Pyramid Adversarial Training Improves ViT Performance

Charles Herrmann* Kyle Sargent* Lu Jiang Ramin Zabih
Huiwen Chang Ce Liu† Dilip Krishnan Deqing Sun
Google Research

Abstract

Aggressive data augmentation is a key component of the strong generalization capabilities of Vision Transformer (ViT). One such data augmentation technique is adversarial training (AT); however, many prior works [28, 45] have shown that this often results in poor clean accuracy. In this work, we present pyramid adversarial training (PyramidAT), a simple and effective technique to improve ViT's overall performance. We pair it with a "matched" Dropout and stochastic depth regularization, which adopts the same Dropout and stochastic depth configuration for the clean and adversarial samples. Similar to the improvements on CNNs by AdvProp [61] (not directly applicable to ViT), our pyramid adversarial training breaks the trade-off between in-distribution accuracy and out-of-distribution robustness for ViT and related architectures. It leads to 1.82% *absolute improvement on ImageNet clean accuracy for the ViT-B model when trained only on ImageNet-1K data, while simultaneously boosting performance on* 7 *ImageNet robustness metrics, by absolute numbers ranging from* 1.76% *to* 15.68%. *We set a new state-of-the-art for ImageNet-C (41.42 mCE), ImageNet-R (53.92%), and ImageNet-Sketch (41.04%) without extra data, using only the ViT-B/16 backbone and our pyramid adversarial training. Our code is publicly available at* `pyramidat.github.io`.

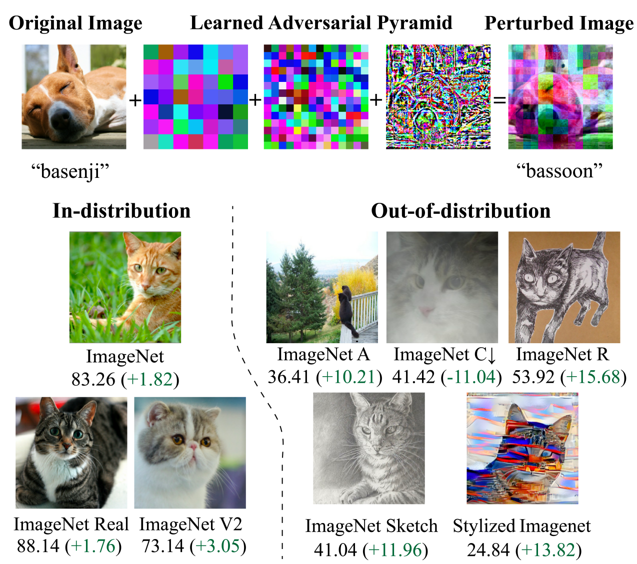

Figure 1. Top: Visualization of our learned multi-scale pyramid perturbations. We show the original image, multiple scales of a perturbation pyramid, and the perturbed image. Bottom: We show thumbnails of in-distribution and out-of-distribution datasets, and the gains from applying our technique on each dataset. (Note that lower is better for ImageNet-C.)

1. Introduction

One fascinating aspect of human intelligence is the ability to generalize from limited experiences to new environments [30]. While deep learning has made remarkable progress in emulating or "surpassing" humans on classification tasks, deep models have difficulty generalizing to out-of-distribution data [31]. Convolutional neural networks (CNNs) may fail to classify images with challenging contexts [22], unusual colors and textures [16, 19, 58] and common or adversarial corruptions [17, 20]. To reliably deploy neural networks on diverse tasks in the real world, we must improve their robustness to out-of-distribution data.

One major line of research focuses on network design. Recently the Vision Transformer (ViT) [14] and its variants [2, 32, 46, 55] have advanced the state of the art on a variety of computer vision tasks. In particular, ViT models are more robust than comparable CNN architectures [36, 38, 49, 49]. With a weak inductive bias and powerful model capacity, ViT relies heavily on strong data augmentation and regularization to achieve better generalization [51, 55]. To further push this envelope, we explore using adversarial training [29, 65] as a powerful regularizer to improve the performance of ViT models.

Prior work [56] suggests that there exists a performance trade-off between in-distribution generalization and robustness to adversarial examples. Similar trade-offs have been

*Equal contribution, ordered alphabetically.
†Currently affiliated with Microsoft Azure AI.

observed between in-distribution and out-of-distribution generalization [45, 65]. These trade-offs have primarily been observed in the context of CNNs [7, 45]. However, recent work has demonstrated the trade-off can be broken. AdvProp [61] achieves this via adversarial training (abbreviated AT) with a "split" variant of Batch Normalization [24] for EfficientNet [53]. In our work, we demonstrate that the trade-off can be broken for the newly introduced vision transformer architecture [14].

We introduce *pyramid adversarial training* (abbreviated as PyramidAT) that trains the model with input images altered at multiple spatial scales, as illustrated in Fig. 1; the pyramid attack is designed to make large edits to the image in a structured, controlled manner (similar to augmenting brightness) and small edits to the image in a flexible manner (similar to pixel adversaries). Using these structured, multi-scale adversarial perturbations leads to significant performance gains compared to both baseline and standard pixel-wise adversarial perturbations. Interestingly, we see these gains for both clean (in-distribution) and robust (out-of-distribution) accuracy. We further enhance the pyramid attack with additional regularization techniques: "matched" Dropout and stochastic depth. Matched Dropout uses the same Dropout configuration for both the regular and adversarial samples in a mini-batch (hence the word matched). Stochastic depth [23, 51] randomly drops layers in the network and provides a further boost when matched and paired with matched Dropout and multi-scale perturbations.

