Revealing the Dark Secrets of Extremely Large Kernel ConvNets on Robustness

Honghao Chen [1 2]   Yurong Zhang [3]   Xiaokun Feng [1 2]   Xiangxiang Chu [4]   Kaiqi Huang [1 2]

Abstract

Robustness is a vital aspect to consider when deploying deep learning models into the wild. Numerous studies have been dedicated to the study of the robustness of vision transformers (ViTs), which have dominated as the mainstream backbone choice for vision tasks since the dawn of 2020s. Recently, some large kernel convnets make a comeback with impressive performance and efficiency. However, it still remains unclear whether large kernel networks are robust and the attribution of their robustness. In this paper, we first conduct a comprehensive evaluation of large kernel convnets' robustness and their differences from typical small kernel counterparts and ViTs on six diverse robustness benchmark datasets. Then to analyze the underlying factors behind their strong robustness, we design experiments from both quantitative and qualitative perspectives to reveal large kernel convnets' intriguing properties that are completely different from typical convnets. Our experiments demonstrate for the first time that pure CNNs can achieve exceptional robustness comparable or even superior to that of ViTs. Our analysis on occlusion invariance, kernel attention patterns and frequency characteristics provide novel insights into the source of robustness. Code available at: https://github.com/Lauch1ng/LKRobust.

1. Introduction

Over the past decade, the evolution of deep learning has been tightly linked with Convolutional Neural Networks (CNNs) (LeCun et al., 1998; Krizhevsky et al., 2012; Simonyan & Zisserman, 2014; He et al., 2016), which have

played a crucial role in propelling the advancement of artificial intelligence. Nonetheless, with the dawn of the 2020s, the advent of Vision Transformers (ViTs) (Dosovitskiy et al., 2020; Liu et al., 2021; Touvron et al., 2021; Wang et al., 2021) has profoundly shaken the prevailing hegemony of CNNs. Facilitated by the self-attention mechanism, ViTs have manifested state-of-the-art performance in a variety of vision tasks such as image classification (Dosovitskiy et al., 2020; Liu et al., 2021), object detection (Carion et al., 2020; Dai et al., 2021; Zhu et al., 2020; Meng et al., 2021), semantic segmentation (Cheng et al., 2021; 2022) and self-supervised learning (Bao et al., 2021; He et al., 2022; Chen et al., 2024b), demonstrating their exceptional properties as the choice of foundation models.

As a versatile and widely used backbone, a key attribute of ViT is its strong robustness. Robustness is a vital aspect to consider when deploying deep learning models into the wild (Paul & Chen, 2022). Specifically, for safety-critical applications such as autonomous vehicles, robots, and healthcare, the learned representations must be robust, and without strong robustness the models can not be practically implemented (Naseer et al., 2021). Numerous studies have been dedicated to the investigation on evaluating the robustness of ViTs (Mahmood et al., 2021; Naseer et al., 2021; Paul & Chen, 2022; Shao et al., 2021; Bhojanapalli et al., 2021). They demonstrate that ViTs possess universal and strong robustness, which excel on various robust datasets (Hendrycks et al., 2021b; Hendrycks & Dietterich, 2019; Hendrycks et al., 2021a; Xiao et al., 2020) and exhibit clear superiority over typical CNNs in dealing with particular disturbances, such as image occlusion and adversarial attack (Naseer et al., 2021). The strong robustness paves the way for ViTs' extensive application across diverse real-world scenarios.

On the other hand, concerted efforts are dedicated to the revival of CNNs. Notably, recent advances (Liu et al., 2022b; Ding et al., 2022; Liu et al., 2022a; Ding et al., 2023; Chen et al., 2024a) in revisiting large kernel designs have revealed that when equipped with large kernel size (e.g., 31×31), pure CNN architecture can perform on par with or even better than state-of-the-art ViTs. Large kernel convolution can substantially increase the model's effective receptive field (Ding et al., 2022), enabling large kernel convnets to have remarkable performances across various vision tasks (Liu et al., 2022b; Woo et al., 2023; Ding et al.,

2022; Liu et al., 2022a; Chen et al., 2023; Lu et al., 2023; Huang et al., 2023), such as classification, segmentation and detection, etc. Nevertheless, it still remains unclear whether large kernel networks are robust and the attribution of their robustness, which are of vital importance and could significantly impact their practical application and development. It is natural to ask: *are large kernel networks inherently robust? How do they differ in terms of robustness from typical CNNs and ViTs?* And if they are robust, *what contributes to their robustness?* In this paper, we delve into the above questions, providing empirical evidence and novel insights to reason about the robustness of large kernel convnets.

To answer the first two questions, we conduct thorough experiments on six diverse and widely-used robustness benchmark datasets. These datasets comprehensively evaluate model robustness from multiple perspectives including natural adversarial (Hendrycks et al., 2021b), common corruptions (Hendrycks & Dietterich, 2019), semantic shifts (Hendrycks et al., 2021a), out-of-domain distribution (Hendrycks et al., 2021b), common perturbations (Hendrycks & Dietterich, 2019) and background dependency (Xiao et al., 2020). In stark contrast to traditional CNNs, large kernel networks exhibit exceptionally strong robustness, comparable or even superior to ViT.

To comprehend the underlying factors for the superior robustness of large kernel networks, we systematically designed nine experiments to provide insights into their robustness from both quantitative and qualitative perspectives. These experiments and visualizations unveil intriguing properties of large kernel convnets, such as occlusion invariance, kernel attention patterns and frequency characteristics, which help understand the reasons behind their robustness. Our contributions can be summarized as follows:

- We explore the robustness of large kernel networks across six ImageNet datasets concerning different types of robustness evaluation and conduct a comprehensive comparison with typical CNNs and ViT. We demonstrate that large kernel convnets' robustness significantly differs from typical CNNs and is comparable or even superior to that of ViT.

- We devise nine experiments to investigate the robustness of large kernel convnets from both quantitative and qualitative perspectives, spreading occlusion invariance, adversarial attack, model perturbations, frequency characteristics and kernel attention pattern, etc.

- Our analysis provides insights into the robustness of large kernel convnets, promoting their application and broader development. Moreover, our study is the first to reveal that pure CNNs can achieve comparable or even superior robustness to ViT, indicating that self-attention is not the only route to strong robustness and making a further step to the revival of CNNs.

