Vision Transformers are Robust Learners

Sayak Paul,[1*] **Pin-Yu Chen** [2*]

[1] Carted [2] IBM Research
sayak@carted.com, pin-yu.chen@ibm.com

Abstract

Transformers, composed of multiple self-attention layers, hold strong promises toward a generic learning primitive applicable to different data modalities, including the recent breakthroughs in computer vision achieving state-of-the-art (SOTA) standard accuracy. What remains largely unexplored is their robustness evaluation and attribution. In this work, we study the robustness of the Vision Transformer (ViT) (Dosovitskiy et al. 2021) against common corruptions and perturbations, distribution shifts, and natural adversarial examples. We use six different diverse ImageNet datasets concerning robust classification to conduct a comprehensive performance comparison of ViT (Dosovitskiy et al. 2021) models and SOTA convolutional neural networks (CNNs), Big-Transfer (Kolesnikov et al. 2020). Through a series of six systematically designed experiments, we then present analyses that provide both quantitative and qualitative indications to explain why ViTs are indeed more robust learners. For example, with fewer parameters and similar dataset and pre-training combinations, ViT gives a top-1 accuracy of 28.10% on ImageNet-A which is 4.3x higher than a comparable variant of BiT. Our analyses on image masking, Fourier spectrum sensitivity, and spread on discrete cosine energy spectrum reveal intriguing properties of ViT attributing to improved robustness. Code for reproducing our experiments is available at https://git.io/J3VO0.

1 Introduction

Transformers (Vaswani et al. 2017) are becoming a preferred architecture for various data modalities. This is primarily because they help reduce inductive biases that go into designing network architectures. Moreover, Transformers have been shown to achieve tremendous parameter efficiency without sacrificing predictive performance over architectures that are often dedicated to specific types of data modalities. Attention, in particular, self-attention is one of the foundational blocks of Transformers. It is a computational primitive that allows us to quantify pairwise entity interactions thereby helping a network learn the hierarchies and alignments present inside the input data (Bahdanau, Cho, and Bengio 2015; Vaswani et al. 2017). These are desirable properties to eliminate the need for carefully designed inductive biases to a great extent.

Although Transformers have been used in prior works (Trinh, Luong, and Le 2019; Chen et al. 2020) it was only until 2020, the performance of Transformers were on par with the SOTA CNNs on standard image recognition tasks (Carion et al. 2020; Touvron et al. 2020; Dosovitskiy et al. 2021). Attention has been shown to be an important element for vision networks to achieve better empirical robustness (Hendrycks et al. 2021). Since attention is a core component of ViTs (and Transformers in general), a question that naturally gets raised here is – *could ViTs be inherently more robust?* If so, *why are ViTs more robust learners?* In this work, we provide an affirmative answer to the first question and provide empirical evidence to reason about the improved robustness of ViTs.

Various recent works have opened up the investigation on evaluating the robustness of ViTs (Bhojanapalli et al. 2021; Shao et al. 2021; Mahmood, Mahmood, and Van Dijk 2021) but with a relatively limited scope. We build on top of these and provide further and more comprehensive analyses to understand why ViTs provide better robustness for semantic shifts, common corruptions and perturbations, and natural adversarial examples to input images in comparison to SOTA CNNs like Big Transfer (BiT) (Kolesnikov et al. 2020). Through a set of carefully designed experiments, we first verify the enhanced robustness of ViTs to common robustness benchmark datasets (Hendrycks and Dietterich 2019; Hendrycks et al. 2020, 2021; Xiao et al. 2021). We then provide quantitative and qualitative analyses to help understand the reasons behind this enhancement. In summary, we make the following contributions:

- We use 6 diverse ImageNet datasets concerning different types of robustness evaluation and conclude that ViTs achieve significantly better performance than BiTs.
- We design 6 experiments, including robustness to masking, energy/loss landscape analysis, and sensitivity to high-frequency artifacts to study ViT's improved robustness.
- Our analysis provides novel insights for robustness attribution of ViT. Moreover, our robustness evaluation and analysis tools are generic and can be used to benchmark and study future image classification models.

2 Related Work

To the best of our knowledge, (Parmar et al. 2018) first explored the use of Transformers (Vaswani et al. 2017) for the task of image super-resolution which essentially belongs to

[*] These authors contributed equally.

the category of image generation. Image-GPT (Chen et al. 2020) used Transformers for unsupervised pre-training from pixels of images. However, the transfer performance of the pre-training method is not on par with supervised pre-training methods. ViT (Dosovitskiy et al. 2021) takes the original Transformers and makes very minimal changes to make it work with images. In fact, this was one of the primary objectives of ViT i.e. to keep the original Transformer architecture as original as possible and then examining how that pans out for image classification in terms of large-scale pre-training. As noted in (Dosovitskiy et al. 2021; Steiner et al. 2021), because of the lesser number of inductive biases, ViT needs to be pre-trained on a relatively larger dataset (such as ImageNet-21k (Deng et al. 2009)) with strong regularization for achieving reasonable downstream performance. Strong regularization is particularly needed in the absence of a larger dataset during pre-training (Steiner et al. 2021).

Multiple variants of Transformers have been proposed to show that it is possible to achieve comparable performance on ImageNet-1k *without* using additional data. DeIT (Touvron et al. 2020) introduces a novel distillation strategy (Hinton, Vinyals, and Dean 2015) to learn a student Transformers-based network from a well-performing teacher network based on RegNets (Radosavovic et al. 2020). With this approach, DeIT achieves 85.2% top-1 accuracy on ImageNet-1k without any external data. T2T-ViT (Yuan et al. 2021) proposes a novel tokenization method enabling the network to have more access to local structures of the images. For the Transformer-based backbone, it follows a deep-narrow network topology inspired by (Zagoruyko and Komodakis 2016). With proposed changes, T2T-ViT achieves 83.3% top-1 accuracy on ImageNet-1k. LV-ViT (Jiang et al. 2021) introduces a new training objective namely token labeling and also tunes the structure of the Transformers. It achieves 85.4% top-1 accuracy on ImageNet-1k. CLIP (Radford et al. 2021) and Swin Transformers (Liu et al. 2021) are also two recent models that make use of Transformers for image recognition problems. In this work, we only focus on ViT (Dosovitskiy et al. 2021).

