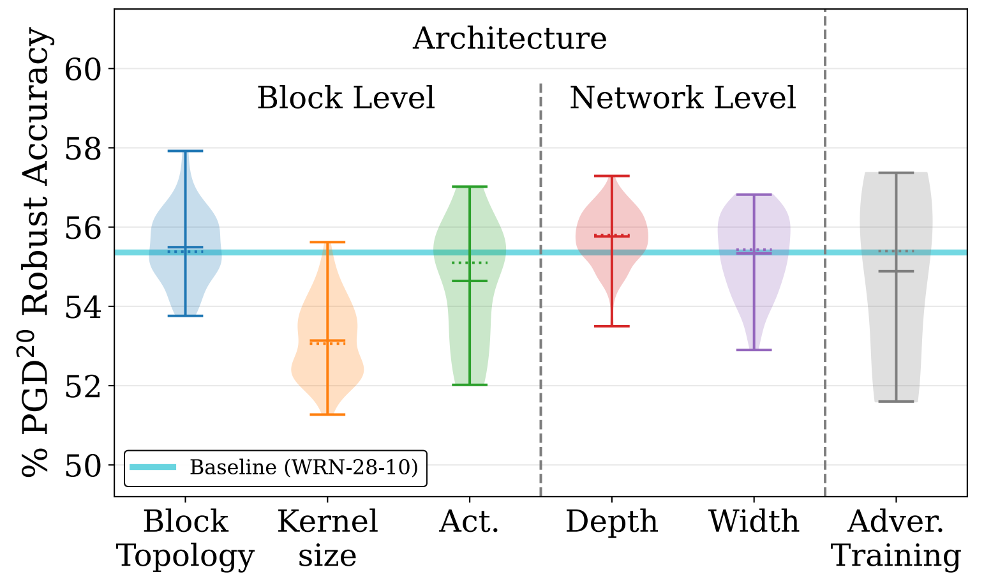
Revisiting Residual Networks for Adversarial Robustness

Shihua Huang[1], Zhichao Lu[2], Kalyanmoy Deb[1], Vishnu Naresh Boddeti[1]