2. Related Work

2.1. Large Kernel ConvNets

Large kernel convolutional networks can be traced back to a few traditional models from the early stages of deep learning (Krizhevsky et al., 2012; Szegedy et al., 2015; 2016). However, it was the rise of small convolutional networks, represented by VGG-Net (Simonyan & Zisserman, 2014) and ResNet (He et al., 2016), that facilitated the success of CNNs. And ever since then, a stack of small kernels (e.g., 1×1 or 3×3) became the mainstream choice for convnet design. Large kernel convolution has received little attention and has even been found to be detrimental to Imagenet performance (Peng et al., 2017). Recently, works represented by ConvNeXt (Liu et al., 2022b) and RepLKNet (Ding et al., 2022) have sparked a revival of large kernel convolutional networks (Trockman & Kolter, 2022; Liu et al., 2022b; Woo et al., 2023; Ding et al., 2022; Liu et al., 2022a; Chen et al., 2024a). They demonstrate that large kernels can effectively enhance model performance, especially in downstream tasks (Ding et al., 2022). Large kernel design has also been introduced into other fields, such as 3D backbone (Chen et al., 2023; Lu et al., 2023), continuous convolutions (Romero et al., 2021b;a), knowledge distillation (Huang et al., 2023) and multi-modal learning (Ding et al., 2023). Despite these advances, what remains largely unexplored is their robustness evaluation and attribution, which could significantly impact their practical application and development. Different from previous works, this paper investigates the robustness of large kernel networks, designing experiments to analyze their intriguing properties from multiple perspectives and making a further step for large kernel's revival.

2.2. Robustness for ViTs

As a versatile backbone, ViTs (Dosovitskiy et al., 2020; Liu et al., 2021) have demonstrated impressive performance across a variety of vision tasks (Dosovitskiy et al., 2020; Liu et al., 2021; Dai et al., 2021; Zhu et al., 2020; Meng et al., 2021; Cheng et al., 2021; 2022). A key advantage of ViT over CNN is its strong robustness, which paves the way for ViTs' extensive application across diverse real-world scenarios. Many works seek to study the robustness of ViTs from different perspectives. Compared with ResNet (He et al., 2016), Bhojanapalli *et al.* (Bhojanapalli et al., 2021) studied improved robustness of ViTs when evaluated on adversarial and natural adversarial examples. Shao *et al.* (Shao et al., 2021) revealed that ViTs have better adversarial robustness over CNNs, attributed largely to their ability to learn highly generalizable high-frequency features, while convolutional layers appear to hinder. Paul *et al.* (Paul & Chen, 2022) further integrates and expands the scope of robustness analysis to elucidate the underlying factors behind Vits' superior
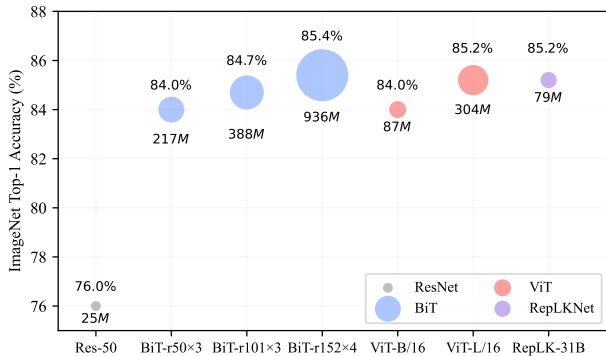
Figure 1. **Model configurations.** We depict the model size and corresponding ImageNet-1k top-1 accuracy. We choose models with similar accuracy and parameter counts (except for resnet-50 as a baseline). All the reported variants were pre-trained on ImageNet-21k and then fine-tuned on ImageNet-1k.

robustness. Raghu *et al.* (Raghu et al., 2021) conducted a detailed analysis of the differences between ViT and CNNs from various perspectives, providing insightful understanding of their robustness disparities. Through comparisons with canonical CNNs, these studies demonstrate ViTs are generally more robust while also inspiring further work to improve the robustness of ViTs (Mao et al., 2022; Zhou et al., 2022; Guo et al., 2023). Building upon and different from previous robustness analysis works for ViTs, this work focuses on large kernel networks (Trockman & Kolter, 2022; Liu et al., 2022b; Woo et al., 2023; Ding et al., 2022; Liu et al., 2022a), which have demonstrated competitive performance with ViTs on mainstream tasks and attracted considerable interest recently. We design systematical experiments to evaluate their robustness and provide insights to analyze the underlying factors.

3. Are Large Kernel ConvNets Robust ?

In this section, we delve into the first two question: whether large kernel convnets are robust learners and how do they differ in terms of robustness from typical CNNs and ViTs. Specifically, we conduct an empirical study comparing large kernel convnets with their typical small kernel counterparts and ViT across six widely accepted robustness datasets. Section 3.1 provides a detailed introduction to the models' configurations, while section 3.2 presents and discusses the datasets and corresponding results.

3.1. Model Configuration

We choose the representative work of large kernel convnets , RepLKNet (Ding et al., 2022), as the primary model for our experiments. RepLKNet is the first work to scale up kernel size to extremely large (i.e., 31×31) and has been applied to various vision tasks (Liu et al., 2022a; Chen et al., 2023; Lu

et al., 2023; Huang et al., 2023). For comparison, we select ViT (Dosovitskiy et al., 2020) as a strong baseline. ViT has been proven to possess significant robustness by numerous studies (Naseer et al., 2021; Paul & Chen, 2022; Shao et al., 2021; Bhojanapalli et al., 2021), demonstrating a clear advantage over traditional convnets. For the comparison with typical convnets, we choose Big Transfer (BiT) (Kolesnikov et al., 2020), which has exceptional performance not only on ImageNet but also on various transfer learning scenarios (Lin et al., 2014; Zhai et al., 2019). Since RepLKNet, BiT and ViT share similar pre-training strategies (such as using larger datasets like ImageNet-21K (Deng et al., 2009), extended pre-training schedules, and so on), they serve as excellent candidates for our comparison purposes. In addition, we add ResNet-50 (He et al., 2016) as a basic baseline. We choose models with similar accuracy and parameter counts (except for resnet-50 as a baseline). The parameter counts and top-1 accuracy on ImageNet-1K of different models are shown in Fig. 1. Note that all the reported variants were initially pre-trained on ImageNet-21K and then fine-tuned on ImageNet-1K.

3.2. Robustness Evaluation on Diverse Datasets

Next, we evaluate the performance of above model variants on six robustness benchmark datasets. These datasets assess the models' robustness from multiple dimensions including: **i)** natural adversarial; **ii)** common corruptions; **iii)** out-of-domain distribution; **iv)** common perturbations; **v)** semantic shifts; **vi)** background dependency. Their specific objectives and venues are summarized in Appendix A.

ImageNet-A (Hendrycks et al., 2021b) is a dataset of real-world adversarially filtered images that fool current ImageNet classifiers. In Fig. 2a, we report the top-1 accuracy of ResNet, BiT, ViT and RepLKNet on the ImageNet-A dataset. In comparison to typical convnets, RepLKNet is shown to outperform ResNet and BiT by large margins. For instance, the top-1 accuracy of RepLKNet-31B is $2.8\times$ higher than BiT-m r101$\times$3. It even slightly surpasses ViT-L, which has a much larger model size, indicating that large kernel convnets possess strong robustness against natural adversarial challenges.

