The Role of ViT Design and Training in Robustness To Common Corruptions

Rui Tian[1,2], Zuxuan Wu[1,2]†, Qi Dai[3], Micah Goldblum[4], Han Hu[3], Yu-Gang Jiang[1,2]

***Abstract*—Vision transformer (ViT) variants have made rapid advances on a variety of computer vision tasks. However, their performance on corrupted inputs, which are inevitable in realistic use cases due to variations in lighting and weather, has not been explored comprehensively. In this paper, we probe the robustness gap among ViT variants and ask how these modern architectural developments affect performance under common types of corruption. Through extensive and rigorous benchmarking, we demonstrate that simple architectural designs such as overlapping patch embedding and convolutional feed-forward networks can promote the robustness of ViTs. Moreover, since the de facto training of ViTs relies heavily on data augmentation, exactly which augmentation strategies make ViTs more robust is worth investigating. We survey the efficacy of previous methods and verify that adversarial noise training is powerful. In addition, we introduce a novel conditional method for generating dynamic augmentation parameters conditioned on input images, which offers state-of-the-art robustness to common corruptions.**

***Index Terms*—Vision Transformer, Common Corruptions, Robustness.**

I. Introduction

ROBUSTNESS to common corruptions has recently attracted attention from the computer vision and machine learning communities [14], [20], [21], [26], [64]. In practice, deployment conditions rarely mirror training data perfectly. For example, practitioners might perform inference under new lighting or weather conditions, and noise levels increase when vision models are deployed on systems with smaller sensors or under suboptimal temperatures. With the emergence of benchmarking datasets such as ImageNet-C [20] and ImageNet-3DCC [30], the brittleness of deep neural networks when facing real-world corruptions has been revealed.

Vision transformers (ViTs) [15] have achieved cutting-edge performance on diverse vision tasks, but it is imperative that we also examine the effects of these architectural improvements in the presence of corruptions that are common in real-world scenarios [1], [39], [41]. In general, studies have revealed the superior robustness of ViTs over convolutional neural networks (CNNs) [1], [37], [41], [43], [50] and showed that the characteristic self-attention mechanisms of ViTs may boost their resistance to corruption [4], [43]. However, these works focused exclusively on vanilla ViT while overlooking the behavior of versatile ViT variants, about which we obtain intriguing findings in this study.

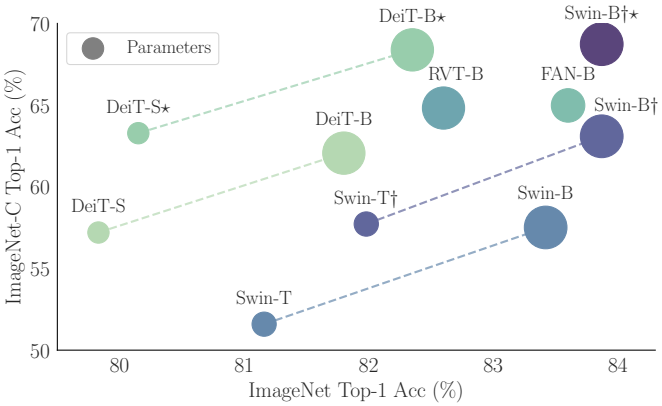


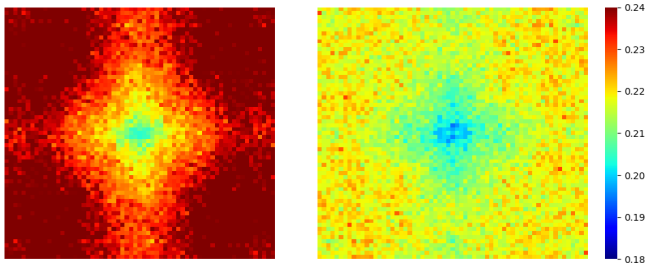
Fig. 1. Trade-off between Top-1 accuracy on ImageNet and ImageNet-C. ⋆ indicates models applied with conditional adversarial noise training. † refers to models with design modifications (*i.e.*, overlapping patch embeddings and convolutional feed-forward networks).

Specifically, we observe that varying ViT designs, which make only small improvements to in-domain test accuracy over vanilla ViTs, can nonetheless make a massive difference in maintaining robustness against corruptions. In particular, PVTv1 [27] and Swin transformer [35] demonstrate promising improvements in in-domain performance on ImageNet [48] but lag far behind on out-of-domain samples with corruptions (*i.e.*, ImageNet-C and ImageNet-3DCC). It is necessary to discover which architectural designs for ViT variants can offer resilience to realist corruptions. We undertake the first such exploration of robustness gaps among ViTs by closely examining the growing body of ViT backbones. Consequently, we reveal that design strategies, including overlapping patch embedding (OPE) and convolutional feed-forward networks (FFNs), are conducive to improving ViT robustness.

In recent work, several effective data augmentation methods have been proposed to address robustness threats to CNNs [22], [26], [40]. In addition, heavy data augmentations, such as Mixup [72], CutMix [70], and RandAugment [10], *etc* have been ingrained in de facto ViT training routines. Therefore, it is essential to explore how such augmentation strategies impact the robustness of ViTs to common corruptions. Specifically, we evaluate the effectiveness of both basic and sophisticated augmentations that have achieved the best performance in previous works, *e.g.*, AugMix [22], PixMix [25], adversarial noise training (ANT) [47], *etc*.

In this paper, we investigate backbone designs and augmentation methods that strengthen the robustness of ViTs

[1]Shanghai Key Lab of Intelligent Information Processing, School of Computer Science, Fudan University, Shanghai, China
[2]Shanghai Collaborative Innovation Center of Intelligent Visual Computing, Shanghai, China
[3]Microsoft Research Asia, Beijing, China
[4]Center for Data Science, New York University, New York, NY, USA
† indicates corresponding author



Swin-T Swin-T + ConvFFN. + OPE.

Fig. 2. Comparison between Fourier heatmaps of different architectural designs. We visualize error rates of models on images corrupted with noise in different frequency spectrums. The central regions of heatmaps indicate the error rate in the low-frequency range.

towards common corruptions. Specifically, we evaluate and analyze the robustness gap among popular ViT variants. We conduct experiments on DeiT [55] and Swin transformer [35], demonstrating that overlapping patch embedding, which captures more local continuity information, can effectively yield performance gains on ImageNet-C and ImageNet-3DCC. Additionally, incorporating depthwise convolutions into FFNs also increases the robustness. Considering Swin-B, combining these two architectural designs decreases the mean correction error (mCE) on ImageNet-C by 12.68 and increases the top-1 accuracy on ImageNet-3DCC by 4.51%. We demonstrate the comparison between Fourier heatmaps [68] of different designs in Fig. 2. In particular, vanilla Swin-T is susceptible to disturbances in the high-frequency domain. However, through minor alterations in architecture, the resilience to corruption in the high-frequency component exhibits a remarkable increase.

In addition, we explore augmentation for ViTs, ranging from de facto training augmentations to those that have achieved exceptional performance in previous studies. Among them, adversarial noise training (ANT) [47] effectively boosts the DeiT-S top-1 accuracy on ImageNet-C from top-1 accuracy on ImageNet-C from 57.20% to 62.74%. Moreover, we introduce a novel conditional augmentation strategy, which build conditional convolutions upon ANT. Consequently, noise parameters can be generated dynamically based on each training sample. Conditional ANT helps DeiT-B achieve a state-of-the-art mCE of 40.68, enables Swin-B to achieve an accuracy of 68.39% on ImageNet-3DCC, and achieves promising performance on other robustness benchmarks.

In summary, our contributions are as follows: 1) We investigate the robustness gap between ViT variants as well as benchmark the influence of different ViT designs and training strategies over robustness towards common corruptions. 2) We provide valuable insights that overlapping OPE and convolutional FFNs strengthen the robustness of a wide range of ViTs. 3) We argue that a one-size-fits-all augmentation strategy is not optimal and propose a novel strategy to automatic augmentation. We incorporate dynamic networks into ANT to encourage a novel input-dependent augmentation technique.

II. Related Work

Common corruptions and perturbations. Real-world visual data suffer from corruption due to bad weather, blurry eyesight, digital distortion, camera noise, *etc*. However, in most cases, advanced models perform training on datasets consisting of clean and pristine samples. To meet practical needs, the ImageNet-C [20] and ImageNet-3DCC [30] datasets, which benchmark safety-critical real-world corruptions, were used to determine the vulnerability of deep models to common corruptions. Thereafter, extensive studies have been proposed to alleviate the performance drop amid common corruptions [5], [22], [26], [33], [52]. DeepAugment [26] distorts images by perturbing image-to-image networks. MoEx [33] increases the stability of models by swapping the mean and standard deviation of image latent features. PixMix [25] uses fractals to create images with structural complexity. RobustMix [42] introduces a new mixing technique by interpolating the low-frequency components of training samples. PaCo [12] makes novel attempts to enhance robustness by rebalancing supervised contrastive learning. Most prior work has focused on improving the robustness of CNNs, but limited efforts have been made toward improving ViTs.

