RBFormer: Improve Adversarial Robustness of Transformer by Robust Bias

Hao Cheng[1]
hcheng046@connect.hkust-gz.edu.cn
Jinhao Duan[2]
jd3734@drexel.edu
Hui Li[3]
hui01.li@samsung.com
Jiahang Cao[1]
jcao248@connect.hkust-gz.edu.cn
Ping Wang[4]
ping.fu@mail.xjtu.edu.cn
Lyutianyang Zhang[5]
lyutiz@uw.edu
Jize Zhang[6]
cejize@ust.hk
Kaidi Xu[2]
kx46@drexel.edu
Renjing Xu[1]
renjingxu@ust.hk

[1] The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China
[2] Drexel University, Philadelphia, USA
[3] Samsung R&D Institute China Xi'an (SRCX), Xi'an, China
[4] Xi'an Jiaotong University, Xi'an, China
[5] University of Washington, Seattle, USA
[6] The Hong Kong University of Science and Technology, Hong Kong SAR

Abstract

Recently, there has been a surge of interest and attention in Transformer-based structures, such as Vision Transformer (ViT) and Vision Multilayer Perceptron (VMLP). Compared with the previous convolution-based structures, the Transformer-based structure under investigation showcases a comparable or superior performance under its distinctive attention-based input token mixer strategy. Introducing adversarial examples as a robustness consideration has had a profound and detrimental impact on the performance of well-established convolution-based structures. This inherent vulnerability to adversarial attacks has also been demonstrated in Transformer-based structures. In this paper, our emphasis lies on investigating the intrinsic robustness of the structure rather than introducing novel defense measures against adversarial attacks. To address the susceptibility to robustness issues, we employ a rational structure design approach to mitigate such vulnerabilities. Specifically, we enhance the adversarial robustness of the structure by increasing the proportion of high-frequency structural robust biases. As a result, we introduce a novel structure called Robust Bias Transformer-based Structure (RBFormer) that shows robust superiority compared to several existing baseline structures. Through a series of extensive experiments, RBFormer outperforms the original structures by a significant margin, achieving an impressive improvement of +16.12% and +5.04% across different evaluation criteria on CIFAR-10 and ImageNet-1k, respectively.

1 Introduction

Convolutional Neural Networks (CNNs) have achieved breakthroughs in many domains [4, 9, 13, 28, 43, 44]. However, adversarial examples [4, 9, 30, 35, 38] as the inherent vulnerability of model structures has been extensively observed in CNNs across diverse contexts. In allusion to alleviating this vulnerability, adversarial training [17] as the most successful robust boosting method has also been proposed. In addition to traditional CNNs, ViT [8] and its subsequent studies [1, 2, 3, 11, 16, 22, 37], which are inspired by the transformer-based architectures [8], have been assumed as a novel base structure for solving various computer vision tasks. Following the continual research about the characteristics of Multi-head Self-Attention (MSA), [7] has found that the overmuch usage of MSA might adversely influence the Transformer performance and lead the entire output to converge exponentially to a rank-1 matrix. However, the Skip-Connections and MLP sub-blocks could mitigate and avoid this rank collapse phenomenon. This finding demonstrates that the MSA is not the most essential factor for the success of the Transformer, but the structure itself is. This phenomenon could quickly transfer to ViT and VMLP [26, 32, 42], which illustrates the structure like ViT but removes attention blocks. About these two new structure, there are various works [13, 25, 27, 29, 40, 41, 46] put their first focus on analyzing the difference between Transformer-based and Convolution-based structures, which ViT and CNN typically represent. They indicate that CNN holds the inductive bias compared with ViT. The core concept of inductive bias is the locality, which will make CNN focus more on low-level local information, but ViT pay more attention to high-level global information. Therefore, the convolution structure could be understood as a low-level or high-frequency structure. For robustness concerns, recently, [18, 21, 23, 27] claim that Transformer-based structures also exist adversarial vulnerability and could also be alleviated by adversarial training. However, the current research in this field primarily focuses on studying basic robust features and enhancing the compatibility between existing defensive methods and established structures. To make a robust improvement based on the structural design level, we may ask:

Can we improve the robustness of original ViT/VMLP by rational structure design?

To pursue the answer to this question, the Transformer-based structures could be divided into three essential components: 1) Embedding, 2) Token-Mixer (TM) block; 3) Classifying MLP (CMLP) block, and two training facilitation techniques: 1) Normalization, 2) Skip-connection. The TM block could be further divided into MSA sub-block and MLP sub-block. Based on prior research, ViT/VMLP models differ from CNNs in their focus on high-level, low-frequency information, leading to a more global representation of images. While CNNs rely on an inductive bias that emphasizes locality, ViT/VMLP models prioritize capturing global context, allowing them to incorporate high-level, low-frequency details directly. This distinction highlights how the inductive bias in CNNs makes them more sensitive to high-level, low-frequency information. In the meantime, [39] analyzes the adversarial training from the Fourier perspective. Therefore, introducing the high-frequency structure is a feasible way to improve adversarial robustness, and it could be called robust bias. Currently, there are two ways of strengthening the inductive bias or locality in ViT/VMLP, except for directly adding convolution operation to structural components [15, 33, 36], introducing multi-hierarchy layer stacking strategies [16, 33, 47], which could boost local or low-level vi-

