Cornell University

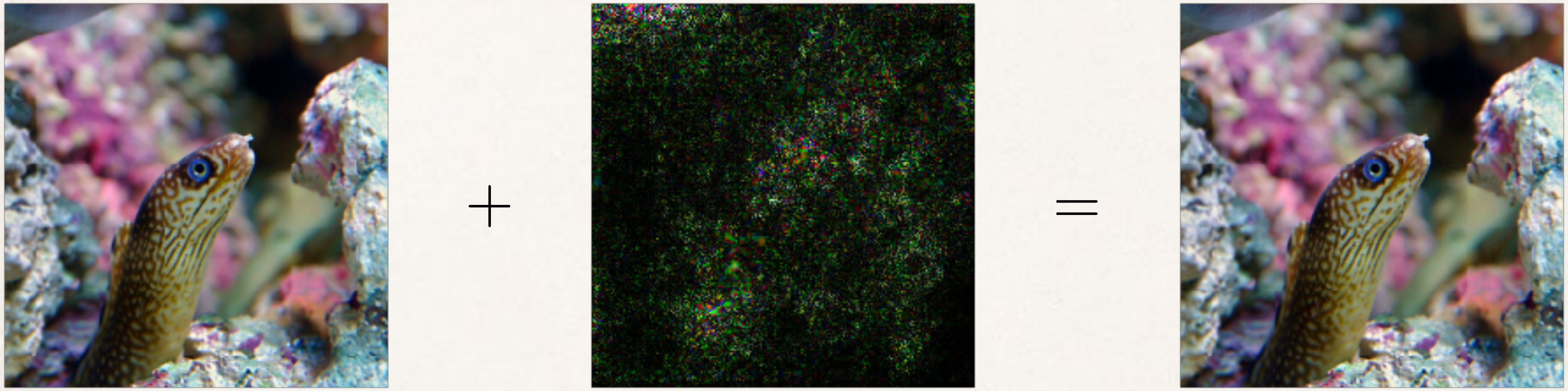
Uber AI

SimBA: Simple Black-box Adversarial Attacks

Chuan Guo, Jacob R. Gardner, Yurong You, Andrew Gordon Wilson, Kilian Q. Weinberger

