Vision Transformers Beat WideResNets on Small Scale Datasets Adversarial Robustness

Juntao Wu, Ziyu Song, Xiaoyu Zhang, Shujun Xie, Longxin Lin, Ke Wang *

State Key Laboratory of Bioactive Molecules and Druggability Assessment, Jinan University, Guangzhou, China.
Guangdong Institute of Smart Education, College of Information Science and Technology, Jinan University, Guangzhou, China
jtitor@stu.jnu.edu.cn, szycs@stu2022.jnu.edu.cn, {xyyzhang,xieshujun}@stu.jnu.edu.cn, {tlinlx, wangke}@jnu.edu.cn

Abstract

For an extensive period, Vision Transformers (ViTs) have been deemed unsuitable for attaining robust performance on small-scale datasets, with WideResNet models maintaining dominance in this domain. While WideResNet models have persistently set the state-of-the-art (SOTA) benchmarks for robust accuracy on datasets such as CIFAR-10 and CIFAR-100, this paper challenges the prevailing belief that only WideResNet can excel in this context. We pose the critical question of whether ViTs can surpass the robust accuracy of WideResNet models. Our results provide a resounding affirmative answer. By employing ViT, enhanced with data generated by a diffusion model for adversarial training, we demonstrate that ViTs can indeed outshine WideResNet in terms of robust accuracy. Specifically, under the Infty-norm threat model with epsilon = 8/255, our approach achieves robust accuracies of 74.97% on CIFAR-10 and 44.07% on CIFAR-100, representing improvements of +3.9% and +1.4%, respectively, over the previous SOTA models. Notably, our ViT-B/2 model, with 3 times fewer parameters, surpasses the previously best-performing WRN-70-16. Our achievement opens a new avenue, suggesting that future models employing ViTs or other novel efficient architectures could eventually replace the long-dominant WRN models.

Introduction

As deep learning continues to advance, neural networks have found widespread applications across diverse fields (Hinton et al. 2012; Krizhevsky, Sutskever, and Hinton 2012). However, these networks facing various challenges, making it increasingly important to ensure the robustness of deployed models and their ability to adapt to different input perturbations. The vulnerability of neural networks has long been recognized, as even minute, imperceptible adversarial perturbations can lead to erroneous predictions (Szegedy et al. 2013; Carlini and Wagner 2017). The concept of adversarial training was pioneered by Goodfellow, Shlens, and Szegedy (2014) to bolster the adversarial robustness of neural networks. Subsequently, Madry et al. (2017) introduced a superior adversarial training method, which has been hailed as one of the most effective approaches for cultivating robust deep neural networks. In the ensuing years, researchers endeavored to enhance models robustness through various means, such as refining loss functions (Zhang et al. 2019; Wang et al. 2019) or reimagining model architectures (Peng et al. 2023). Nevertheless, these advancements failed to address the data aspect, resulting in a stagnation of robust accuracy improvements. The advent of diffusion models marked a turning point, as Gowal et al. (2021) integrated these models into adversarial training, substantially fortifying adversarial robustness. Further advancements were made by Wang et al. (2023), who introduced higher-quality diffusion models, thereby pushing the boundaries of adversarial robustness even further.

Dataset	Method	Clean	AA
CIFAR-10($\ell_\infty, \epsilon = 8/255$)	Rank#1	93.27	71.07
	Ours	**95.76**	**74.97**
CIFAR-10($\ell_2, \epsilon = 0.5$)	Rank#1	95.54	84.97
	Ours	**97.00**	**86.06**
CIFAR-100($\ell_\infty, \epsilon = 8/255$)	Rank#1	75.22	42.67
	Ours	**82.45**	**44.07**

Table 1: A brief summary comparison of test accuracy (%) between our ViT-L/2 models and existing Rank #1 models, which use the WRN-70-16 architecture, as listed in RobustBench (Croce et al. 2020).

Despite these advancements substantially bolstering robust accuracy, the state-of-the-art (SOTA) robust accuracy on small-scale datasets such as CIFAR-10 and CIFAR-100 (Krizhevsky, Hinton et al. 2009) remains firmly under the dominion of WideResNet(WRN) (Zagoruyko and Komodakis 2016). Although there have been endeavors to introduce Vision Transformers (ViT) (Debenedetti, Sehwag, and Mittal 2023) and enhance their robustness, these efforts have yet to surpass the formidable benchmark set by WideResNet. Peng et al. (2023) implemented significant modifications to the WideResNet architecture, achieving improved robust accuracy on CIFAR-10 in comparison to the standard WideResNet. However, this improvement was a mere 0.4% over the existing SOTA, suggesting that further modifications to WideResNet are unlikely to yield substantial gains in robust accuracy. These modifications remain

