Appendix

A Experimental Details and Additional Restuls

A.1 Experimental Detailed Setup

About the particular layer number selection, referring to some current widely used multi-hierarchy structures Swin [16], NesT [47], CVT [33] and etc [15, 42, 45], we adopt 12 as our total number of layers. The simplest OriViT could directly set the layer number to be 12. For a multi-hierarchy stacking structure, we adopt a three-stage hierarchy for CNN-based structures with $\{1,2,9\}$ for each stage. However, since the modification of Swin is on MSA and does not directly introduce convolution operation, We would choose four steps with $\{2,2,6,2\}$ distribution for it. For the ImagePy, the hierarchy distribution could be $\{2,2,8\}$ to maintain the properties of pyramid structures. About the training phase, in CIFAR-10, we respectively validate different structures under natural and adversarial training cases. About ImageNet-1k, we mainly focus on exploring the robustness of adversarial training cases. To generate adversarial examples in CIFAR-10, about natural case, we adopt the $\varepsilon = 1/255, 2/255, 3/255$ with iteration = 10 and step size = 0.01 to attack the original evaluation model its robustness. In the adversarial case, we adopt the $\varepsilon = 8/255$ with the same iteration and step size to generate adversarial examples and do training. For ImageNet-1k, adversarial examples are generated by using $\varepsilon = 4/255$ with iteration = 3 and step size = $2\varepsilon/3$ referring to [31]. For the specific experimental results, Table A1 is the exact values of Fig. 3. Additionally, according to [7] and further study above, the structural design of ViT is the most critical point to achieve better performance. The abuse of the attention block will make the overall structure collapse to the rank-1 matrix. For alleviating this collapse, skip connections are very crucial. MLP could help, but LN plays no role. To verify the impact of this discovery on the design of robust transformer-based structures, we further explore the effect of Skip-Connection or Res as a facilitation technique towards robustness in the natural trained case as Table A2 and adversarial training case as Table A3.

A.2 Experimental Analysis

About the performance effect of skip-connection, MLP, and LN in Transformer[7], the experimental results in Table. A3 and Table. A2 show that the independent existence of skip-connection and MLP has a minor influence on the robustness. On the contrary, the effect of LN seems complex. When just removing LN, there is no performance change or even a little increase in some cases like (5)-(7), (9)-(11), (13)-(15). After removing the skip-connection or MLP, further removing LN will make the structure not converge. For a more detailed statement, there will be the following changes about removing the LN layer:

I. According to the structures (1)-(3), (21)-(23), the performance of these component combinations without LN only has a minor drop compared with the original one;

II. According to structures (5)-(7), (9)-(11), (13)-(15), (17)-(19), some structures could have a better robust performance after removing the LN.

In a word, for the transformer-based structures in CIFAR-10, the existence of the Norm will not help the robustness, but removing it could even promote robustness a little in some cases. Furthermore, in ImageNet-1k, all structures would not converge after removing LN. Sequentially, we could acquire that the LN will 1) play no role or be little harmful to the robust performance of MetFormer and RBFormer structure with small training tasks (small model size or training datasets); 2) guarantee the training convergence under the large model size and extensive datasets. Consequently, based on the analysis above, we would still keep LN in our RBFormer to guarantee successful convergence in our training phase.

B Mathematical Analysis

B.1 ViT Structures Expression

The mathematical representation of ViT structure:

$$\begin{aligned}\mathbf{X}_p =& \mathcal{F}_{\mathrm{DT}}(\mathbf{X}) = [x_p^1; x_p^2; \ldots; x_p^N], \\ & \mathbf{X} \in \mathbb{R}^{H\times W\times C}, \mathbf{X}_p \in \mathbb{R}^{N\times(P^2\cdot C)}\end{aligned} \tag{S1}$$

$$\begin{aligned}\mathbf{Z}_0 =& \mathbf{X}_p + \mathbf{E}_{pos} = [x_p^1 e; x_p^2 e; \ldots; x_p^N e], \\ & \mathbf{Z}_0 \in \mathbb{R}^{N\times(P^2\cdot C)}, \mathbf{E}_{pos} \in \mathbb{R}^{N\times(P^2\cdot C)}\end{aligned} \tag{S2}$$

$$\begin{aligned}\mathbf{Z}_l^{'} =& Res(\mathcal{F}_{\mathrm{TM}}(\mathcal{F}_{\mathrm{LN}}(\mathbf{Z}_{l-1}))), \\ & l = 1, \ldots, L, \quad \mathbf{Z}_l^{'} \in \mathbb{R}^{N\times(P^2\cdot C)}\end{aligned} \tag{S3}$$

$$\begin{aligned}\mathbf{Z}_l =& Res(\mathcal{F}_{\mathrm{MLP}}(\mathcal{F}_{\mathrm{LN}}(\mathbf{Z}_l^{'}))), \\ & l = 1, \ldots, L, \quad \mathbf{Z}_l \in \mathbb{R}^{N\times(P^2\cdot C)}\end{aligned} \tag{S4}$$

$$\mathbf{y} = \mathcal{F}_{\mathrm{CMLP}}(\mathcal{F}_{\mathrm{AVG}}(Z_L)) \tag{S5}$$