June 12, 2019

Adversarial Perturbation

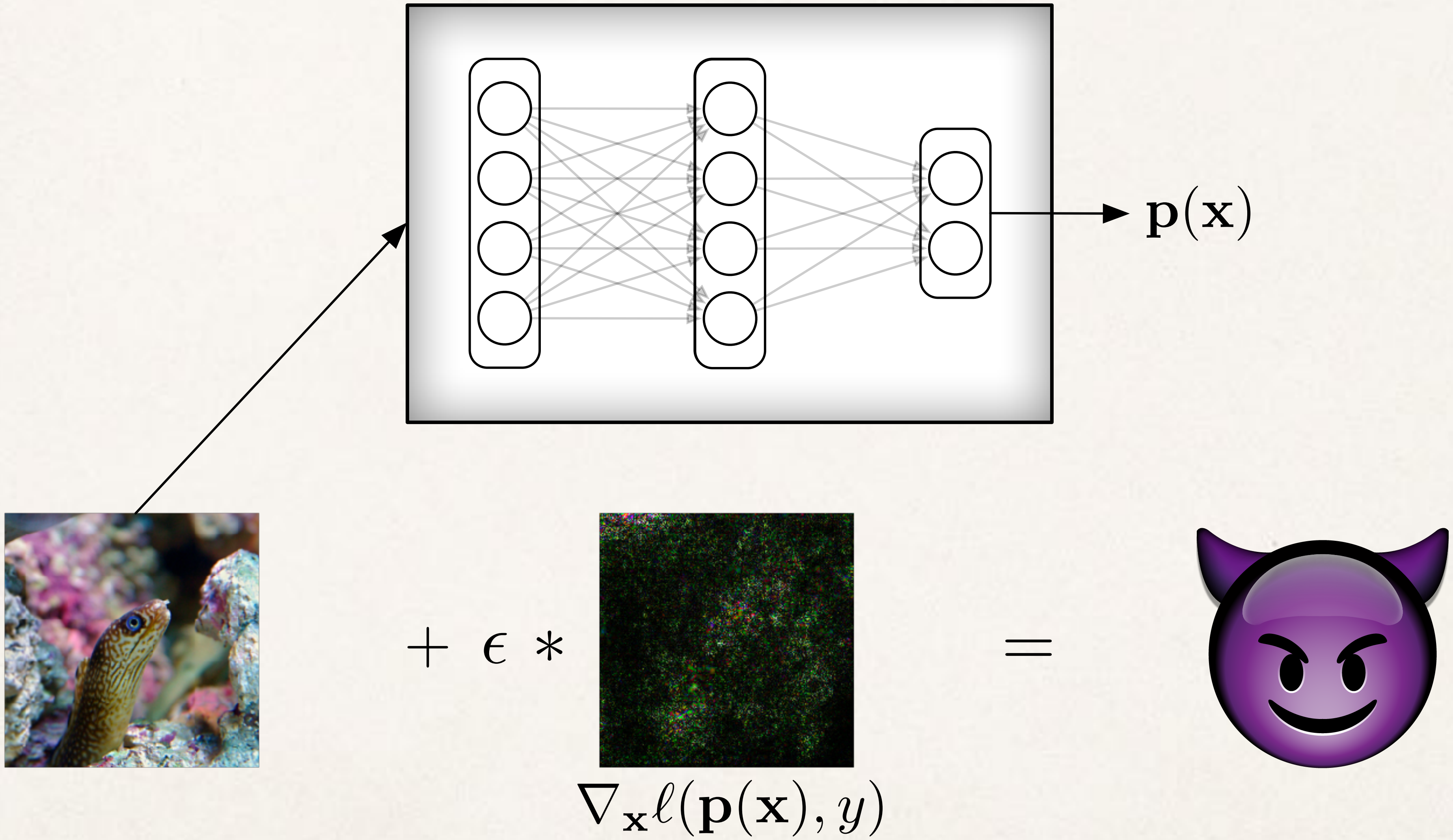


97.75% Eel

99.99% Goldfish

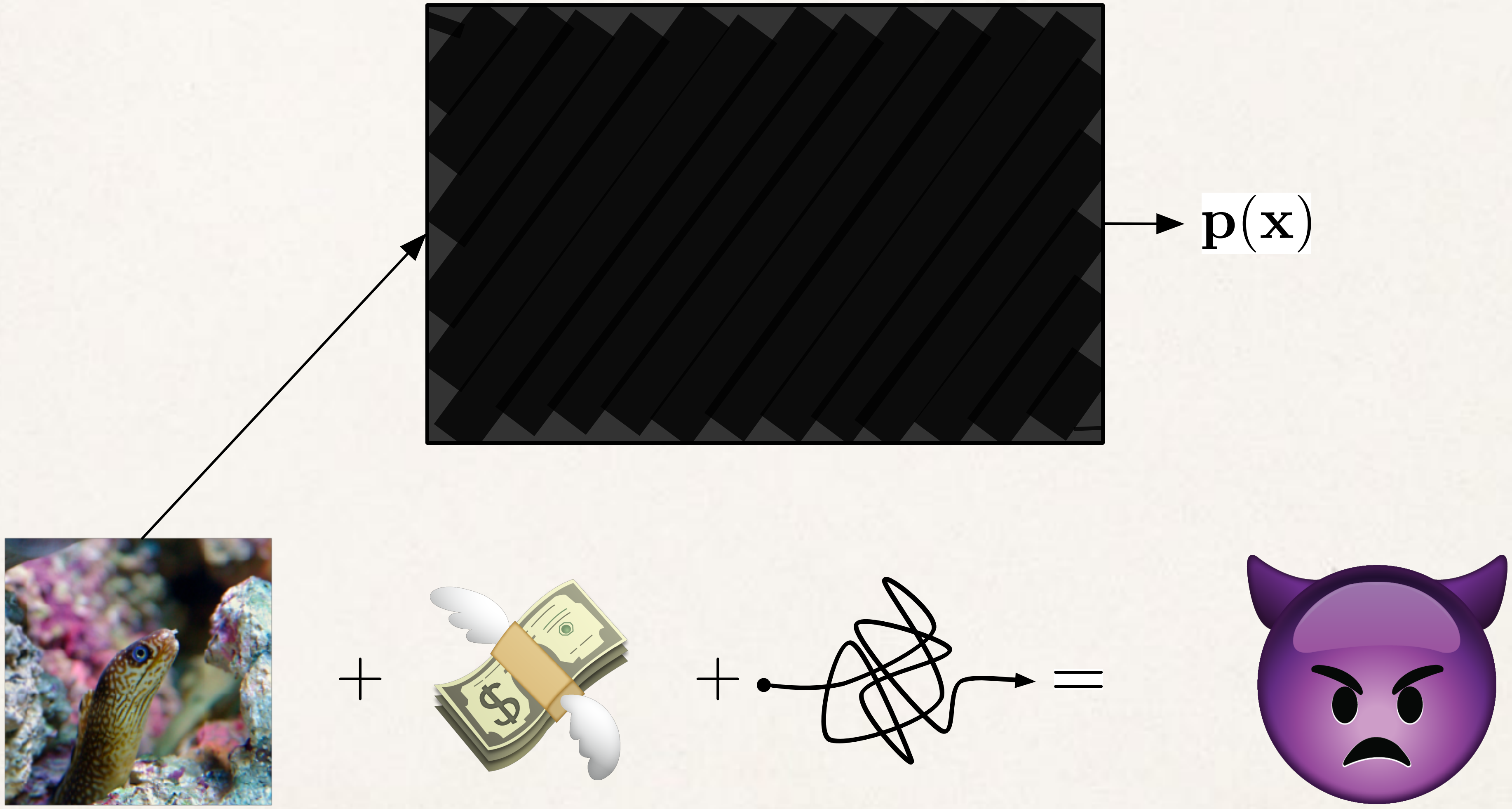
- Small (imperceptible) change in input that alters model decision
- Security implications for critical applications