1 Michigan State University, East Lansing, MI
2 Sun Yat-sen University, China

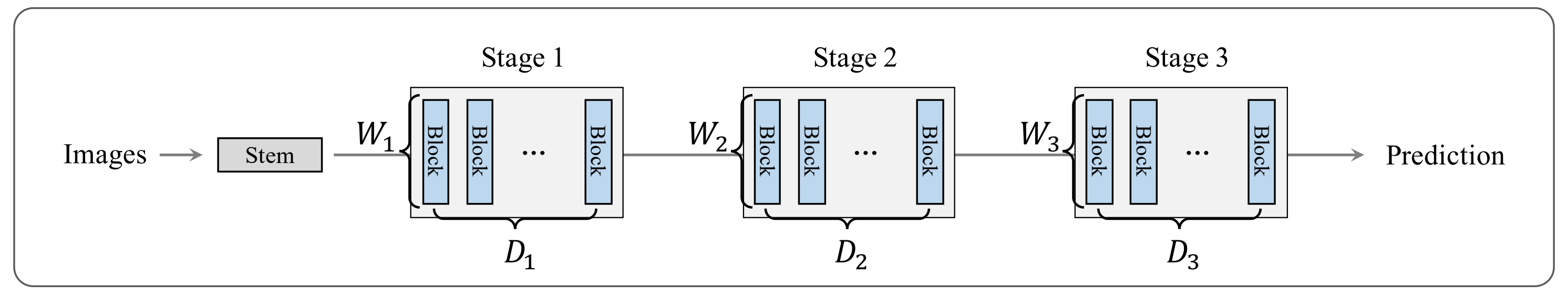
Correspondence: luzhichaocn@gmail.com

Motivation

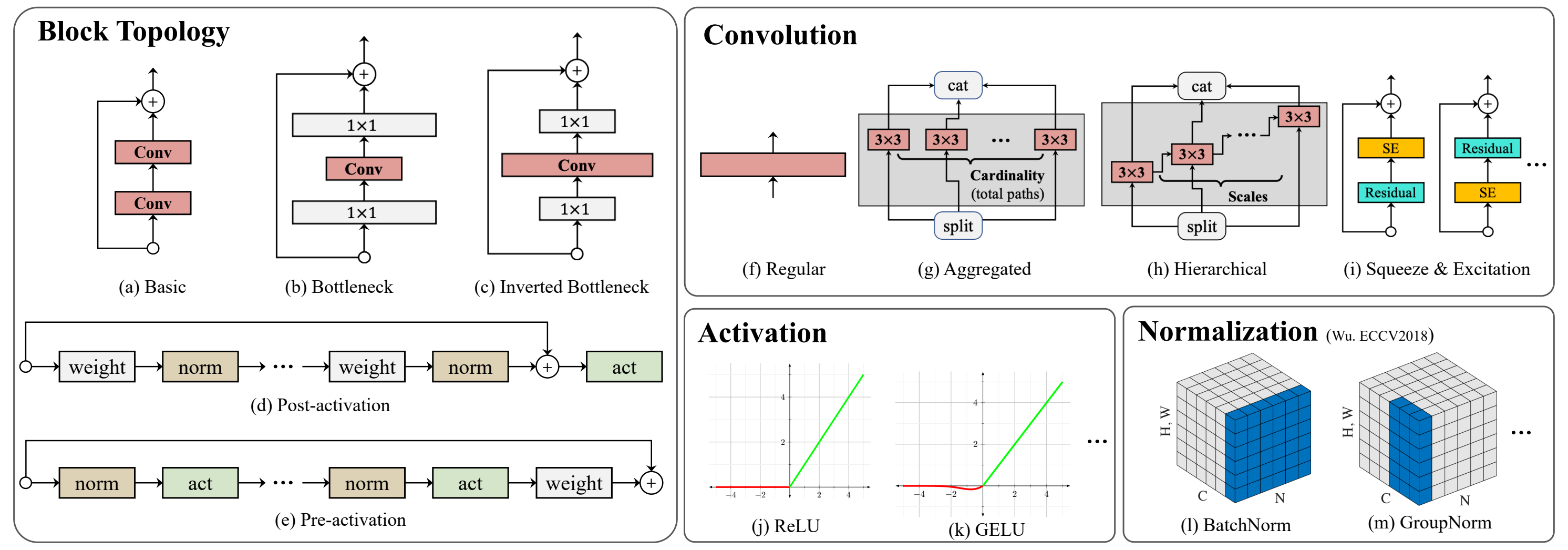


- Existing work on adversarial defenses focuses on better adversarial training.
- Architectural components can impact adversarial robustness as much as different adversarial training methods.

Overview

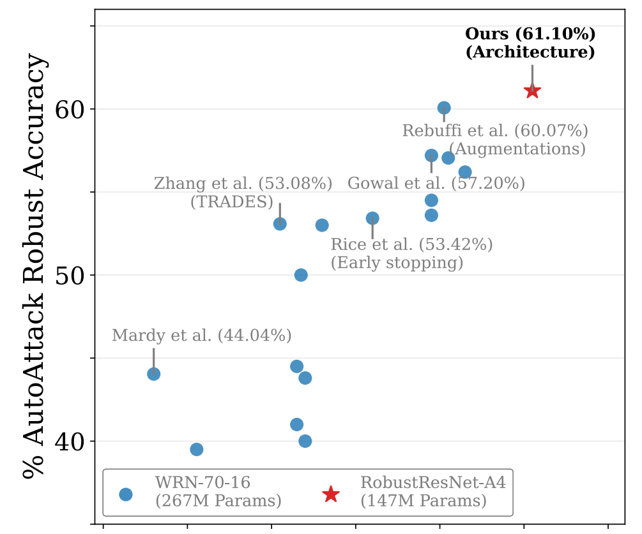


Network scaling level: depth (D_1, D_2, D_3) and width (W_1, W_2, W_3)



Block level: variants of residual blocks and their components, including convolution, activation, kernel size, normalization, etc.

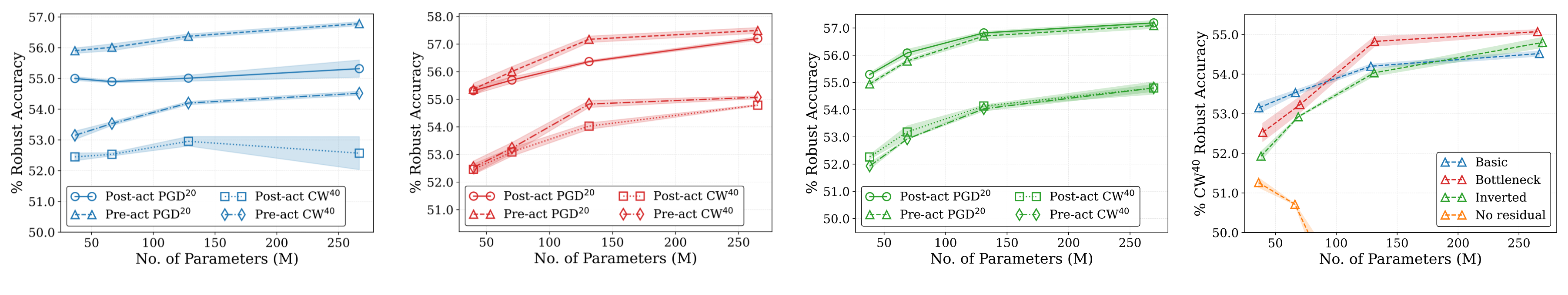
Results



- Our final RobustResNets are based on RobustResBlock (block level) and RobustScaling (network level).
- SoTA performance, ~1 % Autoattack improvement over the second best.
- 2x more compact than others.