Our ablation studies confirm the importance of matched Dropout when used in conjunction with the pyramid adversarial training. They also reveal a complicated interplay between adversarial training, the attack being used, and network capacity. We additionally show that our approach is applicable to datasets of various scales (ImageNet-1K and ImageNet-21K) and for a variety of network architectures such as ViT [14], Discrete ViT [37], ResNet [18], and MLP-Mixer [54]. Our contributions are summarized below:

- To our knowledge, we appear to be the first to demonstrate that adversarial training improves ViT model performance on both ImageNet [12] and out-of-distribution ImageNet robustness datasets [16, 19, 20, 22, 58].
- We demonstrate the importance of matched Dropout and stochastic depth for the adversarial training of ViT.
- We design pyramid adversarial training to generate multi-scale, structured adversarial perturbations, which achieve significant performance gains over non-adversarial baseline and adversarial training with pixel perturbations.
- We establish a new state of the art for ImageNet-C, ImageNet-R, and ImageNet-Sketch without extra data, using only our pyramid adversarial training and the standard ViT-B/16 backbone. We further improve our results by incorporating extra ImageNet-21K data.
- We perform numerous ablations which highlight several elements critical to the performance gains.

2. Related Work

There exists a large body of work on measuring and improving the robustness of deep learning models, in the context of adversarial examples and generalization to non-adversarial but shifted distributions. We define *out-of-distribution accuracy/robustness* to explicitly refer to performance of a model on non-adversarial distribution shifts, and *adversarial accuracy/robustness* to refer to the special case of robustness on adversarial examples. When the evaluation is performed on a dataset drawn from the same distribution, we call this *clean accuracy*.

Adversarial training and robustness The discovery of adversarial examples [52] has stimulated a large body of literature on adversarial attacks and defenses [1, 6, 29, 35, 40, 43, 44, 60]. Of the many proposed defenses, adversarial training [29, 35] has emerged as a simple, effective, albeit expensive approach to make networks adversarially robust. Although some work [56, 65] has suggested a trade-off between adversarial and out-of-distribution robustness or clean accuracy, other analysis [7, 45] has suggested simultaneous improvement is achievable. In [39, 45], the authors note improved accuracy on both clean and adversarially perturbed data, though only on smaller datasets such as CIFAR-10 [27] and SVHN [42], and only through the use of additional data extending the problem to the semi-supervised setting. Similarly in NLP, adversarial training leads to improvement of clean accuracy for machine translation [8, 9].

Most closely related to our work is the technique of [61], which demonstrates the potential of adversarial training to improve both clean accuracy and out-of-distribution robustness. They focus primarily on CNNs and propose split batch norms to separately capture the statistics of clean and adversarially perturbed samples in a mini-batch. At inference time, the batch norms associated with adversarially perturbed samples are discarded, and all data (presumed clean or out-of-distribution) flows through the batch norms associated with clean samples. Their results are demonstrated on EfficientNet [53] and ResNet [18] architectures. However, their approach is not directly applicable to ViT where batch norms do not exist. In our work, we propose novel approaches, and find that properly constructed adversarial training helps clean accuracy and out-of-distribution robustness for ViT models.

Robustness of ViT ViT models have been found to be more adversarially robust than CNNs [41,49], and more importantly, generalize better than CNNs with similar model capacity on ImageNet out-of-distribution robustness benchmarks [49]. While existing works focus on analyzing the cause of ViT's superior generalizability, this work aims at further improving the strong out-of-distribution robustness of the ViT model. A promising approach to this end is data augmentation; as shown recently [51, 55], ViT benefits from strong data augmentation. However, the data augmentation techniques used in ViT [51, 55] are optimized for clean accuracy on ImageNet, and knowledge about robustness is still limited. Different from prior works, this paper focuses on improving both the clean accuracy and robustness for ViT. We show that our technique can effectively complement strong ViT augmentation as in [51]. We additionally verify that our proposed augmentation can benefit three other architectures: ResNet [18], MLP-Mixer [54], and Discrete ViT [37].

Data augmentation Existing data augmentation techniques, although mainly developed for CNNs, transfer reasonably well to ViT models [10, 21, 59]. Other work has studied larger structured attacks [60]. Our work is different from prior work in that we utilize adversarial training to augment ViT and tailor our design to the ViT architecture. To our knowledge, we appear to be the first to demonstrate that adversarial training substantially improves ViT performance in both clean and out-of-distribution accuracies.

3. Approach

We work in the supervised learning setting where we are given a training dataset $\mathcal{D}$ consisting of clean images, represented as x and their labels y. The loss function considered is a cross-entropy loss $L(\theta, x, y)$, where θ are the parameters of the ViT model, with weight regularization f. The baseline models minimize the following loss:

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\Big[L(\theta, \tilde{x}, y) + f(\theta)\Big], \tag{1}$$

where $\tilde{x}$ refers to a data-augmented version of the clean sample x, and we adopt the standard data augmentations as in [51], such as RandAug [10].

3.1. Adversarial Training

The overall training objective for adversarial training [57] is given as follows:

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\Big[\max_{\delta\in\mathcal{P}} L(\theta, \tilde{x}+\delta, y) + f(\theta)\Big], \tag{2}$$

where δ a per-pixel, per-color-channel additive perturbation, and $\mathcal{P}$ is the perturbation distribution. Note that the adversarial image, x^a, is given by $\tilde{x} + \delta$, and we use these two interchangeably below. The perturbation, δ, is computed by optimizing the objective inside the maximization of Eqn. 2. This objective tries to improve the worst-case performance of the network w.r.t. the perturbation; subsequently, the resulting model has lower clean accuracy.