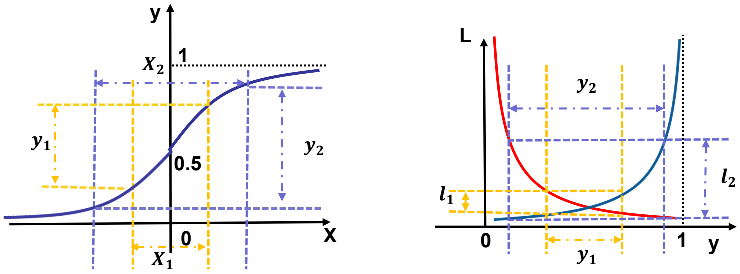

Figure A1: The activation function (left) and loss function (right) map.

The embedding function transforms the original images into the embedding tokens. These generating tokens are corresponding processing objects of the following TM block. Embedding could be divided into two steps. Step 1 is adopted to execute dimension transform in Eq. S1, which could reshape 2D image $x \in \mathbb{R}^{H\times W\times C}$ to a 1D sequence through flattening original 2D patches $x_p \in \mathbb{R}^{N\times(P^2\cdot C)}$. (H,W) is the resolution of the original image, C is the channel number, P is the patch size. And the number of patches will be $N = HW/P^2$. After finishing Step 1, Step 2 in Eq. S2 will add a learnable 1D positional embedding $E_{pos} \in \mathbb{R}^{N\times(P^2\cdot C)}$ to the token vector. The introduction of this positional embedding is due to the loss of related positional information under the patch segmenting phase. Additionally, similar to BERT's [6] [class] token (CLS), ViT also adopts this component to do classification. After generating embedding token $\mathbf{Z}_0$, the next step of the transformer-based structure is to use TM Block to mix those embedding tokens and make the most significant efforts to capture the inner features. As we know before, two sub-blocks mainly constitute the TM Block. The first one is the particular *MSA sub-block* (Eq. S3), and the second is the *MLP sub-block* constituted by two projecting layers with a GELU non-linearity. For the VMLP structures, like the Mixer-MLP [32] and PoolFormer [42], they adopt *MLP sub-block* to replace the initial *MSA sub-block* and thus have two *MLP sub-blocks*. Apart from these two main constituting sub-blocks, Layernorm (LN) and Skip-connection, or Residual (Res), are also adopted in both phases as presented in Eq. S3 and S4. CMLP block is the final main component in Eq. S5 and constitutes two MLP sub-blocks with a GELU non-linearity.

B.2 The Detailed Analysis of Robust Consideration

According to the equation Eq. S1 to Eq. S5 and Fig. A1, we could give a more detailed explanation of the rationalization of robust bias. Inspired by [12, 34, 39], we know that adversarial training leads to the enhancement of model robust through learning adversarial examples generated from the inner high-frequency visual structures. Additionally, the inner maximization process could find more challenging adversarial examples by increasing the proportion of high-frequency structure, and the convolution operation would be a kind of good high-frequency visual structure. Therefore, we further name the convolution operation as robust bias since it could influence the adversarial robustness after adversarial training by modifying its proportion in a whole structure. Apart from using the experimental evaluation to certify the effect of robust bias, we also adopt a simple mathematical analysis here. Fig. A1 is the activation function (Sigmoid as Eq. S6) and loss function (Cross-entropy loss as Eq. S7) for the most straightforward two-class classification problem. Since the Sigmod function is monotonically increasing, when the input of Sigmoid 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. Furthermore, since the inner max in adversarial training is pursuing higher loss value within ℓ_p-ball constraint, higher-frequency information exploration will finally lead to more challenging adversarial examples. Eventually, these harder adversarial examples would facilitate the process of adversarial training and make our proposed robust bias effectively increase the final structure robustness.