Concurrent to our work, there are a few recent works that study the robustness of ViTs from different perspectives. In what follows, we summarize their key insights and highlight the differences from our work. (Shao et al. 2021) showed that ViTs has better robustness than CNNs against adversarial input perturbations. The major performance gain can be attributed to the capability of learning high-frequency features that are more generalizable and the finding that convolutional layers hinder adversarial robustness. (Bhojanapalli et al. 2021) studied improved robustness of ViTs over ResNets (He et al. 2016) against adversarial and natural adversarial examples as well as common corruptions. Moreover, it is shown that ViTs are robust to the removal of almost any single layer. (Mahmood, Mahmood, and Van Dijk 2021) studied adversarial robustness of ViTs through various white-box, black-box and transfer attacks and found that model ensembling can achieve unprecedented robustness without sacrificing clean accuracy against a black-box adversary. This paper shows novel insights that are fundamentally different from these works: (**i**) we benchmark the robustness of ViTs on a wide spectrum of ImageNet datasets (see Table 2), which are the most comprehensive robustness performance benchmarks to date; (**ii**) we design six new experiments to verify the superior robustness of ViTs over BiT and ResNet models.

3 Robustness Performance Comparison on ImageNet Datasets

3.1 Multi-head Self Attention (MHSA)

Here we provide a brief summary of ViTs. Central to ViT's model design is self-attention (Bahdanau, Cho, and Bengio 2015). Here, we first compute three quantities from linear projections ($X \in \mathbb{R}^{N \times D}$): (**i**) **Q**uery = XW_{Q}, (**ii**) **K**ey = XW_{K} , and (**iii**) **V**alue = XW_{V}, where $W_{\mathrm{Q}}, W_{\mathrm{K}}$, and W_{V} are linear transformations. The linear projections (X) are computed from batches of the original input data. Self-attention takes these three input quantities and returns an output matrix ($N \times d$) weighted by attention scores using (1):

$$\mathrm{Attention}(Q, K, V) = \mathrm{Softmax}\left(QK^{\top}/\sqrt{d}\right) V \qquad (1)$$

To enable feature-rich hierarchical learning, h self-attention layers (or so-called ”heads”) are stacked together producing an output of $N \times dh$. This output is then fed through a linear transformation layer that produces the final output of $N \times d$ from MHSA. MHSA then forms the core Transformer block. Additional details about ViT's foundational elements are provided in Appendix A and B.

3.2 Performance Comparison on Diverse ImageNet Datasets for Robustness Evaluation

Baselines In this work, our baseline is a ResNet50V2 model (He et al. 2016) pre-trained on the ImageNet-1k dataset (Russakovsky et al. 2015) except for a few results where we consider ResNet-50 (He et al. 2016)[1]. To study how ViTs hold up with the SOTA CNNs we consider BiT (Kolesnikov et al. 2020). At its core, BiT networks are scaled-up versions of ResNets with Group Normalization (Wu and He 2018) and Weight Standardization (Qiao et al. 2019) layers added in place of Batch Normalization (Ioffe and Szegedy 2015). Since ViT and BiT share similar pre-training strategies (such as using larger datasets like ImageNet-21k (Deng et al. 2009) and JFT-300 (Sun et al. 2017), longer pre-training schedules, and so on) they are excellent candidates for our comparison purposes. So, a question, central to our work is:

> *Where does ViT stand with respect to BiT in terms of robustness under similar parameter and FLOP regime, pre-training setup, and data regimes, and how to attribute their performance difference?*

Even though BiT and ViT share similar pre-training schedules and dataset regimes there are differences that are worth mentioning. For example, ViT makes use of Dropout (Srivastava et al. 2014) while BiT does not. ViT is trained using Adam (Kingma and Ba 2015) while BiT is trained using SGD with momentum. In this work, we focus our efforts

[1] In these cases, we directly referred to the previously reported results with ResNet-50.

Variant	# Parameters (Million)	# FLOPS (Million)	ImageNet-1k Top-1 Acc
ResNet50V2	25.6138	4144.854528	76
BiT m-r50x1	25.549352	4228.137	80
BiT m-r50x3	217.31908	37061.838	84
BiT m-r101x1	44.54148	8041.708	82.1
BiT m-r101x3	387.934888	71230.434	84.7
BiT m-r152x4	936.53322	186897.679	85.39
ViT B-16	86.859496	17582.74	83.97
ViT B-32	88.297192	4413.986	81.28
ViT L-16	304.715752	61604.136	85.15
ViT L-32	306.63268	15390.083	80.99

Table 1: Parameter counts, FLOPS (Floating-Point Operations), and top-1 accuracy (%) of different variants of ViT and BiT. All the reported variants were pre-trained on ImageNet-21k and then fine-tuned on ImageNet-1k.

on the publicly available BiT and ViT models only. Later variants of ViTs have used Sharpness-Aware Minimization (Foret et al. 2021) and stronger regularization techniques to compensate the absence of favored inductive priors (Chen, Hsieh, and Gong 2021; Steiner et al. 2021). However, we do not investigate how those aspects relate to robustness in this work.

Table 1 reports the parameter counts and FLOPS of different ViT and BiT models along with their top-1 accuracy[2] on the ImageNet-1k dataset (Russakovsky et al. 2015). It is clear that different variants of ViT are able to achieve comparable performance to BiT but with lesser parameters.

