Part-Based Models Improve Adversarial Robustness

Chawin Sitawarin[1] Kornrapat Pongmala[1] Yizheng Chen[1]
Nicholas Carlini[2] David Wagner[1]
[1]UC Berkeley [2]Google

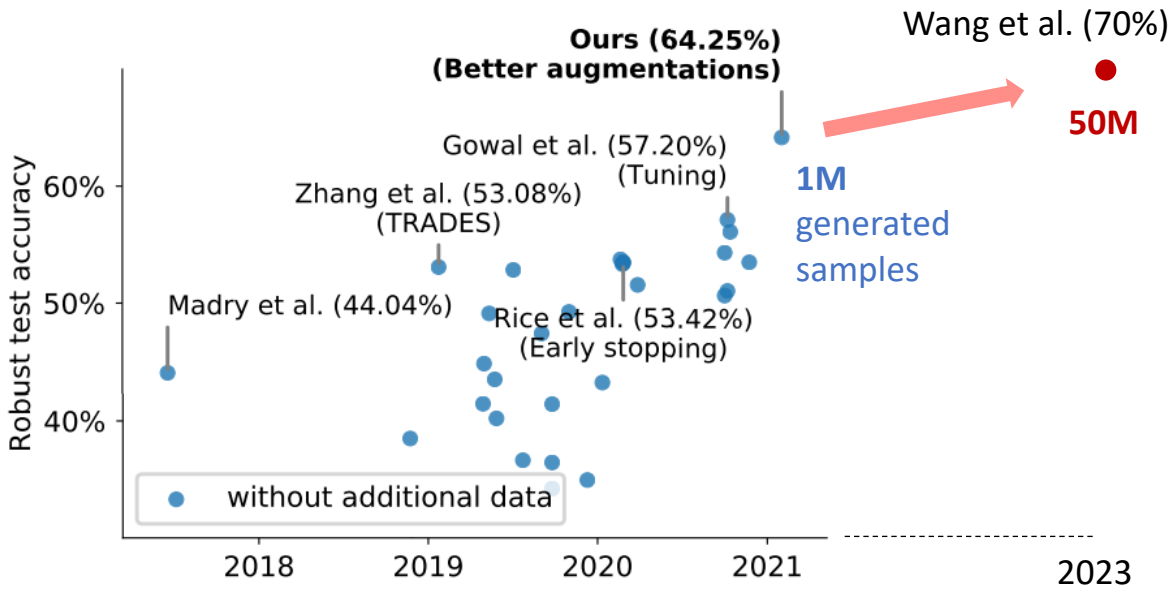
ICLR
International Conference On Learning Representations

Berkeley
UNIVERSITY OF CALIFORNIA

Defense against Adversarial Examples

> Where are we at?