The robustness of ViTs. Recently, multiple attempts have been made to explore improved vision transformers, either by better utilizing contextual information [18], [34], [65], scaling the model size [13], [71] and input resolution [66], or investigating the impact of different training methods on model behavior [57]. Moreover, several studies reveal the robust advantages of ViTs [1], [41], [43], suggesting that self-attention architectures contribute to their superior generalizability over CNNs [1] aand demonstrating that ViTs exhibit less textural bias [41]. Pinto *et al.* [44] investigate the differences in robustness between state-of-the-art CNNs and ViTs. Efforts have also been made to further enhance the robustness of ViTs. Specific ViT architectures have been introduced to improve robustness by redesigning transformer blocks [39], [75] or appending discretized tokens [38]. Another stream of work focuses on extending CNN-based methods to ViTs. For example, Herrmann *et al.* [27] borrows the idea of AdvProp [63] by employing a refined adversarial attack on ViTs in a pyramidal way. Guo *et al.* [17] leverage patch-based adversarial augmentation to achieve improved robustness. Compared to CNNs, these approaches achieve promising performance in relation to common corruptions, yet they overlook the performance gap between different ViT-based variants. Thus, in this paper, we shed light on the impacts of architectural design and training augmentation on the robustness of ViTs.

Learnable data augmentation. Vanilla data augmentation strategies are controlled by hyperparameters, which are often randomly adjusted during training. In contrast, learnable data augmentations [11], [29], [54], [73] aim to derive the best hyperparameters for each training sample for improved training. More specifically, AutoAugment [11] learns an augmentation policy using a reinforcement learning algorithm, while AugMix [22] augments images with stochastic and diverse augmentation methods controlled by the Jensen-Shannon di-

TABLE I
RESULTS OF REPRESENTATIVE VIT BACKBONES ON IMAGENET-C AND IMAGENET-3DCC. THE LISTED MODELS ARE SORTED BY PARAMETERS IN ASCENDING ORDER. COMPARED WITH THEIR COUNTERPARTS, MODELS THAT HAVE INFERIOR PERFORMANCE ON ROBUSTNESS BENCHMARKS ARE HIGHLIGHTED IN BOLD.

Model	IN Acc ↑	IN-C mCE↓	IN-3DC Acc↑
PiT-T [24]	72.84	69.11	47.98
DeiT-T [55]	72.14	71.13	47.44
PVTv1-T [60]	75.00	**79.56**	**46.28**
PVTv2-B1 [61]	78.70	62.65	53.27
DeiT-S [55]	79.83	54.60	57.60
PiT-S [24]	80.98	52.47	58.27
PVTv1-S [60]	79.79	**66.89**	**53.94**
PVTv2-B2 [61]	82.02	52.56	59.06
Swin-T [35]	81.16	**61.96**	**55.90**
PVTv1-M [60]	81.31	**62.39**	**56.49**
Swin-S [35]	83.17	54.92	59.56
PVTv1-L [60]	81.72	**59.86**	**57.80**
PiT-B [24]	82.39	48.16	61.21
PVTv2-B5 [61]	83.77	45.90	62.96
DeiT-B [55]	81.80	48.52	61.45
Swin-B [35]	83.42	**54.45**	**59.93**

vergence. TeachAugment [54] introduces a teacher network to optimize the search space of adversarial augmentations. Another line of work explores adversarial data augmentation. Accordingly, the augmentation parameters are crafted online via min-max optimization. For instance, AugMax [59] achieves a significant performance boost by learning a worst-case combination of random augmentation. Adversarial batch normalization (AdvBN) [53] withstands corruptions by generating the most difficult perturbations on the mean and standard deviation of features. HAT [2] proposes enhancing robustness by perturbing the high-frequency components of inputs adversarially. In contrast to these approaches, we design a sample-specific policy by means of dynamic networks, which further strengthens the efficacy of learnable data augmentation.

Dynamic networks. In contrast to standard neural networks whose architectures and parameters are independent of input samples, dynamic networks adaptively activate parameters or use different parts on a per-sample basis to increase accuracy [6], [67], efficiency [45], [51] or adaptability [76]. One category of dynamic networks focuses on generating parameters based on input samples. In particular, researchers have harnessed convolutions with dynamic parameters [6], [67], [74] to increase their representation capacity with high computational efficiency and to feed selected informative features into models [51]. While the community has extensively designed data augmentation methods, most existing algorithms are implemented identically across different samples. Through the use of dynamic networks, we can generate input-dependent dynamic augmentation parameters.

III. OVERVIEW

Vanilla ViTs have achieved remarkably high performance due to their high capacity to model global relationships among tokenized patches. However, the improved accuracy comes with significantly increased training costs and obstacles to transferring to dense downstream tasks. To reduce the training cost and improve downstream performance, researchers have turned to various backbone designs and training philosophies. Although these approaches have shown increased efficacy and efficiency, their impact on the robustness of vision transformers remains unexplored. This gap in the literature motivates us to investigate underlying architectural designs and augmentation strategies for improving the robustness of ViTs to common corruptions. To this end, we begin by introducing the metrics of the robustness datasets, and then we elaborate on different architectural and augmentation strategies that can influence the robustness of ViTs.

A. Robustness Datasets

We train and evaluate ViTs on ImageNet and test their robustness performance on the ImageNet-C benchmark dataset [20], which provides comprehensive common corruptions in the 2D setting. We also report results on ImageNet-3DCC [30], which includes challenging 3D common corruptions. In particular, we adopt the mean corruption error (mCE) [20] as the metric on ImageNet-C,

$$\text{mCE} = \frac{1}{5 \cdot N} \sum_{c=1}^{N} [(\sum_{s=1}^{5} E_{s,c}) / (\sum_{s=1}^{5} E_{s,c}^{\text{Alex}})], \tag{1}$$

where $N = 15$ denotes the total number of corruption types and S represents the level of corruption severity. E^{Alex} refers to the error rate on AlexNet [32]. In addition, we follow Zhou *et al.* [75] and employ the retention rate, defined as Acc(IN-C)/Acc(IN) to measure the relative resilience of ViTs to common corruptions. Unlike the ImageNet test split, samples in ImageNet-C are shaped in 224^2 and hence are free of resizing preprocessing. Similarly, most corruption categories in ImageNet-3DCC require no resizing, except for corruptions related to video-stream distortions (*i.e.*, bit error, crf compress and abr compress), and images are resized to 224^2 before being fed into networks during inference. In the implementations, we employ `Resize` in the Pytorch implementation. For evaluation on ImageNet-A [23], ImageNet-Rendition [26] and ImageNet-Sketch [58], we follow the common strategy of resizing images to 256^2 followed by cropping the center region size to 224^2.

IV. ARCHITECTURAL DESIGNS

Vision Transformers capture global features by virtue of convolution-free self-attention, which is also believed to contribute to better robustness [4], [43]. While a variety of ViT-based backbones have been recently introduced to improve recognition accuracy or efficiency, whether they are more robust remains unknown. Therefore, we evaluate the prevailing backbones on ImageNet-C and ImageNet-3DCC. As shown in Tab. I, the gap between different ViT variants should not be ignored. Notably, compared with DeiT [55] and PiT [24],

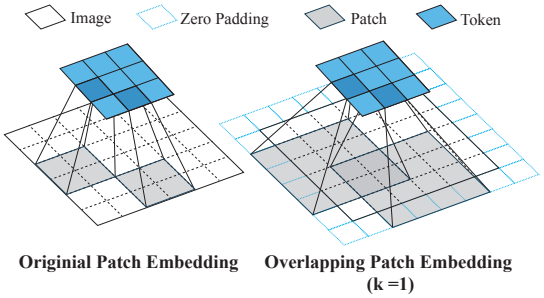


Fig. 3. Comparison between the original patch embedding and the overlapping patch embedding. Given the overlapping width k, the overlapping patch embedding operates within k pixels around the original patch for each token.

TABLE II
PERFORMANCE OF VITS WITH DIFFERENT PATCH EMBEDDING DESIGNS. OPE STANDS FOR OVERLAPPING PATCH EMBEDDING AND k REFERS TO THE OVERLAPPING STEP.

Model	OPE k	FLOPs (G)	IN Acc↑	IN-C mCE↓	Ret. Rate↑
DeiT-S	1	4.62	**80.45**	**52.40**	73.25
	2	4.64	80.40	53.10	72.60
	4	4.68	79.77	52.94	**73.35**
Swin-T	1	4.53	**81.61**	57.74	67.09
	2	4.55	81.59	57.54	67.36
	4	4.62	81.52	**56.11**	**68.83**
	6	4.73	81.32	56.88	68.21
Swin-B	1	15.49	**83.42**	51.60	71.47
	2	15.53	83.11	51.68	71.74
	4	15.62	83.25	50.97	72.30
	6	15.76	83.12	**50.43**	**72.92**

PVTv1 [27] and Swin Transformer [35] exhibit acute weakness on corrupted data.

Notably, the performance differences between PVTv2s and PVTv1s on ImageNet-C and ImageNet-3DCC are considerable. According to Tab. I, PVTv2-B5 outperforms PVTv1-L on ImageNet-C by 13.96 in mCE and by 5.16% in ImageNet-3DCC Top-1 accuracy. Inspired by this encouraging improvement, we start by investigating the differences between these two backbones. On top of PVTv1s, PVTv2s incorporates two minor modifications into the backbone architecture, including (1) *overlapping patch embedding* and (2) *convolutional feed-forward networks*. We hypothesize that these modifications in designs contribute to the leap in robustness performance. To verify their effectiveness, we apply similar modifications to DeiT-S and Swin-T respectively.