In what follows, we compare the performance of ViT and BiT on six robustness benchmark datasets (Hendrycks and Dietterich 2019; Hendrycks et al. 2020, 2021), as summarized in Table 2. These datasets compare the robustness of ViT, BiT and the baseline ResNet50V2 in different perspectives, including (**i**) common corruptions, (**ii**) semantic shifts, (**iii**) natural adversarial examples, and (**iv**) out-of-distribution detection. A summary of the datasets and their purpose is presented in Table 2 for easier reference.

Notably, in these datasets ViT exhibits significantly better robustness than BiT of comparable parameter counts. Section 4 gives the attribution analysis of improved robustness in ViT.

Dataset	Purpose
ImageNet-C (Hendrycks and Dietterich 2019)	Common corruptions
ImageNet-P (Hendrycks and Dietterich 2019)	Common perturbations
ImageNet-R (Hendrycks et al. 2020)	Semantic shifts
ImageNet-O (Hendrycks et al. 2021)	Out-of-domain distribution
ImageNet-A (Hendrycks et al. 2021)	Natural adversarial examples
ImageNet-9 (Xiao et al. 2021)	Background dependence

Table 2: Summary of the studied datasets and their purpose.

ImageNet-C (Hendrycks and Dietterich 2019) consists of 15 types of algorithmically generated corruptions, and each type of corruption has five levels of severity. Along with these, the authors provide additional four types of general corruptions making a total of 19 corruptions. We consider all the 19 corruptions at their highest severity level (5) and report the mean top-1 accuracy in Figure 1 as yielded by the variants of ViT and BiT. We consistently observe a better performance across all the variants of ViT under different parameter regimes. Note that BiT `m-r50x1` and `m-r101x1` have lesser parameters than the lowest variant of ViT (`B-16`) but for other possible groupings, variants of ViT have lesser parameters than that of BiT. Overall, we notice that ViT performs consistently better across different corruptions except for *contrast*. In Figure 2, we report the top-1 accuracy of ViT and BiT on the highest severity level of the contrast corruption. This observation leaves grounds for future research to investigate why this is the case since varying contrast factors are quite common in real-world use-cases. Based on our findings, contrast can be an effective but unexplored approach to studying ViT's robustness, similar to the study of human's vision performance (Hart et al. 2013; Tuli et al. 2021).

In (Hendrycks and Dietterich 2019), mean corruption error (mCE) is used to quantify the robustness factors of a model on ImageNet-C. Specifically, the top-1 error rate is computed for each of the different corruption (c) types $(1 \leq c \leq 15)$ and for each of the severity (s) levels $(1 \leq s \leq 5)$. When error rates for all the severity levels are calculated for a particular corruption type, their average is stored. This process is repeated for all the corruption types and the final value is an average over all the average error rates from the different corruption types. The final score is normalized by the mCE of AlexNet (Krizhevsky, Sutskever, and Hinton 2012).

We report the mCEs for BiT-m `r101x3`, ViT L-16, and a few other models in Table 3. The mCEs are reported for 15 corruptions as done in (Hendrycks and Dietterich 2019). We include two additional models/methods in Table 3 because of the following: (**a**) Noisy Student Training (Xie et al. 2020) uses external data and training choices (such as using RandAugment (Cubuk et al. 2020), Stochastic Depth (Huang et al. 2016), etc.) that are helpful in enhancing the robustness of a vision model; (**b**) DeepAugment and AugMix (Hendrycks et al. 2020; Hendrycks* et al. 2020) are designed explicitly to improve the robustness of models against corruptions seen in ImageNet-C. This is why, to provide a fair ground to understand where BiT and ViT stand in comparison to state-of-the-art, we add these two models. It is indeed interesting to notice that ViT is able to outperform the combination of DeepAugment and AugMix which are specifically designed to provide robustness against the corruptions found in ImageNet-C. As we will discuss in Section 4, this phenomenon can be attributed to two primary factors: (**a**) better pre-training and (**b**) self-attention. It should also be noted that Noisy Student Training (Xie et al. 2020) incorporates various factors during training such as an iterative training procedure, strong data augmentation transformations from RandAugment for noise injection, test-time augmentation, and so on. These factors largely contribute to the improved

[2] Figure 4 of (Kolesnikov et al. 2020) and Table 5 of (Dosovitskiy et al. 2021) were used to collect the top-1 accuracy scores.

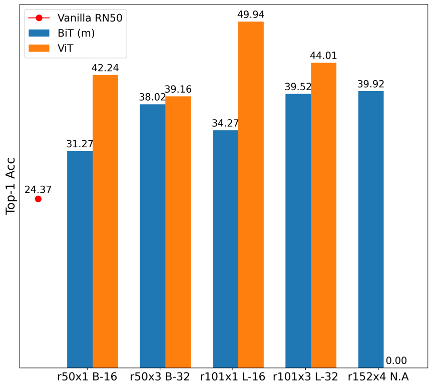

Figure 1: Mean top-1 accuracy scores (%) on the ImageNet-C dataset as yielded by different variants of ViT and BiT.

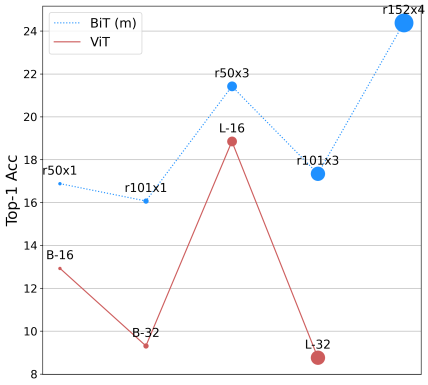

Figure 2: Top-1 accuracy (%) of ViT and BiT for contrast corruption (with the highest severity level) on ImageNet-C.

Model / Method	mCE
ResNet-50	76.7
BiT `m-r101x3`	58.27
DeepAugment+AugMix	53.6
ViT L-16	45.45
Noisy Student Training	28.3

Table 3: mCEs (%) of different models and methods on ImageNet-C (lower is better). Note that Noisy Student Training incorporates additional training with data augmentation for noise injection.