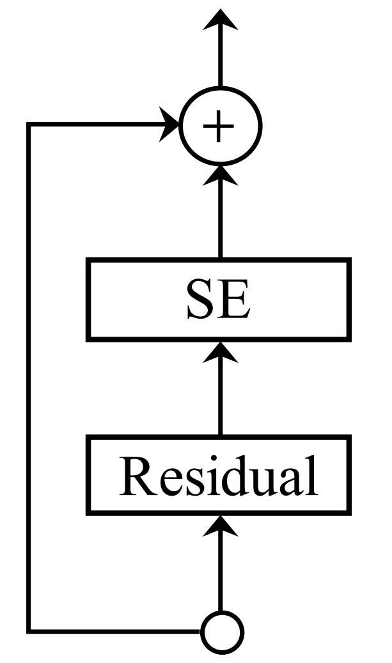
Block Level Design

Impact of Residual Topology

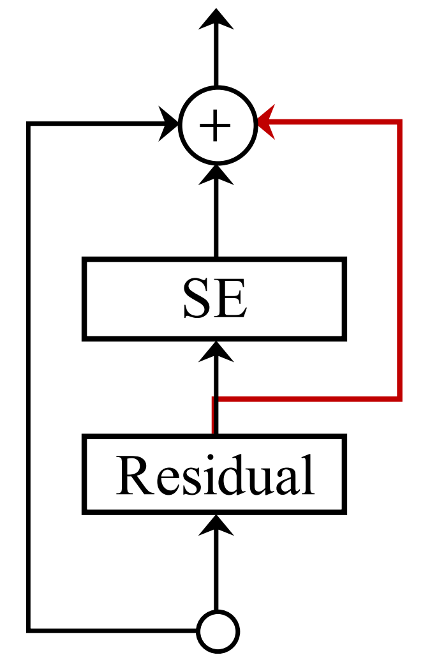


(a) Basic (b) Bottleneck (c) Inverted Bottleneck (d) Comparison among (a) – (c)

