuan Liang, Ruihao Gong, Aishan Liu, Xianglong Liu, and Dacheng Tao. Exploring the relationship between architecture and adversarially robust generalization. *arXiv.org*, 2209.14105, 2022. 3

[47] Rui Tian, Zuxuan Wu, Qi Dai, Han Hu, and Yugang Jiang. Deeper insights into vits robustness towards common corruptions. *arXiv.org*, 2204.12143, 2022. 3

[48] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *Proc. of the International Conference on Machine Learning (ICML)*, 2021. 5, 6, 8

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

[50] Tao Wang, Ruixin Zhang, Xingyu Chen, Kai Zhao, Xiaolin Huang, Yuge Huang, Shaoxin Li, Jilin Li, and Feiyue Huang. Adaptive feature alignment for adversarial training. *arXiv.org*, 2105.15157, 2021. 3

[51] 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. In *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, pages 568–578, 2021. 6

[52] Florian Wenzel, Andrea Dittadi, Peter Vincent Gehler, Carl-Johann Simon-Gabriel, Max Horn, Dominik Zietlow, David Kernert, Chris Russell, Thomas Brox, Bernt Schiele, et al. Assaying out-of-distribution generalization in transfer learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1

[53] Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. *arXiv.org*, abs/2001.03994, 2020. 1

[54] Boxi Wu, Jindong Gu, Zhifeng Li, Deng Cai, Xiaofei He, and Wei Liu. Towards efficient adversarial training on vision transformers. In *Proc. of the European Conference on Computer Vision (ECCV)*, pages 307–325. Springer, 2022. 3

[55] Hanshu Yan, Jingfeng Zhang, Gang Niu, Jiashi Feng, Vincent Tan, and Masashi Sugiyama. Cifs: Improving adversarial robustness of cnns via channel-wise importance-based feature selection. In *Proc. of the International Conference on Machine Learning (ICML)*, pages 11693–11703. PMLR, 2021. 3

[56] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, pages 558–567, 2021. 6

[57] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *Proc. of the International Conference on Machine Learning (ICML)*, pages 7472–7482. PMLR, 2019. 2, 3

[58] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *Proc. of the International Conference on Machine Learning (ICML)*, 2019. 7

[59] Xiaoqin Zhang, Jinxin Wang, Tao Wang, Runhua Jiang, Jiawei Xu, and Li Zhao. Robust feature learning for adversarial defense via hierarchical feature alignment. *Information Sciences*, 560:256–270, 2021. 3

[60] Zhengli Zhao, Dheeru Dua, and Sameer Singh. Generating natural adversarial examples. *Proc. of the International Conference on Learning Representations (ICLR)*, 2018. 2, 5

[61] Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Animashree Anandkumar, Jiashi Feng, and Jose M Alvarez. Understanding the robustness in vision transformers. In *Proc. of the International Conference on Machine Learning (ICML)*, pages 27378–27394. PMLR, 2022. 1, 3, 5, 6