A Extended Description of Robust Architectures

A few research studies have examined the impact of architectural designs on adversarial robustness [14, 20, 38, 50], *e.g.*, RobustArt [41] is the first comprehensive robustness benchmark of architectures and training techniques on ImageNet variants and Jung *et al*. [21] presented the first robustness dataset evaluating a complete NAS search space and demonstrated architectures' impacts on robustness. Among them, Huang *et al*. [20] is the closest to ours, and we thus provide a more detailed comparison with their work. Similar to our work, Huang *et al*. [20] also explored the relationship among depth and width, the SE block, and adversarial robustness through adversarially trained networks. However, our work is markedly different and enhanced in the following aspects:

1. Huang *et al*. assigned a fixed depth and width ratio only for the 3-stage WRN on CIFAR-10. It was an open research question as to how to extrapolate this fixed ratio to networks with more than three stages, such as the commonly used 4-stage residual networks for ImageNet [15, 27, 36]. In contrast, we provide a flexible compound scaling rule that does not place a restriction on the total number of stages, and we verify its generalizability and optimality through extensive experiments on CIFAR-10, CIFAR-100, and ImageNet.
2. Huang *et al*. proposed a specific residual block design using hierarchically aggregated convolution and residual SE. However, we show that such a residual SE is unnecessary due to the negative correlation between reduction ratio r and robustness. Furthermore, our design principles are applicable to both basic and bottleneck residual block designs. This flexibility is advantageous since the basic block is commonly used on CIFAR and the bottleneck block is widely deployed on ImageNet to reduce computational complexity [15, 16].
3. Huang *et al*. found that the adversarial robustness of models with smooth activation functions was sensitive to AT hyperparameters, and that removing BN affine parameters from weight decay was crucial; if the BN affine parameters were not removed, smooth activation functions did not improve performance beyond that of ReLU. This finding contradicts the prevailing consensus in the literature that smooth activation functions significantly improve robustness [2, 51]. In our research, we also find that using smooth activation functions is beneficial to robustness, and removing the BN affine parameters from weight decay is the correct implementation supported by Van [42] and multiple popular code bases and forums.[1,2]
4. Finally, Huang *et al*. only explored on CIFAR-10 and CIFAR-100, leaving no evidence that these findings will extrapolate to the large-scale ImageNet. In contrast, we verified the generalization of all our design principles through extensive experiments over a wide spectrum of dataset scales, AT methods, model parameters, and network design spaces.

A parallel line of related studies leverages NAS to search for optimal robust architectures. Guo *et al*. [14] explored two types of block topologies within the DARTS search space. Subsequent work regulated the NAS loss formulation through the smoothness of the input loss landscape [32]. These compute-intensive NAS frameworks mainly focus on searching for block topology while leaving other factors to manual design, *e.g.*, activation, depth, and width. Furthermore, most searches are conducted on CIFAR-10 since both NAS and AT are already computationally expensive. Besides, current research shows that the NAS-optimized architecture depends on the dataset used [24], thus hindering explorations of network design

[1] minGPT: A PyTorch re-implementation of GPT

[2] PyTorch forum: Weight decay in the optimizers is a bad idea (especially with BatchNorm)

principles that deepen our understanding and generalize to new settings [36]. In Sec. C, we also compare our results with NAS-based networks and demonstrate that our robustified networks exhibit higher robustness.

B Additional Details on Architectural Design Principles

B.1 Full Results for Stem Stage and Residual Block Designs

Here we provide the full results in Table 4 showcasing how various configurations of *stem stage* and *residual block designs* impact the clean and adversarial accuracies over the ResNet-50 baseline, extending the result highlights presented in Figure 2b of the main paper.

For the stem stage, the convolution stem with postponed downsampling operation outperforms the patchify stem. In the patchify stem, we observe a consistent performance improvement while decreasing the stride, and the 4×4 patch with the largest overlapping between neighboring patches (Patch 4, Stride 1) performs almost on par with the postponed downsampling. Finally, decreasing the width from 64 (ResNet-50) to 32 lowers the accuracy due to fewer model parameters, while increasing the width from 64 to 96 significantly boosts both clean and adversarial accuracies with a negligible 0.01M increase in total parameters.