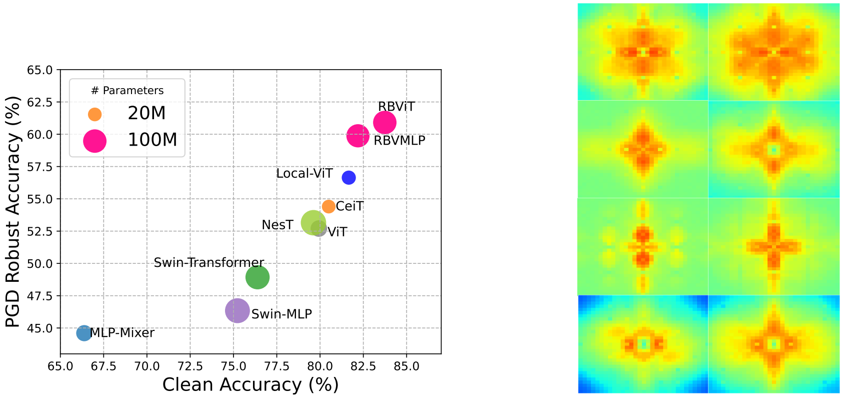

Figure 1: Left: Comparison Results of RBFormer (RBViT/RBVMLP) with current SOTA in clean/robust accuracy and model size. Right: Fourier Heatmap, from top to bottom, ViT/VMLP (internal Left/Right), adding convolution operation to embedding, embedding + block aggregation, and embedding + block aggregation + CMLP

sual structure by amplifying the ability of cross-patch information communication, is another good way. Following these two ways of increasing robust biases to the original structure, we could finally obtain RBFormer with better robustness after the comprehensive evaluation. The evaluation includes employing PGD [17], Auto-Attack [5], frequency heat map [39], and local Lipschitz [14] techniques to assess the clean and robust performance of various structure designs with different robust biases under natural and robust training scenarios. Eventually, we can provide insightful answers to the initial question and offer the following contributions:

- After meticulous analysis and experiments, we demonstrate the effect of two distinct robust biases toward the robustness of Transformer-based structural designs.

- Evaluating various experimental findings leads to a deeper comprehension of robust biases integration characteristics, thereby informing the design of Transformer-based structures. Ultimately, we propose the RBFormer (RBViT/RBVMLP), which exhibits the most robust performance, as depicted in Fig. 2.

- According to the comparison results depicted in Fig. 1, the RBFormer structure outperforms recent popular adopted structures in terms of robustness. Specifically, RBFormer exceeds the original structures by a significant margin of $+16.12\%$ and $+5.04\%$ under various evaluation methods in CIFAR-10 and ImageNet-1k.

2 Exploring Logic and Experimental Setup

As mentioned earlier, this section aims to establish a coherent, logical flow to address the question while outlining the general experimental setup. Firstly, designing Transformer-based structures according to the robust consideration is explored. Then, the original ViT or VMLP structures and our particular implementation are briefly introduced. Finally, we identify convolution operation and multi-hierarchy layer stacking strategy as our two potential available robust biases and further delve into their specific introducing methods.

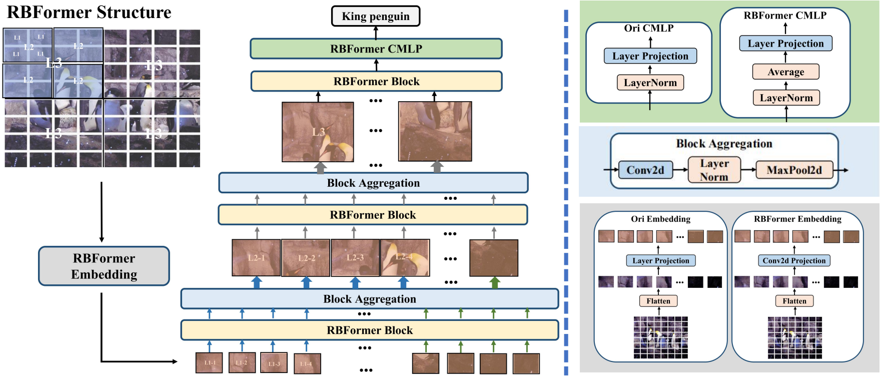

Figure 2: The RBFormer structure can be summarized as follows: (a) The left sub-figure illustrates the general architecture of RBFormer. (b) The right sub-figures depict the specific utilization of the CMLP block, block aggregation, and embedding, showcasing the modifications compared to the original structures

Robust Consideration: The previous research work [39] on the relationship between structural robustness and frequency analysis first inspires us. Through the theoretical analysis of the activation and loss function for the most straightforward classification task and further visual experimental verification, the high-frequency structure could be named the robust bias to facilitate the robustness of Transformer-based models.

