SUPPLEMENTAL MATERIAL

In this supplemental material, we provide more analysis and results in our experiments.

A TARGET MODELS

Summary of the target models investigated in the main text are shown in Table 5. The weights of these models are all publicly available at Paszke et al. (2019); Wightman (2019) such that our experiments can be easily reproduced.

Table 5: Comparison of the target models investigated in the main text.

	ViT backbone				Pretraining	
Model	**Layers**	**Hidden size**	**Attention**	**Params**	**Pretraining dataset**	**Scale**
ViT-S/16	8	786	Self-attention	49M	ImageNet-21K	14M
ViT-B/16	12	786	Self-attention	87M	ImageNet-21K	14M
ViT-L/16	24	1024	Self-attention	304M	ImageNet-21K	14M
ViT-B/16-Res	12	786	Self-attention	87M	ImageNet-21K	14M
T2T-ViT-14	14	384	Self-attention	22M	-	-
T2T-ViT-24	24	512	Self-attention	64M	-	-
DeiT-S/16	12	384	Self-attention	22M	-	-
Dist-DeiT-B/16	12	768	Self-attention	87M	-	-
Swin-S/4	(2,2,18,2)	96	Self-attention	50M	-	-
SEResNet50	-	-	Squeeze-and-Excitation	28M	-	-
ResNeXt-32x4d-ssl	-	-	-	25M	YFCC100M	100M
ResNet50-swsl	-	-	-	26M	IG-1B-Targeted	940M
ResNet18	-	-	-	12M	-	-
ResNet50-32x4d	-	-	-	25M	-	-
ShuffleNet	-	-	-	2M	-	-
MobileNet	-	-	-	4M	-	-
VGG16	-	-	-	138M	-	-

B FREQUENCY FILTERS

We show the design of frequency-filters in Figure 5.

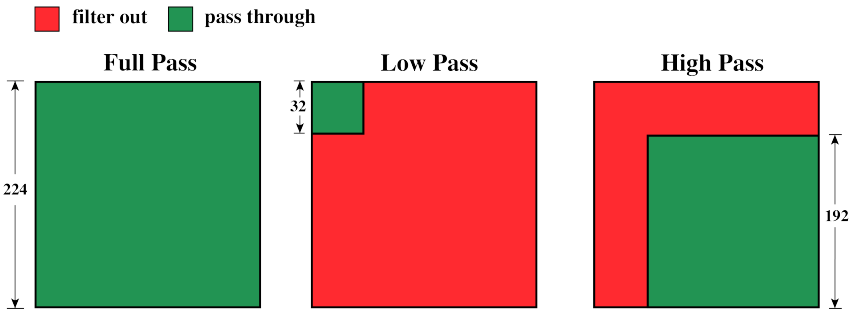

Figure 5: Filters for the frequency-based attack. The frequencies corresponding to the red part are filtered out, and the frequencies corresponding to the green part can pass through. "Full Pass" means all of the frequencies are preserved. "Low Pass" means only low-frequent components are preserved. "High Pass" preserves the high-frequent part.

C TRANSFER ATTACK RESULTS

Transfer attack results using more attack radii are provided in Figure 6.