A. Overlapping Patch Embedding

In vanilla ViTs, the encoding of an image $x \in \mathbb{R}^{C\times H\times W}$ starts with translating it into a sequence of T patches of size $p \times p$. These flattened patches $x_p \in \mathbb{R}^{T\times(p^2\times C)}$ are subsequently mapped to tokens $t \in \mathbb{R}^{T\times D}$ by a fully-connected layer. In DeiT [55], patch embedding is instead instantiated with a convolution whose kernel size and stride are p. Since the stride and kernel size match, the translated kernels are non-overlapping so that patch embedding takes in disjoint sets of pixels. In contrast to non-overlapping patch embedding, CNNs apply convolutions on overlapping spatial areas. Inspired by such a design, overlapping patch embedding (OPE) is introduced by expanding the patch window to the surrounding pixels of the original patch. To obtain an unchanged number of tokens and encode an extended area with neighboring information simultaneously, we perform a convolution with a kernel of $((p+2k)\times(p+2k))$, leave the stride size fixed to p and apply zero padding, as depicted in Fig. 3.

In practice, with p set to 16, we conduct experiments with k varying between 1, 2, 4 and 6 on DeiT and Swin transformers. The results from Tab. II indicate that overlapping patch embedding enhances robustness. Compared to a vanilla DeiT-S with a 54.6 mCE (*c.f.* Tab. I), overlapping patch embedding improves the mCE by **2.2** when $k = 1$. Similarly, we observe decreases of **5.96** and **4.02** in the mCE on Swin-T with $k = 4$ and Swin-B with $k = 6$, respectively.

TABLE III
PERFORMANCE OF VITS WITH DIFFERENT FEED-FORWARD DESIGNS. WHEN CHOOSING W/O CONVOLUTIONS, MODELS ARE IDENTICAL TO THE ORIGINAL SETTINGS. WE SET N TO 32 AND 4 FOR CONDITIONAL DWCONV FOR DEIT-S AND SWIN-T, RESPECTIVELY.

Model	FFN Design	FLOPs (G)	IN Acc↑	IN-C mCE↓	Ret. Rate↑
DeiT-S	w/o Conv	4.61	79.83	54.60	71.65
	DWConv	4.64	79.93	**52.41**	**73.70**
Swin-T	w/o Conv	4.51	81.16	61.96	63.57
	DWConv	4.56	81.83	**55.28**	**69.33**
Swin-B	w/o Conv	15.46	83.42	54.45	68.94
	DWConv	15.58	83.82	**48.36**	**74.60**

B. Convolutional Feed-forward Network

Recently, researchers also couple ViT blocks with convolutions either to improve the efficiency of ViTs [62], [69] or to increase the encoding of spatial information [9], [28]. Depthwise convolution (DWConv) [7] is a particular form of convolution that applies a single filter to each input channel (*i.e.* input depth). We follow PVTv2 by adding a depthwise convolution to each feed-forward block in ViTs.

We denote $\mathcal{F}_i$ as the output of i-th attention layer. Naturally, for ViTs equipped with class token (*e.g.* DeiT), $\mathcal{F}_i$ can be rewritten as $[\mathcal{F}_i^{cls}, \mathcal{F}_i^{img}]$ where $\mathcal{F}_i^{cls}$ represents the class token output of i-th layer while $\mathcal{F}_i^{img}$ refers to the corresponding image token. Otherwise, for ViTs without class token (*e.g.* Swin Transformer), $\mathcal{F}_i$ is equivalent to $\mathcal{F}_i^{img}$. Considering the spatial dimension requirements for convolutions, we proceed to apply convolutions to $\mathcal{F}_i^{img}$ and leave $\mathcal{F}_i^{cls}$ unchanged. The convolution is injected between the two linear layers within the FFN and before the activation layer.

Tab. III shows that convolutional FFNs lead to a decrease of **2.19** in the mCE for DeiT-S. The mCE for Swin-T and Swin-B decreases by **6.68** and **6.09**, respectively. Swin transformers also show an increase in clean accuracy, namely, of 0.67% for Swin-T and 0.4% for Swin-B. Based on these results, we

TABLE IV
PERFORMANCE OF MODELS WITH BOTH OVERLAPPING PATCH EMBEDDING (OPE.) AND CONVOLUTIONAL FFNS ON DIFFERENT ROBUSTNESS BENCHMARKS. THE DROP OR GROWTH COMPARED WITH THE ORIGINAL DESIGN IS DISPLAYED IN BRACKETS.

Model	Param (M)	FLOPs (G)	Conv FFN	OPE. k	IN Acc↑	IN-C mCE↓	Retention Rate↑	IN-3DC Acc↑	IN-R Acc↑	IN-SK Acc↑
DeiT-S	22.4	4.7	DWC.	2	79.99	52.29 (↓2.31)	73.78 (↑2.13)	60.35 (↑1.53)	45.35 (↑3.45)	32.62 (↑3.52)
Swin-T	28.5	4.7	DWC.	4	81.98	54.10 (↓7.86)	70.41 (↑6.84)	60.00 (↑4.69)	45.92 (↑4.66)	32.48 (↑3.43)
Swin-B	88.3	15.7	DWC.	4	83.87	47.25 (↓12.68)	75.22 (↑6.55)	64.24 (↑4.54)	50.33 (↑3.78)	36.93 (↑4.51)

TABLE V
THE ABLATION STUDY OF DESIGNS IN PVTV2-B1. THE ROW INDICATING THE ORIGINAL DESIGN IS HIGHLIGHTED IN GRAY.

OPE	conv FFN	IN Acc↑	IN-C mCE↓	Ret. Rate↑
✓		77.51	67.14	61.05
	✓	78.03	63.46	64.40
✓	✓	**78.70**	**62.65**	**64.56**

TABLE VI
THE ABLATION STUDY OF DESIGNS IN CVT-13. THE ROW INDICATING THE ORIGINAL DESIGN IS HIGHLIGHTED IN GRAY.

OPE	conv FFN	IN Acc↑	IN-C mCE↓	Ret. Rate↑
		80.00	64.99	61.46
✓		**81.40**	57.24	67.64
✓	✓	81.20	**54.79**	**70.25**

believe that adding a convolution to FFNs will result in solid gains in robustness.

C. Combined Designs

The merits of overlapping patch embedding and injecting convolutions into FFNs are their simplicity and high efficacy. Furthermore, we experiment with DeiT and Swin by incorporating these two designs simultaneously. As displayed in Tab. IV, the combined design induces further improvements both on ImageNet-C, ImageNet-3DCC and on other robustness benchmarks, *i.e.*, ImageNet-Rendition [26] and ImageNet-Sketch [58]. Ultimately, the mCE of DeiT-S decreases on ImageNet-C by **2.41**, while that of Swin-T and Swin-B decreases by **7.86** and **12.68**, respectively. In addition, compared to the overall parameters of the original ViTs (*i.e.* 22.1M for DeiT-S, 28.3M for Swin-T and 86.6M for Swin-B), implementing overlapping patch embeddings and FFNs with depthwise convolutions results in a relatively small increase in parameters (*i.e.*, approximately 0.5-2%). Moreover, adding both OPE and convolutional FFNs to DeiT-S results in an increase of only 1.4% in FLOPs while introducing 1.7% more FLOPs on Swin-B. Therefore, the modifications have minimal impacts on the computational costs while greatly boosting the robustness to common corruptions.

Exploration on more ViTs. Regarding the extensive results on DeiT and Swin transformer, we posit that overlapping patch embedding and convolutional feed-forward networks are widely effective for ViT variants. To test this prediction, we conduct additional experiments on other ViTs. In particular, we verify the positive impacts of these two designs on the robustness of PVTv2. As displayed in Tab. V, removing either overlapping patch embedding or convolutional FNNs significantly degrades both clean-domain and robust-domain performance, which further verifies our hypothesis above.

In addition, we can observe a similar trend for CvT [62], a ViT variant that incorporates convolutional projection and adopts overlapping patch embedding in the original setting. As Tab. VI demonstrates, applying nonoverlapping patch embedding to CVT-13 increases the mCE of ImageNet-C by 7.75, while injecting convolutions into FFNs decreases the mCE by 2.45. Hence, we believe that these two minor modifications in design can generate far-reaching improvements in the robustness of ViTs.

V. AUGMENTATION STRATEGIES

Training of ViTs involves strong augmentations (*i.e.*, Mixup [72], CutMix [70], RandAugment [10] and repeated augmentation [3]), whereas how these augmentations impact the robustness performance against common corruptions is uncertain. Furthermore, numerous augmentation techniques have been proposed and proven to be effective in improving robustness. To boost the robustness of ViTs to common corruptions, it is natural to build them upon previous successful strategies. To this end, we experiment with previous advanced augmentation techniques and emphasize adversarial augmentation, *i.e.*, adversarial noise training (ANT) [47]. Furthermore, we propose a conditional augmentation strategy in addition to adversarial augmentation, which effectively strengthens ViTs against common corruption.

A. Basic Augmentations

Bai *et al.* [1] reveal that applying basic augmentations used for training ViTs (i.e., augmentations used in the de facto ViT training schedules) facilitates improvements in the robustness of CNNs. Nevertheless, how basic augmentations contribute to the robustness of ViTs remains unclear. Thus, taking DeiT-S and Swin-T as examples, we analyze the performance changes in ViTs by removing each of these augmentations individually from training. Since de facto training performs Mixup or CutMix alternatively based on a switching probability, when either of them is excluded from training, we apply the other method for the training samples.

As displayed in Tab. VII, repeated augmentation and random erasing have limited influence on the performance on

TABLE VII
PERFORMANCE OF VITS ON IMAGENET AND IMAGENET-C WITH ONE TYPE OF BASIC AUGMENTATION REMOVED FROM TRAINING.