*Corresponding author. Email: wangke@jnu.edu.cn

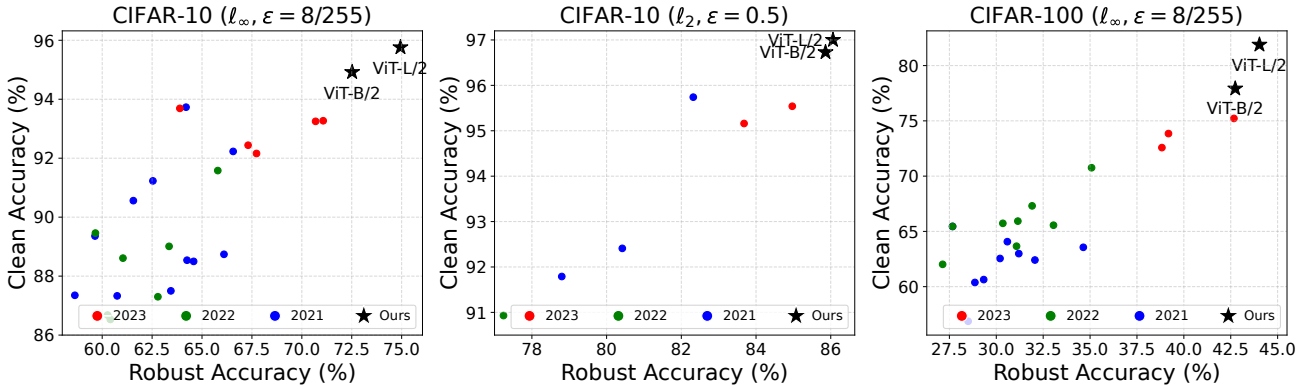

Figure 1: Robust accuracy (against AutoAttack) and clean accuracy of top-rank models in the leaderboard of RobustBench. The publication year of top-rank models is indicated by different colors. Our models use the ViT-B/2 and ViT-L/2 architectures in each setting, and detailed accuracy values are provided in Table 6 .

variations on the WideResNet theme, which continues to uphold its unassailable supremacy on small-scale datasets. As a result, WideResNet has effectively become the quintessential model against which robust accuracy is measured.

This prompts compelling inquiries: **Why did earlier endeavors to adopt ViTs falter? How to make ViTs beat WideResNets on Small Scale Datasets Adversarial Robustness?** Our objective is to address these queries and propose a viable solution to these critical challenges. Intriguingly, the answer may be embedded within RobustBench itself. To this day, the leading models for adversarial robustness on modest datasets such as CIFAR-10 and CIFAR-100 on RobustBench are predominantly WideResNet. Conversely, the top contenders on the ImageNet leaderboard are transformer-based models, not WideResNet. This disparity suggests that transformer models, when afforded ample data, can indeed surpass WideResNet in terms of robust accuracy. Therefore, we propose integrating diffusion models to generate synthetic data into the adversarial training of ViTs, aiming to achieve both superior robust and standard accuracy compared to the current state-of-the-art WideResNet.

Our contributions are as follows:

1. We are the first to demonstrate that Vision Transformers (ViT) can surpass WideResNet in adversarial robustness on small-scale datasets, challenging the prevailing notion that only WideResNet models are suitable for robust adversarial training.

2. Our ViT-B (ViT-Base) model outperforms WRN-70-16 in robust accuracy despite possessing 3x fewer parameters, showcasing the efficiency of training ViTs compared to WRN.

3. Our ViT-L (ViT-Large) model significantly surpasses the best-performing WideResNet in robust accuracy while maintaining a comparable parameter count, highlighting the superior robustness of ViTs.

Related Work

Adversarial Training. Adversarial training has emerged as a key method to enhance neural network robustness against adversarial attacks. Goodfellow, Shlens, and Szegedy (2014) introduced the foundational adversarial training approach using the Fast Gradient Sign Method (FGSM). Building on this, Madry et al. (2017) developed Projected Gradient Descent (PGD) adversarial training, establishing it as a leading technique for improving robustness. Further enhancements, such as Zhang et al. (2019)'s TRADES, optimize the trade-off between standard accuracy and adversarial robustness, offering a more balanced approach.

Diffusion Models. Diffusion models have revolutionized image generation and adversarial training. Ho, Jain, and Abbeel (2020) pioneered denoising diffusion probabilistic models, followed by Song et al. (2020)'s score-based models utilizing stochastic differential equations. Karras et al. (2022) advanced the field with the Elucidating Diffusion Model (EDM), significantly improving generative quality. Integrating these models into adversarial training, Gowal et al. (2021) and Wang et al. (2023) demonstrated substantial improvements in adversarial robustness, underscoring the efficacy of diffusion models in augmenting adversarial training.

Vision Transformer. The Vision Transformer (ViT) represents a groundbreaking shift from convolutional neural networks (CNNs) to self-attention mechanisms. Introduced by Dosovitskiy et al. (2020), ViT captures both global and local image features with unparalleled precision. Despite its advantages, ViT has historically required large datasets to outperform CNNs, which has made it challenging for small-scale datasets. Recent efforts, such as Debenedetti, Sehwag, and Mittal (2023), have aimed to enhance the robustness of ViTs, but they have still not surpassed the benchmarks set by WideResNet.