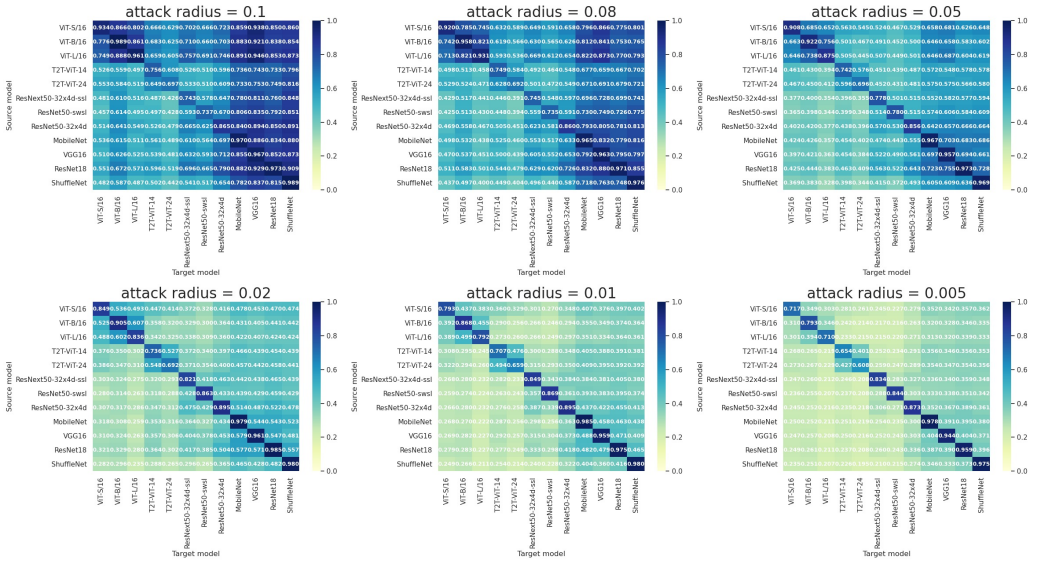

Figure 6: ASR of transfer attack using FGSM with different attack radii. The rows stand for the surrogate models used to generated adversarial examples in the white-box attack approach. The columns stand for the target models. Darker rows correlate to the source models that generate more transferable adversarial examples. While darker columns mean that the target models are more vulnerable to the transfer attack."Res50-ssl" and "Res50-swsl" are in short of "ResNeXt-32x4d-ssl" and "ResNet50-swsl" respectively.

D THE SOURCE OF ADVERSARIAL ROBUSTNESS

In this section we examine the source of the adversarial robustness revealed in our experiments.

The improved robustness of ViT is not caused by insufficient attack optimization. We first demonstrate that the better robustness of ViTs in white-box attacks is not caused by the difficult optimization in ViT by plotting the loss landscape with sufficient attack steps.

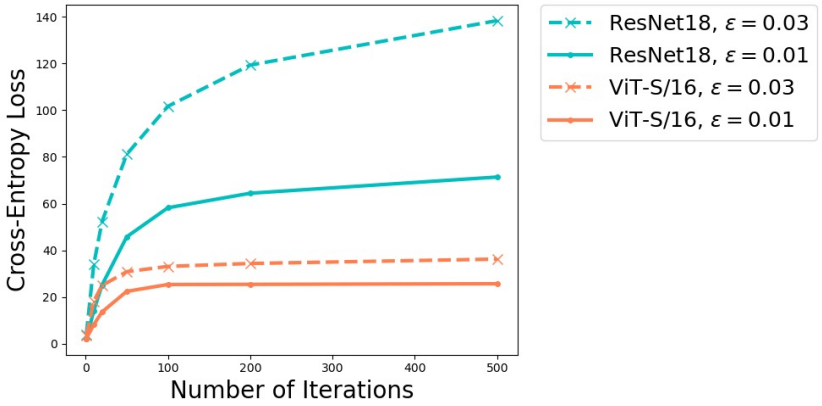

Figure 7: Cross entropy loss versus varying PGD attack steps for ViT-S/16 and RestNet18. The dashed lines corresponds to larger attach radius of 0.03 and the full lines to smaller attack radius of 0.01.

Figure 7 shows the cross entropy loss versus varing PGD attack steps for ViT-S/16 and ResNet18. As shown in the figure, ViT's loss curves converge at a much lower value than RestNet18, suggesting that the improved robustness of ViT is not caused by insufficient attack optimization.