Model	w/o Aug.	IN Acc↑	IN-C mCE↓	Ret. Rate↑
DeiT-S	MixUp	80.20	60.01	65.62 (↓6.03)
	CutMix	78.56	56.84	70.70 (↓0.95)
	RandAug	79.56	57.91	68.46 (↓3.19)
	Rep. Aug	80.41	54.72	71.04 (↓0.62)
	Erasing	80.03	54.55	71.47 (↓0.18)
Swin-T	MixUp	81.42	66.56	58.55 (↓5.02)
	CutMix	80.68	60.24	65.65 (↑2.09)
	RandAug	80.86	65.08	60.78 (↓2.78)
	Erasing	81.25	61.40	64.09 (↑0.52)

TABLE VIII
PERFORMANCE OF DEIT-S WITH DIFFERENT AUGMENTATION METHODS ON IMAGENET AND IMAGENET-C. METHODS WITH THE COLUMN OF RANDAUG TICKLED ARE IMPLEMENTED AND ASSOCIATED WITH THE ORIGINAL RANDAUG TRAINING STRATEGY.

Method	Mixup	Rand Aug	IN Acc↑	IN-C mCE↓	Ret. Rate↑
ThreeAug			80.63	59.55	65.41
AugMix			80.05	56.85	69.17
PixMix	✓	✓	79.76	53.81	72.67
		✓	79.37	54.03	72.63
PRIME	✓	✓	79.26	47.76	79.13
		✓	79.76	48.51	77.72

ImageNet-C, *i.e.*, slight changes are incurred in the retention rate. Mixup, CutMix and RandAugment are robust to different degrees. CutMix is essential for in-domain test accuracy, but minimally improves the retention rate. Notably, Swin-T with CNN-style hierarchical feature maps achieves the best performance without implementing CutMix, which is similar to the performance of ResNet50 [1].

B. Advanced Augmentations

Recently, versatile augmentation-based strategies have been proposed for improving the robustness of deep neural networks, especially classical CNNs. However, whether the advantages of these advanced training techniques persist for ViTs remains unexplored. To shed light on the influence of such augmentations on ViTs' robustness to common corruptions, we tested DeiT-S with five augmentation methods, *i.e.*, Three-Augment [56], AugMix [22], PixMix [25] and PRIME [40]. Three-Augment draws inspiration from the succinct preprocessing of self-supervised learning and hence consists of only grayscale, solarization and Gaussian blur. AugMix incorporates simple augmentation from AutoAugment [11] with consistency loss. PixMix enriches the input distribution by mixing training samples with complex fractals, and PRIME leverages the combination of max-entropy transformations in the spectral, spatial and color domains.

Tab. VIII suggests that ThreeAug improves in-domain performance on the ImageNet validation set but impedes robustness, which is reflected by a plummet of 4.95 in mCE on ImageNet-C. AugMix also fails to enhance resistance to corruption. In contrast, PRIME and PixMix have positive impacts on robustness. In particular, PRIME outperforms ImageNet but substantially outperforms ImageNet-C. Mixup also strongly influences robustness, facilitating a decrease of 1.25 mCE on top of PRIME. In the case of PixMix, the positive influence introduced by Mixup is weakened, as adding Mixup to methods integrated with mixing (*e.g.* PixMix and AugMix) may lead to manifold intrusion [16].

TABLE IX
WE CONDUCT EXTENSIVE EXPERIMENTS OF ADVERSARIAL NOISE TRAINING ON VITS. MODELS OF DEIT-S AND PIT-S ARE INITIALIZED WITH WEIGHTS FROM OFFICIAL OPEN SOURCE. † INDICATES ADDING OVERLAPPING PATCH EMBEDDING AND CONVOLUTIONAL FEED-FORWARD NETWORK TO DEIT-S AND ALL EXPERIMENTS ARE CONDUCTED WITH THE NORM OF NOISE SET TO 80.

Method	IN Acc↑	IN-C mCE↓	Ret. Rate↑	IN-3DC Acc↑
DeiT-S [55]	79.83	54.60	71.65	58.82
+ fine-tune 300 ep.	81.20	52.31	72.66	60.15
+ ANT Gaussian	80.27	48.09	77.87	62.63
+ ANT Speckle	80.35	47.82	78.08	62.55
+ Cond ANT Gaussian	80.18	47.28	78.77	62.79
+ Cond ANT Speckle	80.15	**47.18**	**78.95**	**62.84**
PiT-S [24]	80.98	52.47	72.71	60.64
+ ANT Gaussian	80.77	47.60	77.83	63.63
+ Cond ANT Gaussian	80.80	47.28	78.13	63.54
+ Cond ANT Speckle	80.74	**46.63**	**78.82**	**64.02**
DeiT-S †	79.99	52.29	73.78	60.35
+ Cond ANT Speckle	79.91	**45.33**	**81.00**	**63.91**

TABLE X
THE ABLATION STUDY OF APPLYING DIFFERENT NOISE AUGMENTATION TYPES ON DEIT AND RESNET [19]. W/O NOISE REFERS TO A SUBSET OF IMAGENET-C, WHICH CONTAINS NO NOISE-RELATED CORRUPTIONS.

Model	Method	IN-C Acc↑	w/o Noise Acc↑
DeiT-S	Speckle	62.25 (↑5.05)	60.51 (↑3.46)
	Speckle ANT	62.74 (↑5.54)	60.81 (↑3.76)
	Speckle Cond ANT	63.28 (↑6.08)	61.02 (↑3.97)
ResNet50	Gaussian ANT	51.09 (↑11.9)	47.66 (↑5.36)
	Gaussian Cond ANT	51.30 (↑12.1)	47.63 (↑5.33)
DeiT-B	Speckle ANT	66.83 (↑4.77)	65.33 (↑4.01)
	Speckle Cond ANT	68.38 (↑6.32)	66.46 (↑5.14)

C. Adversarial Noise Augmentation

As indicated by the experiments above, advanced augmentations clearly demonstrate a trade-off between in-domain accuracy and out-of-domain accuracy. In pursuit of a better trade-off as well as deeper insights into a more delicate training approach, we consider approaches focused on applying adversarial learning to achieve improved robustness, *e.g.*, AugMax [59], HAT [2] and *etc*. In particular, Rusak *et al.* [47]

TABLE XI
THE ABLATION STUDY OF REMOVING BASIC AUGMENTATIONS FROM GAUSSIAN ANT FINE-TUNING ON DEIT-S. IN EACH RUN, WE REMOVE ONE SPECIFIC TYPE OF AUGMENTATION.

Augmentation	IN Acc↑	IN-C mCE↓	Ret. Rate↑
Full	80.27	**48.09**	77.87
w/o Mixup	80.65	51.07	74.38
w/o CutMix	79.94	48.62	77.72
w/o RandAug	79.68	52.89	73.73

TABLE XII
THE ABLATION STUDY OF DIFFERENT CONVOLUTION TYPES IN NOISE GENERATOR MODELS ON DEIT-S. $N \times N$ OF COLUMN GEN INDICATES THE KERNEL SIZE OF CONVOLUTIONS.

Noise	Gen	IN Acc↑	IN-C mCE↓	Ret. Rate↑
Gaussian	1×1	80.27	48.09	77.87
Gaussian	3×3	**80.52**	48.18	77.52
Speckle	1×1	80.35	**47.82**	**78.08**
Speckle	3×3	79.80	49.25	77.20

TABLE XIII
THE ABLATION STUDY OF AUGMENTATION SEQUENCES IN GAUSSIAN ANT FINE-TUNING USING PRE-TRAINED DEIT-S, WHERE MIX IS THE ABBREVIATION FOR MIXUP AND CUTMIX.

Strategy	IN Acc↑	IN-C mCE↓	Ret. Rate↑
Early Mix	80.76	50.56	74.85
Full Mix	80.19	48.66	77.25
Separate Mix	80.27	**48.09**	77.87

TABLE XIV
THE ABLATION STUDY OF ATTACK TYPE IN ADVERSARIAL NOISE TRAINING OF GAUSSIAN 1×1. THE ROW INDICATING THE DEFAULT DESIGN IS HIGHLIGHTED IN GRAY.

Attack Type	IN Acc↑	IN-C mCE↓	Ret. Rate↑
non-targeted	80.34	48.47	77.41
targeted	80.27	**48.09**	**77.87**

introduce a novel adversarial framework and make great strides toward improving the robustness of CNNs to common corruptions. Instead of performing traditional adversarial training or learning augmentation parameters adversarially, they attempt to tune random noise with a lightweight neural network and then add the enhanced noise to the images. Afterward, the models are trained against enhanced noise adversarially. The entire framework is denoted as adversarial noise training (ANT). Specifically, Gaussian (Eq. (2a)) noise and speckle noise (Eq. (2b)) are generated to perturb an image x,

$$\mathbf{\Sigma}_1(x) = x + \mathcal{C}_p(\sigma_\delta, \epsilon), \tag{2a}$$
$$\mathbf{\Sigma}_2(x) = x + \mathcal{C}_p(\sigma_\delta \cdot x, \epsilon), \tag{2b}$$

where $\sigma_\delta \in \mathbb{R}^{C\times H\times W}$ is distributed according to $\mathcal{N}(0, \delta^2)$. The clipping function $\mathcal{C}_p$ confines noise to have an L_p norm at most ϵ. Motivated by the efficacy and simplicity of ANT for CNNs, we explore a similar strategy for ViTs (denoted f_θ) with the following optimization objective:

$$\min_\theta \max_\tau \mathop{\mathbb{E}}_{(x,y)\sim D} \mathop{\mathbb{E}}_{\sigma_\delta \sim \mathcal{N}(\mathbf{0},\delta^2)} [\mathcal{L}(f_\theta(x + \mathcal{C}_p(\mathcal{P}_\tau(\sigma_\delta \cdot x), \epsilon), y)],$$
$$\min_\theta \max_\tau \mathop{\mathbb{E}}_{(x,y)\sim D} \mathop{\mathbb{E}}_{\sigma_\delta \sim \mathcal{N}(\mathbf{0},\delta^2)} [\mathcal{L}(f_\theta(x + \mathcal{C}_p(\mathcal{P}_\tau(\sigma_{\delta\cdot x}), \epsilon), y)],$$

where $\mathcal{P}_\tau$ denotes the noise generators which typically consist of 4 1×1 convolution layers with residual connections. Initially, in the inner loop, we focus on improving $\mathcal{P}_\tau$. The optimization goal is to generate noise that is harmful enough to make ViTs produce incorrect predictions. The optimization serves as the attack phase in a similar spirit to adversarial training. Following this step, in the outer loop or defense phase, we optimize ViTs to withstand the effects of the enhanced noise created by the generators. This two-phase cycle is critical for strengthening the ViTs against more challenging noisy inputs.