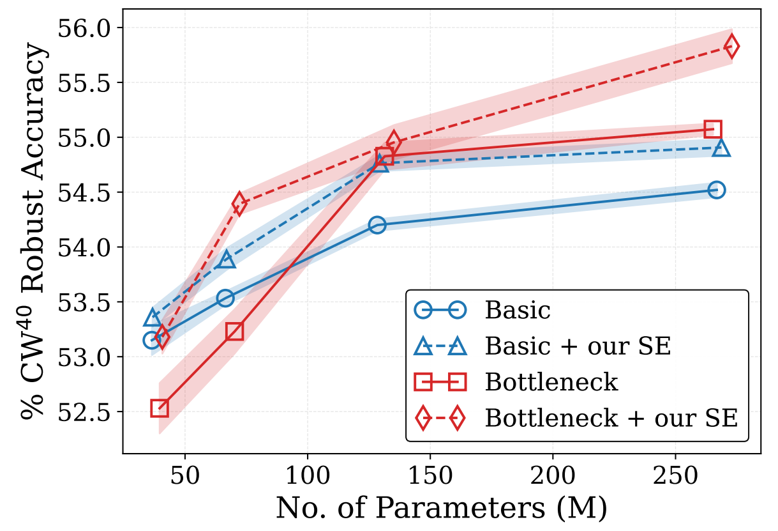
Impact of Squeeze-n-Excitation



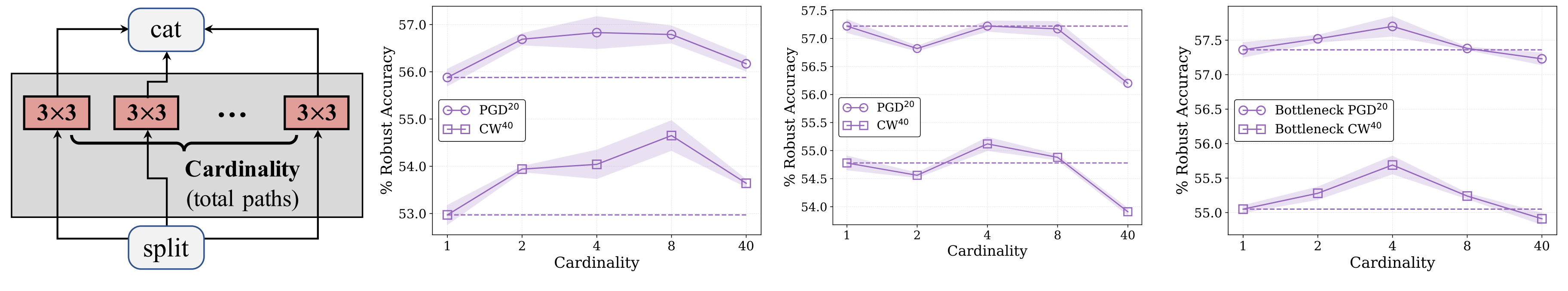
Standard SE



Our SE

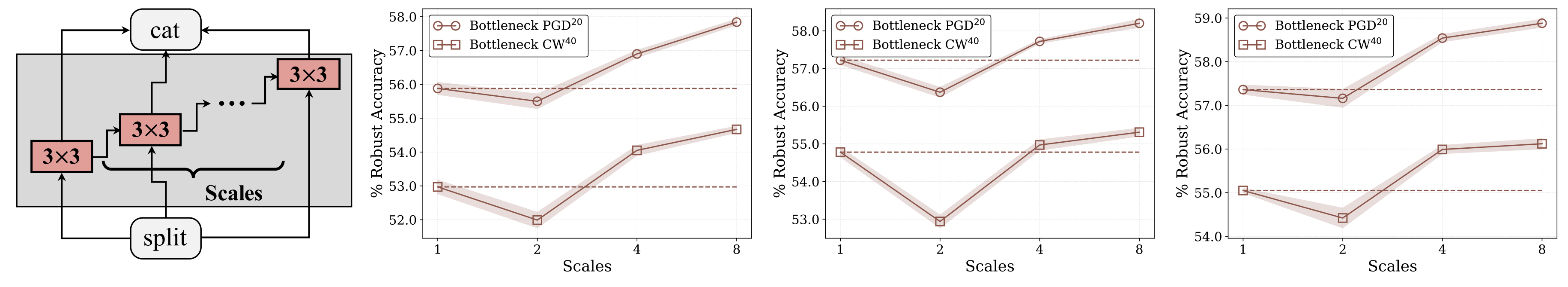


Impact of Aggregated Convolutions



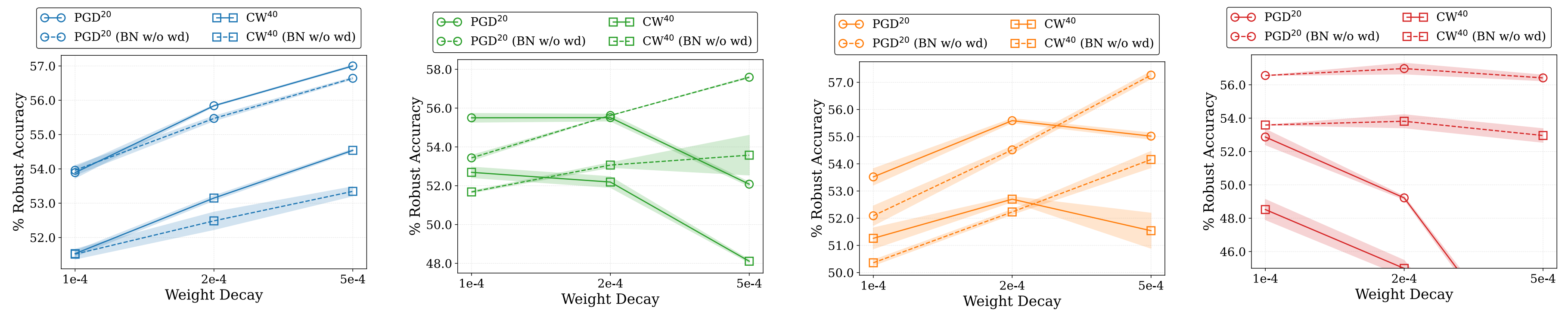
