Part-Based Models Improve Adversarial Robustness

Chawin Sitawarin[1] Kornrapat Pongmala[1]

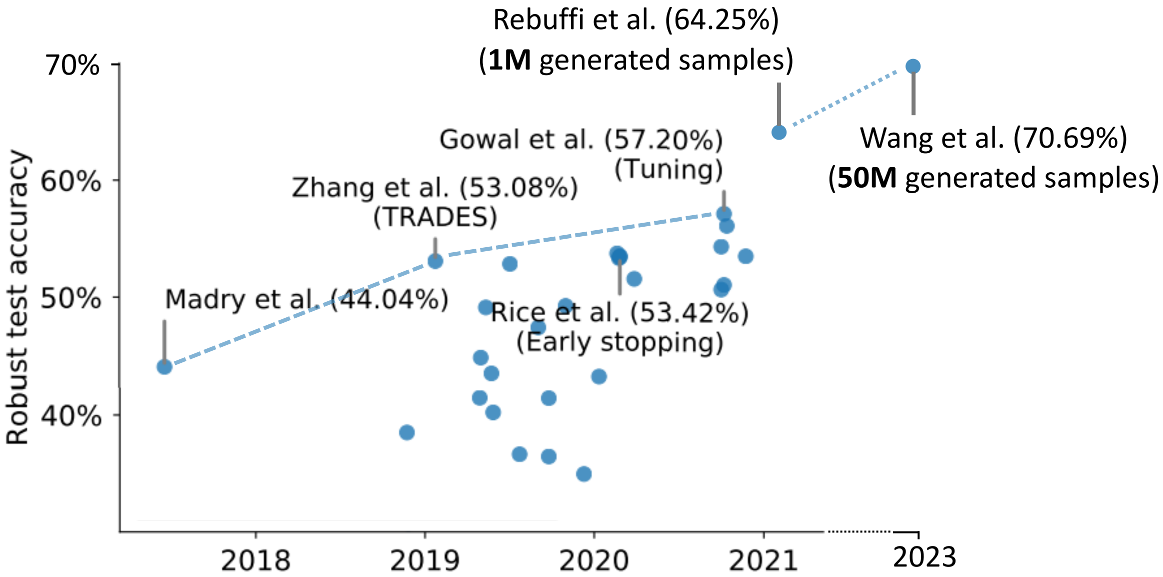
Yizheng Chen[1] Nicholas Carlini[2] David Wagner[1]

[1]UC Berkeley [2]Google

Defense against Adversarial Examples

> Where are we at?

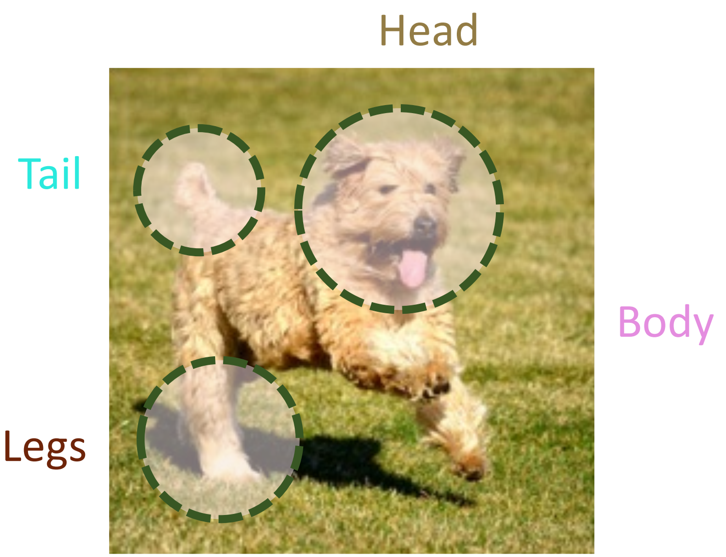
- Adversarial Training [Madry et al., 2018] has been the go-to defense against adversarial examples, but the progress has plateaued.
- Recent works use generated data, but the return is diminishing.

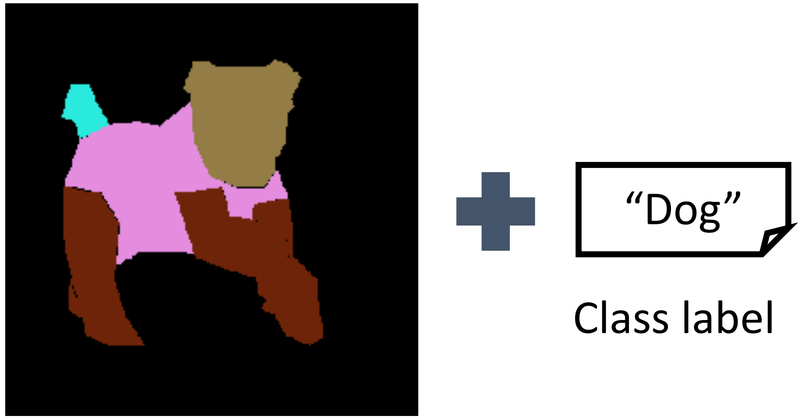


Part-Based Model

> An alternative to “more data”

- We want neural networks to rely on a similar set of features as humans do, i.e., robust features.
- More data and more aggressive augmentation don’t seem to get us there yet. Maybe we should just give the model a hint?
- Leverage richer or fine-grained annotation, specifically part segmentation.