To remedy this, we can train on both clean and adversarial images [17, 29, 61] using the following objective:

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\Big[L(\theta, \tilde{x}, y) + \lambda \max_{\delta\in\mathcal{P}} L(\theta, \tilde{x}+\delta, y) + f(\theta)\Big], \tag{3}$$

This objective uses adversarial images as a form of regularization or data augmentation, to force the network towards certain representations that perform well on out-of-distribution data. These networks exhibit some degree of robustness but still have good clean accuracy. More recently, [61] proposes a split batch norm that leads to performance gains for CNNs on both clean and robust ImageNet test datasets. Note that they do not concern themselves with adversarial robustness, and neither do we in this paper.

3.2. Pyramid Adversarial Training

Pixel-wise adversarial images are defined [29] as $x^a = x + \delta$ where the perturbation distribution P consists of a clipping function $C_{\mathcal{B}_\epsilon}$ that clips the perturbation at each pixel location to be inside the specified ball ($\mathcal{B}_\epsilon$) for a specified l_p-norm [35], with maximal radius ϵ for the perturbation.

Motivation For pixel-wise adversarial images, increasing the value of ϵ or the number of steps of the inner loop in Eqn. 3 eventually causes a drop in clean accuracy (Fig 2). Conceptually, pixel attacks are very flexible and, if given the ability to make large changes (in L_2 distance), can destroy the object being classified; training with these images may harm the network. In contrast, augmentations, like brightness, can lead to large L_2 distances but will preserve the object because they are structured. Our main motivation is to design an attack which has the best of both worlds: a low-magnitude flexible component and a high-magnitude structured component; this attack can lead to large image differences while still preserving the class identity.

Approach We propose pyramid adversarial training (PyramidAT) which generates adversarial examples by perturbing the input image at multiple scales. This attack is more flexible and yet also more structured, since it consists of multiple scales, but the perturbations are constrained at each scale.

$$x^a = C_{\mathcal{B}_1}\Big(\tilde{x} + \sum_{s\in S} m_s \cdot C_{\mathcal{B}_{\epsilon_s}}(\delta_s)\Big), \tag{4}$$

where $C_{\mathcal{B}_1}$ is the clipping function that keeps the image within the normal range, S is the set of scales, m_s is the multiplicative constant for scale s, δ_s is the learned perturbation (with the same shape as x). For scale s, the weights in δ_s are shared for pixels in square regions of size $s \times s$ with top left corner $[s \cdot i, s \cdot j]$ for all discrete $i \in [0, \text{width}/s]$ and

$j \in [0, \text{height}/s]$, as shown in Fig. 1. Note that, similar to pixel AT, each channel of the image is perturbed independently. More details of the parameter settings are given in Section 4 and pseudocode is included in the supplementals.

Setting up the attack For both the pixel and pyramid attacks, we use Projected Gradient Descent (PGD) on a random label using multiple steps [35]. With regards to the loss, we observe that for ViT, maximizing the negative loss of the true label leads to aggressive label leaking [29], *i.e.*, the network learns to predict the adversarial attack and performs better on the perturbed image. To avoid this, we pick a random label and then minimize the softmax cross-entropy loss towards that random label as described in [29].

3.3. "Matched" Dropout and Stochastic Depth

Standard training for ViT models uses both Dropout [50] and stochastic depth [23] as regularizers. During adversarial training, we have both the clean samples and adversarial samples in a mini-batch. This poses a question about Dropout treatment during adversarial training (either pixel or pyramid). In the adversarial training literature, the usual strategy is to run the adversarial attack (to generate adversarial samples) without using Dropout or stochastic depth. However, this leads to a training mismatch between the clean and adversarial training paths when both are used in the loss (Eqn. 3), with the clean samples trained with Dropout and the adversarial samples without Dropout. For each training instance in the mini-batch, the clean branch will only update subsets of the network while the adversarial branch updates the entire network. The adversarial branch updates are therefore more closely aligned with the model performance during evaluation, thereby leading to an improvement of adversarial accuracy at the expense of clean accuracy. This objective function is given below:

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\Big[L(\mathcal{M}(\theta), \tilde{x}, y) + \lambda \max_{\delta\in\mathcal{P}} L(\theta, x^a, y) + f(\theta)\Big], \quad (5)$$

where, with a slight abuse of notation, $\mathcal{M}(\theta)$ denotes a network with a random Dropout mask and a stochastic depth configuration. To address the issue above, we propose adversarial training of ViT with "matched" Dropout, *i.e.*, using the same Dropout configuration for both clean and adversarial training branches (as well as for the generation of adversarial samples). We show through ablation in Section 4 that using the same Dropout configuration leads to the best overall performance for both the clean and robust datasets.

4. Experiments

In this section, we compare the effectiveness of our proposed PyramidAT to non-AT models, and PixelAT models.

4.1. Experimental Setup

Models We focus primarily on ViT-B/16 [14], the baseline ViT with a patch size of 16. We also demonstrate our technique on other network architectures, such as ViT-Ti/16, ResNet [18], MLP-Mixer [54], and the recent Discrete ViT [37].

Datasets We train models on both ImageNet-1K and ImageNet-21K [12, 48]. We evaluate in-distribution performance on 2 additional variants: ImageNet-ReaL [4] which relabels the validation set of the original ImageNet in order to correct labeling errors; and ImageNet-V2 [47] which collects another version of ImageNet's evaluation set. We evaluate out-of-distribution robustness on 6 datasets: ImageNet-A [22] which places the ImageNet objects in unusual contexts or orientations; ImageNet-C [20] which applies a series of corruptions (e.g. motion blur, snow, JPEG, etc.); ImageNet-Rendition [19] which contains abstract or rendered versions of the object; ObjectNet [3] which consists of a large real-world set from a large number of different backgrounds, rotations, and imaging view points; ImageNet-Sketch [58] which contains artistic sketches of the objects; and Stylized ImageNet [16] which processes the ImageNet images with style transfer from an unrelated source image. For brevity, we may abbreviate ImageNet as IM. For all datasets except IM-C, we report top-1 accuracy (where higher is better). For IM-C, we report the standard "Mean corruption error" (mCE) (where lower is better).