For the residual block design, we find that a straightforward application of SE markedly increases all PGD and clean accuracies compared to ResNet-50. In terms of activation, reducing the number of activation layers does not contribute to adversarial robustness. Therefore, we preserve the activations along with all convolutions. Besides, we find that non-parametric smooth activation functions exhibit greater robustness compared to both ReLU and their parametric counterparts.

Table 4: Full results showcasing how various configurations of *stem stage* and *residual block designs* impact the clean and adversarial accuracies over the ResNet-50 baseline, extending the result highlights already presented in Figure 2b of the main paper. All models trained with Fast-AT [46] and evaluated on full ImageNet validation set.

	Config	Clean (%)	PGD^{10}-2 (%)	PGD^{10}-4 (%)	PGD^{10}-8 (%)
	ResNet-50 (baseline)	56.05	42.81	30.59	12.62
Stem stage	Postponed downsampling	57.08 +1.03	45.19 +2.38	33.08 +2.49	14.50 +1.88
	Patch 4, Stride 4	55.40 -0.65	43.45 +0.64	31.68 +1.09	13.80 +1.18
	Patch 2, Stride 2	56.38 +0.33	44.21 +1.40	31.91 +1.32	13.48 +0.86
	Patch 4, Stride 3	55.75 -0.30	44.54 +1.73	32.41 +1.82	13.74 +1.12
	Patch 4, Stride 2	56.58 +0.53	44.60 +1.79	32.83 +2.24	14.03 +1.41
	Patch 4, Stride 1	56.72 +0.67	45.06 +2.25	33.20 +2.61	14.45 +1.83
	Stem width = 32	55.89 -0.16	41.64 -1.17	29.73 -0.86	13.25 +0.63
	Stem width = 96	57.29 +1.24	44.55 +1.47	32.06 +1.47	13.74 +1.12
Residual block design	SE ($r = 4$)	57.83 +1.78	45.09 +2.28	32.64 +2.05	14.01 +1.39
	ReLU-ReLU-0	51.54 -4.51	38.69 -4.12	27.05 -3.54	10.94 -1.68
	ReLU-0-ReLU	53.91 -2.14	41.22 -1.59	29.62 -0.97	12.30 -0.32
	0-ReLU-ReLU	54.81 -1.24	42.10 -0.71	30.34 -0.25	12.86 +0.24
	0-0-ReLU	51.03 -5.02	39.12 -3.69	28.15 -2.44	12.09 -0.53
	0-ReLU-0	47.18 -8.87	34.85 -7.96	24.12 -6.47	9.51 -3.11
	ReLU-0-0	44.21 -11.84	32.34 -10.47	22.24 -8.35	8.77 -3.85
	GELU	57.48 +1.43	45.05 +2.24	33.12 +2.53	14.80 +2.18
	SiLU	58.19 +2.14	46.21 +3.40	34.07 +3.48	14.68 +2.06
	PReLU	55.81 -0.27	42.52 -0.29	30.38 -0.21	12.76 +0.14
	PSiLU	56.38 +0.33	44.90 +2.09	33.76 +3.17	15.40 +2.78
	PSSiLU	57.43 +1.38	44.44 +1.63	32.22 +1.63	13.71 +1.09

B.2 Adversarial Robustness Negatively Correlated with SE Reduction Ratio r on CIFAR-10

Table 5: A hyperparameter sweep of the SE block reduction ratio $r = \{2,4,8,16,32,64\}$ on CIFAR-10. These results show that adversarial robustness is negatively correlated with r, and when $r \geq 32$, the accuracy is inferior to the baseline WRN-22-10, supporting our discovery presented in Sec. 4.3.1. Both AA and 20-step PGD (PGD^{20}) use the same maximum perturbation $\ell_\infty, \varepsilon = 8/255$.