White-box Attacks



- White-box attacks are simple and efficient due to access to gradients

Black-box Attacks



- Black-box attacks are costly and existing approaches are complicated

Black-box Attacks

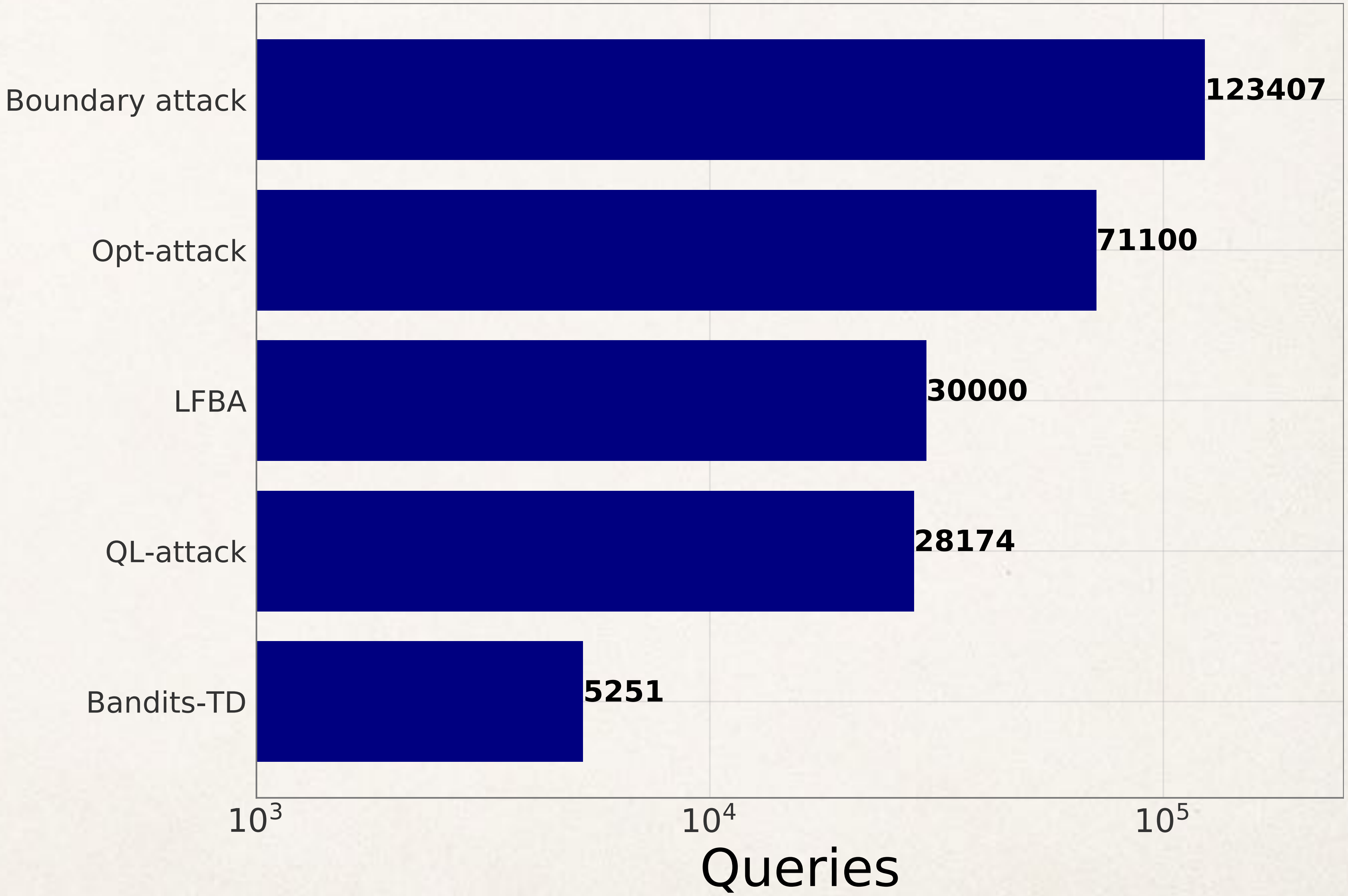
Brendel et al., 2018