Adversarial noise training is effective. For simplicity, we attack ViTs with noise produced by ANT generators and defend against noisy data once per iteration. During the attack process, we assign all the data noise generated by ANT, while we use 50% noisy data together with 50% clean data during the defense phase. We train noise generators with the Adam optimizer and a learning rate of $8e^{-5}$ and retain all augmentation strategies from DeiT training [55] since they improved the performance (*c.f.* Tab. XI). As shown in Tab. IX, ANT decreases the mCE of DeiT-S and a PiT-S by 6.51 and 5.87 on ImageNet-C, respectively. It also outperforms the baseline of further fine-tuning DeiT-S on ImageNet for 300 epochs by 4.22 in mCE.

We build adversarial generators on both Gaussian and speckle noise. The speckle noise amplified by generators is applied to 1×1 pixel regions and hence preserves most of the structural information in the images. As Tab. IX shows, ANT with speckle noise yields the best performance on robustness benchmarks. In addition, we experiment with the baseline of randomly sampled plain speckle noise from distributions of five different standard deviations. The results in Tab. X demonstrate that ANT outperforms the plain noise baseline not only on the full ImageNet-C dataset but also on corruption types excluding noise. Therefore, adversarial noise training enables ViTs to generalize to a wider range of distributions.

Basic augmentations are necessary. Regarding DeiT, since we train ANT on top of the vanilla training recipe, where basic augmentation approaches, *i.e.*, Mixup [72], CutMix [70] and RandAugment [10] are employed to stabilize training and boost clean accuracy. It is necessary to explore the interplay between additive adversarial noise and previously adopted augmentations. The results in Tab. XI indicate that ViTs depend strongly on basic perturbations. The absence of any type of basic augmentation leads to a decrease in performance on both ImageNet and ImageNet-C.

Furthermore, we demonstrate that the sequence of applying Mixup, CutMix, and noise augmentation is also crucial. We

TABLE XV
PERFORMANCE OF VITS ON DIFFERENT ROBUSTNESS BENCHMARKS. † INDICATES SWIN-B EQUIPPED WITH OPE AND CONVOLUTIONAL FFN. TOP-2 STATE-OF-THE-ART RESULTS ARE UNDERSCORED. DEIT-B WITH CONDITIONAL ANT ADOPTS SPECKLE NOISE WITH THE NORM SET TO 100 WHILE SWIN-B EMPLOYS NOISE WITH A NORM OF 80. WE ADOPT OFFICIAL CHECKPOINTS OF RVT-B AND FAN-B IN EVALUATION WHILE REPORTING RESULTS OF DISCRETE VIT AND PYRAMID VIT BY CITING FROM THEIR PAPERS.

Model	IN Acc↑	IN-C mCE↓	Ret. Rate↑	IN-3DC Acc↑	IN-R Acc↑	IN-A Acc↑	IN-SK Acc↑
Discrete ViT [38]	79.48	46.22	-	-	44.77	27.19	34.59
Pyramid ViT [27]	81.71	44.99	-	-	47.66	22.99	36.77
RVT-B [39]	82.60	44.80	78.47	65.18	50.48	**40.89**	34.79
FAN-B [75]	83.60	44.68	77.73	65.95	50.93	39.96	37.73
DeiT-B	81.80	48.52	75.87	62.68	44.66	28.15	31.96
Swin-B †	**83.87**	47.25	75.22	64.24	50.33	**40.45**	36.93
DeiT-B + cond ANT	82.35	**40.68**	**83.04**	66.44	48.63	31.76	36.00
Swin-B † + cond ANT	83.72	40.82	81.43	**68.39**	**54.18**	39.44	**40.34**

investigate 3 different strategies: (1) early mixing: we first implement data normalization, followed by Mixup or CutMix, and finally, noise is added; (2) full mixing: noise is first applied, followed by normalization and Mixup or CutMix along the whole batch; and (3) separate mixing: we first apply noise followed by normalization, while Mixup or CutMix is conducted separately with clean and noisy data. Separate mixing offers the best results, as shown in Tab. XIII.

Following the original setting in ANT [47], we also experiment with noise generators with convolutions of 3×3 kernel size. As displayed in Tab. XII, while the kernel size 3×3 is similar to the 1×1 kernel for Gaussian noise, speckle noise decreases with the 3×3 convolution kernel. Overall, speckle noise with 1×1 kernel generators outperforms its counterparts marginally. In addition, we ablate with the nontargeted attack by randomly choosing one of 1000 labels as the optimization target via de facto adversarial training. Tab. XIV demonstrates that a targeted attack results in better robustness.

D. Conditional Adversarial Augmentation

While ANT is effective, it produces noise regardless of the visual content of the input samples with "one-size-fits-all" parameters. Recently, there has been growing interest in dynamic CNNs that generate sample-specific parameters conditioned on inputs [6], [67], [74]. This development motivates us to explore dynamic techniques in ANT for improved performance on ViTs. We introduce input-dependent adversarial augmentations by means of dynamic networks, which instantiate conditional computation by dynamically generating augmentation parameters individually for each input instance. Notably, improved robustness can be achieved without introducing extra parameters for ViTs and with a moderate increase in training overheads.

Conditional ANT. We aim to generate sample-specific noise conditioned on the input images. Thus, we apply dynamic augmentation weights on top of ANT. To this end, we use CondConv [67] which replaces vanilla convolution layers with conditional convolutions in noise generators. Specifically, an image x is mapped to a space of N expert weights by a routing function $\mathcal{R} : \mathbb{R}^{C\times H\times W} \rightarrow \mathbb{R}^N$. Accordingly, we instantiate the routing function with

$$\mathcal{R}(x) = \texttt{Sigmoid}(\texttt{FC}(\texttt{GlobalAvgPool}(x))). \quad (3)$$

We can expand and express the routing weight as $\mathcal{R}(x) = \{w_1, w_2, \cdots, w_D\}$. Consequently, a customized kernel $\mathcal{W}$ is generated by combining image-specific weights with N learnable kernels.

$$\mathcal{W}(X) = w_1 \cdot K_1 + \cdots + w_D \cdot K_D. \quad (4)$$

Therefore, the noise appended to each image is dynamically tuned. In addition, we incorporate two tactics to alleviate training overhead. On the one hand, we follow Shafahi *et al.* [49] to perform free adversarial training, *i.e.*, the noise generator is optimized based on the current minibatch and then generates noise for the successive minibatch of data. Hence, the ViT and the generator run backward simultaneously. Compared with adversarial training based on the vanilla K-step PGD attack, where the backbone model requires 2K times the feed-forward per minibatch, we readily conduct training with only one-pass feed-forward. On the other hand, the original ANT involves 20% of the samples augmented by randomly drawn previous noise generators. We employ a history generator with a momentum update instead of loading previous checkpoints to reduce I/O loads while remaining aligned with the experience replay setting. With the joint strategies, the training cost is close to the de facto cost without adversarial augmentation, while consistent performance gains can be obtained.

TABLE XVI
RESULTS OF TRAINING DEIT AND SWIN TRANSFORMER FROM SCRATCH WITH CONDITIONAL SPECKLE ANT. ALL EXPERIMENTS ARE CONDUCTED WITH THE NORM OF NOISE SET TO 80.

Model	Cond ANT	IN Acc↑	IN-C mCE↓	Ret. Rate↑	IN-3DC Acc↑
DeiT-S		79.83	54.60	71.65	58.82
	✓	79.94	47.15	79.15	63.28
Swin-B†		83.87	47.25	75.22	64.24
	✓	83.64	**40.02**	**82.17**	**68.69**

Specifically, we fine-tune ViTs via conditional ANT with N set to 16. Our experiments suggest that this dynamic technique achieves state-of-the-art performance. As demonstrated in Tab. IX, training DeiT-S with conditional ANT surpasses the original ANT by 0.65% (Gaussian noise) and 0.54% (speckle noise) in top-1 accuracy on ImageNet-C. In addition, we managed to achieve markedly improved robustness performance via from-scratch training while maintaining a similar clean performance with de facto training. Using PiT-S, conditional ANT with speckle noise yields a boost of 4.76% in accuracy on ImageNet-C. In addition, by applying conditional ANT to DeiT-S, which is equipped with an improved architecture from our work, *i.e.*, overlapping patch embedding and convolutional feed-forward networks, we achieve an mCE of 45.33 on ImageNet-C, which is 4.07 better than that of RVT-S [39].