Our work builds on these advancements by incorporating diffusion model-generated synthetic data into the adversarial training of ViTs, aiming to achieve both superior robust and standard accuracy, addressing the historical challenges faced

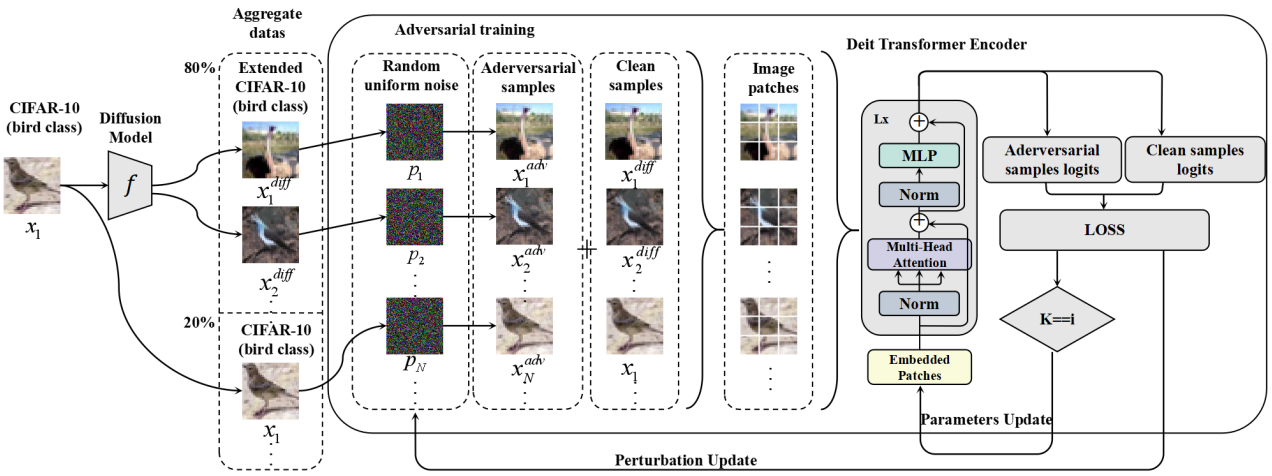

Figure 2: Overview of the proposed adversarial training pipeline. We utilize the Elucidating Diffusion Model (EDM) to generate synthetic data, which is then mixed with the original data. This combined data is subsequently used in the TRADES adversarial training framework to enhance robustness.

by ViTs on small-scale datasets.

Problem Analysis

Why did earlier endeavors to adopt Vision Transformers (ViTs) falter? In large-scale visual pretraining scenarios, such as those exemplified by CLIP (Radford et al. 2021), ViTs consistently outperform convolutional-based models (Cherti et al. 2023; Radford et al. 2021), highlighting their superior capacity for fitting complex visual patterns. ViTs are designed to leverage advanced attention mechanisms to uncover intricate relationships within visual data, offering impressive performance when the training and test data distributions are well-aligned (Dosovitskiy et al. 2020). However, the initial attempts to deploy ViTs in more practical settings often encountered significant challenges.

One major issue lies in their handling of distributional shifts. While the sophisticated architecture of ViTs enables them to capture detailed and nuanced representations of the training data, this complexity also predisposes them to overfitting. Overfitting occurs when a model becomes too tailored to the specifics of the training set, leading to a degradation in performance on unseen data that deviates from the training distribution (Steiner et al. 2021). This problem is exacerbated in scenarios where the data distributions are heterogeneous or exhibit substantial shifts between training and testing phases. In such cases, the inherent strengths of ViTs in learning intricate patterns can paradoxically become a liability, resulting in reduced robustness and generalization capabilities when faced with novel or shifted data characteristics.

Our Solution

How to make ViTs beat WideResNets on Small Scale Datasets Adversarial Robustness? In order for ViTs to perform well in small-scale datasets, we introduce diffusion model to generate data for adversarial training of ViTs. Figure 2 provides an overview of our training pipeline. It illustrates the process of generating synthetic data using the EDM diffusion model, incorporating this data into the TRADES adversarial training framework, and utilizing it within the ViT training to improve robustness and performance.

Our method involves integrating these synthetic data into the TRADES adversarial training process. The enriched dataset mitigates overfitting and enhances the adversarial robustness of ViTs. By aligning the distributions of the training and test datasets more closely, our approach improves the performance of ViTs, enabling them to outperform WRNs in adversarial robustness on small-scale datasets. We use a 10-step TRADES adversarial training process to ensure robust training. Our experimental results demonstrate that ViTs, when trained with synthetic data and TRADES, achieve superior adversarial robustness compared to WRNs, challenging the previously held notion that ViTs are inherently less robust to adversarial attacks on small-scale datasets.

Experiments and Analysis

Experimental Setup

Wang et al. (2023) 's experimental setup has been widely adopted in adversarial training for WRNs and has yielded promising results. As we aim to investigate the performance of ViTs, to ensure a fair comparison, we closely follow their experimental setup, incorporating label smoothing (Szegedy et al. 2016), Exponential Moving Average (EMA), and an 8:2 mix of generated and real data. However, we replace the SGD optimizer with the Lion optimizer (Chen et al. 2024), as ViTs do not perform optimally with SGD. All training is performed using TPU with JAX and then converted back to PyTorch for testing.

Moreover, Wang et al. (2023) performs early stopping as a default trick. They separate the first 1024 images of the training set as a fixed validation set. During every epoch of

AT, they pick the best checkpoint by evaluating robust accuracy under PGD-40 attack on the validation set. In contrast, to simplify our process, we do not use early stopping. Instead, we directly use the last saved checkpoint for testing. Therefore, our training data includes the entire training set without reserving a portion for validation.