Implementation details Following [51], we use a batch size of 4096, a cosine decay learning rate schedule (0.001 magnitude) with linear warmup for the first 10k steps, [34], and the AdamW optimizer [26] in all our experiments. Augmentations and regularizations include RandAug [10] with the default setting of $(2, 15)$, Dropout [50] at probability 0.1, and stochastic depth [23] at probability 0.1. We train with Scenic [11], a Jax [5] library, on DragonFish TPUs.

To generate the pixel adversarial attack, we follow [61]. We use a learning rate of $1/255$, $\epsilon = 4/255$, and attack for 5 steps with SGD. We use PGD [35] to generate the adversarial perturbations. We also experiment with using more recent optimizers [66] to construct the attacks (results are provided in the supplementals). For pyramid attacks, we find using stronger perturbations at coarser scales is more effective than equal perturbation strengths across all scales. By default, we use a 3-level pyramid and use perturbation scale factors $S = [32, 16, 1]$ (a scale of 1 means that each pixel has one learned parameter, a scale of 16 means that each $[16, 16]$ patch has one learned parameter) with multiplicative terms of $m_s = [20, 10, 1]$ (see Eqn. 4). We use a clipping value of $\epsilon_s = 6/255$ for all levels of the pyramid.

4.2. Experimental Results on ViT-B/16

ImageNet-1K Table 1 shows results on ImageNet-1K and robustness datasets for ViT-B/16 models without adversar-

Method	ImageNet	Real	A	C↓	ObjectNet	V2	Rendition	Sketch	Stylized
			Out of Distribution Robustness Test						
ViT [14]	72.82	78.28	8.03	74.08	17.36	58.73	27.07	17.28	6.41
ViT+CutMix [63]	75.49	80.53	14.75	64.07	21.61	62.37	28.47	17.15	7.19
ViT+Mixup [64]	77.75	82.93	12.15	61.76	25.65	64.76	34.90	25.97	9.84
RegViT (RandAug) [51]	79.92	85.14	17.48	52.46	29.30	67.49	38.24	29.08	11.02
+Random Pixel	79.72	84.72	17.81	52.83	28.72	67.17	39.01	29.26	12.11
+Random Pyramid	80.06	85.02	19.15	52.49	29.41	67.81	39.78	30.30	11.64
+PixelAT	80.42	85.78	19.15	47.68	30.11	68.78	45.39	34.40	18.28
+PyramidAT (Ours)	**81.71**	**86.82**	**22.99**	**44.99**	**32.92**	**70.82**	**47.66**	**36.77**	**19.14**
RegViT [51] on 384x384	81.44	86.38	26.20	58.19	35.59	70.09	38.15	28.13	8.36
+Random Pixel	81.32	86.18	25.95	58.69	34.12	69.50	37.66	28.79	9.77
+Random Pyramid	81.42	86.30	27.55	57.31	34.83	70.53	38.12	29.16	9.61
+PixelAT	82.24	87.35	31.23	48.56	37.41	71.67	44.07	33.68	13.52
+PyramidAT (Ours)	**83.26**	**88.14**	**36.41**	**47.76**	**39.79**	**73.14**	**46.68**	**36.73**	**15.00**

Table 1. Main results on ImageNet-1k. All columns reports top-1 accuracy except ImageNet-C which reports mean Corruption Error (mCE) where lower is better. All models are ViT-B/16. The first set of rows show the performance on train and testing on 224×224 images. The second set of rows shows performance by fine-tuning on 384×384 images.

Method	Extra Data	IM-C mCE ↓
DeepAugment+AugMix [19]	✗	53.60
AdvProp [61]	✗	52.90
Robust ViT [38]	✗	46.80
Discrete ViT [37]	✗	46.20
QualNet [25]	✗	42.50
Ours (ViT-B/16 + PyramidAT)	✗	**41.42**
Discrete ViT [37]	✓	38.74
Ours (ViT-B/16 + PyramidAT)	✓	**36.80**

Table 2. Comparison to state of the art for mean Corruption Error (mCE) on ImageNet-C. Extra data is IM-21k.

Method	Extra Data	IM-Rendition
Faces of Robustness [19]	✗	46.80
Robust ViT [38]	✗	48.70
Discrete ViT [37]	✗	48.82
Ours (ViT-B/16 + PyramidAT)	✗	**53.92**
Discrete ViT [37]	✓	55.26
Ours (ViT-B/16 + PyramidAT)	✓	**57.84**

Table 3. Comparison to state of the art for Top-1 on ImageNet-R. Extra data is IM-21k.

ial training, with pixel adversarial attacks and with pyramid adversarial attacks. Both adversarial training attacks use matched Dropout and stochastic depth, and optimize the random target loss. The pyramid attack provides consistent improvements, on both clean and robustness accuracies, over the baseline and pixel adversaries. In Table 1, we also compare against CutMix [63] augmentation. We find that CutMix improves performance over the ViT baseline but cannot improve performance when combined with RandAug. Similar to [33], we find that CutOut [13] does not boost performance on ImageNet for our models.

Method	Extra Data	IM-Sketch
ConViT-B [15]	✗	35.70
Swin-B [32]	✗	32.40
Robust ViT [38]	✗	36.00
Discrete ViT [37]	✗	39.10
Ours (ViT-B/16 + PyramidAT)	✗	**41.04**
Discrete ViT [37]	✓	44.72
Ours (ViT-B/16 + PyramidAT)	✓	**46.03**

Table 4. Comparison to state of the art for Top-1 on ImageNet-Sketch. Extra data is IM-21k.