Model / Method	mFR	mT5D
ResNet-50	58	82
BiT-m `r101x3`	49.99	76.71
AugMix (Hendrycks* et al. 2020)	37.4	NA
ViT L-16	33.064	50.15

Table 4: mFRs (%) and mT5Ds (%) on ImageNet-P dataset (lower is better).

robustness gains achieved by Noisy Student Training.

ImageNet-P (Hendrycks and Dietterich 2019) has 10 types of common perturbations. Unlike the common corruptions, the perturbations are subtly nuanced spanning across fewer number of pixels inside images. As per (Hendrycks and Dietterich 2019) mean flip rate (mFR) and mean top-5 distance (mT5D) are the standard metrics to evaluate a model's robustness under these perturbations. They are reported in Table 4. Since the formulation of mFR and mT5D are more involved than mCE and for brevity, we refer the reader to (Hendrycks and Dietterich 2019) for more details on these two metrics. We find ViT's robustness to common perturbations is significantly better than BiT as well as AugMix.

ImageNet-R (Hendrycks et al. 2020) contains images labelled with ImageNet labels by collecting renditions of ImageNet classes. It helps verify the robustness of vision networks under semantic shifts under different domains. Figure 3 shows that ViT's treatment to domain adaptation is better than that of BiT.

ImageNet-A (Hendrycks et al. 2021) is comprised of natural images that cause misclassifications due to reasons such as multiple objects associated with single discrete categories. In Figure 4, we report the top-1 accuracy of ViT and BiT on the ImageNet-A dataset (Hendrycks et al. 2021). In (Hendrycks et al. 2021), self-attention is noted as an important element to tackle these problems. This may help explain why ViT performs significantly better than BiT in this case. For example, the top-1 accuracy of ViT L-16 is 4.3x higher than BiT-m `r101x3`.

ImageNet-O (Hendrycks et al. 2021) consists of images that belong to different classes not seen by a model during its training and are considered as *anomalies*. For these images, a robust model is expected to output low confidence scores. We follow the same evaluation approach of using *area under the precision-recall curve* (AUPR) as (Hendrycks et al. 2021) for this dataset. In Figure 5, we report the AUPR of the different ViT and BiT models on the ImageNet-O dataset (Hendrycks et al. 2021). ViT demonstrates superior performance in anomaly detection than BiT.

ImageNet-9 (Xiao et al. 2021) helps to verify the background-robustness of vision models. In most cases, the foregrounds of images inform our decisions on what might be present inside images. Even if the backgrounds change, as long as the foregrounds stay intact, these decisions should not be influenced. However, do vision models exhibit a similar kind of treatment to image foregrounds and backgrounds? It turns out that the vision models may break down when the background of an image is changed (Xiao et al. 2021). It may suggest that the vision models may be picking up unnecessary signals from the image backgrounds. In (Xiao et al. 2021) it is also shown that background-robustness can be important for determining models' out of distribution performance. So, naturally, this motivates us to investigate if ViT would have better background-robustness than BiT. We find that is indeed the case (refer to Table 5). Additionally, in Table 6, we report how well BiT and ViT can detect if the foreground of an image is vulnerable[3]. It appears that for this task also, ViT significantly outperforms BiT. Even though we notice ViT's better performance than BiT but it is surprising to see ViT's performance being worse than ResNet-50. We suspect this may be due to the simple tokenization process of ViT to create small image patches that limits the capability to process important local structures (Yuan et al. 2021).

4 Why ViT has Improved Robustness?

In this section, we systematically design and conduct six experiments to identify the sources of improved robustness in ViTs from both qualitative and quantitative standpoints.

4.1 Attention is Crucial for Improved Robustness

In (Dosovitskiy et al. 2021), the authors study the idea of "Attention Distance" to investigate how ViT uses self-attention to integrate information across a given image. Specifically,

[3] For details, we refer the reader to the official repository of the background robustness challenge: https://git.io/J3TUj.

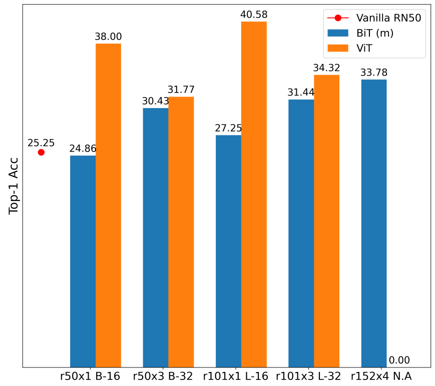

Figure 3: Top-1 accuracy scores (%) on ImageNet-R dataset.

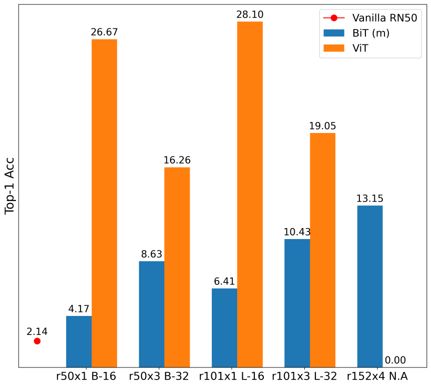

Figure 4: Top-1 accuracy scores (%) on ImageNet-A dataset.

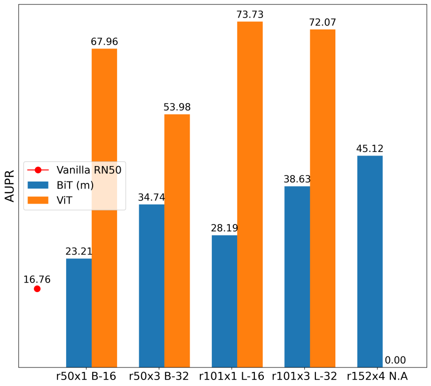

Figure 5: AUPR (higher is better) on ImageNet-O dataset.