(a) Aggregated convolution (b) $D_i = 5, W_i = 12$ (c) $D_i = 7, W_i = 14$ (d) $D_i = 11, W_i = 16$

Impact of Hierarchical Convolutions



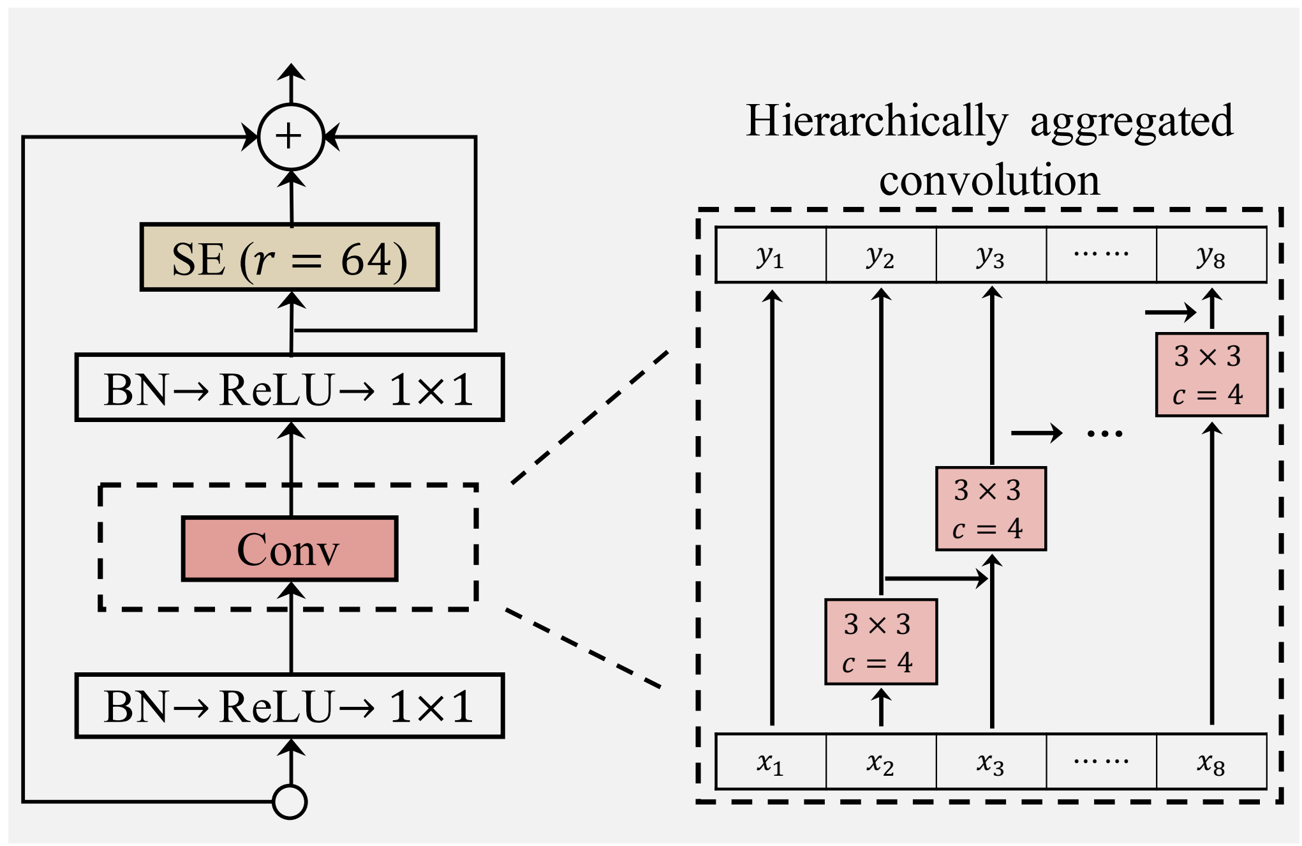
(a) Hierarchical convolution (b) $D_i = 5, W_i = 12$ (c) $D_i = 7, W_i = 14$ (d) $D_i = 11, W_i = 16$