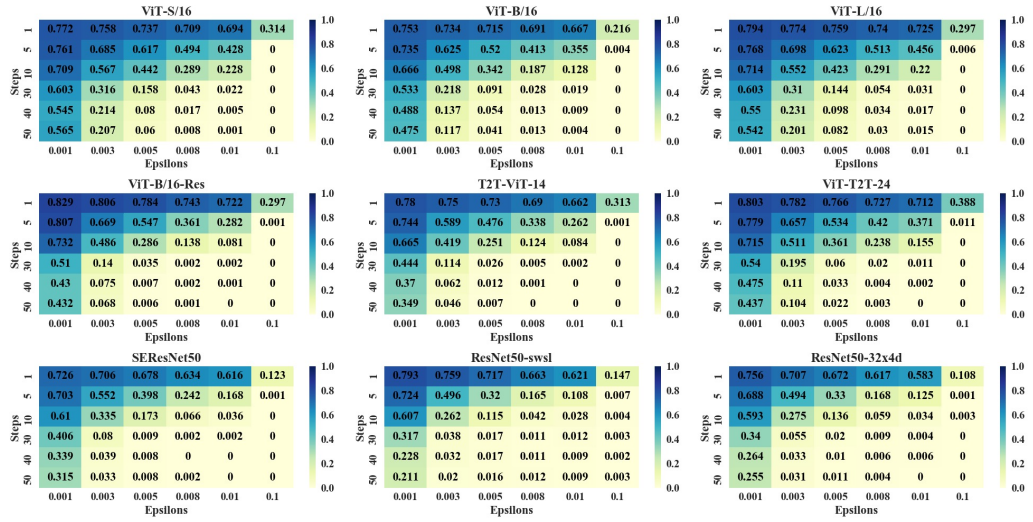

Figure 8: Adversarial accuracy of the target models against PGD attack with different attack radii ("eps") and attack steps ("steps"). When the attack radius and attack steps are increased, the adversarial accuracy of the target model decreases to zero. Darker blocks stand for more robust models against PGD attack.

Figure 8 shows the robust accuracy of more target models against PGD attack with different attack radii ("eps") and attack steps ("steps"). Vision transformers have darker blocks than CNNs', which stands for their superior adversarial robustness against PGD attack.

A Hopfield Network Perspective The equivalence between the attention mechanism in transformers to the modern Hopfield network (Krotov & Hopfield (2016)) was recently shown in Ramsauer et al. (2020). Furthermore, on simple Hopfield network (one layer of attention-like network) and dataset (MNIST), improved adversarial robustness was shown in Krotov & Hopfield (2018). Therefore, the connection of attention in transformers to the Hopfield network can be used to explain the improved adversarial robustness for ViTs.

E Experiments on Cifar-10

We choose the ImageNet as the benchmark because ViTs can hardly converge when training directly on small datasets like Cifar. Therefore, we finetune the ViTs instead. As shown in Table 6, ViT-B/4 performs higher robust accuracy than WideResNet, which is consistent with the trend on ImageNet.

Table 6: Robust accuracy of ViT-B/4 and WideResNet against PGD-10 attack with different attack radii.

Model	**0.001**	**0.003**	**0.01**	**0.03**
ViT-B/4	0.9202	0.6242	0.0994	0.0103
WideResNet	0.7744	0.5923	0.0854	0.0000

F Experiments on SOTA ViT Structures

In this section, we supplement the experimental results of recently proposed SOTA ViTs.

Swin-Trasnformer (Liu et al., 2021) computes the representations with shifted windows scheme which brings greater efficiency by limiting self-attention computation to non-overlapping local windows while also allowing for cross-window connection.

DeiT (Touvron et al., 2021) further improves the ViTs' performance using data augmentation or distillation from CNN teachers with an additional distillation token.

SAM-ViT (Chen et al., 2021) uses sharpness-aware minimization (Foret et al., 2020) to train ViTs from scratch on ImageNet without large-scale pretraining or strong data augmentations.

Table 7 summarizes the information of models investigated in our experiments. The window size of the swin transformers in Table 7 is 7. The pre-trained weights of these models are available in `timm` package.