ImageNet-C (Hendrycks & Dietterich, 2019) comprises 15 categories of algorithmically generated corruptions and an additional four general corruption types, resulting in a total of 19 corruption categories. Each corruption type has five severity levels, ranging from negligible to pulverizing. We evaluate all 19 corruptions at their maximum severity level (5) and depict the mean top-1 accuracy in Fig. 2b. We observed that RepLKNet performs on par with ViT-L and surpasses all typical convolutional networks and ViT-B/16. Considering the notably higher parameter count of ViT-L compared to RepLKNet, this further validates the robustness
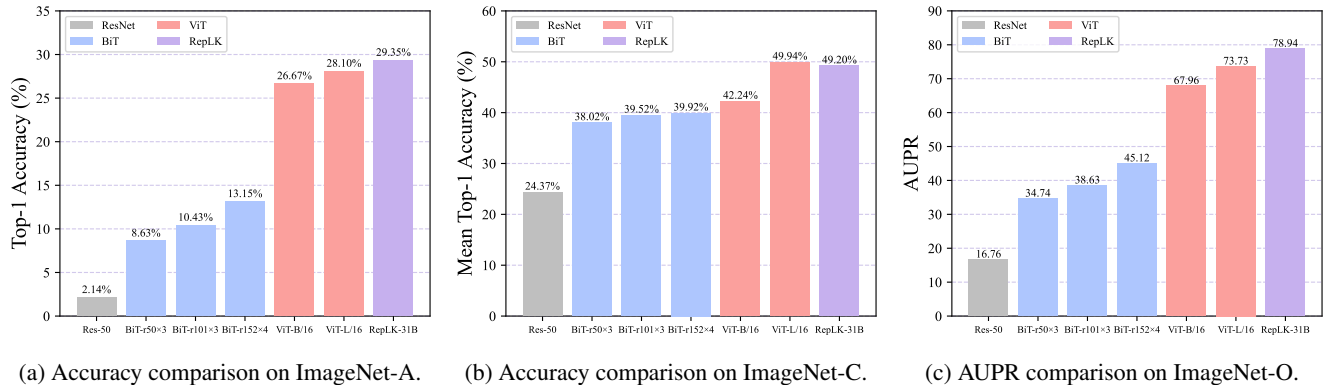
(a) Accuracy comparison on ImageNet-A.

(b) Accuracy comparison on ImageNet-C.

(c) AUPR comparison on ImageNet-O.

Figure 2. **Comparison on ImageNet-A, ImageNet-C, and ImageNet-O.** For ImageNet-A, we report top-1 accuracy; For ImageNet-C, we report mean top-1 accuracy over all the 19 corruptions; For ImageNet-O, we report *area under the precision-recall curve* (AUPR). Note that for all the metrics higher is better. RepLKNet-31B performs on par with or even better than ViT-L on ImageNet-A and ImageNet-O, demonstrating the strong robustness of large kernel convnets.
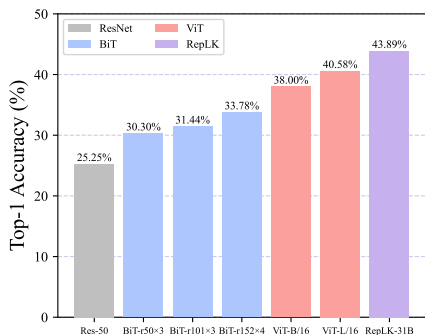

Figure 3. **Accuracy comparison on ImageNet-R.** We report top-1 accuracy for all the model variants. RepLKNet-31B still outperforms ViT-B/L by large margins.

of large kernel convnets.

Moreover, (Hendrycks & Dietterich, 2019) propose mean corruption error (mCE) to quantify the robustness factors of a model on ImageNet-C. We follow the same evaluation approach and report mCE comparison in Table 1. We additionally add DeepAugment (Hendrycks et al., 2021a) and AugMix (Hendrycks et al., 2020), which are specifically aimed at enhancing the model's robustness against corruptions observed in ImageNet-C. Surprisingly, RepLKNet outperforms other methods with a clear gap.

ImageNet-O (Hendrycks et al., 2021b) is a dataset of adversarially filtered examples for ImageNet out-of-distribution detectors. Following (Hendrycks et al., 2021b), we use area under the precision-recall curve (AUPR) as the evaluation metric for ImageNet-O. As shown in Fig. 2c, RepLKNet outperforms all the other model variants by clear margins, demonstrating large kernel convnets' superior robustness in anomaly detection.

ImageNet-P (Hendrycks & Dietterich, 2019) consists of 10 types of common perturbations. ImageNet-P differs from ImageNet-C in that it generates perturbation sequences from each ImageNet validation image. The perturbations are subtly nuanced, affecting a smaller number of pixels within the images. We follow (Hendrycks & Dietterich, 2019) to use mean flip rate (mFR) and mean top-5 distance (mT5D) as the standard metrics to evaluate models' robustness. For brevity, we omit the detailed formulation of mFR and mT5D here. As shown in Table 2, the robustness of RepLKNet has once again been confirmed to be better than BiT, ViT and AugMix as well.

ImageNet-R (Hendrycks et al., 2021a) contains various artistic renditions of 200 classes from the ImageNet-1K dataset. We report top-1 accuracy comparison on ImageNet-R in Fig. 3, exhibiting that RepLKNet's robustness to domain adaptation is better than that of BiT and ViT.

ImageNet-9 (Xiao et al., 2020) helps disentangle the impacts of foreground and background signals on classification. It measures the model's robustness towards background changes. As shown in Table 3, we report four metrics to assess the model's robustness to background changes. *Origin* refers to the original accuracy without modifying background. *Mixed-Same* involves replacing the original background with a random one from **the same** class. *Mixed-Rand* means replacing the original background with a random background from **a random** class. *BG-Gap = Mixed-Same - Mixed-Rand*, which measures how impactful background correlations are when correct-labeled foregrounds exist. The results reveals that RepLKNet surpasses BiT and ViT consistently on the first three metrics. Additionally, the BG-Gap for RepLKNet is much smaller, this suggests that large kernel convnets are less sensitive to background modifications, exhibiting stronger background robustness.

Table 1. mCEs (%) of different models and methods on ImageNet-C (lower is better).

Model / Method	mCE
ResNet-50	76.7
BiT m-r101×3	58.3
DeepAugment + AugMix	53.6
ViT-L/16	45.5
RepLKNet-31B	36.5

Table 2. mFRs (%) and mT5Ds (%) comparison on ImageNet-P dataset (lower is better).

