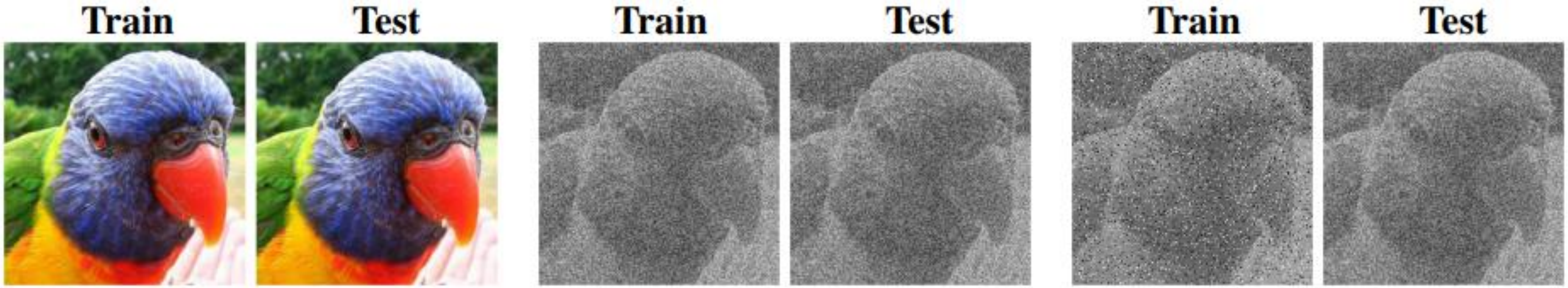
EN.601.482/682 Deep Learning

An Introduction to Adversarial Attacks

Mathias Unberath, PhD
Assistant Professor
Dept of Computer Science
Johns Hopkins University

Generalization



(a) Super-human performance (b) Super-human performance (c) Chance level performance

An easy test: This is an example for "Bird"

Consider this your ***testing*** *stage.* **What is this?**

"Bird", too! Easy right? **So how would a CNN do?**

Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. NeurIPS (pp. 7549-7561).

Evaluation condition	human observers	A1	A2	A3	A4	A5	A6	A7	A8	A9	B1	B2	B3	B4	B5	B6	B7	B8	B9	C1	C2
colour	88.5	96.7	90.6	50.0	83.1	86.1	84.2	90.8	10.4	8.1	97.9	95.4	72.3	93.0	91.1	92.4	94.9	10.2	11.2	95.5	95.9
greyscale	86.6	87.8	95.6	94.1	86.2	93.2	87.8	90.5	10.3	9.8	94.0	96.8	96.2	93.3	95.7	94.3	90.9	11.4	12.8	94.8	95.1
contrast (5%)	47.6	13.1	14.2	89.4	19.6	39.8	17.1	10.2	28.6	29.0	46.3	51.7	95.1	50.5	79.1	59.4	45.2	34.6	37.9	90.9	88.2
low−pass (std=7)	48.5	18.9	16.1	16.4	78.4	11.9	16.0	9.8	6.9	6.6	16.0	18.6	14.4	87.2	20.5	13.8	13.5	7.1	9.3	74.7	74.9
high−pass (std=0.7)	49.8	21.1	24.7	29.9	11.7	92.6	27.7	8.3	10.4	20.6	25.1	22.8	29.2	25.0	94.3	27.5	28.3	18.9	19.8	91.4	90.7
phase noise (90°)	57.4	23.3	28.3	31.2	27.0	46.6	81.4	24.4	7.4	8.9	30.8	31.4	30.6	31.4	43.4	87.4	24.1	7.8	7.6	82.9	82.6
rotation (90°)	78.5	36.5	43.3	39.9	31.8	40.4	37.7	89.0	8.5	8.0	38.5	41.9	40.3	35.2	40.1	40.5	89.0	8.3	8.8	80.1	80.5
salt−and−pepper noise (0.2)	NA	6.1	6.4	5.8	7.9	6.2	6.2	6.4	79.4	6.2	6.2	6.1	6.3	5.4	5.8	5.7	6.2	89.6	6.2	78.6	13.6
uniform noise (0.35)	45.6	6.2	7.3	6.9	9.0	7.3	6.2	6.0	10.2	80.3	84.6	83.3	85.0	84.6	83.7	82.5	83.8	85.4	89.8	11.0	71.5

= manipulation included in training data

16 classes: Chance is at 6.25%

Humans were presented the images for 200 ms

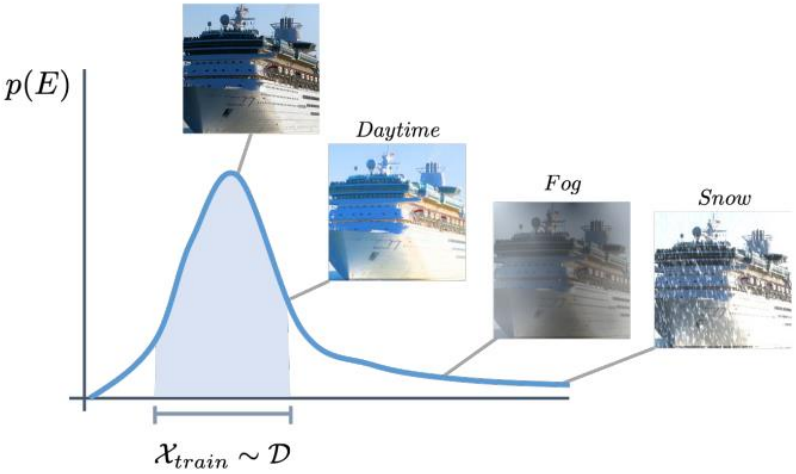
CNNs on the same data they were trained on: **Super-human performance!**