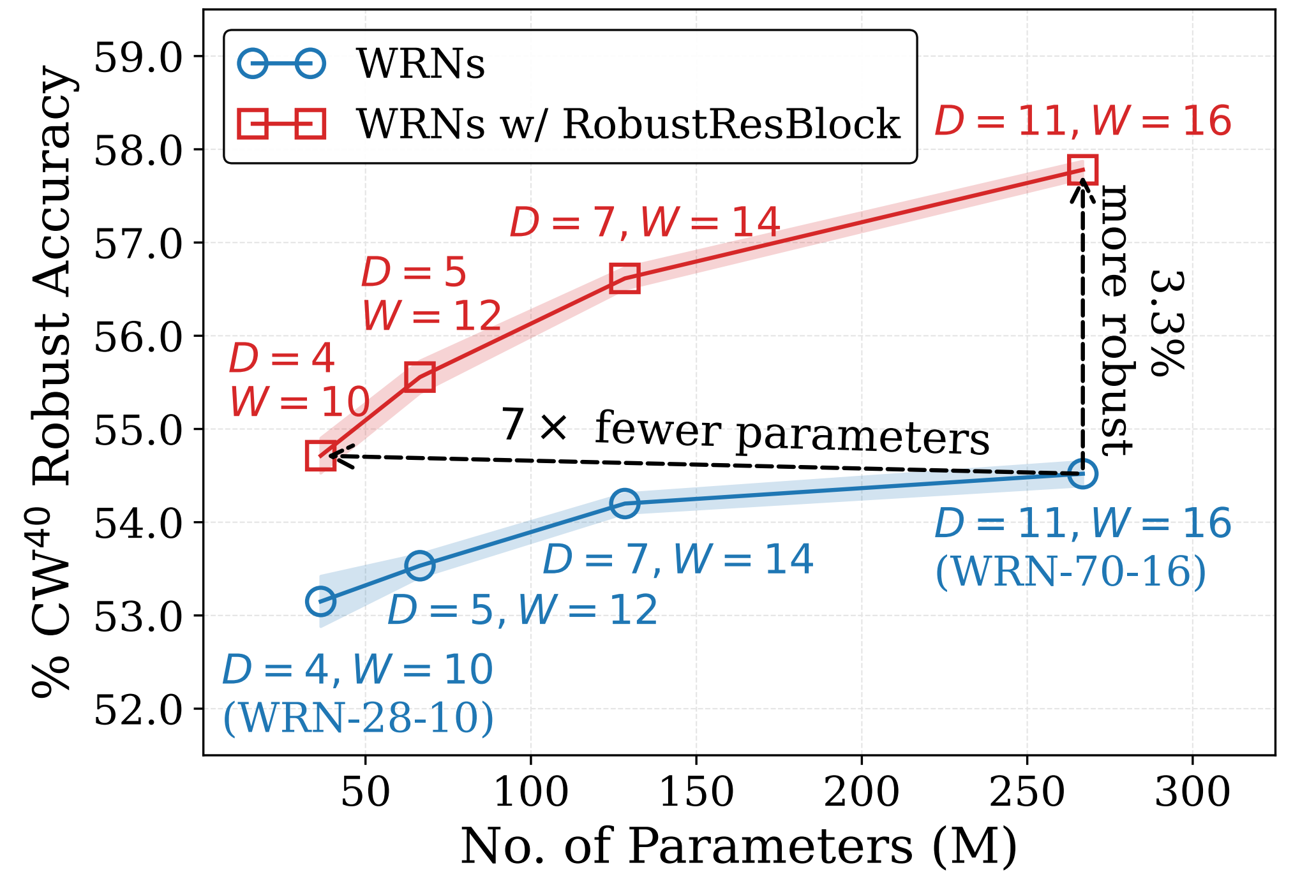
Impact of Activation



(a) ReLU (b) SiLU (c) GELU (d) Softplus

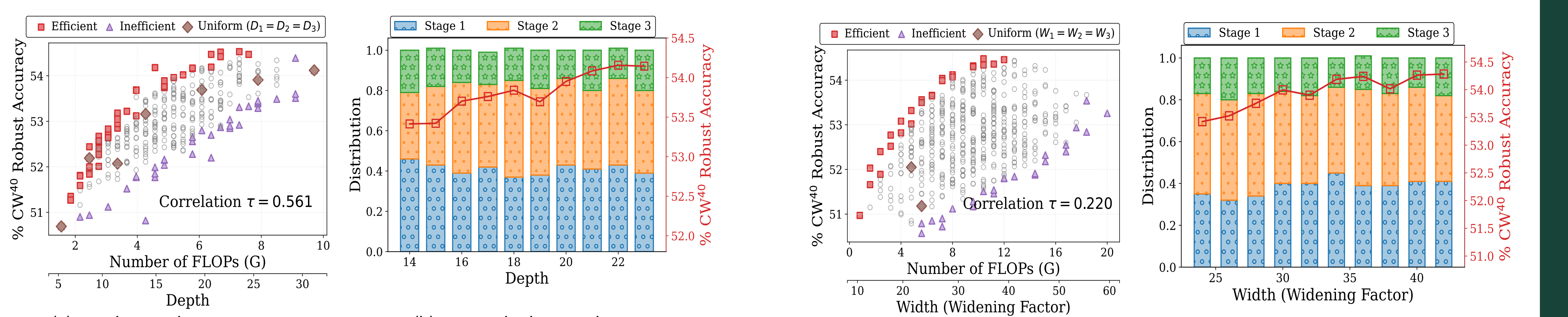
RobustResBlock





Network Level Design

Independent Scaling by Depth or Width



(a) Depth vs. Robustness (b) Top-ranked networks (c) Depth vs. Robustness (d) Top-ranked networks

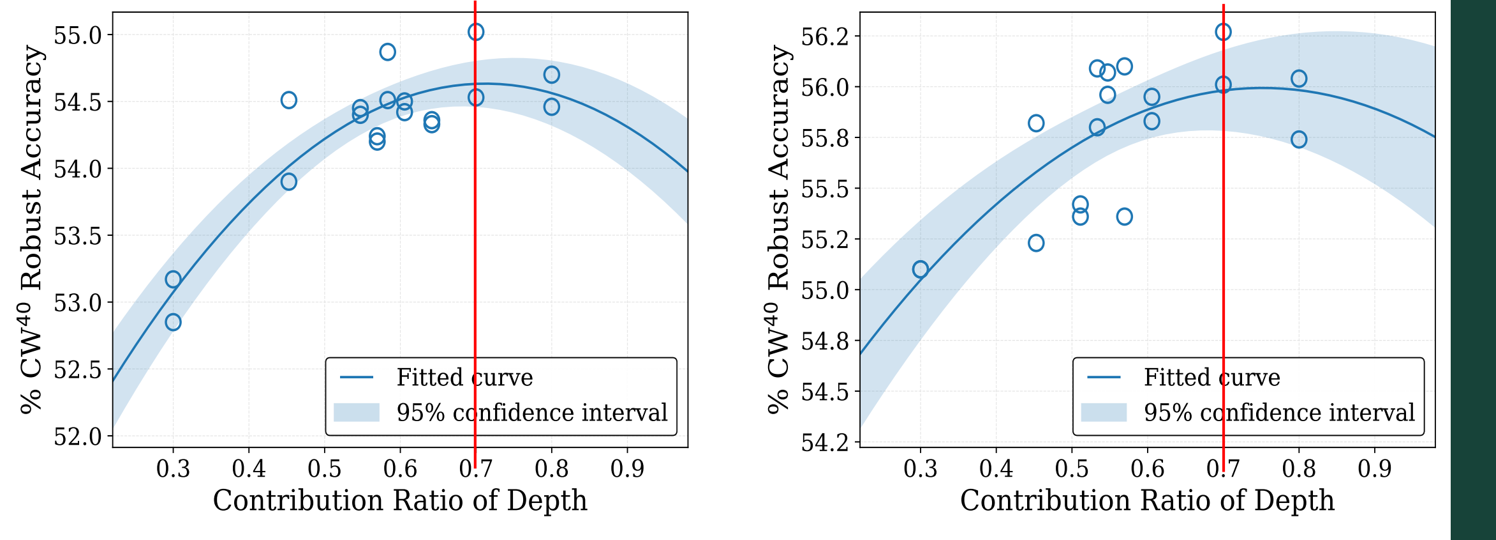
Independent scaling rule: depth@$D_1 : D_2 : D_3 = 2 : 2 : 1$ and width@$W_1 : W_2 : W_3 = 2 : 2.5 : 1$.

Compound Scaling by Depth and Width

RobustScaling setting under the desired FLOP is obtained by solving:

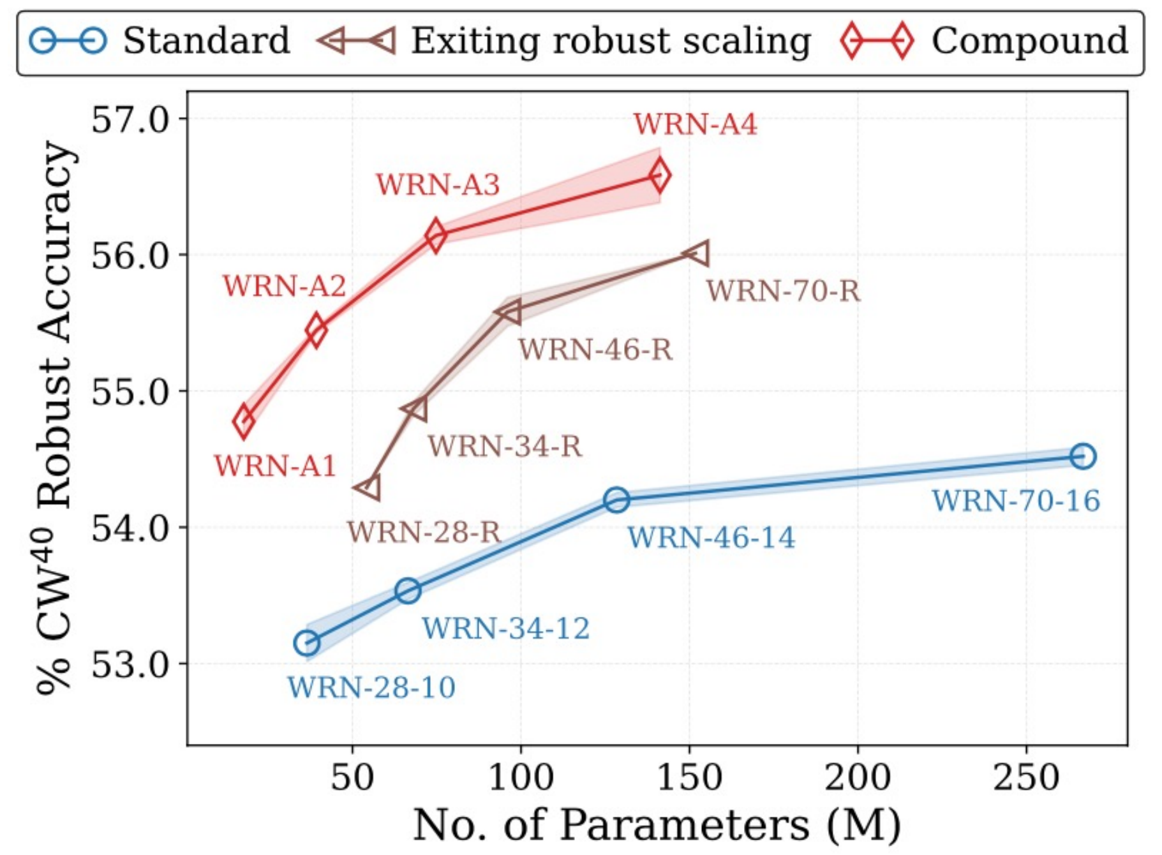
$$FLOPs\left(\sum D_i, \sum W_i\right) \approx \text{ the target}$$

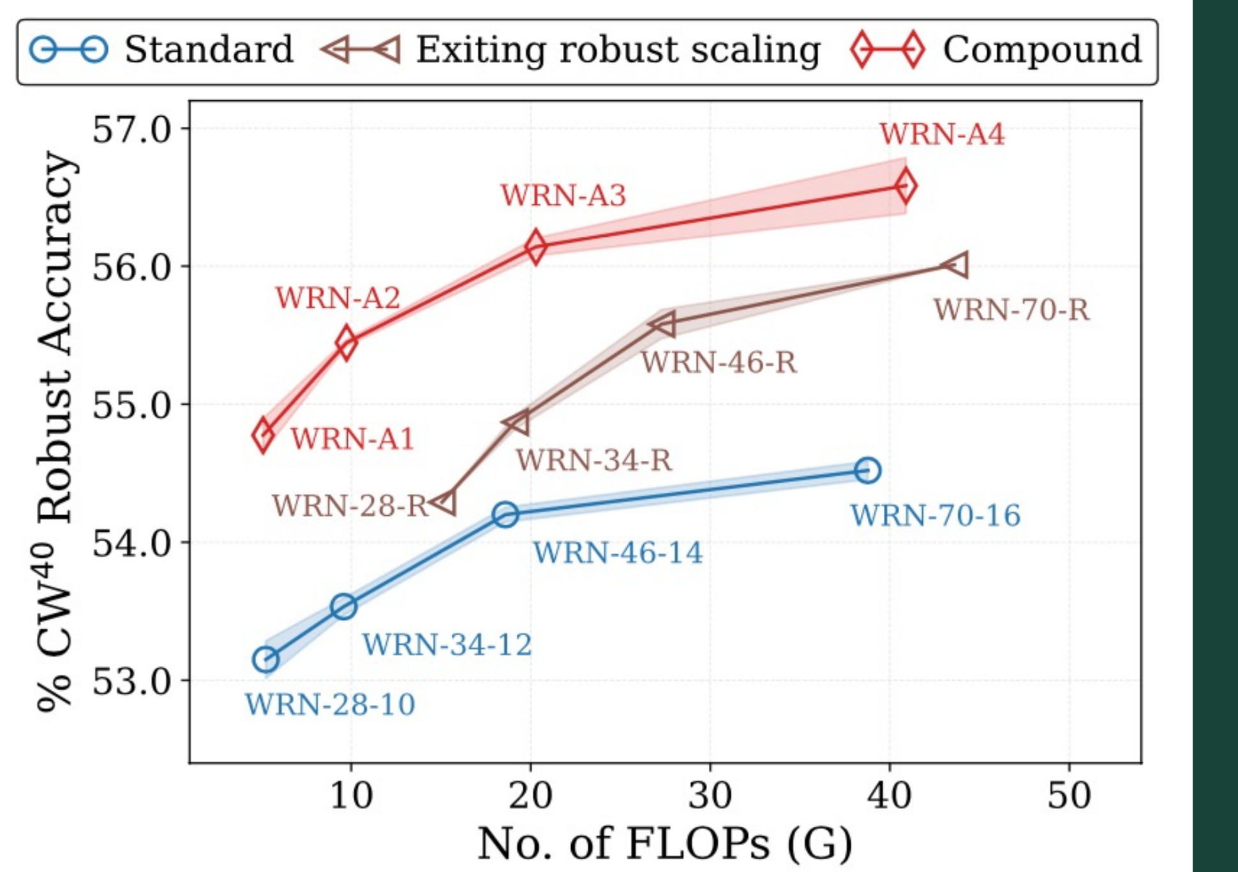
$$r_D = \frac{D_1 + D_2 + D_3}{D_1 + D_2 + D_3 + W_1 + W_2 + W_3} = \frac{2D_3 + 2D_3 + D_3}{2D_3 + 2D_3 + D_3 + 2W_3 + 2.5W_3 + W_3} = 0.7$$