Previous works [12, 34] mention that generating more complicated or worse adversarial examples in the inner maximization would be beneficial to promote the final robust performance. About how to improve this inner optimization, [39] explores the robust characteristic from the frequency domain and claims that adversarial training is a process for moving more structured focuses to high-frequency information. Consequently, the target of finding more challenging adversaries is equivalent to facilitating the ability to concentrate features on high-frequency domains.

Adversarial training [17] as a min-max optimization is mainly pursuing adversarial perturbation δ within ℓ_p-ball constraint Δ_p in the inner maximization process. Inverse Discrete Fourier Transform (IDFT) is $x = IDFT[X] = \frac{1}{N}\sum_k XW_N^{-kn}$, $k = 0, ..., N-1$, where $W = e^{-j\frac{2\pi}{N}}$, x and X is the input in the temporal and frequency domain, N is the transform interval length, all of the parameters are positive values. And IDFT is monotonically increasing. Fig. A1 is the activation function (Sigmoid) and loss function (Cross-entropy loss) for the most straightforward two-class classification problem. Since the Sigmod function is also monotonically increasing, when the input of Sigmod moves to frequency values from X_1 to X_2, the output will also change from y_1 to y_2. In the cross-entropy loss of two labels, the possible value range of loss would also extend from l_1 to l_2. Consequentially, when the input frequency X is more toward the high-value region, this simple classification task will result in a broader range of possible loss values like l_1 to l_2. In a word, higher-frequency information exploration will lead to more intricate adversarial examples since the inner max in adversarial training is pursuing higher loss value within ℓ_p-ball constraint.

Recently, there are various works [13, 25, 27, 46] certify that convolution operation could be used to promote the extracting ability of high-frequency information. Thus, we modify the embedding, block aggregation, and CMLP to a new version with convolution operation presented in Fig. 2. This modification could also be illustrated through straightforward tools Fourier Heatmap [39] for the last feature map. The focus of the structure on high-frequency information would increase when the high-light concentration moves closer to the center.

The right subfigure in Fig. 1 is the Fourier heatmap for the original ViT/VMLP (Left Column/Right Column), and the original ViT/VMLP is inserted convolution operation from top to bottom. The structures will focus more on the center or high-frequency domain when improving the proportion of convolution operation. And the following results in Section 3 and the Appendix will also indicate their better robustness. The convolution operation or high-frequency visual structure could be named the robust bias here. Additionally, we term an assumption: *If we continually improve the proportion of this robust bias, would the whole model be more robust?* To validate this assumption, we would figure out which types of robust bias we can choose and how to design the corresponding experiments in the following sections.

Structural Components: According to the previous introduction of RBFormer, all structures could be preliminarily separated into three main components: Embedding, TM block, and CMLP block. The TM block includes the MSA and MLP sub-block. This subsection will present the particular structural composition of ViT/VMLP and our implementation. The detailed theoretical calculation is in Appendix B.2.

Original Components of ViT/VMLP: (1) Embedding: The embedding function is to transform the original images into the embedding tokens. Embedding could be divided into two steps: Step 1 is to execute dimension transform in Eq. S1, and Step 2 in Eq. S2 will add a learnable 1D positional embedding E_{pos}, which contains the positional information under the patch segment phase, to the token vector. Additionally, similar to BERT [6], ViT also adopts [class] token (CLS) to do classification. (2) TM Block: TM Block mixes the embedding tokens to capture the inner features of input images and mainly consists of two sub-blocks. For ViT, the first is the particular MSA sub-block, and the second is the MLP sub-block, which is constituted by two projecting layers with a GELU non-linearity. And VMLP, such as Mixer-MLP [32] and PoolFormer [42], would have two MLP sub-blocks since the original MSA sub-block would be replaced by MLP sub-block Additionally, two crucial training facilitation techniques, LayerNorm (LN) and Skip-connection or Residual (Res), are adopted in both phases here. (3) CMLP Block: CMLP block is the final main component that constitutes two MLP sub-blocks with a GELU non-linearity. This block would also include LN and Res.

Our Implementation: (1) Embedding: Recent experimental findings [3, 20, 33, 47] indicate that removing the CLS token, which is used in the original ViT, can potentially enhance the performance. This kind of removal eliminates unnecessary components and reduces redundant computational costs. Consequently, we directly average (AVG) all token vectors as the input of the TM block. (2) TM Block: In the TM block, the general structure and components of the MSA and MLP sub-blocks will be retained, preserving similarity with the original ViT/VMLP models. However, a modification was made to the Linear Layer within these two sub-blocks. It will be replaced with either Conv2d or Conv1d, depending on the insertion of the convolution operation, to ensure proper dimensional transformation. About LN and Res, [39] mentions that the normalization operation would play a significant role in analyzing adversarial robustness in the frequency domain. Therefore, we will hold the LN as another main research object and put the Res as the ablation study in the Appendix. (3) CMLP Block: As presented in Fig. 2. Since the CLS is removed in embedding, CMLP comprises an AVG pooling layer to do the average of every patch.