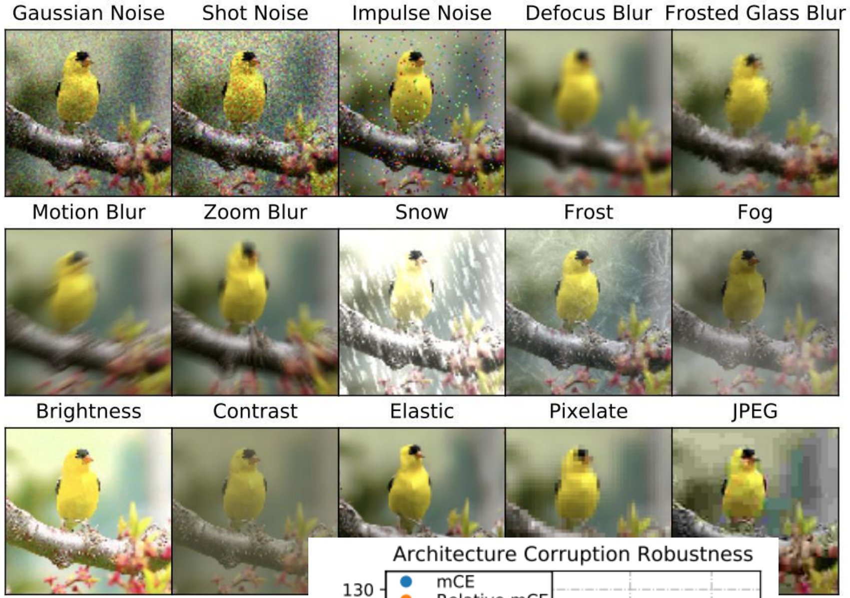
→As bad as chance level on unseen distortions!

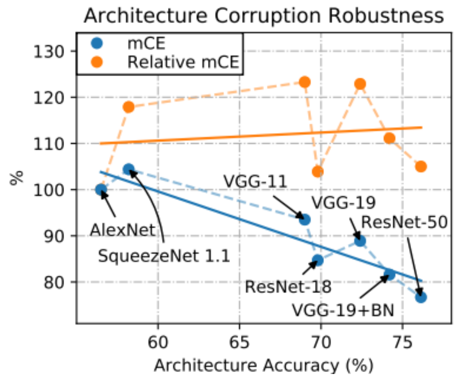
→Slightly better than chance!

Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. NeurIPS (pp. 7549-7561).

“Non-adversarial Robustness”







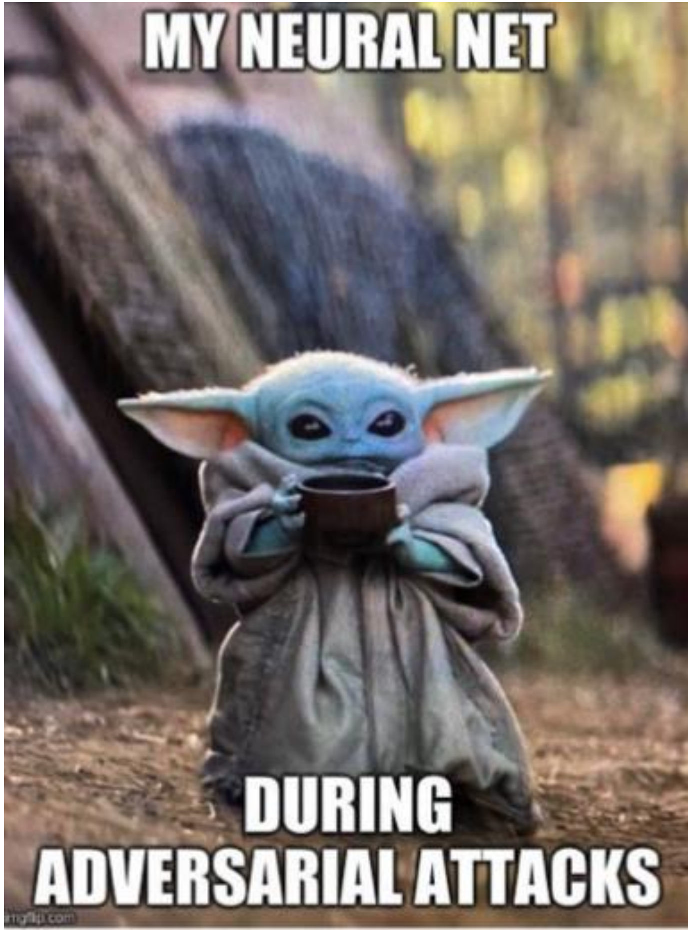
Drenkow, N., Sani, N., Shpitser, I., & Unberath, M. (2021). A systematic review of robustness in deep learning for computer vision: Mind the gap?. *arXiv preprint arXiv:2112.00639*.
Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. arXiv preprint arXiv:1903.12261.

What then is adversarial robustness?