Cheng et al., 2018

Guo et al., 2019

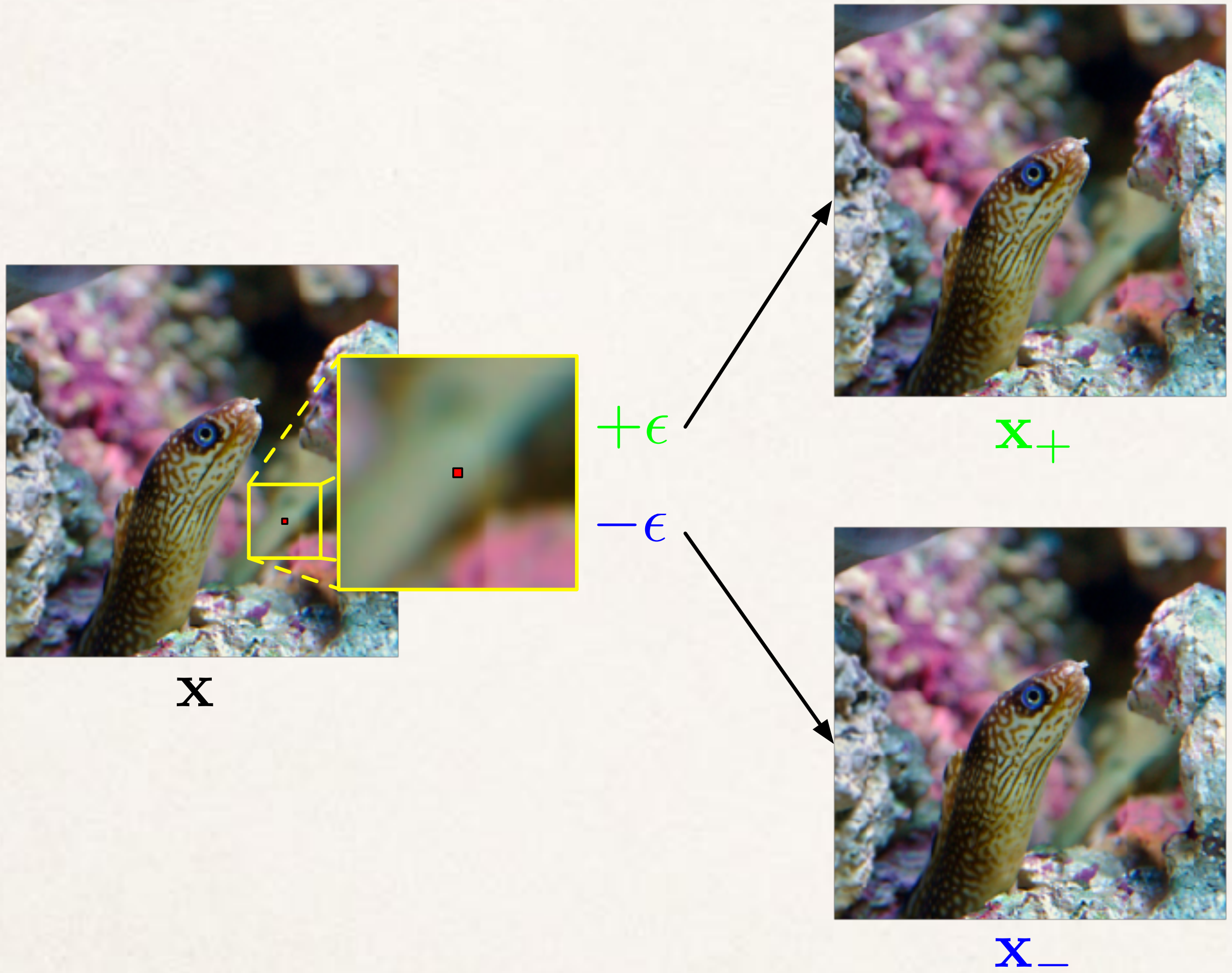
Ilyas et al., 2018

Ilyas et al., 2019

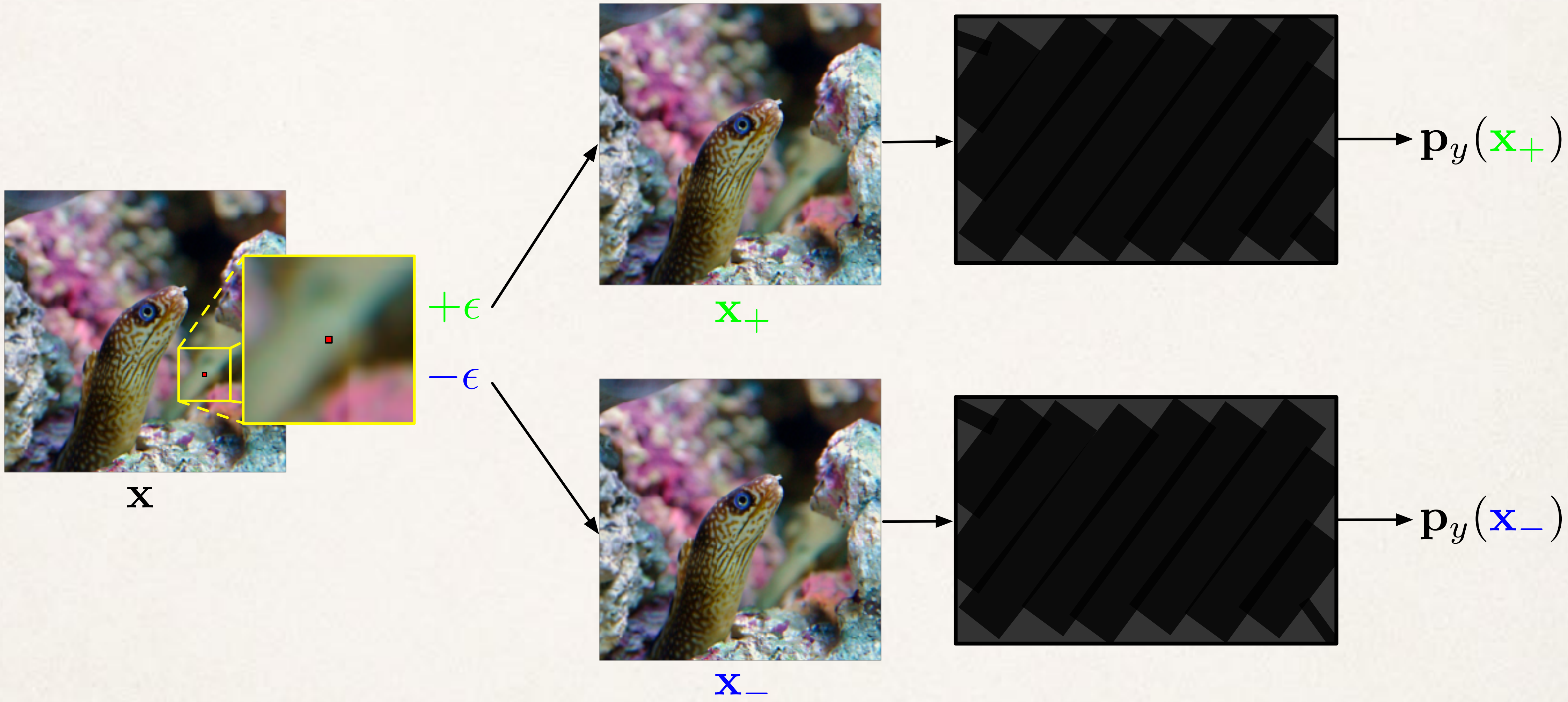


- Black-box attacks are costly and existing approaches are complicated

Simple Black-box Attack (SimBA)