Model / Method	mFR	mT5D
ResNet-50	58.0	82.0
BiT m-r101×3	50.0	76.7
AugMix	37.4	N/A
ViT-L/16	33.1	50.2
RepLKNet-31B	29.2	49.8

Table 3. Top-1 acc comparison on ImageNet-9. BG-Gap=Mixed-Same - Mixed-Rand. It measures how impactful background correlations are when correct-labeled foregrounds exist.

Model	Origin	Mixed-Same	Mixed-Rand	BG-Gap
BiT m-r101×3	94.3	81.2	76.6	4.6
ResNet-50	95.6	86.2	78.9	7.3
ViT-L/16	96.7	88.5	81.7	6.8
RepLKNet-31B	97.3	92.3	88.2	4.1

(a) Random patch-drop.
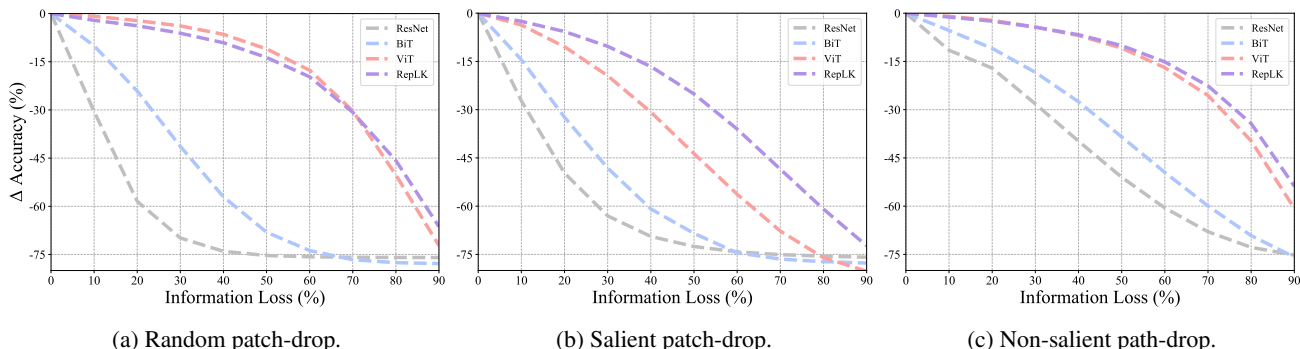(b) Salient patch-drop.
(c) Non-salient path-drop.

Figure 4. **Occlusion robustness comparison under different settings.** We use random drop, salient drop and non-salient drop to evaluate corresponding occlusion robustness. We report top-1 accuracy drop under different information loss ratios, ranging from 10% to 90%. RepLKNet are more robust to extreme occlusion scenarios and more importantly, they outperform ViT for salient occlusion remarkably.

4. Why are Large Kernel ConvNets Robust?

In this section, we thoroughly investigate the robustness of large kernel convnet from both quantitative and qualitative perspectives. To be specific, we systematically design nine distinctive experiments to conduct a comprehensive and in-depth analysis of its key properties like occlusion-invariance and kernel attention patterns, providing multi-dimensional insights into its strong robustness.

4.1. Occlusion Invariance

Occulusion Invariance matters in science and deep learning (Kosmann-Schwarzbach et al., 2011; Kong & Zhang, 2023). It plays a crucial role in human vision's robustness: for instance, we can imagine the occluded parts based on texture or shape even when most of the object is invisible, maintaining consistent judgement under high occlusion ratios. To this end, we delve into the occlusion invariance of large kernel networks under various scenarios, where some or most of the image content is missing.

Occlusion modeling. Formally, we define occlusion in a patch-manner: we firstly divide the input image $X \in \mathbb{R}^{H \times W \times C}$ into non-overlapping flattened patches $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N]$, where $\mathbf{x}_i \in \mathbb{R}^{P^2 C}$ according to the patch size P, and $N = (H \times W)/P^2$ is the number of patches. We choose a subset of the total image patches, $M < N$, and set pixel values of these patches to zero to generate an occluded

Figure 5. **Illustration of patch drop.** We depict an example image of different occlusion types: random, salient and non-salient. The pixel values in the occluded (black) areas are assigned to be zero.

image, denoted as $\mathbf{x}'$. We feed $\mathbf{x}'$ into different models and measure their occlusion invariance by the top-1 accuracy on $\mathbf{x}'$. We conduct experiments with three occlusion types: i) random drop; ii) salient (foreground) drop; 3) non-salient (background) drop. An illustration is provided in Fig. 5.

Random drop. Following ViT (Dosovitskiy et al., 2020), we partition images into 16×16 patches, then for a typical image with resolution 224×224, there will be 196 patches. This design makes a convenient and fair comparison with ViT without altering its input format. We randomly mask different proportions of the image, ranging from 10% to 90%. As shown in Fig. 4a, typical convnets suffer catastrophic degradation when the occlusion ratio reaches 50%, whereas RepLKNet and ViT exhibit clear advantages. Moreover, as the occlusion ratio becomes extremely high, RepLKNet gradually outperforms ViT, indicating that large kernel convnets are more robust to extreme occlusion scenarios.
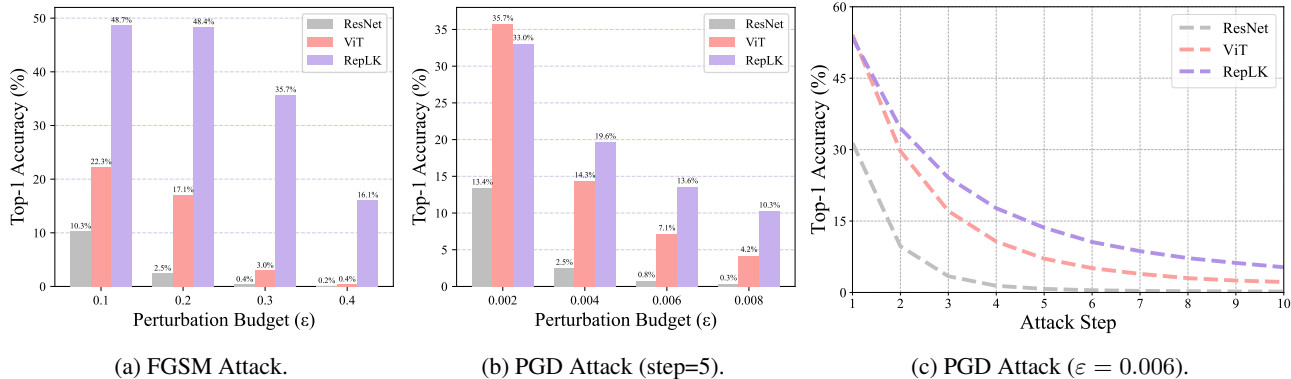
(a) FGSM Attack. (b) PGD Attack (step=5). (c) PGD Attack ($\varepsilon = 0.006$).