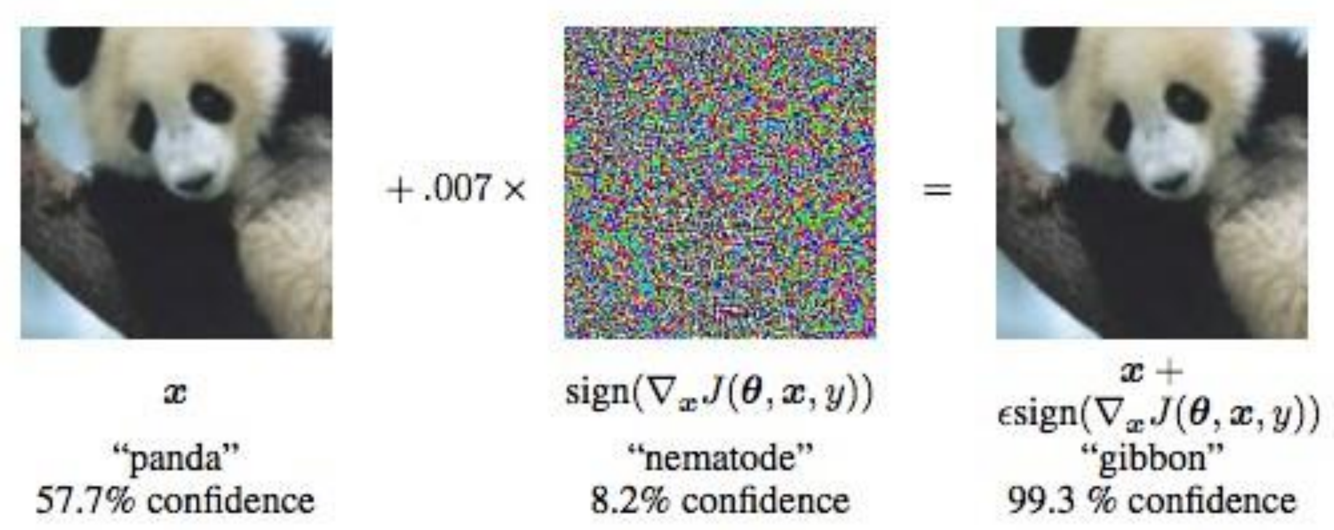
Can we craft a signal that forces networks to fail?

If possible: CNNs are vulnerable!

How can we defend against such attacks?



Adversarial Examples: What are they?



Adversarial Example — Definition

For loss *L* and model *f*, find a small perturbation δ of *x* that maximizes the error:

$$\max_{\delta} L(f(x+\delta))$$

$$\text{s.t. } ||\delta|| \leq \epsilon$$

Adversarial Example — Definition

For loss *L* and model *f*, find a small perturbation δ of *x* that maximizes the error:

$$\max_{\delta} L(f(x+\delta))$$

$$\text{s.t. } ||\delta|| \leq \epsilon$$

Important caveat: Attack needs to be imperceptible to humans!

Intro Adversarial Attacks

Fast Gradient Sign Methods

Adversarial Attack Methods: FGSM

- One method for calculating adversarial examples is to approximate a solution for the optimization problem via gradient descent.
- The Fast Gradient Sign Method (FGSM):

$$x^* = x + \lambda sign(\nabla L(f(x)))$$

- Uses the [element-wise] sign of the gradient for fast calculation.
- Shown to satisfy definition of L_inf norm adversarial attack.

Adversarial Attack Methods: FGSM

- One method for calculating adversarial examples is to approximate a solution for the optimization problem via gradient descent.
- The Fast Gradient Sign Method (FGSM):

Why sign?

$$x^* = x + \lambda sign(\nabla L(f(x)))$$

- Uses the [element-wise] sign of the gradient for fast calculation.
- Shown to satisfy definition of L_inf norm adversarial attack.

Adversarial Attack Methods: I-FGSM

- An extension of FGSM is the iterative version:

$$x_0 = x$$
$$x_{k+1} = x_k + Clip(\lambda sign(\nabla L(f(x))))$$

- Clipping is done so that x_k stays within the epsilon ball around x.
- For large number of iterations (e.g. 200 per sample), this breaks most models.
- Also known as projected gradient descent (PGD).

→ Close to state-of-the-art attacks.

Intro Adversarial Attacks

One-pixel Attacks

One-pixel Attack

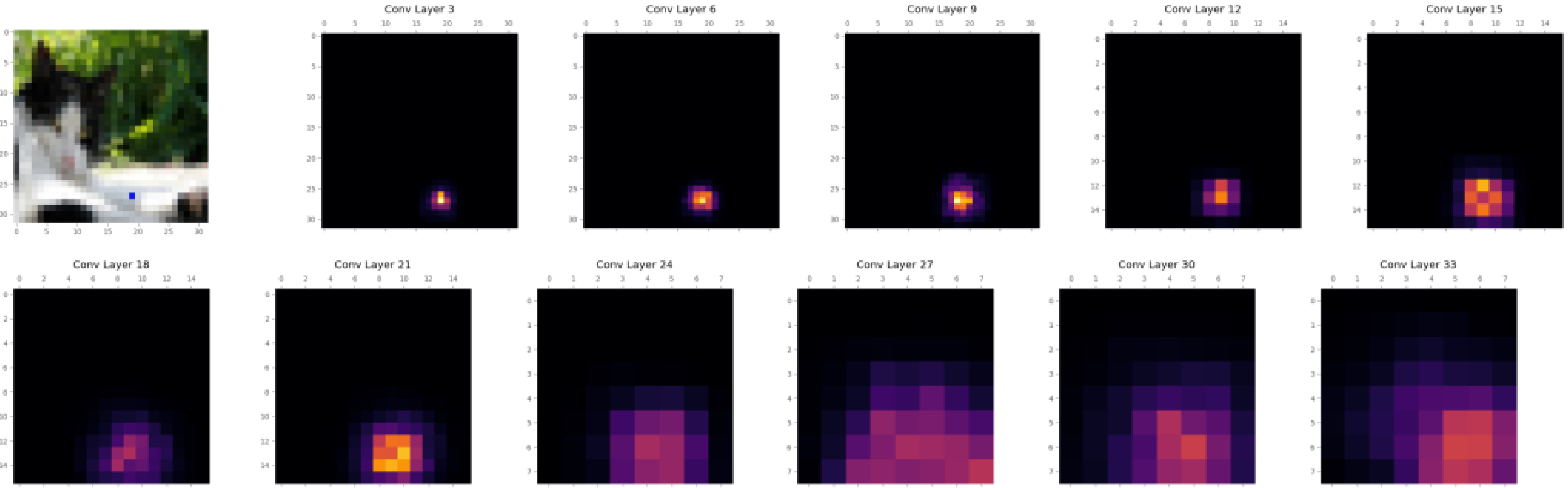
- Do not constrain the overall strength of the perturbation, but

- Constrain the spatial extent
- All the way down to a single pixel!