$$y(x) = Sigmoid(x) = \frac{2}{1+e^{-x}} \tag{S6}$$

$$L(x) = -[ylog(\hat{y}) + (1-y)log(1-\hat{y})] \tag{S7}$$

ViT/VMLP						Clean Accuracy	PGD (4/255)	Auto-Attack (4/255)
Components Combine	Embedding	TM	CMLP	Norm	Stacking Structure			
(b)-Ori	Ori	Ori	Ori	LN	oriViT	58.63/57.93	26.32/23.62	25.98/22.79
(h)	CONV	-	CONV	-	-	61.25/61.12	28.58/27.43	28.22/26.74
(i)-CVT	PCONV	CONV	Ori	-	CNN-based	57.43/56.97	27.40/26.89	26.64/25.98
(j)	-	-	CONV	-	-	61.07/60.61	30.28/29.99	28.77/26.91
(k)-Swin	Ori	WBM+SWBM	Ori	-	Swin-based	59.74/58.98	24.12/21.62	23.58/20.19
(l)	PCONV	-	CONV	-	-	62.37/61.45	28.13/27.24	27.62/25.83
(m)-NT	Ori	Ori	Ori	-	ImagePy	58.37/57.98	27.89/26.59	26.67/25.76
(n)-RB	PCONV	CONV	CONV	-	-	**61.59/60.27**	**32.71/32.09**	**32.25/31.78**

Table A1: The robust performance of our proposed representative structures in ImageNet-1k. All results are shown in **ViT accuracy (%)/VMLP accuracy (%)**. (b) is the original ViT/VMLP [8, 32] (Ori), (h) is the oriViT structure with the most convolution operation, (i) is corresponding to CVT [33]/CVT-based VMLP (CVT) or CNN-based structures, (j) is the CNN-based structure with the most convolution operation, (k) is Swin ViT/MLP [16] (Swin), (l) is the Swin-based structure with the most convolution operation, (m) is the NesT [47]/NesT-based VMLP (NT), and (n) is our final RBViT/RBMLP (RB).

ViT/VMLP						Clean Accuracy	PGD			Auto-Attack		
Elements	Embedding	TM	CMLP	Norm	Skip-Connect		1/255	2/255	3/255	1/255	2/255	3/255
(1)	Ori	Ori	Ori	LN	Res	85.7/81.4	57.1/40.0	52.70/44.06	52.70/44.06	52.70/44.06	52.70/44.06	52.70/44.06
(2)	-	-	-	-	NONE	72.5/80.5	36.9/40.3	11.6/19.0	2.1/7.4	36.7/39.6	10.5/17.6	1.5/6.5
(3)	-	-	-	NONE	Res	84.3/82.1	57.3/44.9	24.5/16.7	8.2/4.9	54.7/43.5	23.4/15.4	6.8
(4)	-	-	-	-	NONE	10/10	10/10	10/10	10/10	10/10	10/10	10/10
(5)	-	-	NONE	LN	Res	78.2/61.4	40.6/39.7	17.8/20.3	4.8/8.4	37.8/38.5	16.5/18.7	3.2/6.9
(6)	-	-	-	-	NONE	70.34/59.4	33.4/39.2	20.3/21.6	5.1/9.9	32.3/38.2	16.4/20.2	7.1/9
(7)	-	-	-	NONE	Res	79.47/61.7	46.8/43.8	27.4/25.8	15.7/13	45.9/42.2	23.5/21.7	10.1/7.2
(8)	-	-	-	-	NONE	45.38/58.9	31.2/41.9	14.5/25.7	6.9/13.5	34.1/39.5	22.5/21.9	9.6/8.7
(9)	-	-	CONV	LN	Res	87.4/85.0	59.8/55.2	27.7/26.2	10.5/9.5	57.8/53.7	25.4/24.6	8.9/7.9
(10)	-	-	-	-	NONE	79.36/10	41.6/10	22.3/10	9.8/10	53.4/51.7	23.7/21.2	8.9/10.2
(11)	-	-	-	NONE	Res	87.4/85.2	59.8/58.9	32.9/31.2	11.2/9.5	53.4/51.7	22.8/21.2	9.2/6.4
(12)	-	-	-	-	NONE	10/10	10/10	10/10	10/10	10/10	10/10	10/10
(13)	CONV	-	Ori	LN	Res	88.1/84.6	46.4/41.9	13.8/10.9	3.0/2.0	42.4/40.1	11.3/9.1	2.4/1.7
(14)	-	-	-	-	NONE	75.8/72.8	19.8/23.1	2.2/4.0	0.2/0.5	18.6/20.2	1.5/2.3	0/0
(15)	-	-	-	NONE	Res	89.2/84.7	56.6/50.7	21.2/17.9	4.9/4.2	55.3/38.2	17.1/16.4	2.7/2.1
(16)	-	-	-	-	NONE	10/10	10/10	10/10	10/10	10/10	10/10	10/10
(17)	-	-	NONE	LN	Res	79.0/68.7	30.6/20.5	5.4/3.1	0.6/0.2	28.7/22.4	6.7/5.4	1.6/1.2
(18)	-	-	-	-	NONE	73.8/67.4	17.1/21.2	1.0/3.2	0.6/0.3	15.4/19.2	1.2/0.6	0/0
(19)	-	-	-	NONE	Res	80.9/68.4	45/38.4	14.5/14.0	2.6/3.4	43.1/36.7	13.2/12.8	2.1/3.2
(20)	-	-	-	-	NONE	76.0/68.4	40.8/35.8	11.8/12.5	1.9/2.7	38.4/34.2	8.4/7.9	1.2/2.4
(21)	-	-	CONV	LN	Res	89.8/88.1	59.2/59.9	24.0/27.9	6.8/9.6	58.7/58.4	22.4/21.9	5.9/6.0
(22)	-	-	-	-	NONE	82.6/10	34.6/10	10.2/6.3	4.3/0	33.4/10	10.0/5.4	3.3/0
(23)	-	-	-	NONE	Res	89.7/88.1	58.2/49.9	17.5/17.2	4.3/5.2	57.2/48.3	16.3/15.9	4.1/3.9
(24)	-	-	-	-	NONE	10/10	10/10	10/10	10/10	10/10	10/10	10/10