Robust Biases Introduction: In this subsection, two potentially available robust biases are declared. One is the convolution operation explained above. The multi-scale hierarchy layer stacking strategy is another implicit robust bias that can facilitate the capturing ability of high-frequency information.

Convolution Operation as a normal replacement of the projection layer used in the ViT/VMLP could be easily introduced into the following components. For embedding, there are two ways of executing it. The first is Convolution Embedding (CONV), which inserts some convolution projection structure before the original embedding. The second Projection Convolution Embedding (PCONV) directly adopts the convolution map to transform the dimension. The MSA sub-block, MLP sub-block, and CMLP block could directly adopt the convolution layer to replace the normal feed forward projection.

Multi-hierarchy Layer Stacking: Four layer stacking strategies inspired by some recent works [15, 16, 36, 47] are explored here and could incorporate convolution operations by using different inserting methods. We introduce them as follows:

I. **The original ViT Structure (OriViT)** directly utilizes the initial ViT structure [8] as the layer stacking strategy. Since the embedding of OriViT is fixed and outputs a 1D token, the PCONV embedding and CONV TM block are unsuitable for OriViT.

II. **CNN-based Structure** would introduce the resolution and channel change process in the multi-hierarchy structure design by imitating the dimension change of CNN [1, 3, 15, 33]. Concretely, the core of this strategy is to directly introduce 2D images as input and keep the 2D dimension in the inner processing step. PCONV embedding, Conv TM block, and CONV MLP block could all be comprehended in this stacking strategy.

III. **Swin-based Structure [16] (Swin)** achieves SOTA performance in various computer vision tasks. Swin modifies the original MSA sub-block to Window-based (WB) MSA and Shifted Window-based (SWB) MSA. WBM is trying to split each patch into a smaller sub-patch further. SWBM would introduce connections across windows and reinforce the ability to use local information. In a word, WB MSA and SWB MSA do not change the essential components of MSA but introduce additional patch splitting and window shift operation. Since the essential components of Swin are just like OriViT except for the TM block, we could introduce convolution operation to each element of it directly. Swin-based VMLP [16] has similar components and convolution operation introducing strategy as Swin-based ViT.

IV. **Image Pyramid Structure (ImagePy)** as another stacking strategy that is inspired by NesT [47]. ImagePy first splits and then aggregates non-overlap image patches in a hierarchy way, and it does not need the cooperation of any component modification. In a word, except for the dimensional transformation limitation of PCONV, ImagePy is very flexible, and the convolution operation could be introduced to components in any way.

3 Experiments and Analysis

In the experiments, we adopt two popular datasets (i.e., CIFAR-10 and ImageNet-1k) to explore more robust structures by gradually increasing the proportion of robust bias. For the specific evaluation, we use ℓ_∞-PGD [17] and Auto-Attack [5], which is an ensemble of white-box and black-box attacks, as the robust validating metrics. In CIFAR-10, we use three ε values for evaluating naturally trained models: $1/255$, $2/255$, and $3/255$. For adversarially trained models, $8/255$ as the most adopted worst value is selected. For ImageNet-1k, we mainly evaluate the structural performance under the robust cases and use $4/255$ as default ε. All models presented in this paper are trained from scratch without any transfer learning strategy. Additionally, when doing the detailed analysis in CIFAR-10, two popular tools (Fourier heatmap [39] and Local Lipschitz [14]) are utilized to help understand the specific impacts of adding robust bias. Fourier heatmap could measure the sensitivity of models when encountering noises in diverse frequency domains. Each pixel of the heatmap is scaled to $[0,1]$ and refers to the error rate after adding frequency noise with the pixel's coordinate as a basis for natural examples. For the Local-Lipschitz constant, as a numerical evaluation index of robustness, the lower value of a structure indicates its smoother and more robust characteristic.

This section is organized as follows: Section 3.1: Experimental setup of adding robust bias structure; Section 3.2: Validate and analyze the particular results and characteristics when inserting different robust bias visual structures to corresponding components. Section 3.3: Compare our RBFormer (RBViT/RBVMLP) with various popular baseline structures.