- Why does this work?

Su, J., Vargas, D. V., & Sakurai, K. (2019). One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation, 23(5), 828-841.

One-pixel Attack



- Local attacks “spread” to global attacks!
- Depends strongly on the position of this pixel

	LENET	RESNET
ORIGINAL ONE-PIXEL ATTACK	59%	33%
ONE-PIXEL ATTACK ON **Random Pixels**	4.9%	3.1%
ONE-PIXEL ATTACK ON **Nearby Pixels**	33.1%	31.3%

Vargas, D. V., & Su, J. (2019). Understanding the one-pixel attack: Propagation maps and locality analysis. arXiv preprint arXiv:1902.02947.

Intro Adversarial Attacks

Black Box Attacks

Black Box Attacks

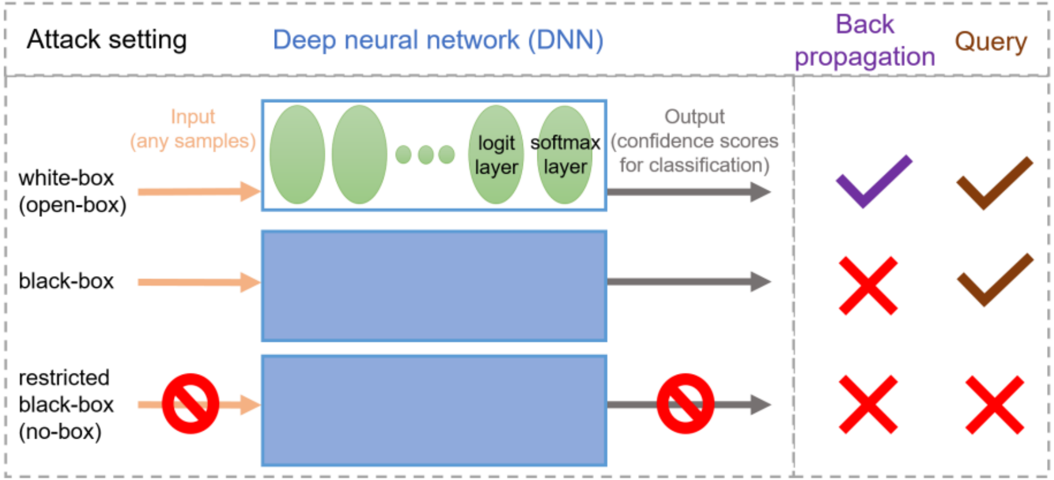


Figure 2: Taxonomy of adversarial attacks to deep neural networks (DNNs). "Back propagation" means an attacker can access the internal configurations in DNNs (e.g., performing gradient descent), and "Query" means an attacker can input any sample and observe the corresponding output.

Black Box Attacks

How would you attack a network if you cannot access the parameters, etc.?

Mahmood, K., Mahmood, R., Rathbun, E., & van Dijk, M. (2021). Back in Black: A Comparative Evaluation of Recent State-Of-The-Art Black-Box Attacks. IEEE Access, 10, 998-1019.

Black Box Attacks

How would you attack a network if you cannot access the parameters, etc.?

- Transfer attacks
  - Access to part of the training set and query access to classifier
  - Idea: Adversary would query the classifier to label the training data
  - Then, train a “synthetic model” on these labels on which to run white-box attacks
  - Hope: Misclassified examples by synthetic model will also be misclassified by classifier

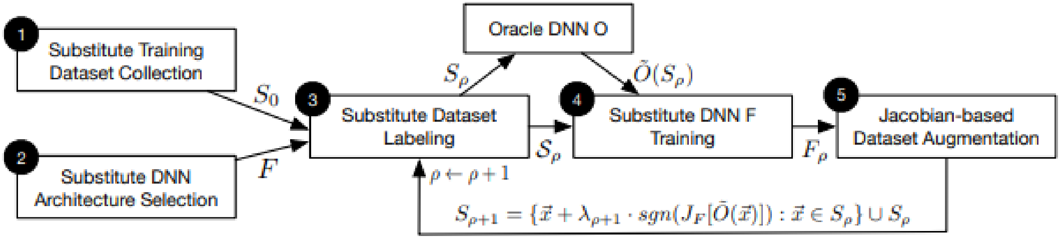


Figure 3: **Training of the substitute DNN** F: the attacker (1) collects an initial substitute training set S_0 and (2) selects an architecture F. Using oracle $\tilde{O}$, the attacker (3) labels S_0 and (4) trains substitute F. After (5) Jacobian-based dataset augmentation, steps (3) through (5) are repeated for several substitute epochs ρ.

Mahmood, K., Mahmood, R., Rathbun, E., & van Dijk, M. (2021). Back in Black: A Comparative Evaluation of Recent State-Of-The-Art Black-Box Attacks. IEEE Access, 10, 998-1019.

Black Box Attacks

How would you attack a network if you cannot access the parameters, etc.?

- Zeroth order optimization
  - Do not need access to the dataset
  - Repeatedly query the black box classifier to craft adversarial noise
  - However, require access to scores (model probabilities or pre-softmax scores)
  - New attempts trying to reduce # queries of strength of perturbation

	CIFAR10					
	Untargeted			Targeted		
	Success Rate	Avg. L_2	Avg. Time (per attack)	Success Rate	Avg. L_2	Avg. Time (per attack)
White-box (C&W)	100 %	0.17980	0.20 min	100 %	0.37974	0.16 min
Black-box (Substitute Model + FGSM)	76.1 %	-	0.005 sec (+ 7.81 min)	11.48 %	-	0.005 sec (+ 7.81 min)
Black-box (Substitute Model + C&W)	25.3 %	2.9708	0.47 min (+ 7.81 min)	5.3 %	5.7439	0.49 min (+ 7.81 min)
Proposed Black-box (ZOO-ADAM)	100 %	0.19973	3.43 min	96.8 %	0.39879	3.95 min
Proposed Black-box (ZOO-Newton)	100 %	0.23554	4.41 min	97.0 %	0.54226	4.40 min

Mahmood, K., Mahmood, R., Rathbun, E., & van Dijk, M. (2021). Back in Black: A Comparative Evaluation of Recent State-Of-The-Art Black-Box Attacks. IEEE Access, 10, 998-1019.