Config	Clean (%)	AA (%)	PGD^{20} (%)
WRN-22-10	83.82	48.38	52.64
WRN-22-10 ($r = 2$)	86.95 +3.13	49.14 +0.76	53.48 +0.84
WRN-22-10 ($r = 4$)	86.75 +2.93	49.13 +0.75	53.52 +0.88
WRN-22-10 ($r = 8$)	86.48 +2.66	49.11 +0.73	53.39 +0.75
WRN-22-10 ($r = 16$)	84.97 +1.15	48.89 +0.51	53.04 +0.40
WRN-22-10 ($r = 32$)	83.61 -0.21	48.20 -0.18	52.24 -0.40
WRN-22-10 ($r = 64$)	82.90 -0.92	47.44 -0.94	51.43 -1.21

C Evaluations on CIFAR-10 & CIFAR-100

This section presents the complete results of our robustified architectures in Table 6 and provides a systematic comparison to SOTA adversarially trained Transformers and NAS-based networks in Table 7, extending the result highlights presented in Sec. 5.2 in the main paper. As the AT recipes for CNNs and Transformers are not compatible with each other, we do not retrain the Transformers and instead, directly extract the results from the literature. In addition, the AT recipe for Transformers requires multiple training tricks built on SAT to boost robustness, *e.g.*, attention random dropping [31], perturbation random masking [31], ε-warmup [6], and larger weight decay [6]. Despite employing all these tricks in training Transformers, our Ra WRN-34-12 trained with "Diff. 1M" is significantly more robust than Swin-B and cross-covariance image Transformers (XCiT)-L12 on both CIFAR-10 and CIFAR-100 even with fewer total parameters. We also compare to NAS-based networks: our Ra WRN-22-10 is significantly more robust than RobNet-large-v2 [14] and performs on par with AdvRush [32] using the same TRADES [54] AT method but with fewer total parameters. Lastly, we compare our robust architectures with Huang *et al.* [19], who have also studied the relationship between robustness and depth and width, and proposed a reconfigured version of WRN-34-12 called WRN-34-R. By using the same SAT methods, both Ra WRN-34-12 and WRN-34-R show greater robustness than the baseline WRN-34-12, but our Ra WRN-34-12 is 1.75 pp and 0.69 pp higher than WRN-34-R in terms of AA and PGD accuracies, respectively.

Table 6: Complete results of adversarial robustness on CIFAR-10 and CIFAR-100 against AA and 20-step PGD (PGD^{20}) with the same maximum perturbation $\ell_\infty, \varepsilon = 8/255$. Applying our principles leads to a consistent 1–3 pp robustness gain across AT methods, parameter budgets, and design spaces, boosting even the SOTA "Diff. 1M" and "Diff. 50M" AT methods proposed by Wang *et al.* [45]. This table extends Table 2 in the main paper by including the results for WRN-22-10.