Figure 6. **Comparison under different adversarial attacks.** (a) For FGSM attack, we apply with increasing perturbation budget ε. (b) For PGD attack, we fix the attack step to 5 while varying the perturbation budget. (c) For PGD attack, we fix ε to 0.006, then gradually increase the attack step. RepLKNet consistently exceeds both ViT and ResNet.

Salient (foreground) drop. Foreground objects play a decisive role in visual recognition, hence investigating robustness against foreground occlusion is of vital importance. Firstly, we use a pretrained DINO (Caron et al., 2021) to detect objects and identify salient regions. Then we mask a subset of patches encompassing the top K% of foreground information. It's important to note that this K% doesn't always equate to the pixel percentage. For instance, an image's 70% foreground data might be contained in just 20% pixels. As shown in Fig. 4b, RepLKNet exhibits remarkable robustness to salient occlusion, even surpassing ViT noticeably. We reckon this is the key reason for its substantial advantage over ViT on background-dependency dataset (i.e., ImageNet-9).

Non-salient (background) drop. Contrary to the above salient setting, we further select patches containing the lowest K% of foreground information and mask them. Fig. 4c shows that RepLKNet also has superior robustness to non-salient occlusion.

4.2. Robustness to Adversarial Attack

Previous studies reveal that even minor adversarial perturbations can substantially alter the decision boundary of neural networks (Ortiz-Jimenez et al., 2020) and investigating adversarial attacks plays a crucial role in enhancing the robustness of models (Madry et al., 2017a). Hence in this subsection, we step further into the performance of large kernel convnets against adversarial attacks.

We use both single-step and multi-step sample specific attacks, Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2014) and Projected Gradient Attack (PGD) (Madry et al., 2017b) respectively. Note that we only compare RepLKNet with ResNet and ViT for this experiment and omit BiT here as it continues to exhibit performance between ResNet and ViT, similar to previous sections' observations.

Table 4. **Comparison under TAIG attack.** We increase the perturbation budget ϵ from 0.03 to 0.1. The attack success rate is reported, note that lower is better.

Model	$\epsilon = 0.03$	$\epsilon = 0.06$	$\epsilon = 0.1$
ResNet-50	47.7	65.2	67.9
BiT-r152×4	44.6	60.2	63.1
ViT-B	62.3	73.3	75.6
ViT-L	54.9	67.9	69.8
RepLKNet-31B	**39.2**	**56.1**	**59.3**

We conduct this experiment with three variants. **i)** For FGSM attack, we apply with an l_∞ perturbation of increasing budget ε, ranging from 0.1 to 0.4. As shown in Fig. 6a, ViT and ResNet becomes comletely destroyed when the budget grows to 0.3, while RepLKNet still holds a 35.7% top-1 accuracy; **ii)** For PGD attack, we first fix the attack step to be 5, varing the perturbation budget ε from 0.002 to 0.008. Results in Fig. 6b exhibit that when the adversarial attacks are strong, RepLKNet outperforms ViT by clear margins; **iii)** Then for PGD attack, we fix the perturbation budget ε to be 0.006, and we gradually increase the step to observe the robustness trend of different models in Fig. 6c. Similarly, RepLKNet consistently exceeds both ViT and ResNet. Note that for all the experiments, we normalize pixel values using default ImageNet mean and standard deviation. The above experiments across three settings point to the same conclusion: large kernel network possesses broader and stronger adversarial robustness, and this could potentially explain why RepLKNet performs better on natural adversarial dataset (i.e., ImageNet-A).

In addition to the classic adversarial attacks FGSM and PGD, we further evaluate with a modern strong adversarial attack TAIG (Huang & Kong, 2022). Specifically, for convnets (i.e., ResNet, BiT, RepLKNet), ViT-B is selected as the surrogate model to generate adversarial examples, for ViTs, RepLKNet-31B is selected as the surrogate model. We
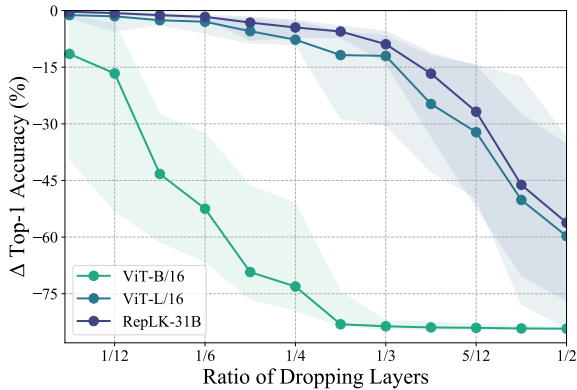
Figure 7. **Lesion study.** Evaluation of RepLKNet and ViT-B/L when $n(n \geq 1)$ blocks are removed from the model after training. For each n, results are from 10 independent samples of n blocks and we show the average accuracy (line) and min/max (shaded area) across samples.

Figure 8. **Robustness to noise frequency.** We measure the decrease in accuracy against frequency-based random noise. ResNet is susceptible to high-frequency noise, while ViT is more affected by low-frequency noise. In a sharp contrast, RepLKNet maintains consistent robustness against noise across all frequencies.

increase the ϵ from 0.03 to 0.1 to gradually increase the attack budget. We report the attack success rate (lower is better). As shown in Table 4, large kernel network still behaves better than typical small kernel convnets and ViTs, showing its inherent robustness against adversarial attacks.

4.3. Robustness to Model Perturbations

Previous studies (Greff et al., 2016; Veit et al., 2016; Bhojanapalli et al., 2021) have shown that layers in residual networks exhibit a large amount of redundancy, and that almost any individual layer can be removed after training without hurting performance. On the other hand, it is observed that identity shortcut is of vital importance especially for networks with very large kernels (Ding et al., 2022). To comprehend the information flow in large kernel convnets, we conduct a lesion study where we remove several blocks from an already trained network during inference, such that information has to flow through the skip connection.

Specifically, we randomly remove $n(n \geq 1)$ blocks and report the changes in top-1 accuracy. For each n, results are from 10 independent samples of n blocks and we show the average accuracy (line) and min/max (shaded area) across samples. Given that the number of blocks in ViT-B is only half of that in RepLKNet and ViT-L, we use the proportion of blocks dropped as the x-axis in Fig. 7. We omit ResNet here as it is much more fragile than ViT-B.