Table A2: Results of the proposed 24 naturally trained structures. All results are shown in **ViT accuracy (%)/VMLP accuracy (%)**. All structures in this table only do not refer to various multi-hierarchy layer stacking and all keep to be OriViT. The specific elements include (1) Embedding (ori/CONV); (2) TM block (Ori); (3) CMLP block (Ori/CONV/NONE); (4) Norm (LN/NONE), and 5)Skip-Connection (Res/NONE).

ViT/VMLP						Clean Accuracy	PGD (8/255)	Auto-Attack (8/255)	Lipschitz Constant
Stru-ctures	Emb-edding	TM	CMLP	Norm	Skip-Connect				
(1)	Ori	Ori	Ori	LN	Res	79.93/66.38	52.70/44.06	51.45/43.10	157.7/164.7
(2)	-	-	-	-	NONE	52.59/53.14	37.10/37.63	35.89/35.66	169.3/168.2
(3)	-	-	-	None	Res	79.88/71.06	52.66/45.56	51.12/44.37	159.2/163.2
(4)	-	-	-	-	NONE	10/10	10/10	10/10	0/0
(5)	-	-	NONE	LN	Res	55.80/49.22	38.69/36.35	37.45/34.98	167.4/172.3
(6)	-	-	-	-	NONE	56.45/50.34	39.26/36.15	39.56/34.96	165.1/173.9
(7)	-	-	-	NONE	Res	58.88/49.89	42.25/36.35	41.76/35.89	152.4/173.3
(8)	-	-	-	-	NONE	46.79/48.70	34.99/36.71	34.14/35.34	180.6/170.3
(9)	-	-	CONV	LN	Res	81.66/77.58	54.69/51.00	53.85/50.88	152.7/157.5
(10)	-	-	-	-	NONE	10/10	10/10	10/10	0/0
(11)	-	-	-	NONE	Res	82.81/78.83	54.79/54.24	53.83/53.69	151.3/152.3
(12)	-	-	-	-	NONE	10/10	10/10	10/10	0/0
(13)	CONV	-	Ori	LN	Res	80.50/75.92	54.40/50.89	53.69/48.99	153.1/162.3
(14)	-	-	-	-	NONE	64.75/62.08	44.49/42.14	43.47/41.81	163.2/166.7
(15)	-	-	-	NONE	Res	82.77/77.86	55.85/53.22	54.98/52.89	151.4/155.8
(16)	-	-	-	-	None	10/10	10/10	10/10	0/0
(17)	-	-	NONE	LN	Res	71.12/56.11	48.35/41.13	47.78/39.89	161.3/168.2
(18)	-	-	-	-	NONE	60.01/54.39	42.40/39.61	41.56/38.44	167.4/171.1
(19)	-	-	-	NONE	Res	75.01/56.25	52.92/39.81	52.10/38.96	158.7/166.9
(20)	-	-	-	-	NONE	64.61/55.80	46.80/41.31	46.90/40.65	163.4/161.5
(21)	-	-	CONV	LN	Res	**82.35/81.42**	**56.41/56.89**	**56.12/57.02**	**140.3/141.26**
(22)	-	-	-	-	NONE	15.53/10	10/10	10/10	0/0
(23)	-	-	-	NONE	Res	80.57/79.25	55.63/53.81	54.23/52.45	146.3/148.5
(24)	-	-	-	-	NONE	10/10	10/10	10/10	0/0

Table A3: Results of the proposed 24 adversarially trained structures in ViT/VMLP. All results are shown in **ViT accuracy (%)/VMLP accuracy (%)**. All structures in this table only do not refer to various multi-hierarchy layer stacking, and all keep to be OriViT. The specific elements include (1) Embedding (ori/CONV), (2) TM block (Ori), (3) CMLP block (Ori/CONV/NONE MLP); (4) Norm (LN/NONE), and (5)Skip-Connection (Res/NONE).