The robustness gains of our technique are preserved through fine-tuning on clean data at higher resolution (384x384), as shown in the second set of rows of Table 1. Further, adversarial perturbations are consistently better than random perturbations on either pre-training or fine-tuning, for both pixel and pyramid models.

State of the art Our model trained on IM-1K sets a new overall state of the art for IM-C [20], IM-Rendition [19], and IM-Sketch [58], as shown in Tables 2, 3, and 4. While we compare all our models under a unified framework in our main experiments, we select the optimal pre-processing, fine-tuning, and Dropout setting for the given dataset when comparing against the state-of-the-art. We also compare against [37] on IM-21K and find that our results still compare favorably.

ImageNet-21K In table 5, we show that our technique maintains gains over the baseline Reg-ViT and pixel-wise attack on the larger dataset IM-21K. Following [51], we

Method	ImageNet	Real	A	C↓	ObjectNet	V2	Rendition	Sketch	Stylized
			Out of Distribution Robustness Test						
ViT-B/16 (512x512)	84.42	88.74	55.77	46.69	46.68	74.88	51.26	36.79	13.44
+PixelAT	84.82	89.10	57.39	43.31	47.53	75.42	53.35	39.07	17.66
+PyramidAT (Ours)	**85.35**	**89.43**	**62.44**	**40.85**	**49.39**	**76.39**	**56.15**	**43.95**	**19.84**

Table 5. Main results from pre-training on ImageNet-21K, fine-tuning on ImageNet-1K. We pre-train with the adversarial technique mentioned (pixel or pyramid), but fine-tune on clean data only.

pre-train on IM-21K and fine-tune on IM-1K at a higher resolution (in our case, 512x512). We apply adversarial training during the pre-training stage only.

4.3. Ablations

ImageNet-1k on other backbones We explore the effects of adversarial training on three other backbones: ResNet [18], Discrete ViT [37], and MLP-Mixer [54]. As shown in Table 6, we find slightly different results. For ResNet, we use the split BN from [61] and show improved performance from PyramidAT. Other ResNet variants (-101, -200) show the same trend and are included in the supplementals. For Discrete ViT, we show that AT with both pixel and pyramid leads to general improvements, though the gain from pyramid over pixel is less consistent than with ViT-B/16. For MLP-Mixer, we observe decreases in clean accuracy but gains in the robustness datasets for PixelAT, similar to what has traditionally been observed from AT on ConvNets. However, with PyramidAT, we observe improvements for all evaluation datasets.

Matched Dropout and Stochastic Depth We study the impact of handling Dropout and stochastic depth for the clean and adversarial update in Table 7. We find that applying matched Dropout for the clean and adversarial update is crucial for achieving simultaneous gains in clean and robust performance. When we eliminate Dropout in the adversarial update ("without Dropout" rows in 7), we observe significant decreases in performance on clean, IM-ReaL, and IM-A; and increases in performance on IM-Sketch and IM-Stylized. This result appears similar to the usual trade-off suggested in [45, 65]. By contrast, carefully handling Dropout and stochastic depth can lead to performance gains in both clean and out-of-distribution datasets.

Pyramid attack setup In Table 8, we ablate the pyramid attacks. Pyramid attacks are consistently better than pixel or patch attacks, while the 3-level pyramid attack tends to have the best overall performance. Note that a 2-level pyramid attack consists of both the pixel and patch attacks. Please refer to the supplementals for comparison on all the metrics.

Network capacity and random augmentation We test the effect of network capacity on adversarial training and, consistent with existing literature [28, 35], find that large capacity is critical to effectively utilizing PixelAT. Specifically, low-capacity networks, like ViT-Ti/16, which already struggle to represent the dataset, can be made worse through PixelAT. Table 9 shows that PixelAT hurts in-distribution performance of the RandAugment 0.4 model but improves out-of-distribution performance. Unlike prior work, we note that this effect depends on both the network capacity and the random augmentation applied to the dataset.

Table 9 shows that a low-capacity network can benefit from adversarial training if the random augmentation is of a small magnitude. Standard training with RandAugment [10] magnitude of 0.4 (abbreviated as RAm=0.4) provides a better clean accuracy than standard training with RAm=0.1; however, PixelAT with the weaker augmentation, RAm=0.1, performs better than either standard training or PixelAT at RAm=0.4. This suggests that the augmentation should be tuned for adversarial training and not fixed based on standard training.

Table 9 also shows that PyramidAT acts differently than PixelAT and can provide in-distribution gains despite being used with stronger augmentation. For these models, we find that for the robustness datasets, PixelAT tends to marginally outperform PyramidAT.

Attack strength Pixel attacks are much smaller in L_2 norm than pyramid attacks. We check that simply scaling up the PixelAT cannot achieve the same performance as PyramidAT in Figure 2. For both ImageNet and ImageNet-C, we show the effect of raising the pixel and pyramid attack strength. While the best PyramidAT performance is achieved at high L_2 perturbation norm, the PixelAT performance degrades beyond a certain norm.