Implementation Details and Hyperparameter Settings. Our ViT implementation aligns with the standard ViT model from timm (Wightman 2019). For adversarial training, we employ TRADES (Zhang et al. 2019) with β set to 3 for CIFAR-10 and 5 for CIFAR-100. The Lion optimizer is configured with $\beta_1 = 0.9$, $\beta_2 = 0.99$, and weight decay (WD) of 0.5, while a cosine annealing learning rate schedule (Loshchilov and Hutter 2016) with a peak learning rate of 1×10^{-4} is used. For image generation, we utilize the Elucidating Diffusion Model (EDM) (Karras et al. 2022) , and we use the dataset generated by Wang et al.

Computational Time. We use TPU v4-256 for training, batch size 1024 for all configurations. For 10-step TRADES (Zhang et al. 2019) adversarial training, the average speed is 10.34 iterations per second (it/s) for ViT-B/2, and 3.18 it/s for ViT-L/2.

Evaluation metrics. We evaluate model robustness against AutoAttack (Croce and Hein 2020), a strong adversarial attack suite comprising multiple attack strategies. Throughout the experiments, we use Clean to represent the model accuracy for clean data, AA to represent AutoAttack accuracy.

Sensitivity Analysis

We test the sensitivity of basic training hyperparameters on CIFAR-10. ViT-B/2 models are trained for 2000 epochs using 50M data generated by EDM. 1024 is the default batch size.

Label Smoothing. Label smoothing (LS) improves generalization and reduces overconfidence in standard training (Szegedy et al. 2016; Hein, Andriushchenko, and Bitterwolf 2019). In adversarial settings, LS helps prevent robust overfitting without external data (Stutz, Hein, and Schiele 2020; Chen et al. 2020). Excessive LS (0.3-0.4) degrades WRN models performance by over-smoothing labels and losing information in the output logits (Pang et al. 2020; Müller, Kornblith, and Hinton 2019). Wang et al. (2023) showed that high LS decreases adversarial robustness in WRN models with synthetic data from EDM. According to the results shown in Table 2(left), LS improves the robustness of ViT but that ViT is not sensitive to high LS. ViT maintains performance even with high LS, indicating better hyperparameter stability than WRN. To demonstrate ViT's insensitivity to LS, we will use $LS = 0.4$ in subsequent experiments.

Effect of β. In the TRADES framework (Zhang et al. 2019), the parameter β plays a pivotal role in balancing robustness and accuracy. An increase in β generally leads to a reduction in clean accuracy while enhancing robust accuracy, as elucidated by Zhang et al. (2019). Through their investigations, Wang et al. (2023) established that for CIFAR-10, a β value of 5 yields the highest robust accuracy for WideResNet. Yet, according to the results shown in Table 2(right), our findings reveal a different trend: ViT exhibits markedly superior performance with $\beta = 3$ rather than $\beta = 5$. This divergence suggests that optimal hyperparameter settings are not universally applicable but must be finely tuned to align with specific model architectures.

LS	Clean	AA
0.0	92.72	70.52
0.1	93.03	**71.07**
0.2	93.05	70.88
0.3	**93.39**	71.01
0.4	**93.39**	70.98

β	Clean	AA
2	95.21	71.22
3	**94.35**	**71.31**
4	94.07	71.06
5	93.39	70.98
6	93.07	70.69

Table 2: Test accuracy (%) with different values of label smoothing (LS)(left) and β in TRADES (right), under the (ℓ_∞, $\epsilon = 8/255$) threat model on CIFAR-10

Mix	AA	Clean
0.0	0.5685	0.7716
0.1	0.6263	0.8889
0.2	0.6673	0.9128
0.4	0.7001	0.9300
0.5	0.7129	0.9328
0.6	0.7227	0.9362
0.7	0.7273	0.9392
0.8	0.7352	0.9417
0.9	**0.7377**	0.9440
1.0	0.7356	**0.9441**

Table 3: Clean accuracy and robust accuracy against PGD-40 with respect to generate and origin data mix ratios (0 means CIFAR-10 training set only , 1 means generated images only). We train ViT-B/2 models against (ℓ_∞, $\epsilon = 8/255$) on CIFAR-10 using 50M generated data.

Mix ratio of original dataset and generated dataset. We conducted additional experiments to explore the impact of different mixing ratios between generated data and original data on Clean and AA (AutoAttack) performance. Due to limitations in computational resources, we used only ViT-B/2 for 1000 epochs of training. Apart from reducing the number of epochs from 2000 to 1000, all other training settings remained consistent with the default. It's worth noting that we used PGD-40 for robustness testing instead of AutoAttack, as PGD-40 is less computationally intensive, simpler to implement, and can be tested directly on TPU. Table 3 shows the clean accuracy and adversarial accuracy for different mixing ratios of generated data and real data. Experiments by Wang et al. (2023) indicate that a mix ratio of 0.8 is optimal for WRNs. However, our experiments show that a ratio of 0.9 is best for ViTs, which further supports our claim that ViTs are data-hungry and require a large amount of data to avoid overfitting.