We compare the gains achieved by conditional augmentation between ResNet and DeiT in Tab. X with noise of the same magnitude. When applied to ResNet50, conditional ANT marginally surpasses ANT by 0.2% in top-1 accuracy on ImageNet-C. In contrast, we observe that conditional ANT helps DeiT-B improve significantly by 1.5%. We posit that ViTs accommodate conditional augmentation well due to their strong innate generalization capability.

State-of-the-art performance. We experiment with the conditional augmentation strategy using DeiT-B and Swin-B. Although a trade-off between performance on clean and corrupted data is demonstrated on DeiT-S, applying conditional ANT on DeiT-B nonetheless improves ImageNet test accuracy by 0.55%. We owe this result to the better capacity of large models and hypothesize that heavy augmentation may catalyze the better performance of DeiT-B. Compared with other state-of-the-art ViTs in Tab. XV, our models demonstrate outstanding performance on a range of robustness benchmarks and are also highly competitive on clean ImageNet recognition.

In addition, we validate the feasibility of training ViTs with adversarial noise from scratch on both DeiT-S and Swin-B. The results in Tab. XVI suggest that conditional ANT not only manages to boost robustness against common corruptions but also maintains moderate clean accuracy, outperforming the counterparts in Tab. VIII in the trade-off between in-domain and out-of-domain performance. Swin-B even achieves a minimum mCE of **40.02** and a peak accuracy of **68.89%** on ImageNet-3DCC.

Furthermore, we claim that conditional ANT can be adapted for a wide range of differentiable augmentations containing random variables. Specifically, the conditional generator can be plugged in to adjust any random variable. Once differentiability is satisfied, adversarial training can be performed by optimizing the predefined conditional generator.

VI. Additional Analysis

According to the quantitative analysis of the designs and training strategies above, ViTs demonstrate promising robustness with overlapping patch embeddings, convolutional feed-forward networks and conditional ANT. To explore the rationale behind these designs and strategies, we investigate their contributions to robustness from the perspective of the Fourier domain and provide more detailed results on specific corruption categories.

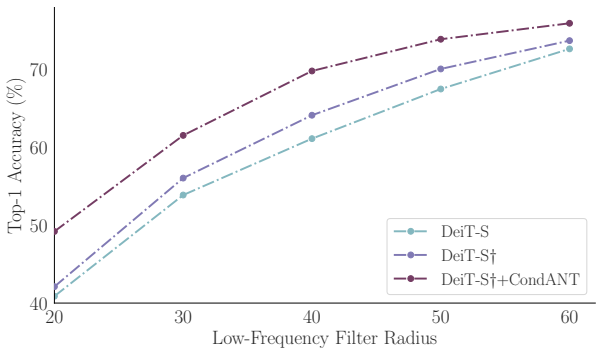


Fig. 4. Performance of different models on low-pass-filtered input. † indicates employing overlapping patch embedding and convolutional feed-forward networks in DeiT.

Fourier-domain robustness. Following Yin *et al.* [68], we explore the robustness of different architectural designs by perturbing models with noise in the Fourier domain and visualizing the error rates corresponding to the Fourier spectrum with heatmaps. While dark red indicates relatively high error rates, dark blue denotes relatively low error rates and hence implies better robustness in the Fourier domain. Fig. 5 shows that Swin-T with OPE improves resilience to high-frequency perturbations (*c.f.* Fig. 2) to an extent, while convolutional FFNs greatly improve the robustness of high-frequency components and even manage to elevate the immunity of the low-frequency part against disturbance.

In addition, we apply low-pass filters to further analyze the performance of the models on low-frequency components. Specifically, we employ discrete Fourier transformation (DCT) to perform Fourier domain filters on input images $x \in \mathbb{R}^{H\times W}$ with a corresponding hyper-parameter r, which refers to the radius of the filter. The low-frequency mask $ML^r \in \{0,1\}$ can be formulated as follows:

$$ML^r_{i,j} = \begin{cases} 1 & \text{if } \sqrt{((i-H/2)^2+(j-W/2)^2)} \le r \\ 0 & \text{otherwise} \end{cases}.$$

Given the DCT function denoted as $\mathcal{F}$ and the inverse discrete fourier transformation as $\mathcal{F}^{-1}$, we can formulate the low-pass filter $\mathcal{PF}^{\text{low}}_r$ as

$$\mathcal{PF}^{\text{low}}_r = \mathcal{F}^{-1}(ML^r \odot \mathcal{F}(x)). \tag{5}$$

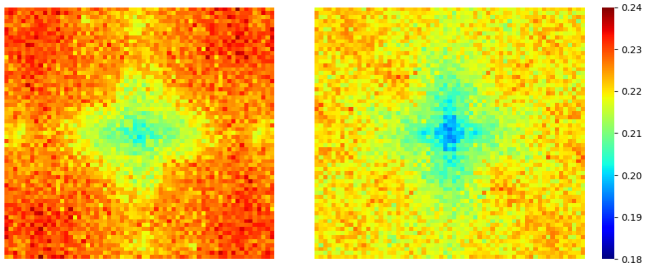
Therefore, we use $\mathcal{PF}^{\text{low}}_r$ with r ranging from 20 to 60 on a subset of ImageNet validation dataset which contains 5000 samples sampled uniformly, and plot the curve of the radius and the corresponding top-1 accuracy. As Fig. 4 shows, the combined architecture improves the capability of modeling low-frequency components, hence resulting in enhanced robustness to high-frequency perturbations. In addition, applying conditional ANT leads to a remarkable increase in performance on the low-frequency components.

Performance for specific corruption types. Specifically, as displayed in Tab. XVII, DeiT-S with overlapping embedding shows superior performance against corruption in terms of noise, brightness, JPEG compression, contrast and pixelation.

TABLE XVII
ERROR RATES OF DEIT-S WITH DIFFERENT DESIGN SETTINGS TOWARDS 15 SPECIFIC CORRUPTION TYPES OF IMAGENET-C.

OPE	FFN	Weather				Blur			
k	Design	Bright.	Fog	Frost	Snow	Defoc.	Glass	Motion	Zoom
0	w/o. Conv	44.93	46.01	46.18	49.86	61.58	71.90	57.86	71.89
1	w/o. Conv	**43.31**	43.27	46.16	48.72	59.65	71.25	56.86	**69.20**
2	w/o. Conv	43.63	46.01	**45.34**	49.33	60.71	71.37	57.88	71.50
4	w/o. Conv	44.28	**43.18**	45.47	50.67	60.92	71.43	56.39	71.41
0	DWC	43.59	45.17	46.34	**47.84**	**59.39**	**70.06**	**54.25**	69.37
2	DWC	43.39	46.42	45.71	47.45	59.64	70.79	55.08	70.48

		Digital				Noise			mCE
		Contra.	Elast.	JPEG	Pixel.	Gauss.	Impulse	Shot	
0	w/o. Conv	42.31	66.58	60.42	59.12	46.29	46.39	47.72	54.60
1	w/o. Conv	40.40	67.30	**56.06**	**52.14**	43.60	43.60	44.53	52.40
2	w/o. Conv	40.25	66.93	57.03	52.87	44.19	44.03	45.36	53.10
4	w/o. Conv	40.91	67.19	56.84	52.75	43.98	43.74	45.00	52.94
0	DWC	40.97	**65.66**	58.53	50.48	44.60	44.52	45.35	52.41
2	DWC	**39.64**	65.80	57.05	51.32	**43.67**	**43.48**	44.44	52.29



Swin-T + OPE. Swin-T + ConvFFN.

Fig. 5. Comparison between Fourier heatmaps of Swin-T with overlapping patch embedding and convolutional FFN.

Additionally, convolutional designs increase the accuracy on corruptions from similar categories. The architecture with combined designs yields the best results on corruption categories of noise.

VII. IMPLEMENTATION DETAILS

For the experiments exploring the architectural designs, we trained ViTs from scratch for 300 epochs with a learning rate of $1e^{-3}$ and a batch size of 1024. We used the Adam-W [36] optimizer with a cosine decay learning rate scheduler. The augmentation and regularization strategies used were identical to those used for the corresponding ViT variants unless otherwise specified. On the other hand, we conducted training with different augmentation strategies in various settings. For basic augmentations, we trained DeiT and Swin transformer for 300 epochs following their original training schedule, although one type of augmentation was removed.

Advanced augmentations. We trained ViTs from scratch and applied advanced augmentations. For AugMix, we optimized it with the JSD loss and removed repeated augmentations. Specifically, we split the training data into three parts and applied augmentation with a magnitude of 5, chain width of 3, and chain depth of 2 via implementation in library [46]. Notably, compared with vanilla ANT, our implementation includes additive augmentation strategies (i.e., color, contrast, brightness, and sharpness). For Three-Augment [56], we follow Touvron *et al.* to remove repeated augmentation and MixUp, replaced RandAugment with a simple combination of grayscale, Gaussian blur and solarization, and employed a color jitter of 0.3 and horizontal flip.

We implement PixMix [25] in its original setting by randomly sampling fractals from datasets released by Hendrycks *et al.* and mixing them with training samples via either addition or multiplication. On the other hand, we employed PRIME [40] with hyperparameters following the original settings on ImageNet.

Adversarial noise augmentation. We optimized ViTs and the noise generator separately during training. Specifically, the learning rate for ViTs is $1e^{-3}$ on the basis of a batch size of 1024, and that for noise generators was fixed to $8e^{-5}$ for all experiments. We employed an Adam [31] optimizer for noise generators, and all augmentation strategies were retained since they improve the final results in Tab. XI. For training with conditional ANT from scratch, we followed AdvProp [63] by casting noise on a duplicated batch of training samples. In this way, we continued to use 50% noisy samples.