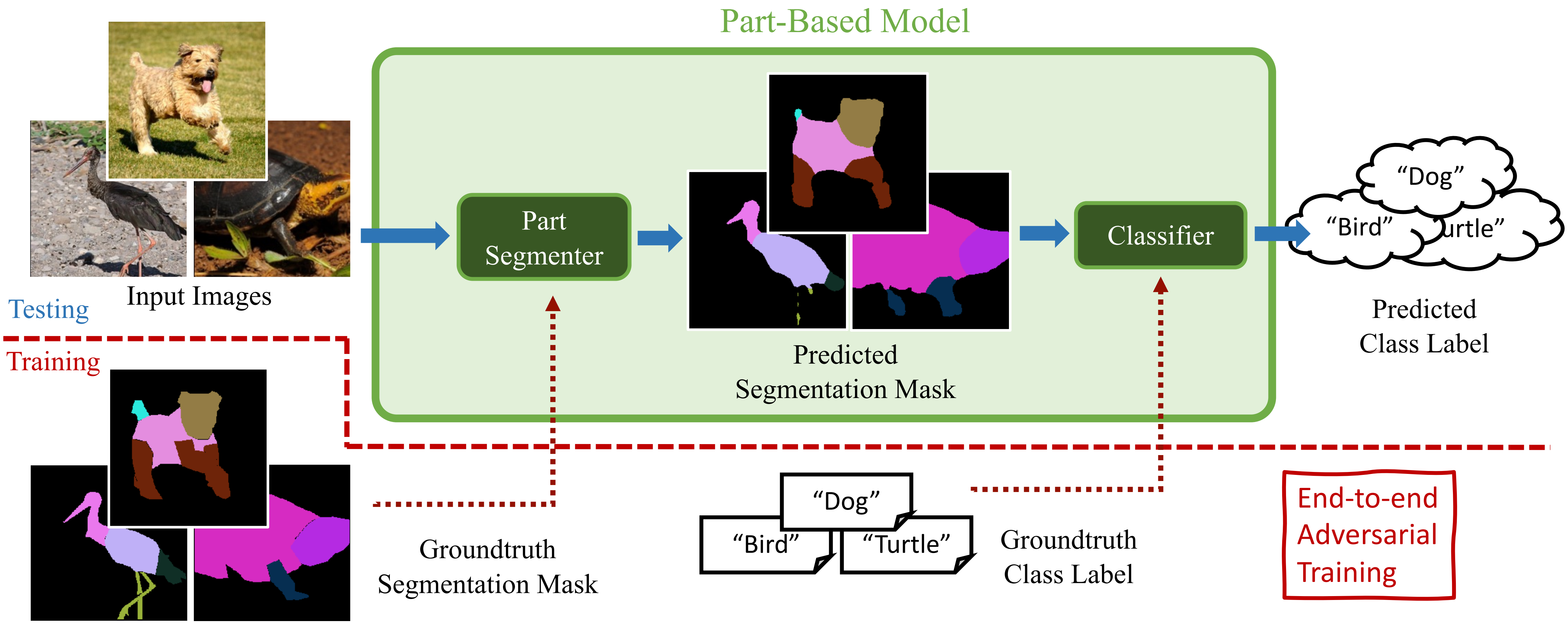


Part segmentation
(fine-grained label)

Class label

Part-Based Model

> Learning robust features with fine-grained labels



Part-Based Model

Input Images

Testing

Training

Part Segmenter

Predicted Segmentation Mask

Classifier

“Dog”

“Bird”

urtle”

Predicted Class Label

Groundtruth Segmentation Mask

“Dog”

“Bird”

“Turtle”

Groundtruth Class Label

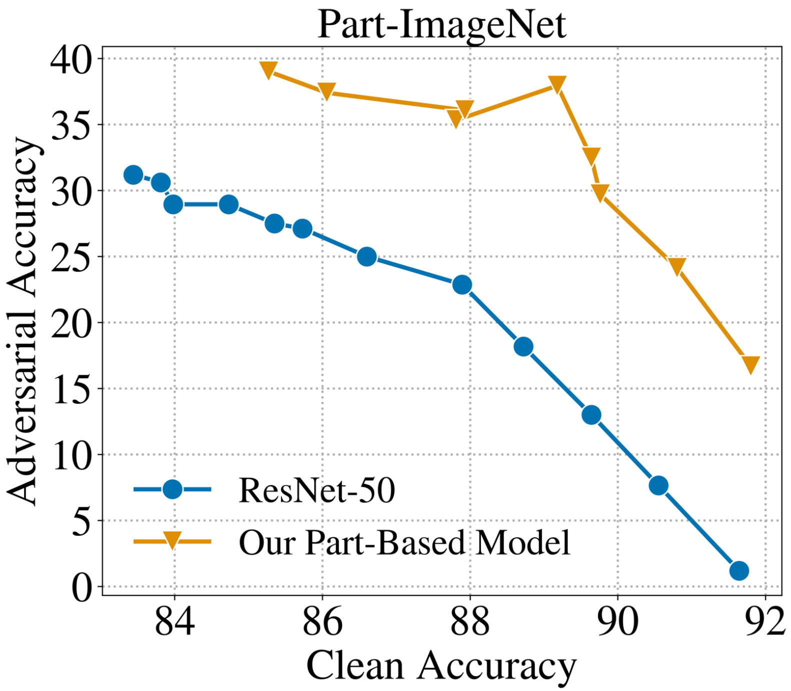
End-to-end Adversarial Training

Part-Based Model

> Learning robust features with fine-grained labels

- Huge improvement (10-17%) on robustness-accuracy trade-off across 3 datasets: PartImageNet, Cityscapes, PASCAL-Part.

- Also improves general robustness by 3-7%: (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**