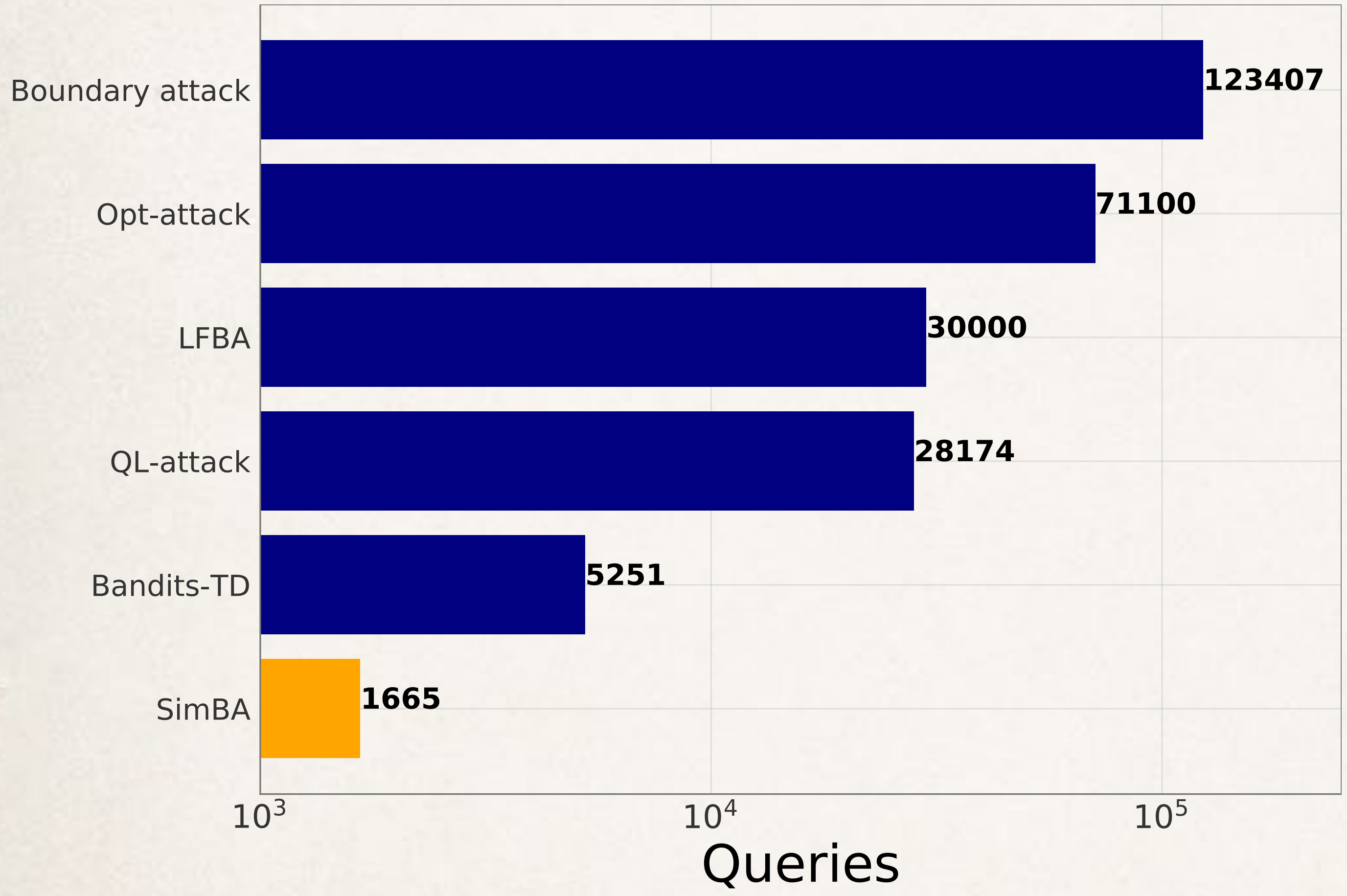
Simple Black-box Attack (SimBA)



$$\mathbf{x} \leftarrow \operatorname{argmin}\{\mathbf{p}_y(\mathbf{x}_