#Param.	Method	Model	CIFAR-10 Clean (%)	CIFAR-10 AA (%)	CIFAR-10 PGD^{20} (%)	CIFAR-100 Clean (%)	CIFAR-100 AA (%)	CIFAR-100 PGD^{20} (%)
26M	SAT	ResNet-50	84.05	49.97	54.37	55.86	23.78	27.48
		Ra ResNet-50	**84.91** +0.86	**50.94** +0.97	**55.19** +0.82	**56.38** +0.52	**24.99** +1.21	**28.84** +1.36
	TRADES	ResNet-50	82.26	49.91	54.50	56.00	25.05	29.91
		Ra ResNet-50	**82.80** +0.54	**51.23** +1.32	**55.44** +0.94	**56.29** +0.29	**25.83** +0.78	**31.87** +1.96
	MART	ResNet-50	77.98	47.17	52.70	53.18	25.35	30.79
		Ra ResNet-50	**79.60** +1.62	**49.19** +2.02	**56.47** +3.77	**53.68** +0.50	**26.97** +1.62	**32.81** +2.02
27M	SAT	WRN-22-10	83.82	48.38	52.64	56.79	23.46	27.08
		Ra WRN-22-10	**84.27** +0.45	**51.30** +2.92	**55.42** +2.78	**57.34** +0.55	**24.27** +0.81	**28.64** +1.56
	TRADES	WRN-22-10	81.81	51.06	55.21	55.48	23.50	29.60
		Ra WRN-22-10	**82.27** +0.46	**51.71** +0.65	**56.20** +0.99	**55.55** +0.07	**24.91** +1.41	**29.78** +0.18
37M	SAT	WRN-28-10	85.44	48.45	53.13	**60.49**	23.64	27.47
		Ra WRN-28-10	**85.52** +0.08	**51.96** +3.51	**56.22** +3.09	59.09 -1.40	**25.14** +1.50	**29.27** +1.80
	TRADES	WRN-28-10	**83.86**	51.79	55.69	55.21	25.47	29.34
		Ra WRN-28-10	83.29 -0.57	**52.10** +0.31	**56.31** +0.62	**55.38** +0.71	**25.68** +0.21	**29.41** +0.07
	MART	WRN-28-10	82.83	50.30	57.00	51.31	25.78	30.06
		Ra WRN-28-10	**82.85** +0.02	**50.81** +0.51	**57.35** +0.35	**51.61** +0.30	**26.11** +0.33	**30.82** +0.76
	Diff. 1M	WRN-28-10	90.61	61.66	66.43	67.26	34.26	39.29
		Ra WRN-28-10	**91.32** +0.71	**65.11** +3.45	**68.93** +2.50	**69.03** +1.77	**37.24** +2.98	**41.59** +2.30
67M	SAT	WRN-34-12	85.92	49.35	53.05	59.08	23.69	27.05
		Ra WRN-34-12	**86.50** +0.58	**51.78** +2.43	**56.04** +2.99	**59.46** +0.38	**25.18** +1.49	**29.49** +2.44
	Diff. 1M	WRN-34-12	91.11	62.83	67.53	68.40	35.67	40.33
		Ra WRN-34-12	**91.75** +0.64	**65.71** +2.88	**69.67** +2.14	**69.75** +1.35	**37.73** +2.06	**42.16** +1.83
267M	SAT	WRN-70-16	86.26	50.19	53.74	60.26	23.99	27.05
		Ra WRN-70-16	**86.72** +0.46	**52.13** +1.94	**56.49** +2.75	**60.42** +0.16	**25.17** +1.18	**29.46** +2.41
	Diff. 1M	WRN-70-16	91.82	65.02	69.10	70.10	37.77	41.95
		Ra WRN-70-16	**92.16** +0.34	**66.33** +1.31	**70.37** +1.27	**70.25** +0.15	**38.73** +0.96	**42.61** +0.66
	Diff. 50M	WRN-70-16	93.25	70.69	73.89	-	-	-
		Ra WRN-70-16	**93.27** +0.02	**71.07** +0.38	**75.28** +1.39	-	-	-

Table 7: A systematic comparison to SOTA adversarially trained Transformers and NAS-based architectures with adversarial robustness on CIFAR-10 and CIFAR-100 against AA and 20-step PGD (PGD^{20}) with the same maximum perturbation $\ell_\infty, \varepsilon = 8/255$. Our robustified architectures (prefixed by Ra) exhibit greater robustness (highlighted in **bold**) than all Transformers and NAS-based architectures compared. The "Diff. 1M" results are extracted from Table 6.

Method	Model	#Param.	CIFAR-10 Clean (%)	CIFAR-10 AA (%)	CIFAR-10 PGD^{20} (%)	CIFAR-100 Clean (%)	CIFAR-100 AA (%)	CIFAR-100 PGD^{20} (%)
Diff. 1M	Ra WRN-28-10	37M	91.32	**65.11**	**68.93**	69.03	**37.24**	**41.59**
	Ra WRN-34-12	67M	91.75	**65.71**	**69.67**	69.75	**37.73**	**42.16**
	Ra WRN-70-16	267M	92.16	**66.33**	**70.37**	70.25	**38.73**	**42.61**
Mo *et al.* [31]	DeiT-S	22M	83.04	48.34	52.52	-	-	-
	Swin-S	50M	84.46	46.17	50.02	-	-	-
	ViT-B/16	86M	84.90	50.03	53.80	-	-	-
	Swin-B	88M	84.16	47.50	51.47	-	-	-
Debenedetti *et al.* [6]	XCiT-S12	26M	90.06	56.14	-	67.34	32.19	-
	XCiT-M12	46M	91.30	57.27	-	69.21	34.21	-
	XCiT-L12	104M	91.73	57.58	-	70.76	35.08	-
TRADES	RobNet-large-v2 [14]	33M	84.57	47.48	52.79	55.27	23.69	29.23
TRADES	AdvRush (7@96) [32]	33M	84.65	52.08	56.23	55.40	25.27	29.40
SAT	WRN-34-R [19]	68M	87.85	50.03	55.35	61.33	25.20	29.02