3.1 Structures with Various Robust Biases

This subsection indicates how to incorporate two robust biases: (1) convolution operation and (2) multi-hierarchy layer stacking strategy to the original ViT/VMLP and finally propose RBFormer (RBViT/RBVMLP). Specifically, for adding convolution operation, according to the explanation of Section 2.1, there would be three components, (1) Embedding, (2) TM Block, including MSA and MLP sub-block, (3) CMLP block, and one technique, (4) LN, that could be inserted. About the multi-hierarchy layer strategy, (1) OriViT [8]; (2) CNN-based structure [10, 33, 36]; (3) Swin [16] and (4) ImagePy [47] would be our research objects. In the specific implementation process, the above-introduced robust biases are not all compatible with each other because of the dimension matching. The concrete component combinations are as follows:

Convolution Operation: To explore the robust influence of inserting convolution operation on each component and the presence or absence of LN, we modify each component according to the presence of convolution-adding degree and LN in Table 1. Additionally, the multi-hierarchy layer stacking strategy, as a structure that can bring that similar utility as the convolution operation, is also included in Table 1. For the specific options: There would be two options in embedding: Original (Ori) and CONV embedding. Since the dimension mismatch issues between CONV TM block/PCONV with the OriViT, we will not consider the CONV TM block and PCONV here. Additionally, accompanied by whether to add the convolution operation, CMLP will have two choices here: Original (Ori) MLP, Convolution (CONV) MLP. Finally, Norm also has two options here: Layernorm (LN) or None. After permutation and combination among all possible choices, the robust performance of Eight structures (a) to (h) are analyzed in Table 1, Fig. 3, and Appendix A

Multi-hierarchy Layer Stacking Robust Bias: After discovering the best structure for introducing convolution operation robust bias, we further focus on multi-hierarchy layer stacking strategies. The robust performance is presented in Table 1, Fig. 3, and Appendix A. I. **OriViT:** We adopt this structure as our basic structure to explore the influence of adding convolution operation in the structure (a) to (h). II. **CNN-based Structure:** For the CNN-based structure, its embedding and TM block should be fixed as PCONV embedding and CONV TM block. For the CMLP block, the original form could be replaced with CONV CMLP to increase its robust bias. Structures (i) and (j) are the evaluation of CNN-based structures. III. **Swin:** The main distinction of Swin is to introduce cyclic shift operation by switching the original TM block to the WB and SWB token-mixer. In this case, the TM block could not be replaced as a CONV block. However, the original embedding and CMLP block could be substituted as PCONV and CONV. The performance of Swin is presented by structures (k) and (l). IV. **ImagePy:** ImagePy is a simple stacking strategy compared with CNN-based structure and Swin without modifying any component. Except for its embedding that is modified to PCONV because of the 2D image dimension, both the form of TM and

ViT/VMLP						Clean Accuracy	PGD (8/255)	Auto-Attack (8/255)	Lipschitz Constant
Components Combine	Embedding	TM	CMLP	Norm	Stacking Structure				
(a)	Ori	Ori	Ori	None	oriViT	79.88/71.06	52.66/45.56	51.12/44.37	159.2/163.2
(b)-Ori	-	-	-	LN	-	79.93/66.38	52.70/44.06	51.45/43.10	157.7/164.7
(c)	-	-	CONV	None	-	82.81/78.83	54.79/54.24	53.83/53.69	151.3/152.3
(d)	-	-	-	LN	-	81.66/77.58	54.69/51.00	53.85/50.88	152.7/157.5
(e)	CONV	-	Ori	None	-	82.77/77.86	55.85/53.22	54.98/52.89	151.4/155.8
(f)	-	-	-	LN	-	80.50/75.92	54.40/50.89	53.69/48.99	153.1/162.3
(g)	-	-	CONV	None	-	80.57/79.25	55.63/53.81	54.23/52.45	146.3/148.5
(h)	-	-	-	LN	-	**82.35/81.42**	**56.41/56.89**	**56.12/57.02**	**140.3/141.3**
(i)-CVT	PCONV	CONV	Ori	-	CNN-based	79.62/77.62	53.15/52.12	52.11/50.21	143.2/146.9
(j)	-	-	CONV	-	-	80.98/79.64	57.67/56.83	57.34/57.06	136.9/138.1
(k)-Swin	Ori	WBM+SWBM	Ori	-	Swin-based	76.39/75.23	48.93/46.34	47.64/45.21	152.0/154.2
(l)	PCONV	-	CONV	-	-	80.08/78.34	52.10/50.92	50.48/50.22	146.3/145.2
(m)-NT	Ori	Ori	Ori	-	ImagePy	76.22/75.94	52.45/51.82	51.28/49.14	158.2/167.9
(n)-RB	PCONV	CONV	CONV	-	-	**83.74/82.19**	**60.91/59.88**	**59.69/59.22**	**89.1/98.7**

Table 1: The performance of our exploring structures under adversarial training in CIFAR-10. All results are shown in **ViT/VMLP accuracy (%)**. This Table includes two robust bias explorations, 1) Structure (a)-(h): the exploration of convolution operation; 2) Structure (h)-(n): the exploration of multi-hierarchy layer stacking strategy. Among (a)-(n), some of them are corresponding to some current typical structures, (b) is the original ViT/VMLP [8, 32] (Ori), (i) is corresponding to CVT [33]/CVT-based VMLP (CVT) or CNN-based structures, (k) is Swin ViT/MLP [16] (Swin), (m) is the NesT [47]/NesT-based VMLP (NT). (n) is our final RBViT/RBMLP (RB).