VIII. CONCLUSION AND DISCUSSION

Currently, deep models have achieved compelling performance on in-domain test datasets, yet they are still susceptible to corrupted real-world visual data. This study revealed that strategies including overlapping patch embedding and convolutional feed-forward networks could facilitate robustness to common corruptions of ViTs. In addition, we benchmarked ViTs on a wide range of augmentation strategies and delved into the adversarial noise training technique. We observed that ViTs are adequately compatible with this strong adversarial augmentation. Moreover, we proposed conditional adversarial augmentation to enable ViTs to achieve state-of-the-art robust-

ness performance. We hope that our explorations provide the research community with better insights.

Acknowledgement This project was supported by NSFC under Grant No. 62102092.

REFERENCES

[1] Bai, Yutong and Mei, Jieru and Yuille, Alan L and Xie, Cihang, "Are Transformers more robust than CNNs?," in *NeurIPS*, 2021. 1, 2, 5, 6

[2] Bai, Jiawang and Yuan, Li and Xia, Shu-Tao and Yan, Shuicheng and Li, Zhifeng and Liu, Wei, "Improving Vision Transformers by Revisiting High-frequency Components," in *ECCV*, 2022. 3, 6

[3] Berman, Maxim and Jégou, Hervé and Vedaldi, Andrea and Kokkinos, Iasonas and Douze, Matthijs, "Multigrain: a unified image embedding for classes and instances," *arXiv preprint arXiv:1902.05509*, 2019. 5

[4] Bhojanapalli, Srinadh and Chakrabarti, Ayan and Glasner, Daniel and Li, Daliang and Unterthiner, Thomas and Veit, Andreas, "Understanding robustness of transformers for image classification," in *ICCV*, 2021. 1, 3

[5] Chen, Guangyao and Peng, Peixi and Ma, Li and Li, Jia and Du, Lin and Tian, Yonghong, "Amplitude-phase recombination: Rethinking robustness of convolutional neural networks in frequency domain," in *ICCV*, 2021. 2

[6] Chen, Yinpeng and Dai, Xiyang and Liu, Mengchen and Chen, Dongdong and Yuan, Lu and Liu, Zicheng, "Dynamic convolution: Attention over convolution kernels," in *CVPR*, 2020. 3, 8

[7] Chollet, François, "Xception: Deep learning with depthwise separable convolutions," in *CVPR*, 2017. 4

[8] Christoph Kamann and Carsten Rother, "Benchmarking the robustness of semantic segmentation models with respect to common corruptions," *IJCV*, 2020.

[9] Chu, Xiangxiang and Tian, Zhi and Zhang, Bo and Wang, Xinlong and Wei, Xiaolin and Xia, Huaxia and Shen, Chunhua, "Conditional positional encodings for vision transformers," *arXiv preprint arXiv:2102.10882*, 2021. 4

[10] Cubuk, Ekin D. and Zoph, Barret and Shlens, Jonathon and Le, Quoc V., "Randaugment: Practical Automated Data Augmentation with a Reduced Search Space," in *CVPR Workshops*, 2020. 1, 5, 7

[11] Cubuk, Ekin D and Zoph, Barret and Mane, Dandelion and Vasudevan, Vijay and Le, Quoc V, "Autoaugment: Learning augmentation policies from data," in *CVPR*, 2019. 2, 6

[12] Cui, Jiequan and Zhong, Zhisheng and Tian, Zhuotao and Liu, Shu and Yu, Bei and Jia, Jiaya, "Generalized parametric contrastive learning," *TPAMI*, 2023. 2

[13] Dehghani, Mostafa and Djolonga, Josip and Mustafa, Basil and Padlewski, Piotr and Heek, Jonathan and Gilmer, Justin and Steiner, Andreas Peter and Caron, Mathilde and Geirhos, Robert and Alabdulmohsin, Ibrahim and others, "Scaling vision transformers to 22 billion parameters," in *ICLR*, 2023. 2

[14] Dong, Yinpeng and Kang, Caixin and Zhang, Jinlai and Zhu, Zijian and Wang, Yikai and Yang, Xiao and Su, Hang and Wei, Xingxing and Zhu, Jun, "Benchmarking robustness of 3d object detection to common corruptions," in *CVPR*, 2023. 1

[15] Dosovitskiy, Alexey and Beyer, Lucas and Kolesnikov, Alexander and Weissenborn, Dirk and Zhai, Xiaohua and Unterthiner, Thomas and Dehghani, Mostafa and Minderer, Matthias and Heigold, Georg and Gelly, Sylvain and Uszkoreit, Jakob and Houlsby, Neil, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *ICLR*, 2021. 1

[16] Guo, Hongyu and Mao, Yongyi and Zhang, Richong, "Mixup as locally linear out-of-manifold regularization," in *AAAI*, 2019. 6

[17] Guo, Yong and Stutz, David and Schiele, Bernt, "Improving Robustness of Vision Transformers by Reducing Sensitivity To Patch Corruptions," in *CVPR*, 2023. 2

[18] Hatamizadeh, Ali and Yin, Hongxu and Heinrich, Greg and Kautz, Jan and Molchanov, Pavlo, "Global context vision transformers," in *ICLR*, 2023. 2

[19] He, Kaiming and Zhang, Xiangyu and Ren, Shaoqing and Sun, Jian, "Deep residual learning for image recognition," in *CVPR*, 2016. 6

[20] Hendrycks, Dan and Dietterich, Thomas, "Benchmarking Neural Network Robustness to Common Corruptions and Perturbations," in *ICLR*, 2018. 1, 2, 3

[21] Hendrycks, Dan and Mu, Norman and Cubuk, Ekin D and Zoph, Barret and Gilmer, Justin and Lakshminarayanan, Balaji, "Augmix: A simple data processing method to improve robustness and uncertainty," in *ICLR*, 2019. 1

[22] Hendrycks, Dan and Mu, Norman and Cubuk, Ekin D and Zoph, Barret and Gilmer, Justin and Lakshminarayanan, Balaji, "Augmix: A simple data processing method to improve robustness and uncertainty," in *ICLR*, 2020. 1, 2, 6

[23] Hendrycks, Dan and Zhao, Kevin and Basart, Steven and Steinhardt, Jacob and Song, Dawn, "Natural Adversarial Examples," in *CVPR*, 2021. [Online]. Available: https://github.com/hendrycks/natural-adv-examples 3

[24] Heo, Byeongho and Yun, Sangdoo and Han, Dongyoon and Chun, Sanghyuk and Choe, Junsuk and Oh, Seong Joon, "Rethinking spatial dimensions of vision transformers," in *ICCV*, 2021. 3, 6

[25] Hendrycks, Dan and Zou, Andy and Mazeika, Mantas and Tang, Leonard and Song, Dawn and Steinhardt, Jacob, "PixMix: Dreamlike Pictures Comprehensively Improve Safety Measures," in *NeurIPS*, 2021. 1, 2, 6, 10

[26] Hendrycks, Dan and Basart, Steven and Mu, Norman and Kadavath, Saurav and Wang, Frank and Dorundo, Evan and Desai, Rahul and Zhu, Tyler and Parajuli, Samyak and Guo, Mike et al., "The many faces of robustness: A critical analysis of out-of-distribution generalization," in *ICCV*, 2021. 1, 2, 3, 5

[27] Herrmann, Charles and Sargent, Kyle and Jiang, Lu and Zabih, Ramin and Chang, Huiwen and Liu, Ce and Krishnan, Dilip and Sun, Deqing, "Pyramid Adversarial Training Improves ViT Performance," in *CVPR*, 2022. 1, 2, 4, 8

[28] Islam, Md Amirul and Jia, Sen and Bruce, Neil DB, "How much Position Information Do Convolutional Neural Networks Encode?," in *ICLR*, 2019. 4

[29] Jing, Mengmeng and Meng, Lichao and Li, Jingjing and Zhu, Lei and Shen, Heng Tao, "Adversarial mixup ratio confusion for unsupervised domain adaptation," *TMM*, 2022. 2

[30] Kar, Oğuzhan Fatih and Yeo, Teresa and Atanov, Andrei and Zamir, Amir, "3D Common Corruptions and Data Augmentation," in *CVPR*, 2022. [Online]. Available: https://github.com/EPFL-VILAB/3DCommonCorruptions 1, 2, 3

[31] Kingma, Diederik P and Ba, Jimmy, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. 10

[32] Krizhevsky, Alex and Sutskever, Ilya and Hinton, Geoffrey E, "Imagenet classification with deep convolutional neural networks," *CACM*, 2017. 3

[33] Li, Boyi and Wu, Felix and Lim, Ser-Nam and Belongie, Serge and Weinberger, Kilian Q., "On Feature Normalization and Data Augmentation," in *CVPR*, 2021. 2

[34] Li, Yehao and Yao, Ting and Pan, Yingwei and Mei, Tao, "Contextual transformer networks for visual recognition," *TPAMI*, 2022. 2

[35] Liu, Ze and Lin, Yutong and Cao, Yue and Hu, Han and Wei, Yixuan and Zhang, Zheng and Lin, Stephen and Guo, Baining, "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," in *ICCV*, 2021. 1, 2, 3, 4

[36] Loshchilov, Ilya and Hutter, Frank, "Decoupled Weight Decay Regularization," in *ICLR*, 2018. 10

[37] Mahmood, Kaleel and Mahmood, Rigel and van Dijk, Marten, "On the Robustness of Vision Transformers to Adversarial Examples," in *ICCV*, 2021. 1