D Evaluations on ImageNet

This section presents an extended discussion of the ImageNet results in Sec. 5.3 in the main paper. Table 8 provides a controlled comparison of our robustified architectures to SOTA CNNs and Transformers using Fast-AT method [46]. We present the robustified ResNet-50, ResNet-101, and WRN-101-2 and make the following observations:

1. Our robustified architectures consistently demonstrate a 4–9 pp gain in robustness across different model parameters and design spaces. Furthermore, increasing the total number of parameters in general leads to higher robustness.
2. Under a fixed model capacity, our Ra ResNet-50 outperforms the baseline ResNet-50 and ResNeXt-50 32×4d [52], and Ra ResNet-101 outperforms ResNet-101 and RegNetX-8GF [36].
3. Compared to models with larger parameters, our Ra ResNet-50 is more robust than ResNet-152 and WRN-50-2, and even WRN-101-2 despite having 4.85× fewer parameters. Similarly, our Ra WRN-101-2 outperforms the baseline WRN-101-2 and achieves SOTA performance under the Fast-AT method.
4. Transformers such as Swin-T [26] and Transformer-based architectures such as ConvNeXt-T [27] exhibit lower robustness when employing Fast-AT. The phenomenon can be attributed to the differences in optimizers, learning rates, and data augmentation, where most Transformer-related architectures use AdamW [28], tiny learning rates, and heavy data augmentation.

Then, we provide a systematic comparison of our SAT-trained robust architectures with CNNs and Transformers that utilize specifically optimized AT methods in Table 9. By applying our design principles, the robustified architecture achieves a similar or even superior level of robustness compared to the Transformers that utilize additional training tricks to enhance their robustness. For example, under similar total parameters, ResNet-50 and ResNet-101 are less robust than XCiT-S12 and XCiT-M12, respectively, but the robustified Ra ResNet-50 and Ra ResNet-101 show higher clean and AA accuracies. Additionally, there is no sign of saturation when scaling up the total parameters, as Ra WRN-101-2 remains markedly more robust than XCiT-L12 with the same 104M parameters.

Table 8: Our robustified architectures (Ra) consistently demonstrate a 4–9 pp gain in robustness over SOTA CNNs and Transformers using Fast-AT method [46], across different model parameters, design spaces, and attack budgets.

Model	#Param.	Clean (%)	PGD^{10}-2 (%)	PGD^{10}-4 (%)	PGD^{10}-8 (%)
Ra ResNet-50	26M	62.02	51.47	39.65	18.97
Ra ResNet-101	46M	64.40	53.97	42.06	20.98
Ra WRN-101-2	104M	66.08	55.52	43.81	22.50
SqueezeNet 1.1	1M	0.10	0.10	0.10	0.10
MobileNet V2	4M	41.60	31.23	21.89	8.94
EfficientNet-B0	5M	48.78	37.74	26.90	10.92
ShuffleNet V2 2.0×	7M	49.99	0.01	0.01	0.02
DenseNet-121	8M	52.29	40.06	28.72	12.23
ResNet-18	12M	46.59	35.05	24.64	9.95
RegNetX-3.2GF	15M	57.26	45.74	33.85	15.37
RegNetY-3.2GF	19M	59.15	47.09	34.82	15.51
EfficientNetV2-S	21M	57.64	45.89	33.48	14.03
ResNeXt-50 32×4d	25M	57.33	45.46	33.08	14.45
ResNet-50	26M	56.09	42.66	30.43	12.61
Swin-T	28M	38.83	28.08	18.49	6.20
ConvNeXt-T	29M	21.35	15.39	10.51	4.07
DenseNet-161	29M	59.80	47.60	35.35	15.77
EfficientNet-B5	30M	55.90	44.80	33.26	14.53
RegNetY-8GF	39M	63.61	52.26	40.15	19.21
RegNetX-8GF	40M	60.26	48.98	36.89	17.22
ResNet-101	45M	58.04	45.72	33.90	15.93
ResNet-152	60M	61.55	48.50	35.85	15.87
WRN-50-2	69M	60.66	46.99	34.10	15.37
WRN-101-2	127M	61.63	49.10	36.23	16.14

Table 9: By applying our design principles, the robustified architecture achieves a similar or even superior level of robustness compared to the Transformers that utilize additional training tricks to enhance their robustness. The Ra results are extracted from Table 3 in the main paper.