Model Size Scaling. In standard training, models with more parameters are generally considered to achieve better performance compared to smaller models. This notion

Case	Arch	Epoch	Clean	AA
Previous Rank #1	WRN-70	2000	93.27	71.07
Baseline	ViT-B/2	2000	94.35	71.31
Model Scaling	ViT-L2	2000	94.75	72.17
Epoch Scaling	ViT-B/2	6000	94.92	72.52
Both Scaling	ViT-L2	10000	**95.76**	**74.97**

Table 4: Test accuracy (%) with different scaling strategies for Vision Transformers (ViT). Under the (ℓ_∞, $\epsilon = 8/255$) threat model on CIFAR-10

is widely supported in the context of adversarial robustness as well. For instance, WRN-70-16 consistently outperforms WRN-28-10. Therefore, we investigated the impact of scaling model size on performance (Xie and Yuille 2019; Madry et al. 2017). According to the results shown in Table 4, larger models can further improve both the clean accuracy and robust accuracy of ViTs.

Training Epoch. In the realm of standard training, ViTs are often deemed to necessitate a more protracted training period than their convolutional counterparts to reach their peak performance (Dosovitskiy et al. 2020; Touvron, Cord, and Jégou 2022). To explore this notion, we delved into the effects of extending the number of training epochs. Our investigations reveal that prolonging the training duration can indeed yield significant enhancements in both clean accuracy and robust accuracy for ViTs, underscoring the benefits of sustained training for optimizing these models.

Ablation Study of Optimizer

WRNs typically use SGD, which is less suited for ViTs (Xiao et al. 2021). We used the Lion optimizer and conducted ablations to ensure that performance gains were not due to switching from SGD to Lion. Results show that WRN does not benefit from Lion, confirming the advantage of ViTs in this context.

WRN. For WRN ablation experiments, we follow Wang et al. (2023) with a batch size of 2048, $\beta = 5.0$, label smoothing of 0.1, EMA decay of 0.995, and 2400 epochs. Due to computational constraints, we use WRN-28-10 instead of WRN-70-16, as the results are deemed representative. We initially applied the Lion optimizer settings used for ViTs (lr 1×10^{-4}, WD 0.5), but encountered instability. We reduced the lr to 5×10^{-5} and doubled the WD, which resolved the instability.

Arch	Optimizer	Clean	AA
WRN-28-10 (Wang's)	SGD	92.44	67.31
WRN-28-10 (Ours)	Lion	92.47	67.42
ViT-B/2	SGD	75.06	✗
ViT-B/2	Lion	94.35	71.31

Table 5: Comparison of WRN-28-10 and ViT-B/2 models on clean and adversarial accuracy using SGD and Lion optimizers.

ViT. We also conduct ablation experiments on ViT to demonstrate that the SGD optimizer is not suitable for ViT compared to the Lion optimizer. For the SGD optimizer, we follow Wang et al. (2023) for the learning rate settings, while keeping the other settings consistent with the default settings for Lion, i.e. $\beta = 0.3$ and label smoothing (LS) of 0.4.

While the Lion optimizer contributed to performance gains, the improved adversarial robustness of WRN-28-10 is not solely due to it. Table 5 shows that although Lion provided a slight improvement in WRN-28-10's accuracy, the overall robustness improvements are attributable to multiple factors beyond just replacing SGD with Lion. For ViT-B/2, the Lion optimizer significantly enhanced performance and stability compared to SGD, highlighting its key role in optimizing ViTs.

Vision Transformers Beat WideResNets

To benchmark against WRN-70-16, we chose ViT-B and ViT-L as our comparison models. As can be seen from the table, regardless of different datasets or different training epochs, our ViT models achieve SOTA performance in both clean and adversarial accuracy. Specifically, under the (ℓ_∞, $\epsilon = 8/255$) and (ℓ_2, $\epsilon = 0.5$) threat models on CIFAR-10. And under the (ℓ_∞, $\epsilon = 8/255$) threat model on CIFAR-100. Our ViT models achieve SOTA performance in both clean and adversarial accuracy. Moreover,WRN-70-16 has 38.81G FLOPs,while our ViT-B/2 has 21.85G. Over 2000 epochs, ViT-B/2 achieved 71.31%, compared to WRN's 70.69%, showing ViT's clear advantage.

Remark for Table 6. Under the (ℓ_∞, $\epsilon = 8/255$) threat model on CIFAR-10, even when using a ViT-B/2 model that is 3x smaller than WRN-70-16, our ViT-B/2 achieves better robust accuracy comparable to that of Peng et al. (2023) who use RaWRN, while our clean accuracy improves significantly +1.08%. After applying longer training of 6000 epochs, our ViT-B/2 model surpasses the previous best result with a large margin (clean accuracy $+1.65\%$, robust accuracy $+1.45\%$). **Our ViT-L/2 is the first adversarially trained model to achieve clean accuracy over** 95.5% **and robust accuracy over** 74.5% **.**

Enhanced Spatial Information Capture in ViT Compared to WRN. EigenCAM (Muhammad and Yeasin 2020) is a visualization technique that highlights the regions in an image where the model places significant focus. As shown in Figure 3, ViT focuses its attention across the entire semantic object, while WRN tends to concentrate on the center of the image. This indicates that ViT has genuinely learned the semantic information of the image, capturing a more holistic understanding. In contrast, WRN appears to rely on memorizing specific parts of the image without truly grasping the underlying semantics. The heatmaps show that ViT can effectively recognize and utilize relationships across different parts of the input, leading to superior performance in capturing intricate details and structures. WRN's narrower focus restricts its ability to perceive extensive relationships, resulting in less effective feature extraction and spatial awareness.