Intro Adversarial Attacks

Defending Against Attacks

Feature Denoising

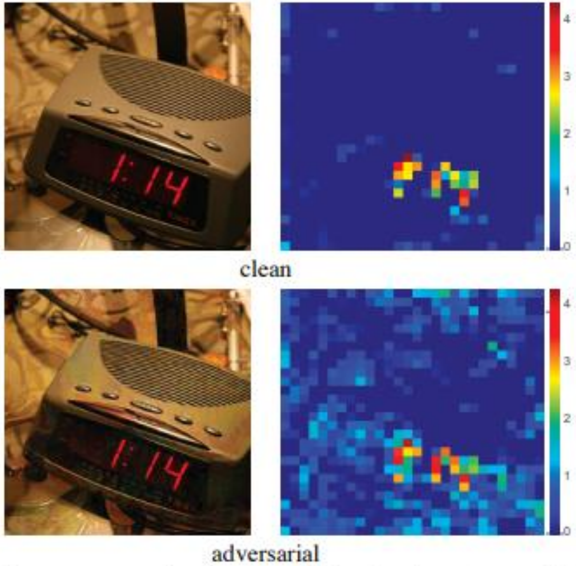


Figure 1. Feature map in the res_3 block of an ImageNet-trained ResNet-50 [9] applied on a clean image (top) and on its adversarially perturbed counterpart (bottom). The adversarial perturbation was produced using PGD [16] with maximum perturbation $\epsilon = 16$ (out of 256). In this example, the adversarial image is incorrectly recognized as "space heater"; the true label is "digital clock".

- Randomly selected feature map in ResNet-50
- Adversarial perturbations…
  - Are **small** in pixel space
  - Result in **substantial noise** in feature space
- Feature denoising might be a way of increasing adversarial robustness?!

Xie, C., Wu, Y., Maaten, L. V. D., Yuille, A. L., & He, K. (2019). Feature denoising for improving adversarial robustness. CVPR (pp. 501-509).

Feature Denoising

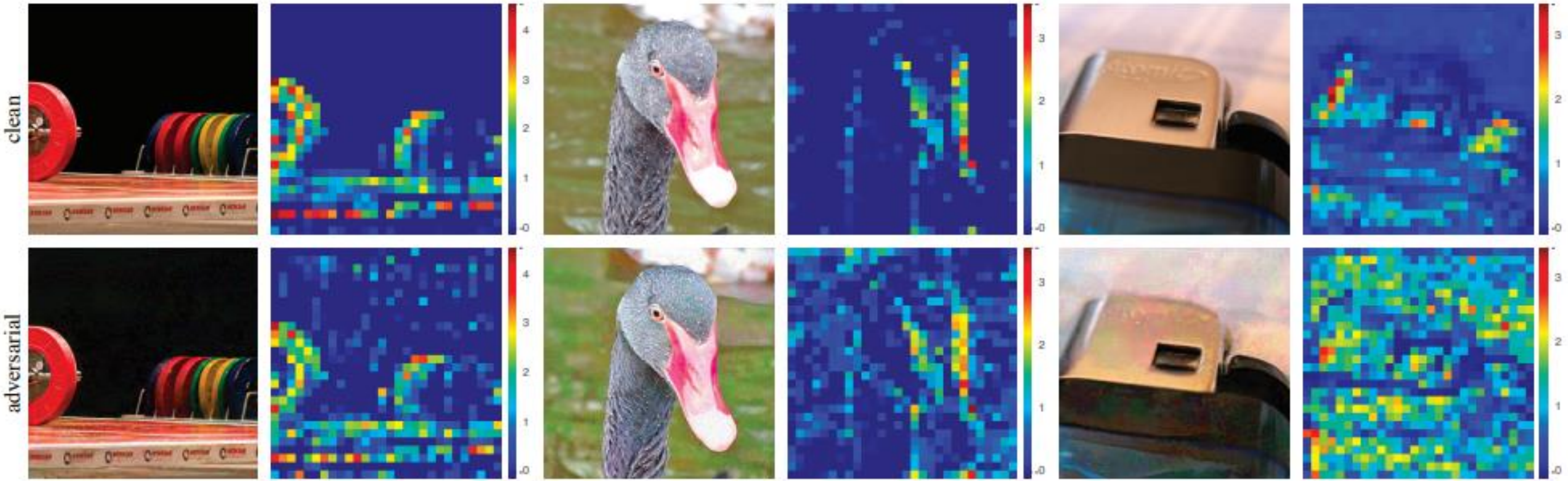
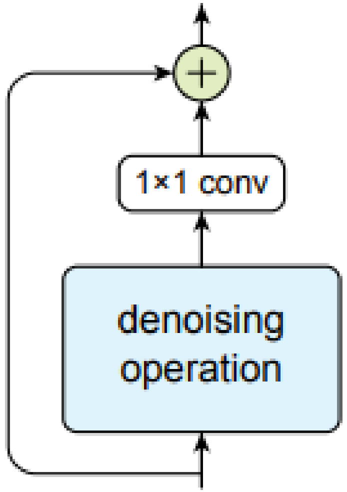


Figure 2. More examples similar to Figure 1. We show feature maps corresponding to clean images (top) and to their adversarial perturbed versions (bottom). The feature maps for each pair of examples are from the same channel of a res_3 block in the same ResNet-50 trained on clean images. The attacker has a maximum perturbation $\epsilon = 16$ in the pixel domain.