[38] Mao, Chengzhi and Jiang, Lu and Dehghani, Mostafa and Vondrick, Carl and Sukthankar, Rahul and Essa, Irfan, "Discrete Representations Strengthen Vision Transformer Robustness," in *ICLR*, 2022. 2, 8

[39] Mao, Xiaofeng and Qi, Gege and Chen, Yuefeng and Li, Xiaodan and Duan, Ranjie and Ye, Shaokai and He, Yuan and Xue, Hui, "Towards robust vision transformer," in *CVPR*, 2022. 1, 2, 8, 9

[40] Modas, Apostolos and Rade, Rahul and Ortiz-Jiménez, Guillermo and Moosavi-Dezfooli, Seyed-Mohsen and Frossard, Pascal, "PRIME: A Few Primitives Can Boost Robustness to Common Corruptions," in *ECCV*, 2022. 1, 6, 10

[41] Naseer, Muhammad Muzammal and Ranasinghe, Kanchana and Khan, Salman H and Hayat, Munawar and Shahbaz Khan, Fahad and Yang, Ming-Hsuan, "Intriguing properties of vision transformers," in *NeurIPS*, 2021. 1, 2

[42] Ngnawe, Jonas and NJIFON, Marianne ABEMGNIGNI and Heek, Jonathan and Dauphin, Yann, "Robustmix: Improving Robustness by Regularizing the Frequency Bias of Deep Nets," in *NeurIPS Workshop*, 2022. 2

[43] Paul, Sayak and Chen, Pin-Yu, "Vision transformers are robust learners," in *AAAI*, 2022. 1, 2, 3

[44] Pinto, Francesco and Torr, Philip HS and K. Dokania, Puneet, "An impartial take to the cnn vs transformer robustness contest," in *ECCV*, 2022. 2

[45] Rao, Yongming and Zhao, Wenliang and Liu, Benlin and Lu, Jiwen and Zhou, Jie and Hsieh, Cho-Jui, "DynamicViT: Efficient Vision Transformers with Dynamic Token Sparsification," in *NeurIPS*, 2021. 3

[46] Wightman, Ross, "PyTorch Image Models," *GitHub repository*, GitHub, 2019. [Online]. Available: https://github.com/rwightman/pytorch-image-models 10

[47] Rusak, Evgenia and Schott, Lukas and Zimmermann, Roland S and Bitterwolf, Julian and Bringmann, Oliver and Bethge, Matthias and Brendel, Wieland, "A simple way to make neural networks robust against diverse image corruptions," in *ECCV*, 2020. 1, 2, 5, 6, 8

[48] Russakovsky, Olga and Deng, Jia and Su, Hao and Krause, Jonathan and Satheesh, Sanjeev and Ma, Sean and Huang, Zhiheng and Karpathy, Andrej and Khosla, Aditya and Bernstein, Michael and others, "Imagenet large scale visual recognition challenge," *IJCV*, 2015. 1

[49] Shafahi, Ali and Najibi, Mahyar and Ghiasi, Mohammad Amin and Xu, Zheng and Dickerson, John and Studer, Christoph and Davis, Larry S and Taylor, Gavin and Goldstein, Tom, "Adversarial training for free!," in *NeurIPS*, 2019. 8

[50] Shao, Rulin and Shi, Zhouxing and Yi, Jinfeng and Chen, Pin-Yu and Hsieh, Cho-Jui, "On the Adversarial Robustness of Vision Transformers," *arXiv preprint arXiv:2103.15670*, 2021. 1

[51] Sharma, Vivek and Diba, Ali and Neven, Davy and Brown, Michael S and Van Gool, Luc and Stiefelhagen, Rainer, "Classification-driven dynamic image enhancement," in *CVPR*, 2018. 3

[52] Shu, Manli and Shen, Yu and Lin, Ming C. and Goldstein, Tom, "Adversarial Differentiable Data Augmentation for Autonomous Systems," in *ICRA*, 2021. 2

[53] Shu, Manli and Wu, Zuxuan and Goldblum, Micah and Goldstein, Tom, "Encoding Robustness to Image Style via Adversarial Feature Perturbations," in *NeurIPS*, 2021. 3

[54] Suzuki, Teppei, "TeachAugment: Data Augmentation Optimization Using Teacher Knowledge," in *CVPR*, 2022. 2, 3

[55] Touvron, Hugo and Cord, Matthieu and Douze, Matthijs and Massa, Francisco and Sablayrolles, Alexandre and Jégou, Hervé, "Training data-efficient image transformers & distillation through attention," in *ICML*, 2021. 2, 3, 4, 6, 7

[56] Touvron, Hugo and Cord, Matthieu and Jégou, Hervé, "Deit iii: Revenge of the vit," in *ECCV*, Springer, 2022. 6, 10

[57] Walmer, Matthew and Suri, Saksham and Gupta, Kamal and Shrivastava, Abhinav, "Teaching matters: Investigating the role of supervision in vision transformers," in *CVPR*, 2023. 2

[58] Wang, Haohan and Ge, Songwei and Lipton, Zachary and Xing, Eric P., "Learning Robust Global Representations by Penalizing Local Predictive Power," in *NeurIPS*, 2019. [Online]. Available: https://github.com/HaohanWang/ImageNet-Sketch 3, 5

[59] Wang, Haotao and Xiao, Chaowei and Kossaifi, Jean and Yu, Zhiding and Anandkumar, Anima and Wang, Zhangyang, "Augmax: Adversarial composition of random augmentations for robust training," in *NeurIPS*, 2021. 3, 6

[60] Wang, Wenhai and Xie, Enze and Li, Xiang and Fan, Deng-Ping and Song, Kaitao and Liang, Ding and Lu, Tong and Luo, Ping and Shao, Ling, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *ICCV*, 2021. 3

[61] Wang, Wenhai and Xie, Enze and Li, Xiang and Fan, Deng-Ping and Song, Kaitao and Liang, Ding and Lu, Tong and Luo, Ping and Shao, Ling, "PVTv2: Improved Baselines with Pyramid Vision Transformer," *CVMJ*, 2022. 3

[62] Wu, Haiping and Xiao, Bin and Codella, Noel and Liu, Mengchen and Dai, Xiyang and Yuan, Lu and Zhang, Lei, "Cvt: Introducing convolutions to vision transformers," in *ICCV*, 2021. 4, 5

[63] Xie, Cihang and Tan, Mingxing and Gong, Boqing and Wang, Jiang and Yuille, Alan L and Le, Quoc V, "Adversarial examples improve image recognition," in *CVPR*, 2020. 2, 10

[64] Xing, Weiwei and Yao, Jie and Liu, Zixia and Liu, Weibin and Zhang, Shunli and Wang, Liqiang, "Contrastive JS: A Novel Scheme for Enhancing the Accuracy and Robustness of Deep Models," *TMM*, 2022. 1

[65] Yao, Ting and Li, Yehao and Pan, Yingwei and Wang, Yu and Zhang, Xiao-Ping and Mei, Tao, "Dual vision transformer," *TPAMI*, 2023. 2

[66] Yao, Ting and Li, Yehao and Pan, Yingwei and Mei, Tao, "HIRI-ViT: Scaling Vision Transformer with High Resolution Inputs," *TPAMI*, 2024. 2

[67] Yang, Brandon and Bender, Gabriel and Le, Quoc V and Ngiam, Jiquan, "Condconv: Conditionally parameterized convolutions for efficient inference," in *NeurIPS*, 2019. 3, 8

[68] Yin, Dong and Gontijo Lopes, Raphael and Shlens, Jon and Cubuk, Ekin Dogus and Gilmer, Justin, "A fourier perspective on model robustness in computer vision," in *NeurIPS*, 2019. 2, 9

[69] Yuan, Kun and Guo, Shaopeng and Liu, Ziwei and Zhou, Aojun and Yu, Fengwei and Wu, Wei, "Incorporating convolution designs into visual transformers," in *ICCV*, 2021. 4

[70] Yun, Sangdoo and Han, Dongyoon and Oh, Seong Joon and Chun, Sanghyuk and Choe, Junsuk and Yoo, Youngjoon, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *ICCV*, 2019. 1, 5, 7

[71] Zhai, Xiaohua and Kolesnikov, Alexander and Houlsby, Neil and Beyer, Lucas, "Scaling vision transformers," in *CVPR*, 2022. 2

[72] Zhang, Hongyi and Cisse, Moustapha and Dauphin, Yann N and Lopez-Paz, David, "mixup: Beyond empirical risk minimization," in *ICLR*, 2018. 1, 5, 7

[73] Zhang, Jiansong and Chen, Kejiang and Qin, Chuan and Zhang, Weiming and Yu, Nenghai, "Distribution-preserving-based automatic data augmentation for deep image steganalysis," *TMM*, 2021. 2

[74] Zhang, Yikang and Zhang, Jian and Wang, Qiang and Zhong, Zhao, "Dynet: Dynamic convolution for accelerating convolutional neural networks," *arXiv preprint arXiv:2004.10694*, 2020. 3, 8

[75] Zhou, Daquan and Yu, Zhiding and Xie, Enze and Xiao, Chaowei and Anandkumar, Animashree and Feng, Jiashi and Alvarez, Jose M, "Understanding the robustness in vision transformers," in *ICML*, 2022. 2, 3, 8

[76] Zhu, Mingjian and Han, Kai and Wu, Enhua and Zhang, Qiulin and Nie, Ying and Lan, Zhenzhong and Wang, Yunhe, "Dynamic Resolution Network," in *NeurIPS*, 2021. 3