CMLP block should be similar to OriViT. More detailed information, like layer number selection, distribution of each hierarchy, and others, are all offered in Appendix A.

3.2 Results Analysis

According to the adversarial robust performance of CIFAR-10 and ImageNet-1k in Table 1, Fig. 3 and Table A1, we could conclude some interesting findings about the robust characteristic of transformer-based structures and propose our RBFormer:

After comparing the performance of the (a)-(h) structure in Table 1, we obtain the conclusion that improving the proportion of convolution operation in ViT/VMLP-based models could availably boost the robustness of corresponding models. Undoubtedly, the presence of LN has less impact on robustness. Comparing the results among structure (a) to (h) in CIFAR-10, the robust results of ViT/VMLP under PGD attack and Auto-Attack with $\varepsilon = 4/255$ has at most 3.71%/12.83% and 4.67%/12.92% enhancement. And the Lipschitz-constant value comes into a 17.4/23.4 decrease (the lower, the better in robust case). Additionally, when ignoring the presence of LN, we adopt the Fourier heatmap on structures (b), (d), (f), and (h). As mentioned in Section 1.1, after adding more convolution operations from (b) to (h), the Fourier map concentrates more on the central zone, which indicates this model could capture more high-frequency information. The robust performance also increases from (a) to (h). Therefore, convolution operation as a high-frequency information-capturing structure or robust bias could promote robustness. In ImageNet-1k, the structure (a) and (h) in the left sub-figure in Fig. 3 could also conclude a similar observation. After ensuring the positive effect of adding convolution operation, we target the oriViT, CNN-based, Swin, and ImagePy. The robust influence of these four strategies is our primary purpose in this subsection. According to the PGD attack and Auto-Attack with $\varepsilon = 8/255$, as well as Lipschitz constant value from structures (h) to (n) in Table 1, Fig. 3 and Table A1, we could acquire two observations about changing layer stacking strategy: (1) In each kind of layer stacking strategy, adding convolution operation to any components could generate a positive effect on improving robustness; (2) Not any layer stacking strategy could successfully introduce robust bias to boost robustness, like Swin in structures (k) and (l). The suitable layer stacking strategy should be a significant consideration when designing transformer-based structures. We also utilize the Fourier heatmap to analyze the characteristics of four strategies with the highest proportion of convolution operation (structure h, j, l, and n). According to the experimental results in CIFAR-10, following the partial enhancement of clean accuracy, structure (n) could generate 8.21%/15.82, 8.24%/16.12% robust accuracy improvement under PGD adversarial examples and Auto-Attack, 27.0/30.2 decrease in Lipschitz constant, and concentrating more on the central zone in the Fourier heatmap as Fig. 3. In ImageNet-1k, compared with the original ViT/VMLP (b), structure (n) could generate 4.13%/4.66%, 4.03%/5.04% improvement under PGD attack and Auto-Attack. Therefore, structure (n) would be our target RBFormer (RBViT/RBVMLP).

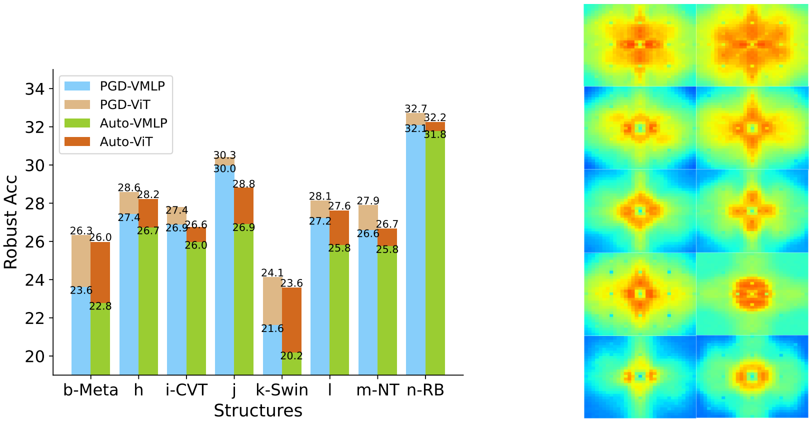

Figure 3: Left: ImageNet-1k results for ViT/VMLP structure. PGD-VMLP/PGD-ViT and Auto-VMLP/Auto-ViT represent the robust accuracy under PGD and auto ($\varepsilon = 8/255$) attack. Right: Fourier heatmaps of Ori ViT/VMLP, some representative structures with convolution operation, and our RBFormer (b, h, j, l, n) from top to bottom.