Model	**Original**	**Mixed-Same**	**Mixed-Rand**	**BG-Gap**
BiT-m `r101x3`	94.32	81.19	76.62	4.57
ResNet-50	95.6	86.2	78.9	7.3
ViT L-16	96.67	88.49	81.68	6.81

Table 5: Top-1 accuracy (%) of ImageNet-9 dataset and its different variants. ”BG-Gap” is the gap between ”Mixed-Same” and ”Mixed-Rand”. It measures how impactful background correlations are in presence of correct-labeled foregrounds.

Model	**Challenge Accuracy (%)**
BiT-m `r101x3`	3.78
ViT L-16	20.02
ResNet-50	22.3

Table 6: Performance on detecting vulnerable image foregrounds from ImageNet-9 dataset.

they analyze the average distance covered by the learned attention weights from different layers. One key finding is that in the lower layers some attention heads attend to almost the entirety of the image and some heads attend to small regions. This introduces high variability in the attention distance attained by different attention heads, particularly in the lower layers. This variability gets roughly uniform as the depth of the network increases. This capability of building rich relationships between different parts of images is crucial for contextual awareness and is different from how CNNs interpret images as investigated in (Raghu et al. 2021).

Since the attention mechanism helps a model learn better contextual dependencies we hypothesize that this is one of the attributes for the superior performance ViTs show on three robustness benchmark datasets. To this end, we study the performance of different ImageNet-1k models that make use of attention in some form (spatial, channel, or both)[4]. These models include EfficientNetV2 (Tan and Le 2021) with Global Context (GC) blocks (Cao et al. 2020), several ResNet variants with Gather-Excite (GE) blocks (Hu et al. 2018) and Selective Kernels (SK) (Li et al. 2019). We also include a ViT S/16 model pre-trained on ImageNet-1k for a concrete comparison. We summarize our findings in Table 7. The results suggest that adding some form of attention is usually a good design choice especially when robustness aspects are concerned as there is almost always a consistent improvement in performance compared to that of a vanilla ResNet-50. This is also suggested by Hendrycks et al. (Hendrycks et al. 2021) but only in the context of ImageNet-A. We acknowledge that the models reported in Table 7 differ from the corresponding ViT model with respect to their training configurations, regularization in particular. But exploring how regularization affects the robustness aspects of a model is not the question we investigate in this work.

Self-attention constitutes a fundamental block for ViTs. So, in a realistic hope, they should be able to perform even better when they are trained in the right manner to compensate for the absence of strong inductive priors as CNNs. We confirm this in Table 7 (last row). Note that the work on AugReg (Steiner et al. 2021) showed that it is important to incorporate stronger regularization to train better performing ViTs in the absence of inductive priors and larger data regimes. More experiments and attention visualizations showing the connection between attention and robustness are presented in Appendix C.

4.2 Role of Pre-training

ViTs yield excellent transfer performance when they are pre-trained on larger datasets (Dosovitskiy et al. 2021; Steiner et al. 2021). This is why, to better isolate the effects of pre-training with larger data regimes we consider a ViT B/16 model but trained with different configurations and assess their performance on the same benchmark datasets as used in Section 4.1. These configurations primarily differ in terms of the pre-training dataset. We report our findings in the Table 8. We notice that the model pre-trained on ImageNet-1k performs worse than the one pre-trained on ImageNet-21k and then fine-tuned on ImageNet-1k.

Observations from Table 8 lead us to explore another questions i.e., under similar pre-training configurations how do the ViT models stand out with respect to BiT models. This further helps to validate which architectures should be preferred

[4] We used implementations from the `timm` library for this.

Model	**# Parameters (Million)**	**# FLOPS (Million)**	**ImageNet-A (Top-1 Acc)**	**ImageNet-R (Top-1 Acc)**	**ImageNet-O (AUPR)**
ResNet-50	25.6138	4144.854528	2.14	25.25	16.76
EfficientV2 (GC)	13.678	1937.974	7.389285	32.701343	20.34
ResNet-L (GE)	31.078	3501.953	5.1157087	29.905242	21.61
ResNet-M (GE)	21.143	3015.121	4.99335	29.345	22.1
ResNet-S (GE)	8.174	749.538	2.4682036	24.96156	17.74
ResNet18 (SK)	11.958	1820.836	1.802681	22.95351	16.71
ResNet34 (SK)	22.282	3674.5	3.4683768	26.77625	18.03
Wide (4x) ResNet-50 (SK)	27.48	4497.133	6.0972147	28.3357	20.58
ViT S/16	**22**	**4608.338304**	**6.39517**	**26.11397**	**22.50**

Table 7: Complexity and performance of different attention-fused models on three benchmark robustness datasets. All models reported here operate on images of size 224 × 224.

Pre-training	**ImageNet-A (Top-1 Acc)**	**ImageNet-R (Top-1 Acc)**	**ImageNet-O (AUPR)**
ImageNet-1k	8.630994	28.213835	26.25
ImageNet-21k	21.746947	41.815233	54.61

Table 8: Performance of the ViT B/16 model on three benchmark datasets.

Masking Factor	**Top-1 Acc (BiT)**	**Top-1 Acc (ViT)**
0	79	83
0.05	76	82.3
0.1	75	81.4
0.2	72.4	77.9
0.5	52	60.4

Table 9: Mean top-1 accuracy (%) of BiT (`m-r101x3`) and ViT (`L-16`) with different masking factors.

	ResNet-50	**BiT-m r101x3**	**ViT L-16**
P=10	21.8	13.9	6.7
P=25	30.2	14.8	7
P=50	40.4	16.4	7.6
P=90	58.9	23	13.1
P=95	63.6	24.9	15.1

Table 10: Different percentiles (P) of the error matrix computed from Fourier analysis (Figure 6).

for longer pre-training with larger datasets as far as robustness aspects are concerned. This may become an important factor to consider when allocating budgets and resources for large-scale experiments on robustness. Throughout Section 3 and the rest of Section 4, we show that ViT models significantly outperform similar BiT models across six robustness benchmark datasets that we use in this work. We also present additional experiments in Appendix D by comparing ViTs to BiTs of similar parameters.