4.4. Analysis and Discussions

Qualitative results Following [14], we visualize the learned pixel embeddings (filters) of models trained normally, with pixel adversaries, and with pyramid adversaries in Fig. 3. We observe that the PixelAT model tends to tightly "snap" its attention to the perceived object, disregarding the majority of the background. While this may appear to be a desirable behavior, this kind of focusing can be suboptimal for the in-distribution datasets (where the background can provide valuable context) and prone to errors for out-of-distribution datasets. Specifically, the PixelAT model may

Method	ImageNet	Real	A	C↓	ObjectNet	V2	Rendition	Sketch	Stylized
			Out of Distribution Robustness Test						
ResNet-50 [18] (our run)	76.70	83.11	4.49	74.90	26.47	64.31	36.24	23.44	6.41
+PixelAT	77.37	84.11	6.03	66.88	27.80	65.59	41.75	27.04	8.13
+PyramidAT	**77.48**	**84.22**	**6.24**	**66.77**	**27.91**	**65.96**	**43.32**	**28.55**	**8.83**
MLP-Mixer [54] (our run)	78.27	83.64	10.84	58.50	25.90	64.97	38.51	29.00	10.08
+PixelAT	77.17	82.99	9.93	57.68	24.75	64.03	44.43	33.68	**15.31**
+PyramidAT	**79.29**	**84.78**	**12.97**	**52.88**	**28.60**	**66.56**	**45.34**	**34.79**	14.77
Discrete ViT [37] (our run)	79.88	84.98	18.12	49.43	29.95	68.13	41.70	31.13	15.08
+PixelAT	80.08	85.37	16.88	48.93	**30.98**	68.63	**48.00**	**37.42**	**22.34**
+PyramidAT	**80.43**	**85.67**	**19.55**	**47.30**	30.28	**69.04**	46.72	37.21	19.14

Table 6. Pyramid adversarial training improves the performance of ResNet, MLP-Mixer, and Discrete ViT. On MLPMixer, pixel attacks degrade clean performance but improve robustness, similar to the traditionally observed effect of adversarial training.

Method	ImageNet	Real	A	C↓	ObjectNet	V2	Rendition	Sketch	Stylized
			Out of Distribution Robustness Test						
PixelAT with matched Dropout	**80.42**	**85.78**	**19.15**	**47.68**	**30.11**	**68.78**	45.39	34.40	18.28
PixelAT without Dropout	79.35	84.67	15.27	51.45	29.46	67.01	**47.83**	**35.77**	**18.75**
PyramidAT with matched Dropout	**81.71**	**86.82**	**22.99**	**44.99**	**32.92**	**70.82**	47.66	36.77	19.14
PyramidAT without Dropout	79.43	85.13	14.13	54.70	29.67	67.40	**52.34**	**40.25**	**22.34**

Table 7. Matched Dropout leads to better performance on in-distribution datasets than AT without Dropout.

Method	IM	A	C↓	Rend.	Sketch
Pixel	80.42	19.15	47.68	45.39	34.40
Patch	81.20	21.33	50.30	42.87	33.75
2-level Pyramid	81.65	22.79	45.27	47.00	36.71
3-level Pyramid	**81.71**	22.99	**44.99**	47.66	36.77
4-level Pyramid	81.66	**23.21**	45.29	**47.68**	**37.41**

Table 8. Pyramid structure ablation. This shows the effect of the layers of the pyramid. Adding coarser layers with larger magnitudes typically improves performance. Patch attack is a 1-level pyramid with shared parameters across a patch of size 16×16.

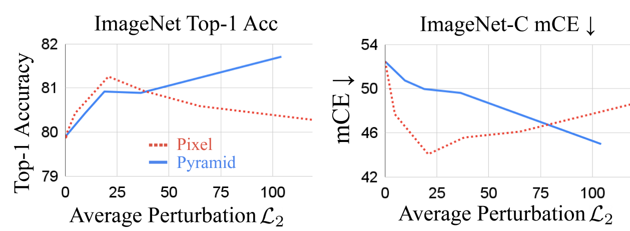

Figure 2. Performance on clean and robust data as a function of perturbation size. Pyramid performance increases as perturbation size is increased, while pixel performance with large perturbation size is poor

Method	IM	A	C↓	Rend	Sketch
Ti/16 RAm=0.1	63.58	4.80	79.23	23.66	12.54
+PixelAT	64.66	4.39	74.54	**32.52**	**17.65**
+PyramidAT	**65.49**	**5.16**	**74.30**	29.18	16.55
Ti/16 RAm=0.4	64.27	4.69	78.10	24.99	13.47
+PixelAT	62.78	4.05	77.67	**29.75**	**16.35**
+PyramidAT	**65.61**	**4.80**	**74.72**	28.89	16.14

Table 9. Results on Ti/16 with lower random augmentation. RAm is the RandAugment [10] magnitude – larger means stronger augmentation; both have RandAugment number of transforms = 1. The strength of the random augmentation affects whether PixelAT improves clean accuracy; in contrast, PyramidAT provides consistent gains over the baseline.

under-estimate the size or shape of the object and focus on a part of the object and not the whole. This can be problematic for fine-grained classification when the difference between two classes comes down to something as small as the stripes or subtle shape cues (tiger shark vs great white); or texture and context (green mamba vs vine snake). Figure 4 shows the heat maps for the average attention on images in the evaluation set of ImageNet-A. We observe that PyramidAT tends to more evenly spread its attention across the entire image than both the baseline and PixelAT.

Figure 5 demonstrates the difference in representation between the baseline, PixelAT, and PyramidAT models. The pixel attacks on the baseline and PixelAT have a small amount of structure but appear to consist of mostly texture-level noise. In contrast, the pixel level of the PyramidAT shows structures from the original image: the legs and back

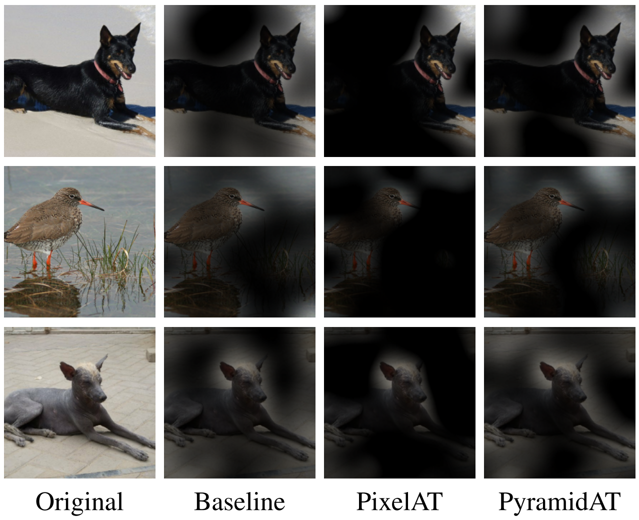

Figure 3. Visualizations of the attention for different models. PixelAT focuses aggressively on the perceived object. However, if the object is not identified correctly, this focus can be detrimental, as shown above where large parts of the object are discarded. PyramidAT uses a more global perspective and considers context.