We observe that ViT-B is highly sensitive to layer removal, with the model's accuracy nearing zero after removing just $\frac{1}{3}$ of the blocks. In contrast, RepLKNet and ViT-L can still maintain an accuracy of around 75% at that dropping ratio. Moreover, we notice that despite having a smaller capacity, RepLKNet consistently outperforms ViT-L in terms of accuracy under the same block dropping ratios. This sug-
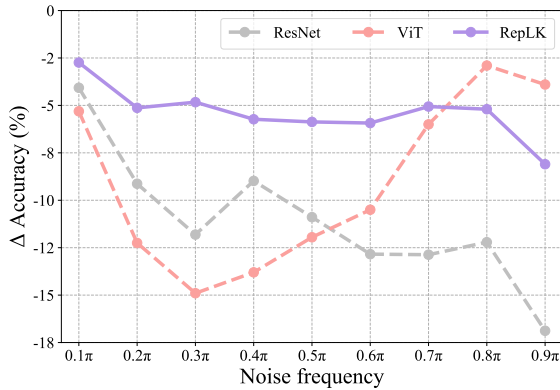
gests that large kernel networks are more robust to model perturbations, which may explain their robustness against common perturbations (i.e., ImageNet-P).

4.4. Robustness to Noise Frequency

To deepen the understanding of large kernel convnet's strong robustness, we further analyze its robustness in frequency-domain. In particular, we subject the model to random noise attacks at varying frequencies and evaluate the accuracy drop. We normalize the frequency to be between 0.0π (center) and 1.0π (boundary). We use a frequency window size of 0.1π for frequency-based noise.

As shown in Fig. 8, ResNet is highly susceptible to high-frequency noise, while ViT exhibits poorer performance against low-frequency noise. This phenomenon aligns well with previous study (Park & Kim, 2022). In contrast, RepLKNet consistently demonstrates robustness against noise across all frequency bands. For instance, for noise across 0.1π to 0.8π frequency range, RepLKNet consistently maintains the accuracy loss within 6%, consistently outperforming ResNet and ViT.

4.5. Large Kernel Size is the Key

Since RepLKNet, BiT, and ViT share similar pre-training strategies, the primary distinction between them lies in the use of large kernel convolutions. Therefore, we proceed to explore the impact of kernel size on robustness. Specifically, we replace all large kernel convolutions in RepLKNet-31B with 3x3 small kernel convolutions (the same as kernel size used in BiT and ResNet), while maintaining the same data augmentation and training schedule. We train the modified model, RepLKNet-3B, and evaluate its robustness across multiple above-mentioned tasks. As shown in Table 5, re-

Table 5. **Ablation on kernel size.** We replace all large kernel convolutions in RepLKNet-31B with 3x3 small kernels. The metrics are the same as Section 3&4. Reducing the kernel size from 31 to 3 significantly degrades the robustness of RepLKNet across various metrics.

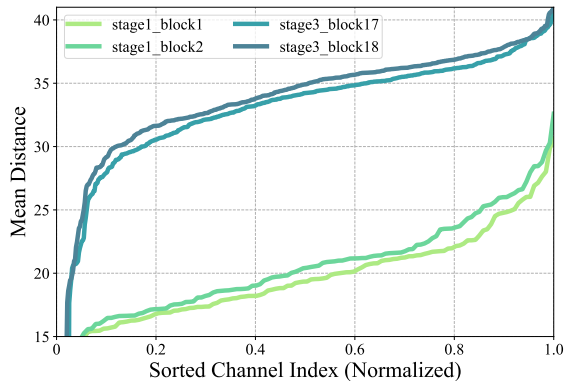| Model	ImageNet-A	ImageNet-C	ImageNet-R	ImageNet-O	Salient-Drop-50%	Noise-0.7π
RepLKNet-31B	29.4 (+2.7)	49.2 (+7.0)	43.9 (+5.9)	78.9 (+10.9)	-25.1 (+19.3)	-5.0 (+1.6)
ViT-B	26.7 (− −)	42.2 (− −)	38.0 (− −)	68.0 (− −)	-44.4 (− −)	-6.6 (− −)
RepLKNet-3B	14.6 (-12.1)	39.1 (-3.1)	32.9 (-5.1)	51.6 (-16.4)	-61.8 (-17.4)	-10.6 (-4.0)

Figure 9. **Kernel attention distance of different layers.** We calculate RepLKNet's sorted kernel attention distance in the first two layers of stage-1 and the last two layers of stage-3. It also tends to aggregate both local and global information at shallow layers, while focusing more on global information at deeper layers.
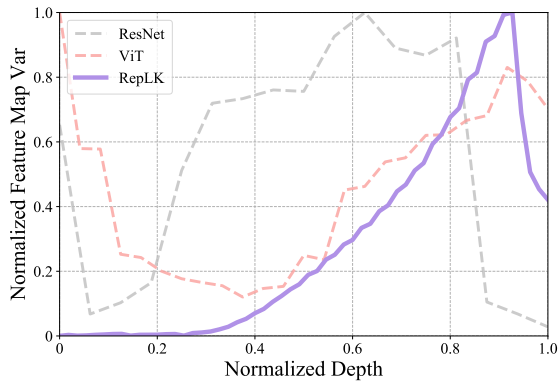

Figure 10. **Feature map variance change curve.** We measure the normalized feature map variance given the same input. Compared with ResNet and ViT, RepLKNet has much lower feature map variance in the very early layers and the overall curve changes in a simpler and more stable manner.

ducing the kernel size from 31 to 3 significantly degrades the robustness of RepLKNet across various metrics, leading to inferior performance compared to ViT-B. This finding highlights the critical role of large kernel convolution in enhancing model robustness.

4.6. Local and Global Kernel Attention

Then we dive into the kernel attention pattern to explore the property of large kernels. Facilitated by the self-attention mechanism, ViT can aggregate global information through the global interactions between all tokens. However, global attention is not always optimal. Raghu *et al.* (Raghu et al., 2021) revealed that powerful ViTs consider both local and global information at shallow layers, mainly focusing on global information at deeper layers. This is in contrast to typical CNNs (Simonyan & Zisserman, 2014; He et al., 2016), which are hardcoded to attend only locally in all layers. Since large kernel convnets have large receptive field comparable to ViTs (Ding et al., 2022; Liu et al., 2022a), we conjecture that they may exhibit similar attention patterns with ViT, attending to both local and global information at shallow layers while mainly global in deep layers.

To verify, we calculate the average kernel attention distance of different layers in RepLKNet. Specifically, for each

kernel, we measure the Euclidean distance from every kernel position to the center of the kernel, we weight this distance by the absolute value of the corresponding parameter, and then we normalize the sum of all parameters to get this kernel's attention distance. Since each kernel has multiple channels, we sort them by the kernel attention distance.

We depict the kernel attention distance of the first two layers of stage1 and the last two layers of stage3 in Fig. 9, since they have different channel dims, we normalize the channel index for a better comparison. Interestingly, RepLKNet also tends to aggregate both local and global information at shallow layers, while focusing more on global information at deeper layers. This essentially suggests that simultaneously aggregating local and global information can more effectively capture different levels of information in images, thereby resulting in more powerful and robust performance.