Dataset	Architecture	Method	Params	Generated	Batch	Epoch	Clean	AA
CIFAR-10 ($\ell_\infty, \epsilon = 8/255$)	WRN-28-10	Pang et al. (2022)	36M	1M	512	400	88.10	61.51
	WRN-28-10	Gowal et al. (2021)	36M	100M	1024	2000	87.50	63.38
	WRN-28-10	Wang et al. (2023)	36M	20M	2048	2400	92.44	67.31
	ViT-B/2	**Ours**	85M	50M	1024	2000	**94.35**	**71.31**
			85M	50M	1024	6000	**94.92**	**72.52**
	WRN-70-16	Pang et al. (2022)	267M	1M	512	400	88.57	63.74
	WRN-70-16	Rebuffi et al. (2021)	267M	1M	1024	800	88.54	64.20
	WRN-70-16	Gowal et al. (2021)	267M	100M	1024	2000	88.74	66.11
	WRN-70-16	Wang et al. (2023)	267M	50M	1024	2000	93.25	70.69
	RaWRN-70-16	Peng et al. (2023)	267M	50M	1024	2000	93.27	71.07
	ViT-L/2	**Ours**	302M	50M	1024	2000	**94.75**	**72.17**
			302M	50M	1024	10000	**95.76**	**74.94**
CIFAR-10 ($\ell_2, \epsilon = 0.5$)	WRN-28-10	Pang et al. (2022)	36M	1M	512	400	90.83	78.10
	WRN-28-10	Rebuffi et al. (2021)	36M	1M	1024	800	91.79	78.69
	WRN-28-10	Wang et al. (2023)	36M	50M	2048	1600	95.16	83.63
	ViT-B/2	**Ours**	85M	50M	1024	2000	**96.53**	**85.40**
			85M	50M	1024	6000	**96.73**	**85.86**
	WRN-70-16	Rebuffi et al. (2021)	267M	1M	1024	800	92.41	80.42
	WRN-70-16	Wang et al. (2023)	267M	50M	1024	2000	95.54	84.86
	ViT-L/2	**Ours**	302M	50M	1024	2000	**96.82**	**85.67**
			302M	50M	1024	6000	**97.00**	**86.06**
CIFAR-100 ($\ell_\infty, \epsilon = 8/255$)	WRN-28-10	Pang et al. (2022)	36M	1M	512	400	62.08	31.40
	WRN-28-10	Rebuffi et al. (2021)	36M	1M	1024	800	62.41	32.06
	WRN-28-10	Wang et al. (2023)	36M	50M	2048	1600	72.58	38.83
	ViT-B/2	**Ours**	85M	50M	1024	2000	**77.92**	**42.74**
	WRN-70-16	Pang et al. (2022)	267M	1M	512	400	63.99	33.65
	WRN-70-16	Rebuffi et al. (2021)	267M	1M	1024	800	63.56	34.64
	WRN-70-16	Wang et al. (2023)	267M	50M	1024	2000	75.22	42.67
	ViT-L/2	**Ours**	302M	50M	1024	6000	**82.45**	**44.07**

Table 6: Test accuracy (%) of clean images and under AutoAttack (AA) on CIFAR-10 and CIFAR-100. We highlight our results in bold whenever the value represents an improvement relative to the strongest baseline, and we underline them whenever the value achieves new SOTA result under the threat model.

Recent Work and Discussion

Upon the culmination of our rigorous experiments, the apex position on RobustBench was steadfastly held by Peng et al. (2023). However, a remarkable shift transpired in July, when Bartoldson et al. ascended to the summit.

Bartoldson et al. (2024) harnessed an advanced generative model, yielding synthetic data of superior quality, as evidenced by a lower FID score of 1.65 compared to EDM's 1.824, and a significantly higher number of unique samples (300 million versus Wang's 50 million). A lower FID score indicates that the generated data more closely mirrors the test data distribution, while a larger pool of unique samples helps mitigate overfitting. This dual advantage enhances the alignment of out-of-distribution data with in-distribution data, thereby boosting model performance on the test set. Additionally, Bartoldson et al. (2024) utilized a marginally larger model than WRN-70-16, extended the training duration to 10,000 epochs from Wang's 2,000 epochs, and reaped the benefits of superior synthetic data quality and quantity, culminating in a robust accuracy of 73.71% on CIFAR-10.

We did not include Bartoldson et al. (2024) in Table 6 due to their limited experimental scope, restricted data availability, and different focus of improvements.

- **Limited Experiment Scope:** Bartoldson et al. (2024) limited their experiments to CIFAR-10 ($\ell_\infty, \epsilon = 8/255$) and did not extend their work to CIFAR-100 or other datasets. This limitation in scope prevented a broader evaluation of their approach across different datasets.
- **Restricted Data Availability:** At the time of our manuscript submission, Bartoldson et al. (2024) had not open-sourced their dataset, unlike Wang et al. (2023). Without access to their data, we were unable to re-train using their synthetic samples.