Table 7: SOTA ViT models investigated in our experiments.

Model	Layers	Hidden size	Heads	Params
Deit-T/16 (Touvron et al., 2021)	12	192	3	6M
Deit-S/16 (Touvron et al., 2021)	12	384	6	22M
Deit-B/16 (Touvron et al., 2021)	12	768	12	87M
Dist-Deit-T/16 (Touvron et al., 2021)	12	192	3	6M
Dist-Deit-S/16 (Touvron et al., 2021)	12	384	6	22M
Dist-Deit-B/16 (Touvron et al., 2021)	12	768	12	87M
ViT-SAM-B/16 (Chen et al., 2021)	12	768	12	87M
ViT-SAM-B/32 (Chen et al., 2021)	12	768	12	88M
Swin-T/4 (Liu et al., 2021)	(2,2,6,2)	96	(3,6,12,24)	28M
Swin-S/4 (Liu et al., 2021)	(2,2,18,2)	96	(3,6,12,24)	50M
Swin-B/4 (Liu et al., 2021)	(2,2,18,2)	128	(4,8,16,32)	88M
Swin-L/4 (Liu et al., 2021)	(2,2,18,2)	192	(6,12,24,48)	197M

Table 8: Robust accuracy (%) of ViTs described in Table 7 against 40-step PGD attack with different attack radii, and also the clean accuracy ("Clean"). A model is considered to be more robust if the robust accuracy is higher.

Model	Clean	0.001	0.003	0.005	0.01
Deit-T/16	72.3	36.8	8.3	2.6	0.3
Deit-S/16	77.7	48.9	17.6	7.1	1.1
Deit-B/16	81.3	46.6	14.3	6.0	0.9
Dist-Deit-T/16	74.4	40.6	5.7	0.7	0.2
Dist-Deit-S/16	79.3	52.4	15.1	4.3	0.3
Dist-Deit-B/16	81.8	55.6	17.7	4.5	0.4
ViT-SAM-B/16	76.7	63.4	37.0	20.1	3.8
ViT-SAM-B/32	63.8	53.2	32.3	19.7	3.1
Swin-T/4	78.8	33.5	6.0	1.2	0.1
Swin-S/4	81.8	40.0	12.4	3.2	0.2
Swin-B/4	82.3	38.8	11.1	4.1	0.3
Swin-L/4	84.2	38.7	11.1	2.9	0.4

Table 8 shows the clean and robust accuracy of ViTs in Table 7 against 40-step PGD attack with different radii. And results for AutoAttack are shown in Table 9. Swin-transformers introduce shifted windows scheme that limit self-attention computation to non-overlapping local windows, which harms the robustness as Tokens-to-Token scheme according to the above results.

Table 9: Robust accuracy (%) of ViTs described in Table 7 against AutoAttack with different attack radii, and also the clean accuracy ("Clean"). A model is considered to be more robust if the robust accuracy is higher.

Model	**Clean**	**0.001**	**0.003**	**0.005**	**0.01**
Deit-T/16	72.3	23.4	0.5	0.0	0.0
Deit-S/16	77.7	30.2	1.2	0.0	0.0
Deit-B/16	81.3	20.4	0.3	0.1	0.0
Dist-Deit-T/16	74.4	31.1	0.8	0.1	0.0
Dist-Deit-S/16	79.3	43.1	3.7	0.2	0.0
Dist-Deit-B/16	81.8	42.7	3.4	0.2	0.0
ViT-SAM-B/16	76.7	59.8	26.0	8.4	0.1
ViT-SAM-B/32	63.8	48.9	23.6	9.7	0.8
Swin-T/4	78.8	6.8	0.1	0.0	0.0
Swin-S/4	81.8	7.9	0.1	0.0	0.0
Swin-B/4	82.3	2.4	0.1	0.0	0.0
Swin-L/4	84.2	4.3	0.1	0.0	0.0