4.3 ViT Has Better Robustness to Image Masking

In order to further establish that attention indeed plays an important role for the improved robustness of ViTs, we conduct the following experiment:

- Randomly sample a common set of 1000 images from the ImageNet-1k validation set.
- Apply Cutout (DeVries and Taylor 2017) at four different levels: $\{5,10,20,50\}\%$ and calculate the mean top-1 accuracy scores for each of the levels with BiT (`m-r101x3`) and ViT (`L-16`)[5]. In Cutout, square regions from input images are randomly masked out. It was originally proposed as a regularization technique.

Table 9 reports that ViT is able to consistently beat BiT when square portions of the input images have been randomly masked out. Randomness is desirable here because ViT can utilize global information. If we fixate the region of masking it may be too restrictive for a ViT to take advantage of its ability to utilize global information. Note that the ViT variant (`L-16`) we use in this experiment is shallower than the BiT variant (`m-r101x3`). This may suggest that attention indeed is the strong force behind this significant gain.

4.4 Fourier Spectrum Analysis Reveals Low Sensitivity for ViT

A common hypothesis about vision models is that they can easily pick up the spurious correlations present inside input data that may be imperceptible and unintuitive to humans (Jo and Bengio 2017; Hendrycks and Dietterich 2019). To measure how ViT holds up with this end of the bargain, we conduct a Fourier analysis (Yin et al. 2019) of ViT, BiT, and our baseline ResNet-50. The experiment goes as follows:

- Generate a Fourier basis vector with varying frequencies.
- Add the basis vector to 1000 randomly sampled images from the ImageNet-1k validation set.
- Record error-rate for every perturbed image and generate a heatmap of the final error matrix.

For additional details on this experiment, we refer the reader to (Yin et al. 2019). In Figure 6, it is noticed that both ViT and BiT stay robust (have low sensitivity) to most of the regions present inside the perturbed images while the baseline ResNet50V2 loses its consistency in the high-frequency regions. The value at location (i, j) shows the error rate on data perturbed by the corresponding Fourier basis noise.

The low sensitivity of ViT and BiT may be attributed to the following factors: (**a**) Both ViT and BiT are pre-trained on a larger dataset and then fine-tuned on ImageNet-1k. Using a larger dataset during pre-training may be acting as a regularizer here (Kolesnikov et al. 2020). (**b**) Evidence also suggests that increased network width has a positive effect on model robustness (Hendrycks and Dietterich 2019; Hendrycks et al. 2021). To get a deeper insight into the heatmaps shown in Figure 6, in Table 10, we report error-rate percentiles for the three models under consideration. For a more robust model, we should expect to see lower numbers across all the five different percentiles reported in Table 10 and we confirm that is indeed the case. This may also help explain the better behavior of BiT and ViT in this experiment.

4.5 Adversarial Perturbations of ViT Has Wider Spread in Energy Spectrum

In (Ortiz-Jimenez et al. 2020), it is shown that small adversarial perturbations can change the decision boundary of neural networks (especially CNNs) and that adversarial training (Madry et al. 2018) exploits this sensitivity to induce

[5] We use these two variants because they are comparable with respect to the number model parameters.

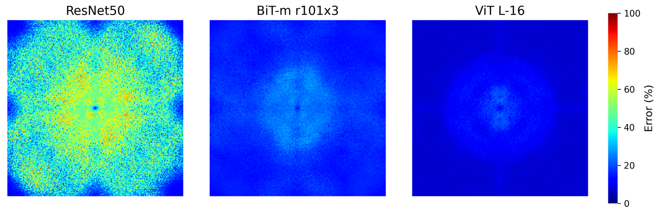

Figure 6: Sensitivity heatmap of 2D discrete Fourier transform spectrum (Yin et al. 2019). The low-frequency/high-frequency components are shifted to the center/corner of the spectrum.

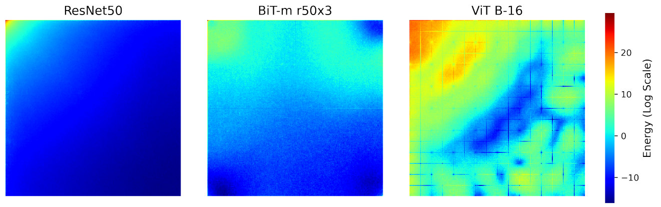

Figure 7: Spectral decomposition of adversarial perturbations generated using DeepFool (Moosavi-Dezfooli, Fawzi, and Frossard 2016). The top-left/bottom-right quadrants denote low-frequency/high-frequency regions.

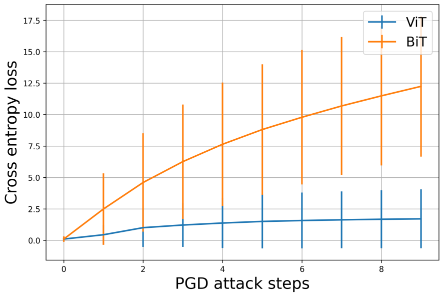

Figure 8: Loss progression (mean and standard deviation) ViT (L-16) and BiT-m (`r101x3`) during PGD attacks (Madry et al. 2018).

robustness. Furthermore, CNNs primarily exploit discriminative features from the low-frequency regions of the input data. Following (Ortiz-Jimenez et al. 2020), we conduct the following experiment on 1000 randomly sampled images from the ImageNet-1k validation set with ResNet-50, BiT-m `r50x3`, and ViT B-16[6]:

- Generate small adversarial perturbations (δ) with DeepFool (Moosavi-Dezfooli, Fawzi, and Frossard 2016) with a step size of 50[7].
- Change the basis of the perturbations with discrete cosine transform (DCT) to compute the energy spectrum of the perturbations.