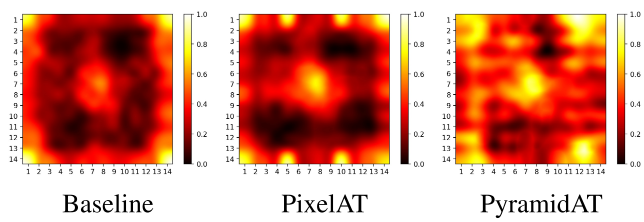

Figure 4. Averaged attentions on ImageNet-A: PyramidAT models attend to more of the image than the baseline or PixelAT.

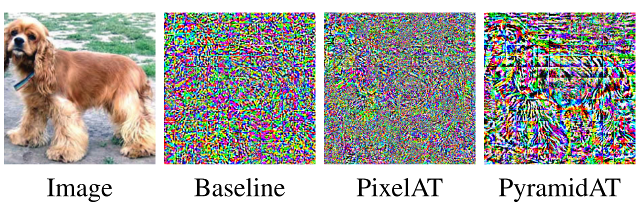

Figure 5. Visualizations of attacks: a pixel attack on a baseline ViT; a pixel attack on a PixelAT ViT; and the pixel level of a pyramid attack on a PyramidAT ViT. The pixel attack on the baseline exhibits low amounts of structure and can perturb the label with small changes. The pixel level on the PyramidAT model makes larger changes to the structure; this suggests that the representation is robust to semi-random noise and focuses primarily on structures.

of the dog. This suggests that the representation for the PyramidAT model focuses on shape and is less sensitive to texture than the baseline model.

Analysis of attacks Inspired by [62], we analyze the pyramid adversarial training from a frequency perspective. For this analysis, all visualizations and graphs are averaged

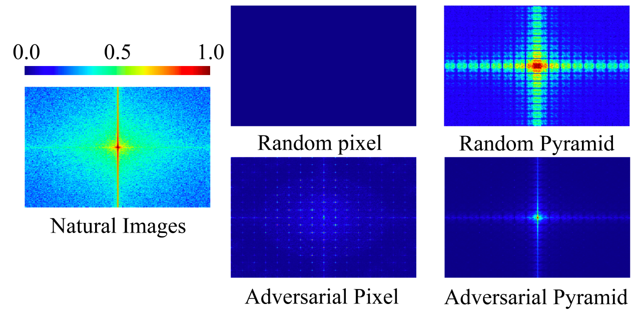

Figure 6. Heatmaps of fourier spectrum for various perturbations.

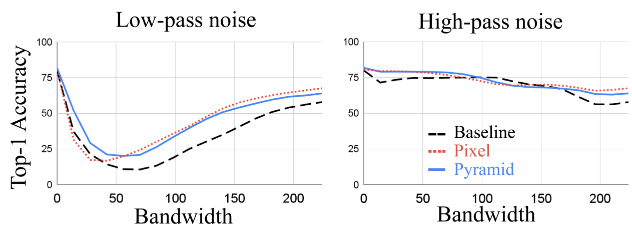

Figure 7. Model performance when inputs are corrupted with low-pass/high-pass filtered noise. The L_2 norm of the filtered noise is held constant as the bandwidth is increased.

over the entire ImageNet validation set. Figure 6 shows a Fourier heat map of random and adversarial versions of the pixel and pyramid attacks. While random pixel noise is evenly concentrated over all frequencies, adversarial pixel attack tends to concentrate in the lower frequencies. Random pyramid shows a bias towards low frequency as well, a trend which is amplified in the adversarial pyramid. To further explore this, we replicate an analysis from [62], where low-pass- and high-pass-filtered random noise is added to test data to perturb a classifier. Figure 7 gives the result for our baseline, pixel, and pyramid adversarially trained models. While pixel and pyramid models are generally more robust than the baseline, the pyramid model is more robust than the pixel model to low-frequency perturbations.

Limitations The cost of our technique is increased training time. A k-step PGD attack requires k forward and backward passes for each step of training. Note that this limitation holds for any adversarial training and the inference time is the same. Without adversarial training, more training time does not improve the baseline ViT-B/16.

5. Conclusion

We have introduced pyramid adversarial training, a simple and effective data augmentation technique that substantially improves the performance of ViT and MLP-Mixer architectures on in-distribution and a number of out-of-distribution ImageNet datasets.