Feature Denoising

- If adversarial attacks produce feature noise…
- Then, denoising might be a solution for increasing robustness
- Introducing: **Denoising block**
  - Can be any feature layer
  - Denoising can be any denoiser (mean, non-local means)
- Residual design
  - Denoising reduces noise, but
  - May also affect signal!



Feature Denoising - Training

- Attacker
  - Iterative gradient-based attacker (similar I-FSGM)
  - L_inf norm: Maximally allowed change per pixel (16)
  - Attacker labels are chosen at random

- Training with adversarial images
  - Images in batch are attacked using attacker
  - Updates ONLY based on attacked images
  - Training time is considerably longer (n-steps for iterative attack in every batch)

- Evaluation
  - Targeted attacks under white-box setting with L_inf varying ε
  - ImageNet evaluation: baselines are ResNet-101/152

Feature Denoising - Results

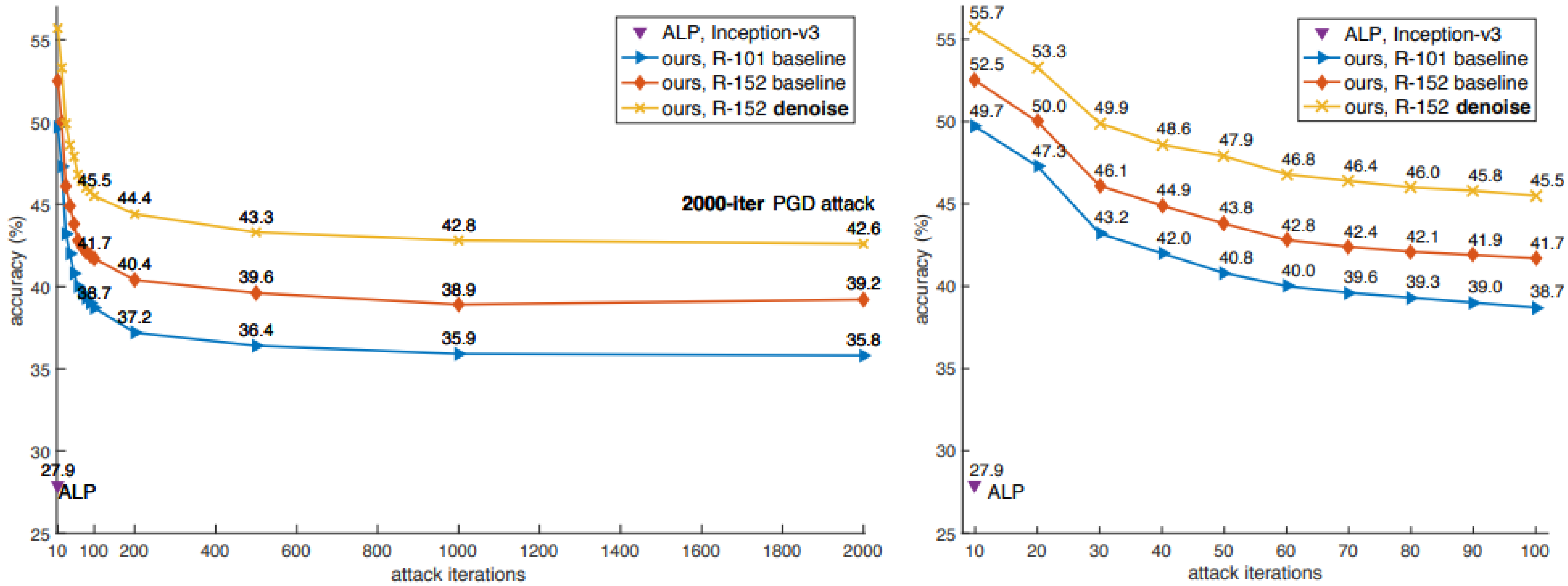


Figure 6. **Defense against white-box attacks on ImageNet**. The left plot shows results against a white-box PGD attacker with 10 to **2000** attack iterations. The right plot zooms in on the results with 10 to 100 attack iterations. The maximum perturbation is $\epsilon = 16$.

Certified Robustness

- **So far**: Robustness demonstrated empirically
- **Now**: Would be appealing to have certified robustness