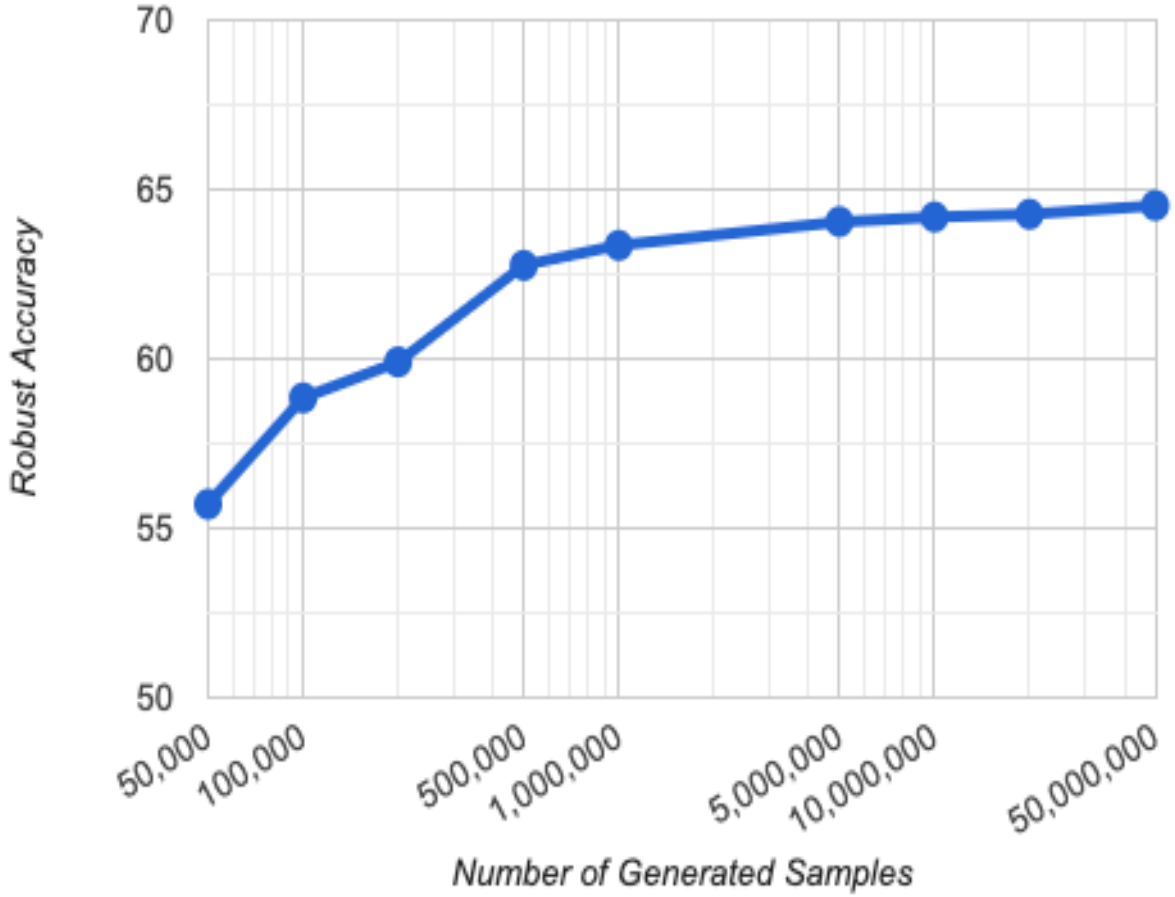
- Adversarial Training [Madry et al., 2018] has been the go-to defense against adversarial examples.
- Recent works rely on synthetic data from advanced generative models.



Defense against Adversarial Examples

> Are we done?

- The improvement plateaus…
- Linear improvement requires exponentially more computes.
- Large model + more data + Adversarial Training = way too expensive!
- We are probably not done yet!



[Wang et al., 2023]

Part-Based Model

> An alternative to “more data”

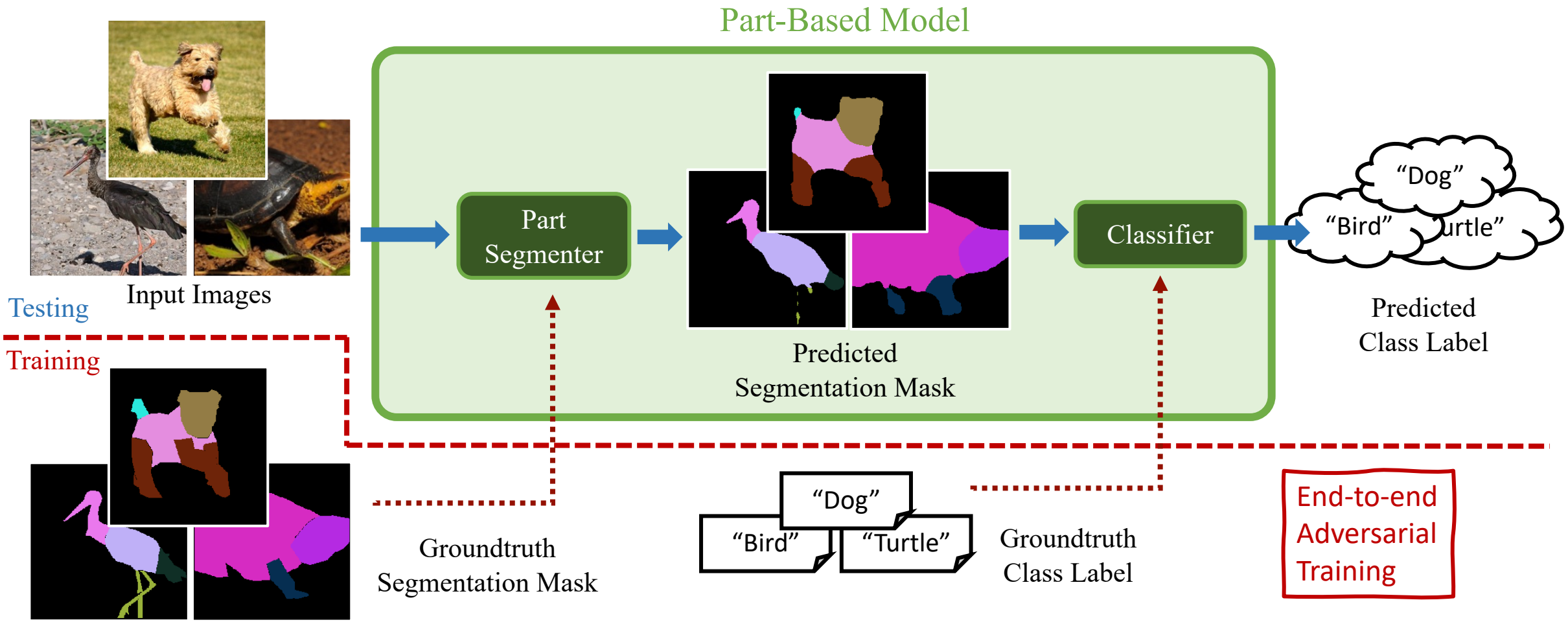
- We want neural networks to rely on a similar set of features as we do, i.e., robust features.
- Can we achieve this by not relying on lots of data? Maybe just give the model a hint!
- Leverage richer or fine-grained annotation, specifically part segmentation.