References

[1] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning (ICML)*, pages 274–283. PMLR, 2018. 2

[2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEit: BERT pre-training of image transformers. In *International Conference on Learning Representations (ICLR)*, 2022. 1

[3] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32. Curran Associates, Inc., 2019. 4

[4] Lucas Beyer, Olivier J. Henaff, Alexander Kolesnikov, Xiaohua Zhai, and Aaron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2002.05709*, 2020. 4

[5] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. 4

[6] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE symposium on security and privacy (SP)*, pages 39–57. IEEE, 2017. 2

[7] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. 2

[8] Yong Cheng, Lu Jiang, and Wolfgang Macherey. Robust neural machine translation with doubly adversarial inputs. *arXiv preprint arXiv:1906.02443*, 2019. 2

[9] Yong Cheng, Lu Jiang, Wolfgang Macherey, and Jacob Eisenstein. Advaug: Robust adversarial augmentation for neural machine translation. *arXiv preprint arXiv:2006.11834*, 2020. 2

[10] Ekin Dogus Cubuk, Barret Zoph, Jon Shlens, and Quoc Le. Randaugment: Practical automated data augmentation with a reduced search space. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 18613–18624. Curran Associates, Inc., 2020. 3, 4, 6, 7

[11] Mostafa Dehghani, Alexey Gritsenko, Anurag Arnab, Matthias Minderer, and Yi Tay. Scenic: A JAX library for computer vision research and beyond. *arXiv preprint arXiv:2110.11403*, 2021. 4

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 4

[13] Terrance Devries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout. *CoRR*, abs/1708.04552, 2017. 5

[14] 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 *International Conference on Learning Representations (ICLR)*, 2021. 1, 2, 4, 5, 6

[15] 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 *International Conference on Machine Learning (ICML)*, pages 2286–2296. PMLR, 2021. 5

[16] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*, 2019. 1, 2, 4

[17] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015. 1, 3

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 2, 3, 4, 6, 7

[19] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 4, 5

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

[21] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019. 3

[22] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 4

[23] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *European conference on computer vision*, pages 646–661. Springer, 2016. 2, 4

[24] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, pages 448–456. PMLR, 2015. 2

[25] Insoo Kim, Seungju Han, Ji-won Baek, Seong-Jin Park, Jae-Joon Han, and Jinwoo Shin. Quality-agnostic image recognition via invertible decoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12257–12266, June 2021. 5

[26] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *International Conference on Learning Representations (ICLR)*, 2015. 4

[27] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 2

[28] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016. 1, 6

[29] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016. 1, 2, 3, 4

[30] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017. 1

[31] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2018. 1

[32] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, October 2021. 1, 5

[33] Raphael Gontijo Lopes, Dong Yin, Ben Poole, Justin Gilmer, and Ekin Dogus Cubuk. Improving robustness without sacrificing accuracy with patch gaussian augmentation. *ArXiv*, abs/1906.02611, 2019. 5

[34] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2017. 4

[35] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2018. 2, 3, 4, 6

[36] Kaleel Mahmood, Rigel Mahmood, and Marten van Dijk. On the robustness of vision transformers to adversarial examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7838–7847, October 2021. 1

[37] Chengzhi Mao, Lu Jiang, Mostafa Dehghani, Carl Vondrick, Rahul Sukthankar, and Irfan Essa. Discrete representations strengthen vision transformer robustness. In *International Conference on Learning Representations (ICLR)*, 2022. 2, 3, 4, 5, 6, 7

[38] Xiaofeng Mao, Gege Qi, Yuefeng Chen, Xiaodan Li, Ranjie Duan, Shaokai Ye, Yuan He, and Hui Xue. Towards robust vision transformer. *CoRR*, abs/2105.07926, 2021. 1, 5

[39] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018. 2

[40] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Jonathan Uesato, and Pascal Frossard. Robustness via curvature regularization, and vice versa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9078–9086, 2019. 2

[41] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021. 2

[42] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2011. 2

[43] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE symposium on security and privacy (SP)*, pages 582–597. IEEE, 2016. 2

[44] Chongli Qin, James Martens, Sven Gowal, Dilip Krishnan, Krishnamurthy Dvijotham, Alhussein Fawzi, Soham De, Robert Stanforth, and Pushmeet Kohli. Adversarial robustness through local linearization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. 2

[45] Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C. Duchi, and Percy Liang. Understanding and mitigating the tradeoff between robustness and accuracy. *CoRR*, abs/2002.10716, 2020. 1, 2, 6

[46] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12179–12188, October 2021. 1

[47] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning (ICML)*, pages 5389–5400. PMLR, 2019. 4

[48] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 4

[49] Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. On the adversarial robustness of visual transformers. *CoRR*, abs/2103.15670, 2021. 1, 2, 3

[50] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. 4

[51] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *CoRR*, abs/2106.10270, 2021. 1, 2, 3, 4, 5

[52] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 2

[53] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning (ICML)*, pages 6105–6114. PMLR, 2019. 2

[54] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021. 2, 3, 4, 6, 7

[55] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning (ICML)*, pages 10347–10357. PMLR, 2021. 1, 3

[56] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy, 2019. 1, 2

[57] Abraham Wald. Statistical decision functions which minimize the maximum risk. *Annals of Mathematics*, pages 265–280, 1945. 3

[58] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 10506–10518, 2019. 1, 2, 4, 5

[59] Haotao Wang, Chaowei Xiao, Jean Kossaifi, Zhiding Yu, Anima Anandkumar, and Zhangyang Wang. Augmax: Adversarial composition of random augmentations for robust training. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021. 3

[60] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. *arXiv preprint arXiv:1801.02612*, 2018. 2, 3

[61] Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L Yuille, and Quoc V Le. Adversarial examples improve image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 819–828, 2020. 1, 2, 3, 4, 5, 6

[62] Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. 8

[63] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6023–6032, 2019. 5

[64] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 5

[65] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7472–7482. PMLR, 09–15 Jun 2019. 1, 2, 6

[66] Juntang Zhuang, Tommy Tang, Yifan Ding, Sekhar C Tatikonda, Nicha Dvornek, Xenophon Papademetris, and James Duncan. Adabelief optimizer: Adapting stepsizes by the belief in observed gradients. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 18795–18806, 2020. 4