+), \mathbf{p}_y(\mathbf{x}_-), \mathbf{p}_y(\mathbf{x})\}$$

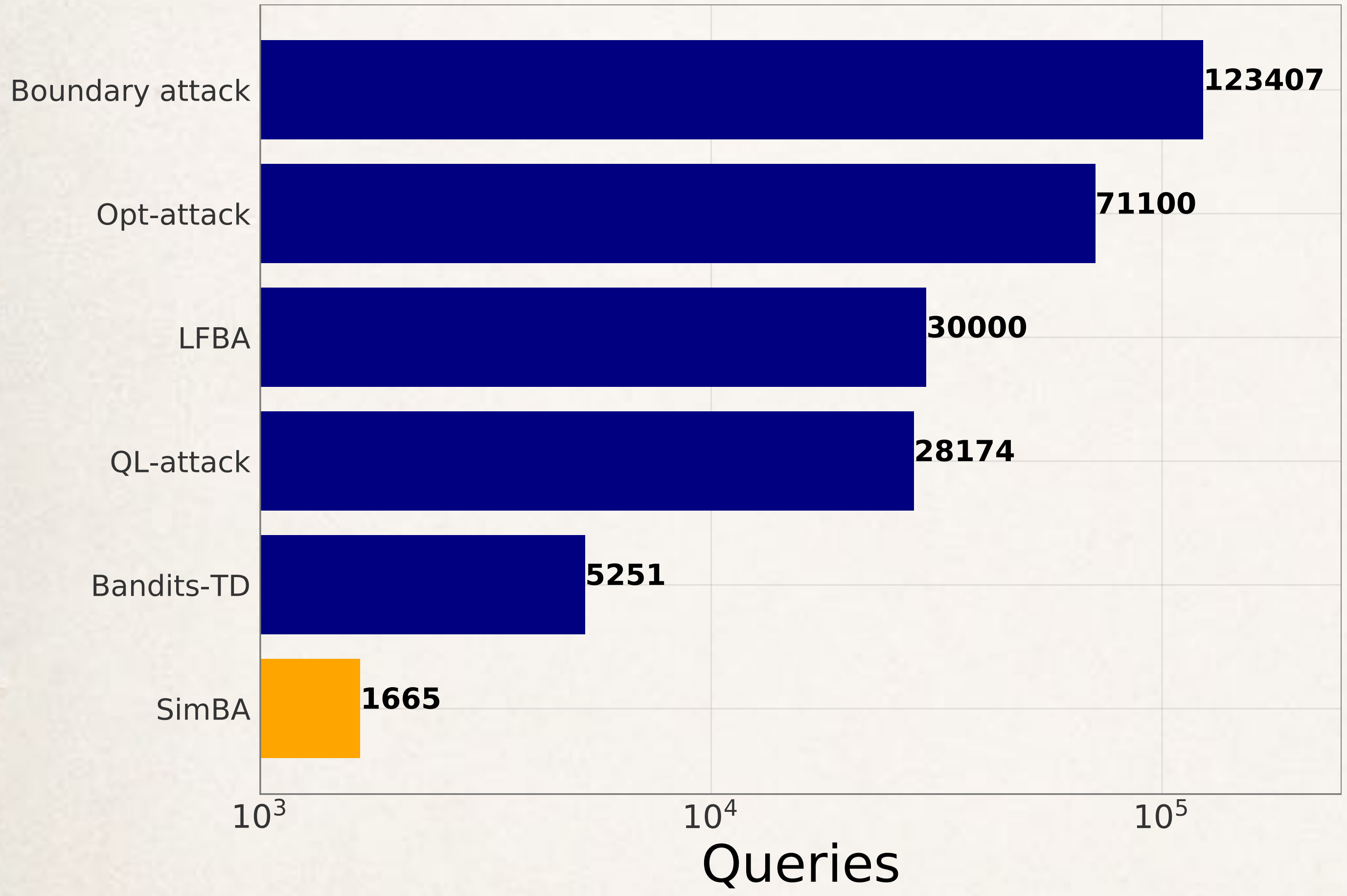
- Can be implemented in ~20 lines of code!