This experiment aims to confirm that ViT's perturbations will spread out the whole spectrum, while perturbations of ResNet-50 and BiT will be centered only around the low-frequency regions. This is primarily because ViT has the ability to better exploit information that is only available in a global context. Figure 7 shows the energy spectrum analysis. It suggests that to attack ViT, (almost) the entire frequency spectrum needs to be affected, while it is less so for BiT and ResNet-50.

4.6 ViT Has Smoother Loss Landscape to Input Perturbations

One way to attribute the improved robustness of ViT over BiT is to hypothesize ViT has a smoother loss landscape with respect to input perturbations. Here we explore their loss landscapes based on a common set of 100 ImageNet-1k validation images that are correctly classified by both models. We apply the multi-step projected gradient descent (PGD) attack (Madry et al. 2018) with an ℓ_∞ perturbation budget of $\epsilon = 0.002$ when normalizing the pixel value range to be between $[-1, 1]$[8] (refer to Appendix J for details on hyperparameters). Figure 8 shows that the classification loss (cross entropy) of ViT increases at a much slower rate than that of BiT as one varies the attack steps, validating our hypothesis of smoother loss landscape to input perturbations.

In summary, in this section, we broadly verify that ViT can yield improved robustness (even with fewer parameters/FLOPS in some cases). This indicates that the use of Transformers can be orthogonal to the known techniques to improve the robustness of vision models (Balaji, Goldstein, and Hoffman 2019; Carmon et al. 2019; Xie et al. 2020).

5 Conclusion

Robustness is an important aspect to consider when deploying deep learning models into the wild. This work provides a comprehensive robustness performance assessment of ViTs using 6 different ImageNet datasets and concludes that ViT significantly outperforms its CNN counterpart (BiT) and the baseline ResNet50V2 model. We further conducted 6 new experiments to verify our hypotheses of improved robustness in ViT, including the use of large-scale pre-training and attention module, the ability to recognize randomly masked images, the low sensibility to Fourier spectrum domain perturbation, and the property of wider energy distribution and smoother loss landscape under adversarial input perturbations. Our analyses and findings show novel insights toward understanding the source of robustness and can shed new light on robust neural network architecture design.

Acknowledgements

We are thankful to the Google Developers Experts program[9] (specifically Soonson Kwon and Karl Weinmeister) for providing Google Cloud Platform credits to support the experiments. We also thank Justin Gilmer (of Google), Guillermo Ortiz-Jimenez (of EPFL, Switzerland), and Dan Hendrycks (of UC Berkeley) for fruitful discussions.

[6] For computational constraints we used smaller BiT and ViT variants for this experiment.

[7] Rest of the hyperparameters are same as what is specified https://git.io/JEhpG.

[8] We follow the PGD implementation from https://adversarial-ml-tutorial.org/introduction/.

[9] https://developers.google.com/programs/experts/

References

Abnar, S.; and Zuidema, W. 2020. Quantifying Attention Flow in Transformers. In *Annual Meeting of the Association for Computational Linguistics*, 4190–4197.

Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *International Conference on Learning Representations*.

Balaji, Y.; Goldstein, T.; and Hoffman, J. 2019. Instance adaptive adversarial training: Improved accuracy tradeoffs in neural nets. *arXiv preprint arXiv:1910.08051*.

Bhojanapalli, S.; Chakrabarti, A.; Glasner, D.; Li, D.; Unterthiner, T.; and Veit, A. 2021. Understanding Robustness of Transformers for Image Classification. *arXiv preprint arXiv:2103.14586*.

Cao, Y.; Xu, J.; Lin, S.; Wei, F.; and Hu, H. 2020. Global Context Networks. arXiv:2012.13375.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European Conference on Computer Vision*, 213–229. Springer.

Carmon, Y.; Raghunathan, A.; Schmidt, L.; Duchi, J. C.; and Liang, P. S. 2019. Unlabeled Data Improves Adversarial Robustness. In *Advances in Neural Information Processing Systems*, volume 32.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging Properties in Self-Supervised Vision Transformers. *arXiv preprint arXiv:2104.14294*.

Chen, M.; Radford, A.; Child, R.; Wu, J.; Jun, H.; Luan, D.; and Sutskever, I. 2020. Generative Pretraining From Pixels. In *International Conference on Machine Learning*, volume 119, 1691–1703.

Chen, X.; Hsieh, C.-J.; and Gong, B. 2021. When Vision Transformers Outperform ResNets without Pretraining or Strong Data Augmentations. arXiv:2106.01548.

Cubuk, E. D.; Zoph, B.; Shlens, J.; and Le, Q. V. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 3008–3017.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition,*, 248–255.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186.

DeVries, T.; and Taylor, G. W. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Foret, P.; Kleiner, A.; Mobahi, H.; and Neyshabur, B. 2021. Sharpness-aware Minimization for Efficiently Improving Generalization. In *International Conference on Learning Representations*.

Geirhos, R.; Rubisch, P.; Michaelis, C.; Bethge, M.; Wichmann, F. A.; and Brendel, W. 2019. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*.

Hart, B. M.; Schmidt, H. C. E. F.; Klein-Harmeyer, I.; and Einhäuser, W. 2013. Attention in natural scenes: contrast affects rapid visual processing and fixations alike. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1628): 20130067.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Identity Mappings in Deep Residual Networks. In Leibe, B.; Matas, J.; Sebe, N.; and Welling, M., eds., *European Conference on Computer Vision*, 630–645. Springer.

Hendrycks, D.; Basart, S.; Mu, N.; Kadavath, S.; Wang, F.; Dorundo, E.; Desai, R.; Zhu, T.; Parajuli, S.; Guo, M.; Song, D.; Steinhardt, J.; and Gilmer, J. 2020. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. *arXiv preprint arXiv:2006.16241*.

