Appendix

A. Additional Results of Robustness Analysis on Designed Components

Here we will show the remaining results of robustness analysis in section **??**. As the each component has been discussed detailly, we only give a summary of the results in appendix. We report the additional results of robustness analysis in Table 2, 1, 4, 3, 5 and 6 respectively, where each table presents the results of one or some components. The detailed architecture of models used in robustness analysis on stage distribution is shown in Table 7. Although each robustness benchmark is consistent on the overall trend, we still find some special cases. For example, in Table 4, the V6 version of stage distribution poorly performs on adversarial robustness, but achieves best results on IN-A and IN-R datasets, showing the superior generalization power. Another case is the token-to-token embedder in Table 3. Compared with original linear embedder, token-to-token embedder obtains better results on IN-C, IN-A, IN-R and IN-SK datasets. However, under PGD attacker, it only gets the robust accuracy of 4.7%. The above phenomenon also indicates that using only several robustness benchmarks is biased and cannot get a comprehensive assessment result. Therefore, we advocate that the works about model robustness in future should consider multiple benchmarks. For validating the generality of the proposed techniques, we show the robustness evaluation results when trained on other ViT architectures and larger datasets (ImageNet-22k) in Table 5 and 6.

Heads	1	2	4	6	8	12
Acc	69.0	71.7	73.1	73.4	**73.9**	73.5
FGSM	17.6	21.4	22.8	24.6	**25.2**	24.7
PGD	4.3	6.1	7.1	7.7	**8.2**	8.0
IN-C (↓)	79.5	72.9	69.0	68.5	**67.7**	68.2
IN-A	5.1	6.9	8.2	8.3	**8.9**	8.4
IN-R	28.1	32.9	33.9	34.1	**34.2**	33.7
IN-SK	15.9	20.4	21.4	21.6	**22.0**	21.1

Table 1. Additional results of robustness analysis on different head number.

B. Feature Visualization

In general understanding, intra-class compactness and inter-class separability are crucial indicators to measure the effectiveness of a model to produce discriminative and robust features. We use t-Distributed Stochastic Neighbor Embedding (t-SNE) to visualize the feature sets extracted by ResNet50, DeiT-Ti, Swin-T, PVT-Ti and our RVT respectively. The features are produced on validation set of ImageNet and ImageNet-C. We randomly selected 10 classes for better visualization. As shown in Figure 1, features extracted by our RVT is the closest to the intra-class compactness and inter-class separability. It's confirmed from the side that our RVT does have the stronger robustness and classification performance.

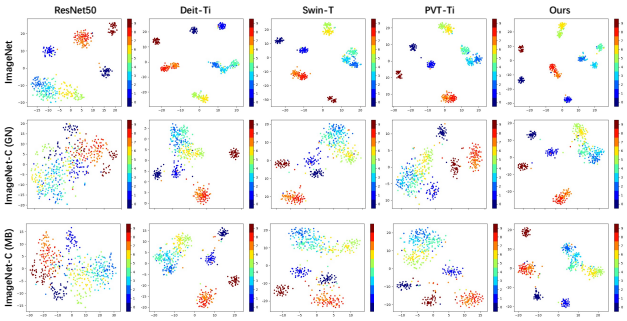

Figure 1. **t-SNE** visualization of features produced by different models.

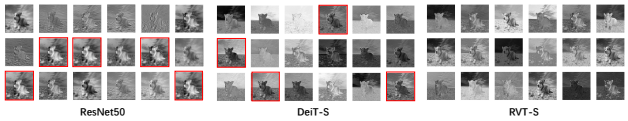

Figure 2. Feature visualization of ResNet50, DeiT-S and our proposed RVT-S trained on ImageNet. Red boxes highlight the feature maps with high similarity.

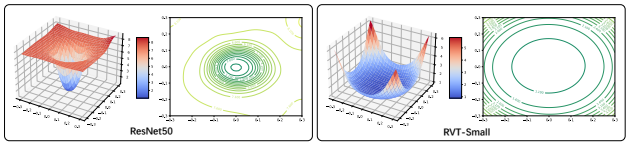

Figure 3. Loss landscape of ResNet50 and RVT-S.

We also visualize the feature maps of ResNet50, DeiT-S and our proposed RVT-S in Figure 2. Visualized features are extracted on the 5th layer of the models. The result shows ResNet50 and DeiT-S contain a large part of redundant features, highlighted by red boxes. While our RVT-S reduces the redundancy and ensures the diversity of features, reflecting the stronger generalization ability.

C. Loss Landscape Visualization

Loss landscape geometry has a dramatic effect on generalization and trainability of the model. We visualize the loss surfaces of ResNet50 and our RVT-S in Figure 3. RVT-S has a flatter loss surfaces, which means the stability under input changes.

Position Embeddings	Acc	FGSM	PGD	IN-C (↓)	IN-A	IN-R	IN-SK
none	68.3	15.8	3.6	82.4	5.2	24.3	12.0
learned absolute Pos.	72.2	**22.3**	**6.2**	**71.1**	7.3	**32.6**	**20.3**
sin-cos absolute Pos.	72.0	21.9	5.9	71.9	7.0	31.4	20.2
learned relative Pos. [**?**]	71.8	22.3	6.1	71.6	**7.6**	32.5	18.6
input-conditioned Pos. [**?**]	**72.4**	21.5	5.3	72.5	6.8	31.0	18.0

Table 2. Additional results of robustness analysis on different position encoding methods.