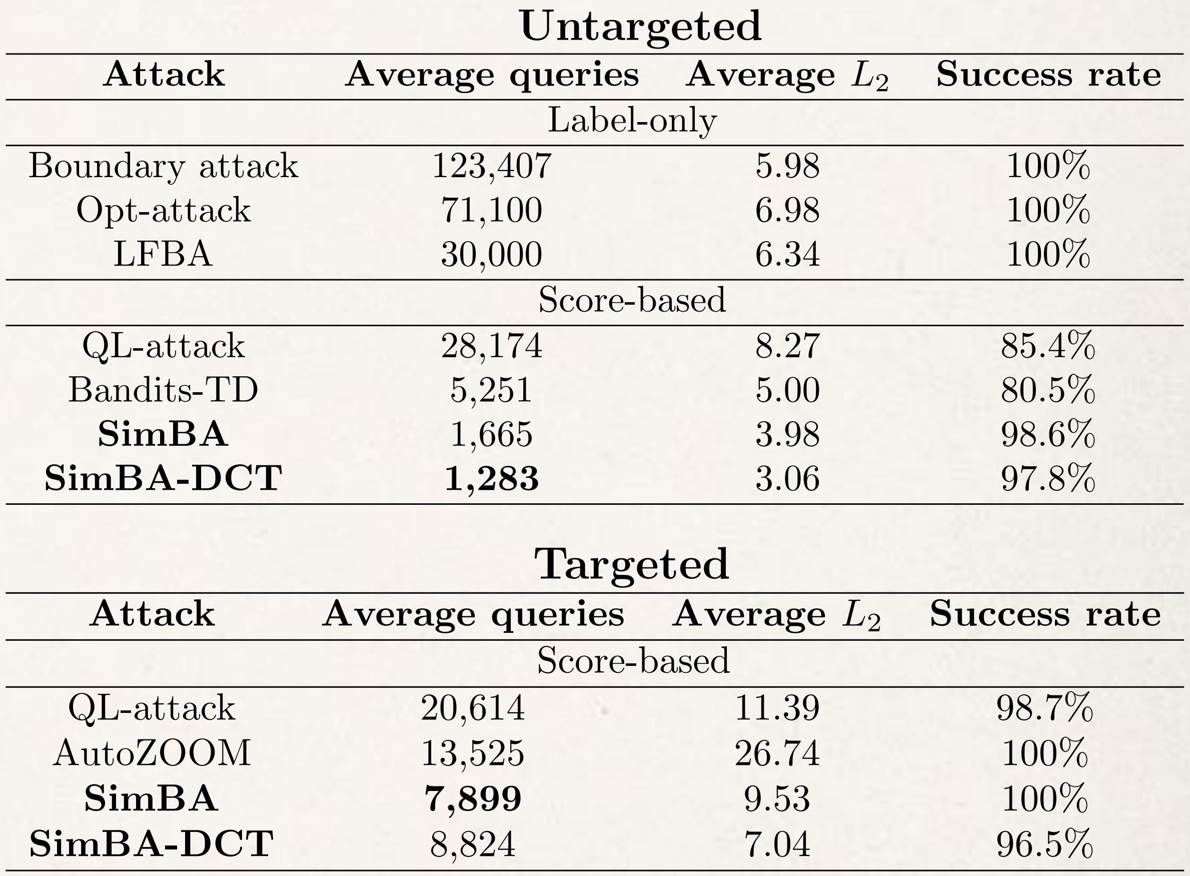
Evaluation



- ImageNet classification with ResNet-50 model
- Drastically improved performance compared to previous SOTA