Hendrycks, D.; and Dietterich, T. G. 2019. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *International Conference on Learning Representations*.

Hendrycks, D.; and Gimpel, K. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Hendrycks*, D.; Mu*, N.; Cubuk, E. D.; Zoph, B.; Gilmer, J.; and Lakshminarayanan, B. 2020. AugMix: A Simple Method to Improve Robustness and Uncertainty under Data Shift. In *International Conference on Learning Representations*.

Hendrycks, D.; Zhao, K.; Basart, S.; Steinhardt, J.; and Song, D. 2021. Natural Adversarial Examples. *Conference on Computer Vision and Pattern Recognition*.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. In *NeurIPS Deep Learning and Representation Learning Workshop*.

Hu, J.; Shen, L.; Albanie, S.; Sun, G.; and Vedaldi, A. 2018. Gather-Excite: Exploiting Feature Context in Convolutional Neural Networks. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Huang, G.; Sun, Y.; Liu, Z.; Sedra, D.; and Weinberger, K. Q. 2016. Deep networks with stochastic depth. In *European conference on computer vision*, 646–661. Springer.

Ioffe, S.; and Szegedy, C. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *International Conference on Machine Learning*, volume 37, 448–456.

Jiang, Z.; Hou, Q.; Yuan, L.; Zhou, D.; Jin, X.; Wang, A.; and Feng, J. 2021. Token labeling: Training a 85.5% top-1 accuracy vision transformer with 56m parameters on imagenet. *arXiv preprint arXiv:2104.10858*.

Jo, J.; and Bengio, Y. 2017. Measuring the tendency of cnns to learn surface statistical regularities. *arXiv preprint arXiv:1711.11561*.

Kingma, D.; and Ba, J. 2015. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Kolesnikov, A.; Beyer, L.; Zhai, X.; Puigcerver, J.; Yung, J.; Gelly, S.; and Houlsby, N. 2020. Big Transfer (BiT): General Visual Representation Learning. In *European Conference on Computer Vision*, 491–507.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25.

Li, X.; Wang, W.; Hu, X.; and Yang, J. 2019. Selective Kernel Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 510–519.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. arXiv:2103.14030.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. *International Conference on Learning Representations*.

Mahmood, K.; Mahmood, R.; and Van Dijk, M. 2021. On the Robustness of Vision Transformers to Adversarial Examples. *arXiv preprint arXiv:2104.02610*.

Moosavi-Dezfooli, S.-M.; Fawzi, A.; and Frossard, P. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2574–2582.

Ortiz-Jimenez, G.; Modas, A.; Moosavi, S.-M.; and Frossard, P. 2020. Hold me tight! Influence of discriminative features on deep network boundaries. In *Advances in Neural Information Processing Systems*, volume 33, 2935–2946.

Parmar, N.; Vaswani, A.; Uszkoreit, J.; Kaiser, L.; Shazeer, N.; Ku, A.; and Tran, D. 2018. Image Transformer. In *International Conference on Machine Learning*, volume 80, 4055–4064.

Qiao, S.; Wang, H.; Liu, C.; Shen, W.; and Yuille, A. 2019. Micro-Batch Training with Batch-Channel Normalization and Weight Standardization. *arXiv preprint arXiv:1903.10520*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020.

Radosavovic, I.; Kosaraju, R.; Girshick, R.; He, K.; and Dollar, P. 2020. Designing Network Design Spaces. In *IEEE Conference on Computer Vision and Pattern Recognition*, 10425–10433.

Raghu, M.; Unterthiner, T.; Kornblith, S.; Zhang, C.; and Dosovitskiy, A. 2021. Do Vision Transformers See Like Convolutional Neural Networks? arXiv:2108.08810.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3): 211–252.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *IEEE International Conference on Computer Vision*, 618–626.

Shao, R.; Shi, Z.; Yi, J.; Chen, P.-Y.; and Hsieh, C.-J. 2021. On the Adversarial Robustness of Visual Transformers. *arXiv preprint arXiv:2103.15670*.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(56): 1929–1958.

Steiner, A.; Kolesnikov, A.; Zhai, X.; Wightman, R.; Uszkoreit, J.; and Beyer, L. 2021. How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers. arXiv:2106.10270.

Sun, C.; Shrivastava, A.; Singh, S.; and Gupta, A. 2017. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. In *IEEE International Conference on Computer Vision*, 843–852.

Tan, M.; and Le, Q. 2021. EfficientNetV2: Smaller Models and Faster Training. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 10096–10106. PMLR.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2020. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*.

Trinh, T. H.; Luong, M.-T.; and Le, Q. V. 2019. Selfie: Self-supervised pretraining for image embedding. *arXiv preprint arXiv:1906.02940*.

Tuli, S.; Dasgupta, I.; Grant, E.; and Griffiths, T. L. 2021. Are Convolutional Neural Networks or Transformers more like human vision? Accepted at CogSci 2021.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30.

Wu, Y.; and He, K. 2018. Group Normalization. In *European Conference on Computer Vision*, 3–19.

Xiao, K.; Engstrom, L.; Ilyas, A.; and Madry, A. 2021. Noise or Signal: The Role of Image Backgrounds in Object Recognition. *International Conference on Learning Representations*.

Xie, Q.; Luong, M.-T.; Hovy, E.; and Le, Q. V. 2020. Self-Training With Noisy Student Improves ImageNet Classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Yin, D.; Gontijo Lopes, R.; Shlens, J.; Cubuk, E. D.; and Gilmer, J. 2019. A Fourier Perspective on Model Robustness in Computer Vision. In *Advances in Neural Information Processing Systems*, volume 32.

Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Tay, F. E.; Feng, J.; and Yan, S. 2021. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv preprint arXiv:2101.11986*.

Zagoruyko, S.; and Komodakis, N. 2016. Wide residual networks. *arXiv preprint arXiv:1605.07146*.