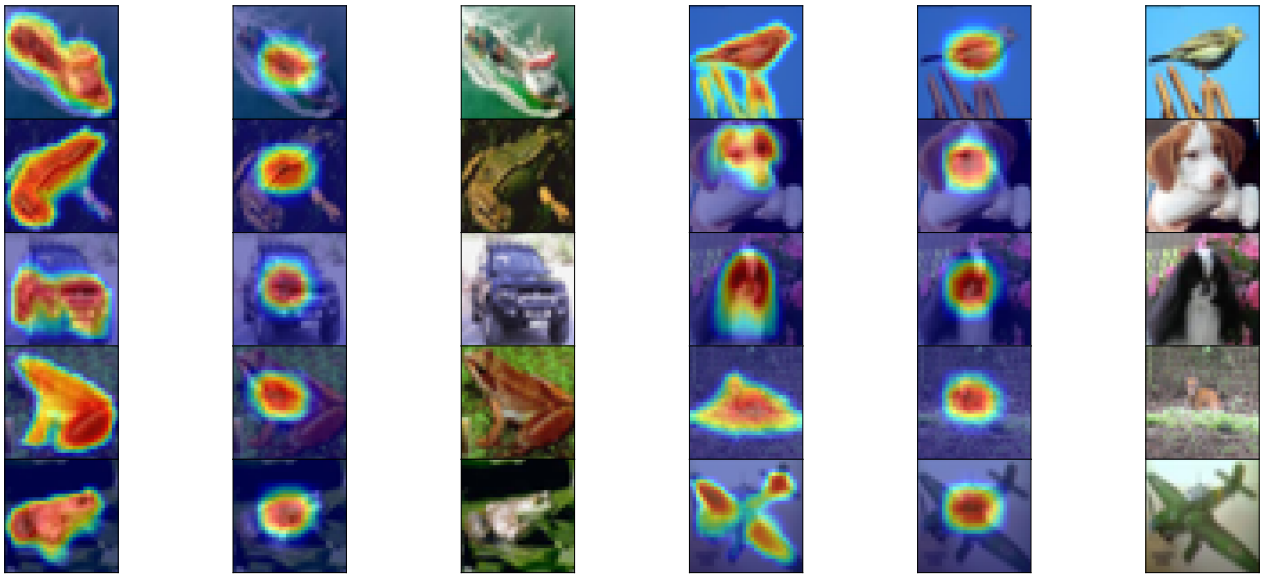

Figure 3: EigenCAM (Muhammad and Yeasin 2020) visualization results on the test dataset. For each group of three images: the left image is ViT under AutoAttack, the middle image is WRN under AutoAttack, and the right image is the original image.

- **Different Focus of Improvements:** Bartoldson et al. (2024) concentrated on enhancing synthetic data quality and increasing the number of unique samples to bolster robustness. In contrast, our improvements are model-centric. By the time we recognized their top position on RobustBench, we had completed most of our experiments and were unable to utilize their data. Therefore, for the sake of fairness, we still compare our method with those employing Wang's EDM 50M or previous for adversarial training.

Nevertheless, we extended the training of ViT-L/2 to 10,000 epochs. Our ViT-L/2, with fewer unique samples, lower quality synthetic data, and fewer parameters compared to Bartoldson et al. (2024)'s WideResNet-94-16, achieved robust accuracy of 74.97%. This represents a significant improvement of 1.26% over Bartoldson et al. (2024) and a clean accuracy increase of 2.08%. This strongly suggests that our ViT outperforms WRN, even under less favorable conditions. Bartoldson et al. (2024) essentially adhered to the WRN framework, whereas we innovatively utilized ViT. It is crucial to emphasize that our improvements do not conflict with those of Bartoldson et al. (2024). **They enhanced data quality, while we focused on model advancements.** We believe that with access to their synthetic data, our model's adversarial robustness could be further amplified.

Conclusion and Future Work

Our experiments demonstrate that Vision Transformers (ViTs) can achieve superior adversarial robustness and clean accuracy compared to state-of-the-art WideResNets (WRNs). Specifically, our ViT-L/2 model significantly surpasses previous benchmarks, achieving a robust accuracy of 74.97% on CIFAR-10. This represents a substantial improvement in adversarial robustness, highlighting the effectiveness of ViTs in enhancing model performance under adversarial conditions compared to existing WRN models.

While our study demonstrates the effectiveness of ViTs , there are several areas for future work. We did not explore other advanced transformer models such as Swin Transformer, which could potentially yield better results. Additionally, the training process for our ViT models was computationally expensive, highlighting the need to reduce training time without compromising performance. The substantial computational resources required for our experiments, especially for larger models like ViT-L/2, may limit accessibility and reproducibility for researchers with fewer resources. Our future studies will focus on enhancing computational efficiency, exploring advanced models, and evaluating performance across diverse datasets.