3.3 Comparison Results of RBFormer with SOTA Baselines

In this section, we mainly focus on the robust enhancement of our RBViT/RBVMLP (n) with some popularly used ViT [8]/Mixer-MLP [32] (b), CVT/CVT-based VMLP [33] (i), Swin ViT/VMLP [16] (k), NesT/NesT-based VMLP [47] (m) in Table 1, Fig. 3 and Table A1. We first use PGD and Auto-Attack to compare our RBViT/RBVMLP with the other ViT/VMLP-based structures. Among these four values, our RBViT/RBVMLP could achieve at most 16.12%, 9.01%, 14.01%, and 9.56% improvement for CIFAR-10. And for ImageNet-1k in Fig. 3 and Table A1, RBViT/RBVMLP could attain at most 5.04%, 5.04%, 11.59%, and 6.02 % enhancement. Additionally, in CIFAR-10, we also further adopt Lipschitz Constant (Lower values mean better robustness) to evaluate the robustness after comparing RBViT/RBVMLP (n) with those four ViT/VMLP-based structures (b), (i), (k), and (m), it can earn at most 68.6, 54.1, 62.9, 69.2 value decrease that means better robustness. We also adopt the left figure in Fig. 1 to illustrate the superiority of our RBFormer compared with ViT, CeiT [45], Local-ViT [15], Mixer-MLP, NesT, CVT, Swin Transformer and Swin MLP [16]. Our RBFormer (RBViT/RBMLP) could obtain the best robust performance compared with other baseline models. Additionally, in contrast to [19, 24] that claim to specialize in improving the general robustness of Transformer-based structures, Table 2 shows that RBFormer also maintains performance advantages in both clean and adversarial cases.

Metric (%)	ViT/VMLP			
	CIFAR-10		ImageNet-1k	
	Clean Acc	Adv Acc	Clean Acc	Adv Acc
Mao et al. [19]	83.13/**83.44**	52.79/47.13	60.46/58.81	24.23/21.67
Qin et al. [24]	82.76/81.17	53.61/45.65	61.14/59.43	26.55/24.52
RBFormer	**83.74**/82.19	**60.91/59.88**	**61.59/60.27**	**32.71/32.09**

Table 2: Comparing RBFormer with benchmarks under clean and adversarial case in CIFAR-10 and ImageNet-1k

3.4 The Affinity for Sparse Algorithms

RBFormer is realized through the process of rational component analysis and structure redesign, but instead, simply adding new parameters to increase the redundancy. This process involves only some modification in the dimensional transformation operation and does not significantly introduce computational complexity. Additionally, RBFormer could adopt various model compression methods, such as model pruning, quantization, and sparse training, to further reduce our parameter numbers and model size without sacrificing robust performance. The most straightforward and least technically advantageous irregular magnitude pruning is adopted in Fig. 4, the robust performance could be maintained under a low percentage of remaining non-zero weights.

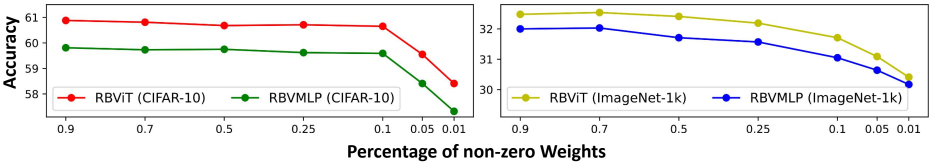

Figure 4: Applying irregular pruning to RBViT/RBVMLP in CIFAR-10 (Left) and ImageNet-1k (Right).

4 Conclusion

In this paper, we perform an exhaustive examination of the critical components that notably influence the performance of Transformer-based structures. Our analysis systematically explores the impact of each component on robust vulnerability by experimenting with various combinations. Furthermore, we delve into a study of adversaries in the frequency domain, identifying robust biases that could potentially enhance adversarial robustness. Our proposed RBFormer integrates a carefully selected mix of these robust biases. Through rigorous experimental validation, we affirm that the RBFormer surpasses robust SOTA baselines, including currently prevalent structures and some methods that could be used to enhance general robustness. Therefore, introducing robust biases leads to a noticeable enhancement in overall performance, as evidenced by our study results.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6836–6846, 2021.

[2] Philipp Benz, Soomin Ham, Chaoning Zhang, Adil Karjauv, and In So Kweon. Adversarial robustness comparison of vision transformer and mlp-mixer to cnns. *The British Machine Vision Conference (BMVC)*, 2021.

[3] Shuo Chen, Tan Yu, and Ping Li. Mvt: Multi-view vision transformer for 3d object recognition. *The British Machine Vision Conference (BMVC)*, 2021.

[4] Hao Cheng, Kaidi Xu, Zhengang Li, Pu Zhao, Chenan Wang, Xue Lin, Bhavya Kailkhura, and Ryan Goldhahn. More or less (mol): Defending against multiple perturbation attacks on deep neural networks through model ensemble and compression. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 645–655. IEEE, 2022.

[5] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning (ICLR)*, pages 2206–2216. PMLR, 2020.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pretraining of deep bidirectional transformers for language understanding. *North American Chapter of the Association for Computational Linguistics(NACCL)*, 2019.

[7] Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International Conference on Machine Learning (ICML)*, pages 2793–2803. PMLR, 2021.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*, 2021.