Model	#Param.	Clean (%)	AA (%)	PGD^{100}-2 (%)	PGD^{100}-4 (%)	PGD^{100}-8 (%)
Ra ResNet-50	26M	70.17	44.14	60.06	47.77	21.77
Ra ResNet-101	46M	71.88	46.26	61.89	49.30	23.01
Ra WRN-101-2	104M	73.44	48.94	63.49	51.03	25.31
PoolFormer-M12 [6]	22M	66.16	34.72	-	-	-
DeiT-S [2]	22M	66.50	35.50	-	40.32	-
ResNet50 + GELU [2]	26M	67.38	35.51	40.27	-	-
ResNet50 + DiffPure [33]	26M	67.79	40.39	-	-	-
XCiT-S12 [6]	26M	72.34	41.78	-	-	-
ConvNeXt-T [38]	29M	67.60	41.60	-	-	-
XCiT-M12 [6]	46M	74.04	45.24	-	-	-
WRN-50-2 [37]	69M	68.41	38.14	55.86	41.24	16.29
WRN-50-2 + DiffPure [33]	69M	71.16	44.39	-	-	-
Vit-B/16 [31]	86M	69.10	34.62	-	37.52	-
Swin-B [31]	88M	74.36	38.61	-	40.87	-
XCiT-L12 [6]	104M	73.76	47.60	-	-	-

E Architecture Details

Table 10 contains the details of all the robustified architectures mentioned in the paper. For depth and width, we present the list of total depths and widths in each stage and compute the corresponding WD ratio. For the stem stage, we use the convolution stem stage with postponed downsampling operation and increase the output channels to 96. Regarding the design of the residual block, we append the SE block ($r = 4$) to the 3×3 convolution layer and replace ReLU with SiLU.

Table 10: Details of all the robustified architectures mentioned in the paper.

Model	#Param.	Depth	Width	WD ratio	Stem width	SE	Activation
Ra ResNet-50	26M	[5, 8, 13, 1]	[36, 72, 140, 270]	8.99	96	$r = 4$	SiLU
Ra WRN-22-10	27M	[13, 15, 2]	[120, 240, 480]	12.62	96	$r = 4$	SiLU
Ra WRN-28-10	37M	[14, 16, 3]	[128, 256, 512]	12.57	96	$r = 4$	SiLU
Ra ResNet-101	46M	[7, 11, 18, 1]	[42, 84, 166, 328]	7.62	96	$r = 4$	SiLU
Ra WRN-34-12	67M	[18, 20, 5]	[144, 288, 576]	11.20	96	$r = 4$	SiLU
Ra WRN-101-2	104M	[7, 11, 18, 1]	[64, 128, 252, 504]	11.59	96	$r = 4$	SiLU
Ra WRN-70-16	267M	[30, 31, 10]	[216, 432, 864]	10.57	96	$r = 4$	SiLU

F Training curves and convergence rate

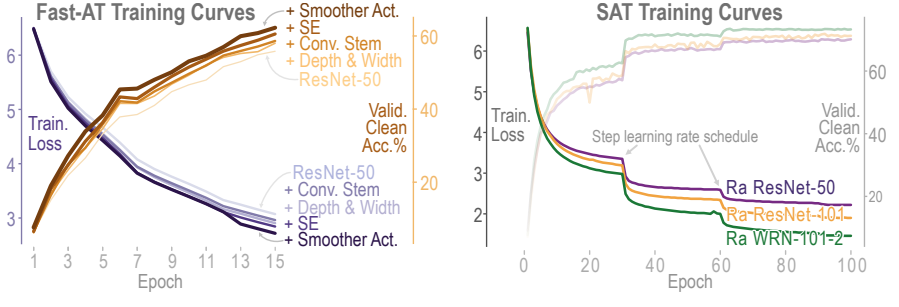

Figure 3: Visualization of the Fast-AT curves of individual architectural modifications (Table 1), and the SAT curves of the final robustified model (Table 3). We observed that a lower training loss leads to a higher robustness as expected and no catastrophic overfitting occurs during training.