4.7. Stable Feature Map Variance

Further, we investigate the stability of the feature maps of large kernel networks. Specifically, we feed a batch of ImageNet validation images into different networks, the batch size is set to 64 and we use regular validation augmentation (only center crop and normalization). Then we calculate the normalized feature map variance layer by layer.

Table 6. **More results of large kernel convnets.** We add ConvNeXt as another large kernel convnet, and scale up the model size to large to demonstrate the scaling properties. *Note that ImageNet-C's image size is only 224×224.

Model	Kernel Size	ImageNet	ImageNet-A	ImageNet-R	ImageNet-C*
BiT-r152×4	3×3	85.4	13.2	33.8	39.9
ViT-B	N/A	84.0	26.7	38.0	42.2
ViT-L	N/A	85.2	28.1	40.6	49.9
ConvNeXt-B	7×7	85.8	33.9	45.5	53.2
ConvNeXt-L	7×7	86.6	38.7	47.6	56.5
RepLKNet-31B	31×31	85.2	29.4	43.9	49.2
RepLKNet-31L	31×31	86.6	39.6	49.1	52.6

As shown in Fig. 10, RepLKNet differs from the other two networks in two distinct aspects: **i)** It is very stable in the early stages. For instance, when the normalized depth is smaller than 0.4, the variance of RepLKNet's feature maps remains at a very low level, while ViT and ResNet tend to have a large variance from the very beginning; **ii)** The variance change in a simple and coherent manner. RepLKNet rises and falls slowly, changing in a simple and stable way throughout the process, while ResNet and ViT experience sharp fluctuations.

4.8. More Large Kernel ConvNets' Results

In order to further verify the effect of large kernel convolution on robustness and its scaling properties, we add another modern strong large kernel convnet, ConvNeXt (Liu et al., 2022b), and increase the model size of RepLKNet and ConvNeXt from base to large for comparison. As shown in Table 6, ConvNeXt also demonstrates strong robustness, and its robustness is further improved when the model size increases. As for RepLKNet-31L, we use the 384×384 pretrained model from the official release, and it achieves consistent improvements on ImageNet, ImageNet-A and ImageNet-R. For ImageNet-C, this dataset's images are all 224×224, evaluating a 384×384 model on this dataset will cause inconsistent input distribution, thus the improvement on this dataset is less significant.

4.9. How Large Can Make Strong Robustness?

While ConvNeXt and RepLKNet have different large kernel sizes (i.e., 7×7 v.s. 31×31), they both exhibit strong robustness. It is natural to ask: *how large kernel size can make strong robustness?* To answer this question, we conduct ablation studies by gradually increasing kernel sizes.

Specifically, we train ConvNext-Tiny with different kernel sizes in a 120 epoch schedule on ImageNet-1K (for computation constraints, it is not affordable for us to conduct this ablation on ImageNet-21K). The only difference is the kernel size. As shown in Table 7, there are three observations: **i)** scaling up kernels can bring consistent improvements

Table 7. **Ablation of kernel sizes on robustness.** We gradually increase the kernel size of ConvNeXt-T from 3 to 51. Scaling up kernel size can make consistent improvements on robustness. Note that ImgN is short for ImageNet.

Model	Kernel	ImgN	ImgN-A	ImgN-R
ConvNeXt-3	3×3	79.4	5.20	28.87
ConvNeXt-7	7×7	80.7	7.76	29.86
ConvNeXt-13	13×13	81.3	9.89	30.97
ConvNeXt-31	31×31	81.4	10.20	31.34
ConvNeXt-51	51×51	81.6	10.71	31.77

both on ImageNet and robustness benchmarks; **ii)** basically, scaling up to 13×13 can make a favorable robustness, but continuing scaling up kernel size to 51×51 can bring further robustness ; **iii)** Although scaling up to extremely large can not bring significant improvements on ImageNet, it can bring better improvements on robustness.

5. Limitation

Although we provide a comprehensive and in-depth empirical analysis of the strong robustness of large kernel convnets from multiple quantitative and qualitative perspectives, given the black box nature of deep learning, we struggle to provide direct theoretical proofs. Another limitation is due to computational constraints, we conduct ablation of kernel sizes on ImageNet-1K but not ImageNet-21K.

6. Conclusion

In this study, we thoroughly investigate the robustness of large kernel convnet, validating its strong robustness across six widely-used robustness benchmark datasets. Then we comprehensively analyze the source for its exceptional robustness from both quantitative and qualitative perspectives. Our research and analysis provide novel insights into the source of robustness, potentially promoting large kernel convnet's application and development in the future.

Acknowledgements

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Bao, H., Dong, L., Piao, S., and Wei, F. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Bhojanapalli, S., Chakrabarti, A., Glasner, D., Li, D., Unterthiner, T., and Veit, A. Understanding robustness of transformers for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10231–10241, 2021.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Chen, H., Chu, X., Ren, Y., Zhao, X., and Huang, K. Pelk: Parameter-efficient large kernel convnets with peripheral convolution. *arXiv preprint arXiv:2403.07589*, 2024a.

Chen, H., Kong, X., Zhang, X., Zhao, X., and Huang, K. Ddae: Towards deep dynamic vision bert pretraining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1037–1045, 2024b.

Chen, Y., Liu, J., Zhang, X., Qi, X., and Jia, J. Largekernel3d: Scaling up kernels in 3d sparse cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13488–13498, 2023.

Cheng, B., Schwing, A., and Kirillov, A. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34: 17864–17875, 2021.

Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., and Girdhar, R. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1290–1299, 2022.

Dai, X., Chen, Y., Yang, J., Zhang, P., Yuan, L., and Zhang, L. Dynamic detr: End-to-end object detection with dynamic attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2988–2997, 2021.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Ding, X., Zhang, X., Han, J., and Ding, G. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11963–11975, 2022.

Ding, X., Zhang, Y., Ge, Y., Zhao, S., Song, L., Yue, X., and Shan, Y. Unireplknet: A universal perception large-kernel convnet for audio, video, point cloud, time-series and image recognition. *arXiv preprint arXiv:2311.15599*, 2023.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Greff, K., Srivastava, R. K., and Schmidhuber, J. Highway and residual networks learn unrolled iterative estimation. *arXiv preprint arXiv:1612.07771*, 2016.

Guo, Y., Stutz, D., and Schiele, B. Robustifying token attention for vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17557–17568, 2023.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J., and Lakshminarayanan, B. Augmix: A simple method to improve robustness and uncertainty under data shift. In *International conference on learning representations*, volume 1, pp. 5, 2020.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021a.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15262–15271, 2021b.

Huang, T., Yin, L., Zhang, Z., Shen, L., Fang, M., Pechenizkiy, M., Wang, Z., and Liu, S. Are large kernels better teachers than transformers for convnets? *arXiv preprint arXiv:2305.19412*, 2023.