[9] Jinhao Duan, Quanfu Fan, Hao Cheng, Xiaoshuang Shi, and Kaidi Xu. Improve video representation with temporal adversarial augmentation. *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2023.

[10] Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *International Conference on Machine Learning (ICML)*, pages 2286–2296. PMLR, 2021.

[11] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 244–253, 2019.

[12] Sven Gowal, Krishnamurthy Dj Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. Scalable verified training for provably robust image classification. In *Proceedings of the*

IEEE/CVF International Conference on Computer Vision (ICCV), pages 4842–4851, 2019.

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

[14] Hanxun Huang, Yisen Wang, Sarah Erfani, Quanquan Gu, James Bailey, and Xingjun Ma. Exploring architectural ingredients of adversarially robust deep neural networks. *Advances in Neural Information Processing Systems (NeruIPS)*, 34, 2021.

[15] Yawei Li, Kai Zhang, Jiezhang Cao, Radu Timofte, and Luc Van Gool. Localvit: Bringing locality to vision transformers. *arXiv preprint arXiv:2104.05707*, 2021.

[16] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021.

[17] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations (ICLR)*, 2018.

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

[19] Chengzhi Mao, Lu Jiang, Mostafa Dehghani, Carl Vondrick, Rahul Sukthankar, and Irfan Essa. Discrete representations strengthen vision transformer robustness. *International conference on machine learning (ICLR)*, 2022.

[20] Xiaofeng Mao, Gege Qi, Yuefeng Chen, Xiaodan Li, Ranjie Duan, Shaokai Ye, Yuan He, and Hui Xue. Towards robust vision transformer. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

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

[22] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3163–3172, 2021.

[23] Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. *Association for the Advancement of Artificial Intelligence (AAAI)*, 2021.

[24] Yao Qin, Chiyuan Zhang, Ting Chen, Balaji Lakshminarayanan, Alex Beutel, and Xuezhi Wang. Understanding and improving robustness of vision transformers through patch-based negative augmentation. *Advances in Neural Information Processing Systems*, 35:16276–16289, 2022.

[25] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021.

[26] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14335–14345, 2021.

[27] Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. On the adversarial robustness of vision transformers. *arXiv preprint arXiv:2103.15670*, 2021.

[28] Xiaoshuang Shi, Fuyong Xing, Kaidi Xu, Pingjun Chen, Yun Liang, Zhiyong Lu, and Zhenhua Guo. Loss-based attention for interpreting image-level prediction of convolutional neural networks. *IEEE Transactions on Image Processing*, 30:1662–1675, 2020.

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

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

[31] Shiyu Tang, Ruihao Gong, Yan Wang, Aishan Liu, Jiakai Wang, Xinyun Chen, Fengwei Yu, Xianglong Liu, Dawn Song, Alan Yuille, et al. Robustart: Benchmarking robustness on architecture design and training techniques. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TAPMI)*, 2021.

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

[33] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22–31, 2021.

[34] Cihang Xie, Mingxing Tan, Boqing Gong, Alan Yuille, and Quoc V Le. Smooth adversarial training. *arXiv preprint arXiv:2006.14536*, 2020.

[35] Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang, and Xue Lin. Adversarial t-shirt! evading person detectors in a physical world. In *European conference on computer vision (ECCV)*, pages 665–681. Springer, 2020.

[36] Yufei Xu, Qiming Zhang, Jing Zhang, and Dacheng Tao. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021.

[37] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5791–5800, 2020.

[38] Shaokai Ye, Kaidi Xu, Sijia Liu, Hao Cheng, Jan-Henrik Lambrechts, Huan Zhang, Aojun Zhou, Kaisheng Ma, Yanzhi Wang, and Xue Lin. Adversarial robustness vs. model compression, or both? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 111–120, 2019.

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

[40] Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021.

[41] Tan Yu, Xu Li, Yunfeng Cai, Mingming Sun, and Ping Li. Rethinking token-mixing mlp for mlp-based vision backbone. *The British Machine Vision Conference (BMVC)*, 2021.

[42] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[43] Chenxi Yuan and Mohsen Moghaddam. Attribute-aware generative design with generative adversarial networks. *IEEE Access*, 8:190710–190721, 2020.

[44] Chenxi Yuan, Jinhao Duan, Nicholas J Tustison, Kaidi Xu, Rebecca A Hubbard, and Kristin A Linn. Remind: Recovery of missing neuroimaging using diffusion models with application to alzheimer's disease. *medRxiv*, pages 2023–08, 2023.

[45] Kun Yuan, Shaopeng Guo, Ziwei Liu, Aojun Zhou, Fengwei Yu, and Wei Wu. Incorporating convolution designs into visual transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 579–588, 2021.

[46] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *The British Machine Vision Conference (BMVC)*, 2016.

[47] Zizhao Zhang, Han Zhang, Long Zhao, Ting Chen, and Tomas Pfister. Aggregating nested transformers. *arXiv preprint arXiv:2105.12723*, 2021.