Part segmentation
(fine-grained label)

Class label

Part-Based Model

> Learning robust features with fine-grained labels

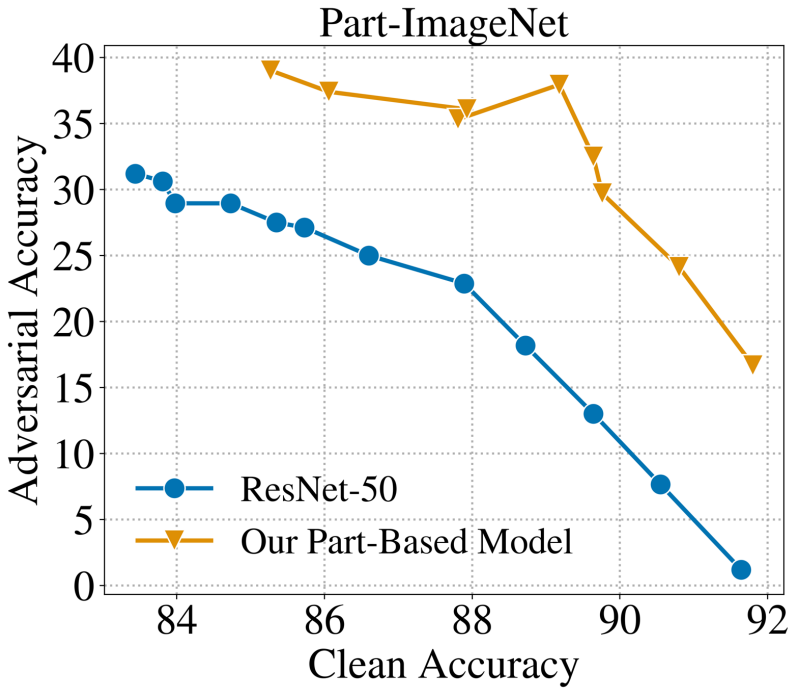


Part-Based Model

> Learning robust features with fine-grained labels

- Huge improvement on robustness-accuracy trade-off across 3 datasets: PartImageNet, Cityscapes, PASCAL-Part.

- Also improves general robustness (1) common corruption, (2) shape-texture bias, and (3) background-foreground bias.



Takeaway: Richer auxiliary task/label is a promising alternative to improving adversarial and general robustness.

Models	Corruptions	Texture Bias	Background Bias
ResNet-50	82.3	40.6	58.6
Part Model	**85.8**	**45.7**	**65.1**