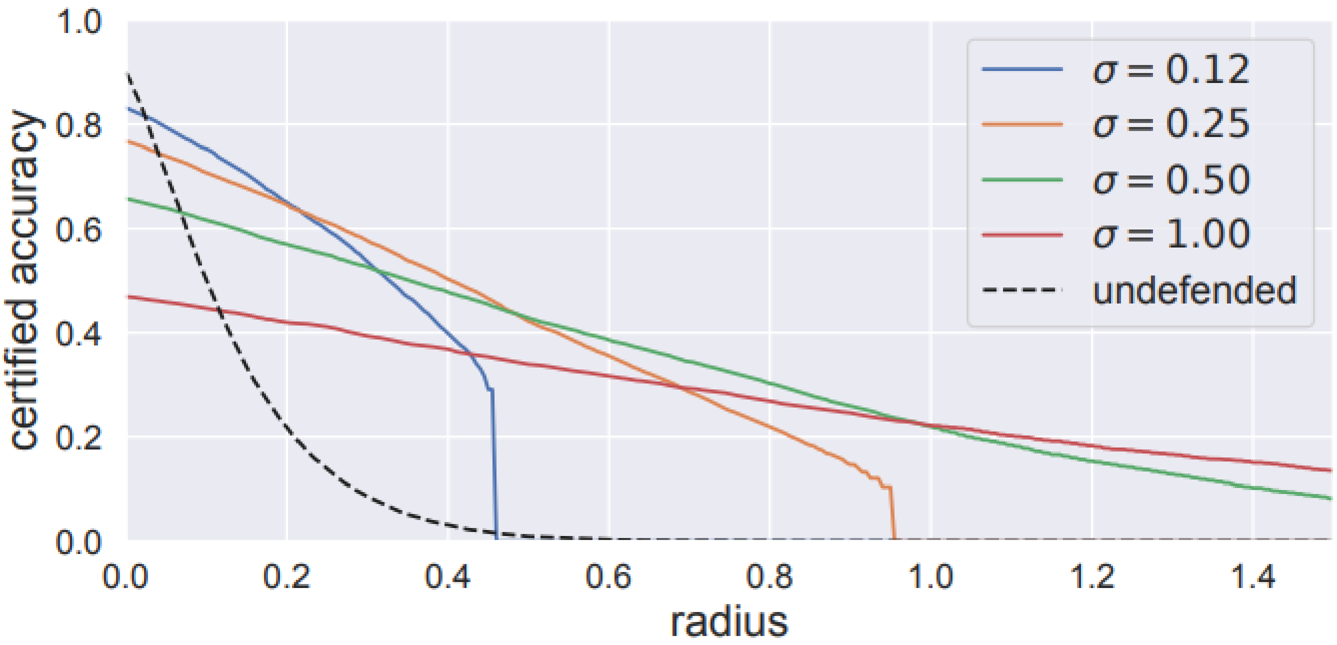
- **Enter**: Certified defenses

A classifier is said to be *certifiably robust* if for any input x, one can easily obtain a guarantee that the classifier's prediction is constant within some set around x, often an ℓ_2 or ℓ_∞ ball.

- Randomized smoothing
  - Create a smoothed classifier g from base classifier f
  - g returns the most likely class f would give if input x was perturbed with Gaussian noise
  - Results in an epsilon ball around x with radius r within which accuracy can be certified

Cohen, J., Rosenfeld, E., & Kolter, Z. (2019, May). Certified adversarial robustness via randomized smoothing. ICML (pp. 1310-1320). PMLR.

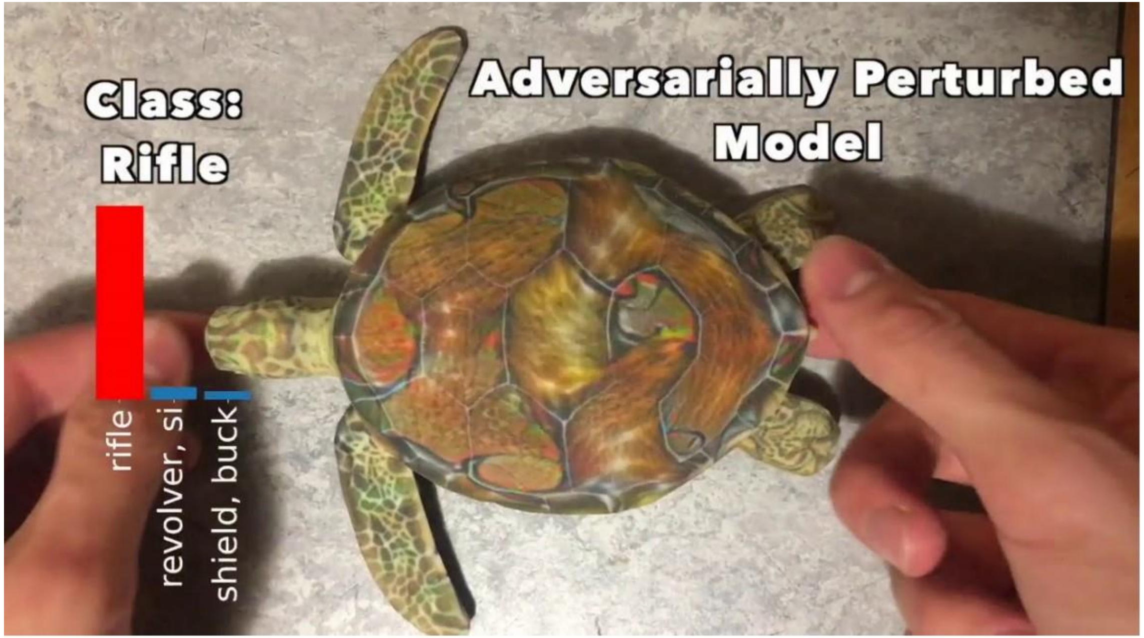
Certified Robustness



Intro Adversarial Attacks

Physical Attacks

Adversarial Examples in the Physical World: Object Patterns



Video: https://www.labsix.org/physical-objects-that-fool-neural-nets/

Adversarial Examples in the Physical World: Autonomous Vehicles

Figure : Before: Stop sign; After: 45 mph sign

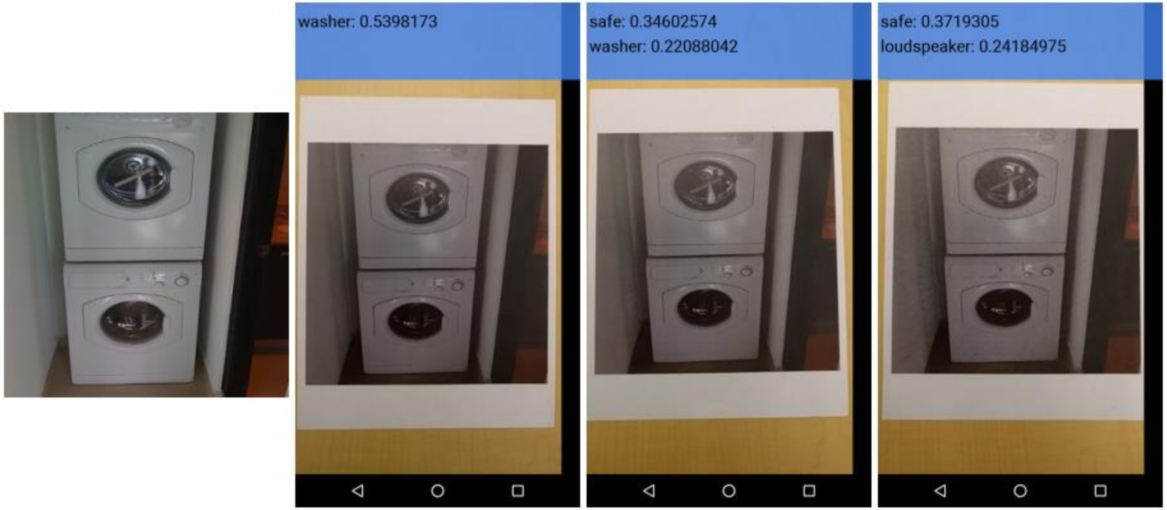
Physical Adversarial Attacks

How would you create a physical adversarial attack?