Evaluation





Untargeted

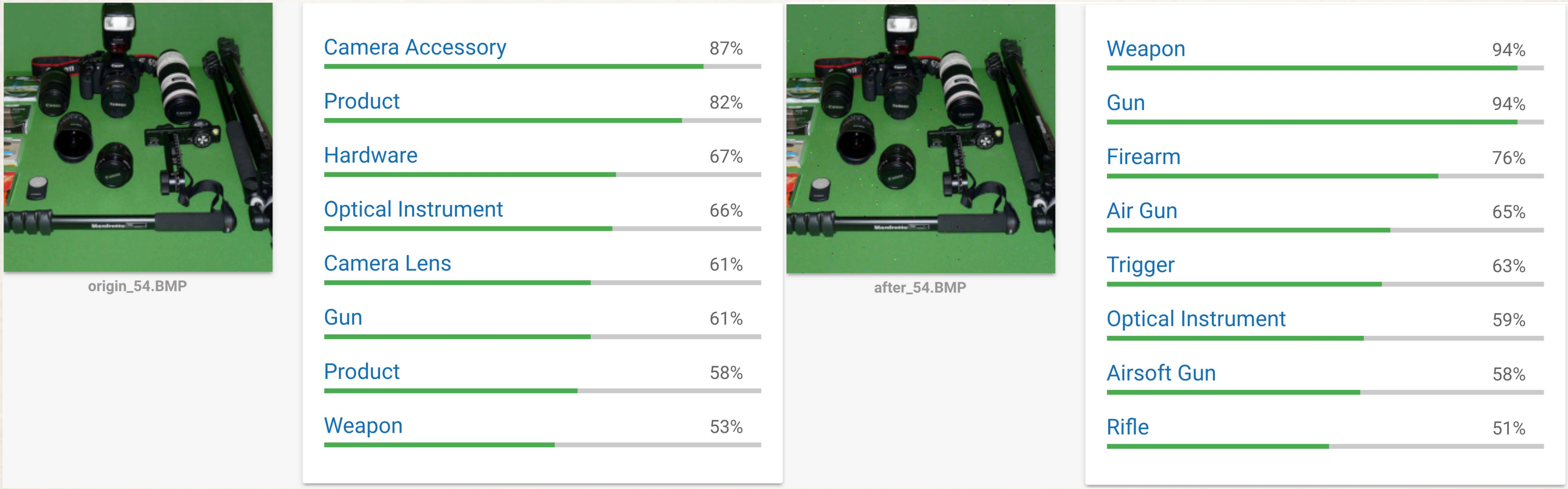
Attack	**Average queries**	**Average L_2**	**Success rate**
	Label-only		
Boundary attack	123,407	5.98	100%
Opt-attack	71,100	6.98	100%
LFBA	30,000	6.34	100%
	Score-based		
QL-attack	28,174	8.27	85.4%
Bandits-TD	5,251	5.00	80.5%
SimBA	1,665	3.98	98.6%
SimBA-DCT	**1,283**	3.06	97.8%

Targeted

Attack	**Average queries**	**Average L_2**	**Success rate**
	Score-based		
QL-attack	20,614	11.39	98.7%
AutoZOOM	13,525	26.74	100%
SimBA	**7,899**	9.53	100%
SimBA-DCT	8,824	7.04	96.5%

- ImageNet classification with ResNet-50 model
- Drastically improved performance compared to previous SOTA

Attacking Google Cloud Vision



origin_54.BMP

Camera Accessory	87%
Product	82%
Hardware	67%
Optical Instrument	66%
Camera Lens	61%
Gun	61%
Product	58%
Weapon	53%

after_54.BMP

Weapon	94%
Gun	94%
Firearm	76%
Air Gun	65%
Trigger	63%
Optical Instrument	59%
Airsoft Gun	58%
Rifle	51%

- Generated using 5000 queries ($10 cost)
- 70% success rate across 50 images

Collaborators

Jacob R. Gardner[2]

Yurong You[1]

Andrew Gordon Wilson[1]

Kilian Q. Weinberger[1]

Poster session: June 12 (today) 6:30-9:00 PM @ Pacific Ballroom #70

[1] Cornell University

[2] Uber AI Labs