Patch Emb.			Local	Conv.						
Linear	Conv.	T2T	SA	FFN	CLS	PGD	IN-C (↓)	IN-A	IN-R	IN-SK
✓					✓	6.2	71.1	7.3	32.6	20.3
	✓				✓	6.8	69.2	8.3	33.6	21.1
		✓			✓	4.7	69.6	10.1	36.7	23.8
✓			✓		✓	9.0	76.9	4.8	28.7	16.6
✓				✓	✓	12.7	65.0	8.4	39.0	31.9
✓						12.0	70.0	7.4	32.5	20.2

Table 3. Additional results of robustness analysis on different patch embeddings, locality of attention, convolutional FFN and the replacement of CLS token.

Var.	[S_1, S_2, S_3, S_4]	Acc	FGSM	PGD	IN-C (↓)	IN-A	IN-R	IN-SK
V1	[0, 0, 12, 0]	72.2	22.3	6.2	71.1	7.3	32.6	20.3
V2	[0, 0, 10, 2]	74.8	**24.3**	**6.8**	**66.9**	8.8	35.5	21.9
V3	[0, 2, 10, 0]	73.8	22.0	5.1	76.4	8.2	33.6	21.1
V4	[0, 2, 8, 2]	76.4	22.3	4.5	71.5	10.3	36.8	**23.9**
V5	[2, 2, 8, 0]	73.4	17.0	2.3	76.8	9.0	33.2	20.7
V6	[2, 2, 6, 2]	**76.4**	17.5	1.9	71.6	**11.2**	**36.8**	23.1

Table 4. Additional results of robustness analysis on stage distribution.

Models	Acc	FGSM	PGD	IN-C (↓)	IN-A	IN-R	IN-SK
DeiT-Ti	72.2	22.3	6.2	71.1	7.3	32.6	20.2
DeiT-Ti*	74.4	29.9	9.1	67.9	8.1	34.9	23.1
ConViT-Ti	73.3	24.7	7.5	68.4	8.9	35.2	22.4
ConViT-Ti*	74.4	30.7	9.6	65.6	9.4	37.0	25.2
PiT-Ti	72.9	20.4	5.1	69.1	6.2	34.6	21.6
PiT-Ti*	74.3	27.7	7.9	66.7	7.1	36.6	24.0
DeiT-S	79.9	40.7	16.7	54.6	18.9	42.2	29.4
DeiT-S*	80.6	42.3	18.8	53.1	20.5	43.5	31.3
ConViT-S	81.5	41.0	17.2	49.8	24.5	45.4	33.1
ConViT-S*	81.8	42.3	18.7	49.1	25.6	46.1	34.2
PiT-S	80.9	41.0	16.5	52.5	21.7	43.6	30.8
PiT-S*	81.4	42.2	18.3	51.4	23.3	44.6	32.3

Table 5. Additional results of position-aware attention scaling and patch-wise augmentation on other ViT architectures.

Models	Acc	FGSM	PGD	IN-C (↓)	IN-A	IN-R	IN-SK
DeiT-B	83.20	47.21	24.89	45.50	38.01	52.37	**39.54**
RVT-B	83.57	53.67	30.45	44.26	41.00	49.67	35.01
RVT-B*	**83.80**	**55.40**	**33.86**	**42.99**	**42.27**	**52.63**	38.43

Table 6. RVT pre-trained on ImageNet-22K and finetuned on ImageNet-1K.

	Output Size	Layer Name	DeiT-Ti (V1)	V4	V5
Stage 1	$\frac{H}{4} \times \frac{W}{4}$	Patch Embedding	$C_1 = 192$	$C_1 = 96$	$C_1 = 48$
		Transformer Encoder	-	-	$\begin{bmatrix} H_1 = 48 \\ N_1 = 1 \\ C_1 = 48 \end{bmatrix} \times 2$
Stage 2	$\frac{H}{8} \times \frac{W}{8}$	Pooling Layer	-	-	$k = 2 \times 2$
		Transformer Encoder	-	$\begin{bmatrix} H_2 = 48 \\ N_2 = 2 \\ C_2 = 96 \end{bmatrix} \times 2$	$\begin{bmatrix} H_2 = 48 \\ N_2 = 2 \\ C_2 = 96 \end{bmatrix} \times 2$
Stage 3	$\frac{H}{16} \times \frac{W}{16}$	Pooling Layer	-	$k = 2 \times 2$	$k = 2 \times 2$
		Transformer Encoder	$\begin{bmatrix} H_2 = 64 \\ N_2 = 3 \\ C_2 = 192 \end{bmatrix} \times 12$	$\begin{bmatrix} H_3 = 64 \\ N_3 = 3 \\ C_3 = 192 \end{bmatrix} \times 8$	$\begin{bmatrix} H_3 = 64 \\ N_3 = 3 \\ C_3 = 192 \end{bmatrix} \times 6$
Stage 4	$\frac{H}{32} \times \frac{W}{32}$	Pooling Layer	-	$k = 2 \times 2$	$k = 2 \times 2$
		Transformer Encoder	-	$\begin{bmatrix} H_3 = 64 \\ N_3 = 6 \\ C_3 = 384 \end{bmatrix} \times 2$	$\begin{bmatrix} H_4 = 64 \\ N_4 = 6 \\ C_4 = 384 \end{bmatrix} \times 2$

Table 7. Detailed architecture of models used in robustness analysis on stage distribution. C, H and N represent the total feature dimension, feature dimension of each head and head number respectively. Only V4 and V5 are listed as examples. The other versions of the model can be generalized by V4 and V5.