Huang, Y. and Kong, A. W.-K. Transferable adversarial attack based on integrated gradients. *arXiv preprint arXiv:2205.13152*, 2022.

Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., and Houlsby, N. Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pp. 491–507. Springer, 2020.

Kong, X. and Zhang, X. Understanding masked image modeling via learning occlusion invariant feature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6241–6251, 2023.

Kosmann-Schwarzbach, Y., Schwarzbach, B. E., and Kosmann-Schwarzbach, Y. *The Noether Theorems*. Springer, 2011.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Liu, S., Chen, T., Chen, X., Chen, X., Xiao, Q., Wu, B., Pechenizkiy, M., Mocanu, D., and Wang, Z. More convnets in the 2020s: Scaling up kernels beyond 51x51 using sparsity. *arXiv preprint arXiv:2207.03620*, 2022a.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976–11986, 2022b.

Lu, T., Ding, X., Liu, H., Wu, G., and Wang, L. Link: Linear kernel for lidar-based 3d perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1105–1115, 2023.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017a.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017b.

Mahmood, K., Mahmood, R., and Van Dijk, M. On the robustness of vision transformers to adversarial examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7838–7847, 2021.

Mao, X., Qi, G., Chen, Y., Li, X., Duan, R., Ye, S., He, Y., and Xue, H. Towards robust vision transformer. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 12042–12051, 2022.

Meng, D., Chen, X., Fan, Z., Zeng, G., Li, H., Yuan, Y., Sun, L., and Wang, J. Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3651–3660, 2021.

Naseer, M. M., Ranasinghe, K., Khan, S. H., Hayat, M., Shahbaz Khan, F., and Yang, M.-H. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34:23296–23308, 2021.

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

Park, N. and Kim, S. How do vision transformers work? *arXiv preprint arXiv:2202.06709*, 2022.

Paul, S. and Chen, P.-Y. Vision transformers are robust learners. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 36, pp. 2071–2081, 2022.

Peng, C., Zhang, X., Yu, G., Luo, G., and Sun, J. Large kernel matters–improve semantic segmentation by global convolutional network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4353–4361, 2017.

Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., and Dosovitskiy, A. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34:12116–12128, 2021.

Romero, D. W., Bruintjes, R.-J., Tomczak, J. M., Bekkers, E. J., Hoogendoorn, M., and van Gemert, J. C. Flexconv: Continuous kernel convolutions with differentiable kernel sizes. *arXiv preprint arXiv:2110.08059*, 2021a.

Romero, D. W., Kuzina, A., Bekkers, E. J., Tomczak, J. M., and Hoogendoorn, M. Ckconv: Continuous kernel convolution for sequential data. *arXiv preprint arXiv:2102.02611*, 2021b.

Shao, R., Shi, Z., Yi, J., Chen, P.-Y., and Hsieh, C.-J. On the adversarial robustness of vision transformers. *arXiv preprint arXiv:2103.15670*, 2021.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

Trockman, A. and Kolter, J. Z. Patches are all you need? *arXiv preprint arXiv:2201.09792*, 2022.

Veit, A., Wilber, M. J., and Belongie, S. Residual networks behave like ensembles of relatively shallow networks. *Advances in neural information processing systems*, 29, 2016.

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., and Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 568–578, 2021.

Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I. S., and Xie, S. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16133–16142, 2023.

Xiao, K., Engstrom, L., Ilyas, A., and Madry, A. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*, 2020.

Zhai, X., Puigcerver, J., Kolesnikov, A., Ruyssen, P., Riquelme, C., Lucic, M., Djolonga, J., Pinto, A. S., Neumann, M., Dosovitskiy, A., et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019.

Zhou, D., Yu, Z., Xie, E., Xiao, C., Anandkumar, A., Feng, J., and Alvarez, J. M. Understanding the robustness in vision transformers. In *International Conference on Machine Learning*, pp. 27378–27394. PMLR, 2022.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., and Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

A. Details of Robustness Datasets

ImageNet-A (Hendrycks et al., 2021b) is a dataset of real-world adversarially filtered images that fool current ImageNet classifiers. Specifically, it comprises a 200-class subset of ImageNet-1K's 1000 classes, covering most broad categories. These images, which should be correctly classified, are instead misclassified by a ResNet-50 with high confidence into incorrect categories. They cause consistent classification mistakes across various models due to scene complications encountered in the long tail of scene configurations and by exploiting classifier blind spots.

ImageNet-C (Hendrycks & Dietterich, 2019) comprises 15 categories of algorithmically generated corruptions and an additional four general corruption types, resulting in a total of 19 corruption categories. Each corruption type has five severity levels, ranging from negligible to pulverizing. This range allows the benchmark to provide a comprehensive assessment of each corruption type. These corruptions are all applied to the ImageNet validation images, so the total data volume for ImageNet-C is $19 \times 5 \times 50\text{k} = 4750\text{k}$ images.

ImageNet-O (Hendrycks et al., 2021b) is a dataset of adversarially filtered examples for ImageNet out-of-distribution detectors. It comprises of 200 categories from ImageNet-22K that are not included in ImageNet-1K. It contains *anomalies* of unforeseen classes for which a robust model is expected to output low-confidence predictions.

ImageNet-P (Hendrycks & Dietterich, 2019) consists of 10 types of common perturbations. ImageNet-P differs from ImageNet-C in that it generates perturbation sequences from each ImageNet validation image. The perturbations are subtly nuanced, affecting a smaller number of pixels within the images. To offset the increase in dataset size and evaluation time caused by each sequence containing over 30 frames, it contains only 10 common perturbations.

ImageNet-R (Hendrycks et al., 2021a) contains various artistic renditions of 200 classes from the ImageNet-1K dataset. Contrary to the original ImageNet, which discouraged such images as the annotators were instructed to collect "photos only, no paintings, no drawings, etc.", ImageNet-R adopts the opposite approach. Aiming at verifying the robustness of vision networks under semantic shifts under different domains.

ImageNet-9 (Xiao et al., 2020) helps disentangle the impacts of foreground and background signals on classification. Human vision exhibits a high degree of robustness to background changes, maintaining consistent decisions as long as the foreground remains the same. However, for most vision models, alterations in the background can significantly impact the model's output and accuracy. Hence, we delve further into the robustness towards background changes.

Table 8. Summary of the studied robustness benchmark datasets on their objectives and corresponding venues.

Dataset	Objective	Venue
ImageNet-A	natural adversarial	CVPR′21
ImageNet-C	common corruptions	ICLR′19
ImageNet-O	out-of-domain distribution	CVPR′21
ImageNet-P	common perturbations	ICLR′19
ImageNet-R	semantic shifts	ICCV′21
ImageNet-9	background dependency	ICLR′21