(a) 5G FLOPs (b) 20G FLOPs

RobustScaling





Comparison To State-of-the-Art

Model	#P (M)	#F (G)	CIFAR-10 Clean	CIFAR-10 PGD^{20}	CIFAR-10 CW^{40}	CIFAR-10 AutoAttack	CIFAR-100 Clean	CIFAR-100 PGD^{20}	CIFAR-100 CW^{40}	CIFAR-100 AutoAttack
WRN-28-10	36.5	5.20	$84.62_{\pm0.06}$	$55.90_{\pm0.21}$	$53.15_{\pm0.33}$	$51.66_{\pm0.29}$	$56.30_{\pm0.28}$	$29.91_{\pm0.40}$	$26.22_{\pm0.23}$	$25.26_{\pm0.06}$
RobNet-large-v2	33.3	5.10	$84.57_{\pm0.16}$	$52.79_{\pm0.08}$	$48.94_{\pm0.13}$	$47.48_{\pm0.04}$	$55.27_{\pm0.02}$	$29.23_{\pm0.15}$	$24.63_{\pm0.11}$	$23.69_{\pm0.19}$
AdvRush (7@96)	32.6	4.97	$84.95_{\pm0.12}$	$56.99_{\pm0.08}$	$53.27_{\pm0.03}$	$52.90_{\pm0.11}$	$56.40_{\pm0.09}$	$30.40_{\pm0.21}$	$26.16_{\pm0.03}$	$25.27_{\pm0.02}$
RACL (7@104)	32.5	**4.93**	$83.91_{\pm0.32}$	$55.98_{\pm0.15}$	$53.22_{\pm0.08}$	$51.37_{\pm0.11}$	$56.09_{\pm0.08}$	$30.38_{\pm0.03}$	$26.65_{\pm0.02}$	$25.65_{\pm0.10}$
RobustResNet-A1 (ours)	**19.2**	5.11	**85.46** (↑ **0.5**)	**58.74** (↑ **1.8**)	**55.72** (↑ **2.6**)	**54.42** (↑ **1.5**)	**59.34** (↑ **2.9**)	**32.70** (↑ **2.3**)	**27.76** (↑ **1.1**)	**26.75** (↑ **1.1**)
WRN-34-12	66.5	**9.60**	$84.93_{\pm0.24}$	$56.01_{\pm0.28}$	$53.53_{\pm0.15}$	$51.97_{\pm0.09}$	$56.08_{\pm0.41}$	$29.87_{\pm0.23}$	$26.51_{\pm0.11}$	$25.47_{\pm0.10}$
WRN-34-R	68.1	19.1	$85.80_{\pm0.08}$	$57.35_{\pm0.09}$	$54.77_{\pm0.10}$	$53.23_{\pm0.07}$	$58.78_{\pm0.11}$	$31.17_{\pm0.08}$	$27.33_{\pm0.11}$	$26.31_{\pm0.03}$
RobustResNet-A2 (ours)	**39.0**	10.8	**85.80** (↑ **0.0**)	**59.72** (↑ **2.4**)	**56.74** (↑ **2.0**)	**55.49** (↑ **2.3**)	**59.38** (↑ **0.6**)	**33.0** (↑ **1.8**)	**28.71** (↑ **1.4**)	**27.68** (↑ **1.4**)
WRN-46-14	128	**18.6**	$85.22_{\pm0.15}$	$56.37_{\pm0.18}$	$54.19_{\pm0.11}$	$52.63_{\pm0.18}$	$56.78_{\pm0.47}$	$30.03_{\pm0.07}$	$27.27_{\pm0.05}$	$26.28_{\pm0.03}$
RobustResNet-A3 (ours)	**75.9**	19.9	**86.79** (↑ **1.6**)	**60.10** (↑ **3.7**)	**57.29** (↑ **3.1**)	**55.84** (↑ **3.2**)	**60.16** (↑ **3.4**)	**33.59** (↑ **3.6**)	**29.58** (↑ **2.3**)	**28.48** (↑ **2.2**)
WRN-70-16	267	**38.8**	$85.51_{\pm0.24}$	$56.78_{\pm0.16}$	$54.52_{\pm0.16}$	$52.80_{\pm0.14}$	$56.93_{\pm0.61}$	$29.76_{\pm0.17}$	$27.20_{\pm0.16}$	$26.12_{\pm0.24}$
RobustResNet-A4 (ours)	**147**	39.4	**87.10** (↑ **1.6**)	**60.26** (↑ **3.5**)	**57.9** (↑ **3.4**)	**56.29** (↑ **3.5**)	**61.66** (↑ **4.7**)	**34.25** (↑ **4.5**)	**30.04** (↑ **2.8**)	**29.00** (↑ **2.9**)

Conclusions

- Architectural design significantly affects adversarial robustness.
- Residual block advancements for standard ERM training translate well to improve adversarial robustness under adversarial training.
- **RobustResNets** are proposed based on the observations from block and network levels.
- **RobustResNets** achieves better adversarial robustness while being more compact than state-of-the-art solutions.