Acknowledgments

This work was supported in part by National Key Research and Development Program of China (Grant No. 2022YFC3303200), National Key Program of National Natural Science of China (Grant No. 82430108), Guang-Dong Basic and Applied Basic Research Foundation(Grant No. 2023A1515012111), Shenzhen-Hong Kong-Macao Science and Technology Plan Project (Category C Project: SGDX20210823103537030), Key Laboratory of Smart Education of Guangdong Higher Education Institutes, JinanUniversity(2022LSYS003).

We would also like to thank TPU Research Cloud (TRC) program, for supporting our computing needs.

References

Bartoldson, B. R.; Diffenderfer, J.; Parasyris, K.; and Kailkhura, B. 2024. Adversarial Robustness Limits via Scaling-Law and Human-Alignment Studies. *arXiv preprint arXiv:2404.09349*.

Carlini, N.; and Wagner, D. 2017. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, 3–14.

Chen, T.; Zhang, Z.; Liu, S.; Chang, S.; and Wang, Z. 2020. Robust overfitting may be mitigated by properly learned smoothening. In *International Conference on Learning Representations*.

Chen, X.; Liang, C.; Huang, D.; Real, E.; Wang, K.; Pham, H.; Dong, X.; Luong, T.; Hsieh, C.-J.; Lu, Y.; et al. 2024. Symbolic discovery of optimization algorithms. *Advances in Neural Information Processing Systems*, 36.

Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; and Jitsev, J. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2818–2829.

Croce, F.; Andriushchenko, M.; Sehwag, V.; Debenedetti, E.; Flammarion, N.; Chiang, M.; Mittal, P.; and Hein, M. 2020. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*.

Croce, F.; and Hein, M. 2020. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, 2206–2216. PMLR.

Debenedetti, E.; Sehwag, V.; and Mittal, P. 2023. A light recipe to train robust vision transformers. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 225–253. IEEE.

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

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

Gowal, S.; Rebuffi, S.-A.; Wiles, O.; Stimberg, F.; Calian, D. A.; and Mann, T. A. 2021. Improving robustness using generated data. *Advances in Neural Information Processing Systems*, 34: 4218–4233.

Hein, M.; Andriushchenko, M.; and Bitterwolf, J. 2019. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 41–50.

Hinton, G.; Deng, L.; Yu, D.; Dahl, G. E.; Mohamed, A.-r.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T. N.; et al. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6): 82–97.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Karras, T.; Aittala, M.; Aila, T.; and Laine, S. 2022. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35: 26565–26577.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

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

Loshchilov, I.; and Hutter, F. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

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

Muhammad, M. B.; and Yeasin, M. 2020. Eigen-cam: Class activation map using principal components. In *2020 international joint conference on neural networks (IJCNN)*, 1–7. IEEE.

Müller, R.; Kornblith, S.; and Hinton, G. E. 2019. When does label smoothing help? *Advances in neural information processing systems*, 32.

Pang, T.; Lin, M.; Yang, X.; Zhu, J.; and Yan, S. 2022. Robustness and accuracy could be reconcilable by (proper) definition. In *International Conference on Machine Learning*, 17258–17277. PMLR.

Pang, T.; Yang, X.; Dong, Y.; Su, H.; and Zhu, J. 2020. Bag of tricks for adversarial training. *arXiv preprint arXiv:2010.00467*.

Peng, S.; Xu, W.; Cornelius, C.; Hull, M.; Li, K.; Duggal, R.; Phute, M.; Martin, J.; and Chau, D. H. 2023. Robust principles: Architectural design principles for adversarially robust cnns. *arXiv preprint arXiv:2308.16258*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rebuffi, S.-A.; Gowal, S.; Calian, D. A.; Stimberg, F.; Wiles, O.; and Mann, T. 2021. Fixing data augmentation to improve adversarial robustness. *arXiv preprint arXiv:2103.01946*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

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 preprint arXiv:2106.10270*.

Stutz, D.; Hein, M.; and Schiele, B. 2020. Confidence-calibrated adversarial training: Generalizing to unseen attacks. In *International Conference on Machine Learning*, 9155–9166. PMLR.

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

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Touvron, H.; Cord, M.; and Jégou, H. 2022. Deit iii: Revenge of the vit. In *European conference on computer vision*, 516–533. Springer.

Wang, Y.; Zou, D.; Yi, J.; Bailey, J.; Ma, X.; and Gu, Q. 2019. Improving adversarial robustness requires revisiting misclassified examples. In *International conference on learning representations*.

Wang, Z.; Pang, T.; Du, C.; Lin, M.; Liu, W.; and Yan, S. 2023. Better diffusion models further improve adversarial training. In *International Conference on Machine Learning*, 36246–36263. PMLR.

Wightman, R. 2019. PyTorch Image Models. https://github.com/rwightman/pytorch-image-models.

Xiao, T.; Singh, M.; Mintun, E.; Darrell, T.; Dollár, P.; and Girshick, R. 2021. Early convolutions help transformers see better. *Advances in neural information processing systems*, 34: 30392–30400.

Xie, C.; and Yuille, A. 2019. Intriguing properties of adversarial training at scale. *arXiv preprint arXiv:1906.03787*.

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

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E.; El Ghaoui, L.; and Jordan, M. 2019. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, 7472–7482. PMLR.