Physical Adversarial Attacks

How would you create a physical adversarial attack?

- Create white box attack
- Print adversarial picture
- Take image with phone
- ???
- Profit (or not)



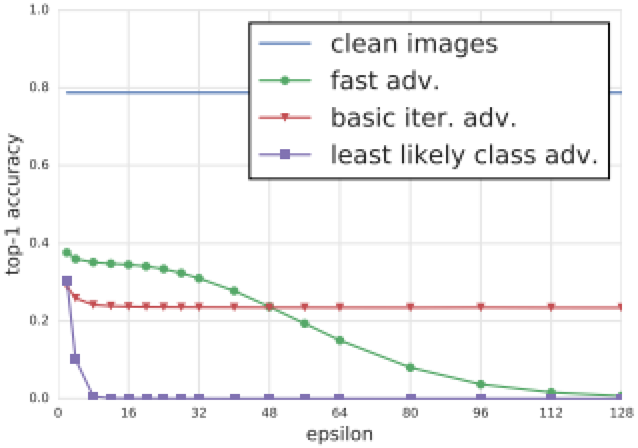
(a) Image from dataset (b) Clean image (c) Adv. image, $\epsilon = 4$ (d) Adv. image, $\epsilon = 8$

Kurakin, A., Goodfellow, I. J., & Bengio, S. (2018). Adversarial examples in the physical world. In *Artificial intelligence safety and security* (pp. 99-112). Chapman and Hall/CRC.

Physical Adversarial Attacks

How would you create a physical adversarial attack?

- Create white box attack
- Print adversarial picture
- Take image with phone
- ???
- Profit (or not)



(a) Printout (b) Photo of printout (c) Cropped image

Kurakin, A., Goodfellow, I. J., & Bengio, S. (2018). Adversarial examples in the physical world. In *Artificial intelligence safety and security* (pp. 99-112). Chapman and Hall/CRC.

Physical Adversarial Attacks

How would you create a physical adversarial attack?

- Model physical variation (augmentation)
- Put constraints on attack
  - Cannot change background
  - Limits on “imperceptibility”
  - Must be “fabricable”
- Introduce masks
  - Optimize position/value of masks

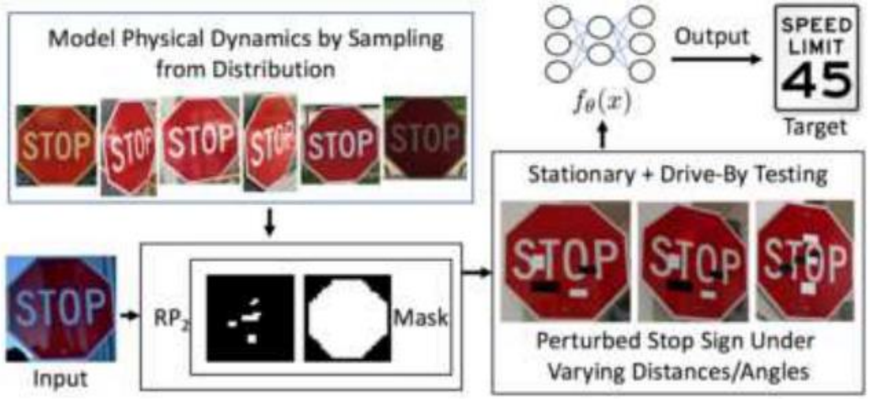


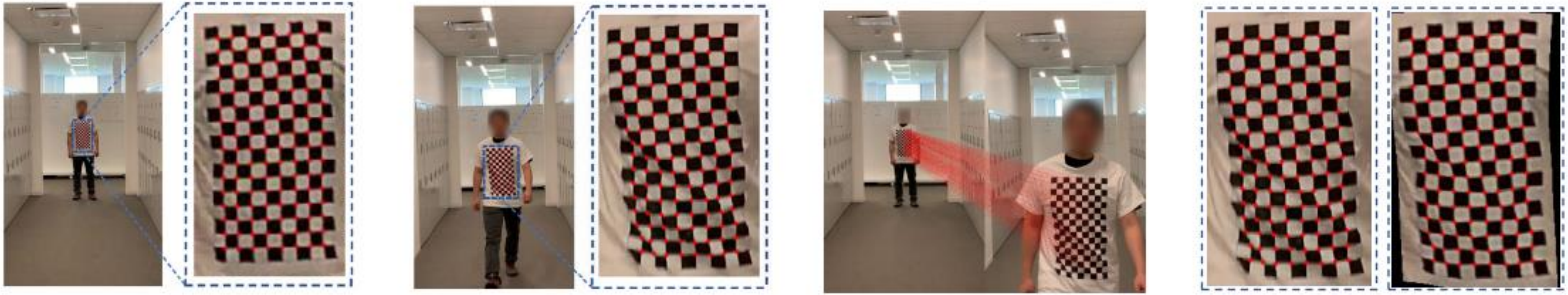
Figure 2: RP_2 pipeline overview. The input is the target Stop sign. RP_2 samples from a distribution that models physical dynamics (in this case, varying distances and angles), and uses a mask to project computed perturbations to a shape that resembles graffiti. The adversary prints out the resulting perturbations and sticks them to the target Stop sign.

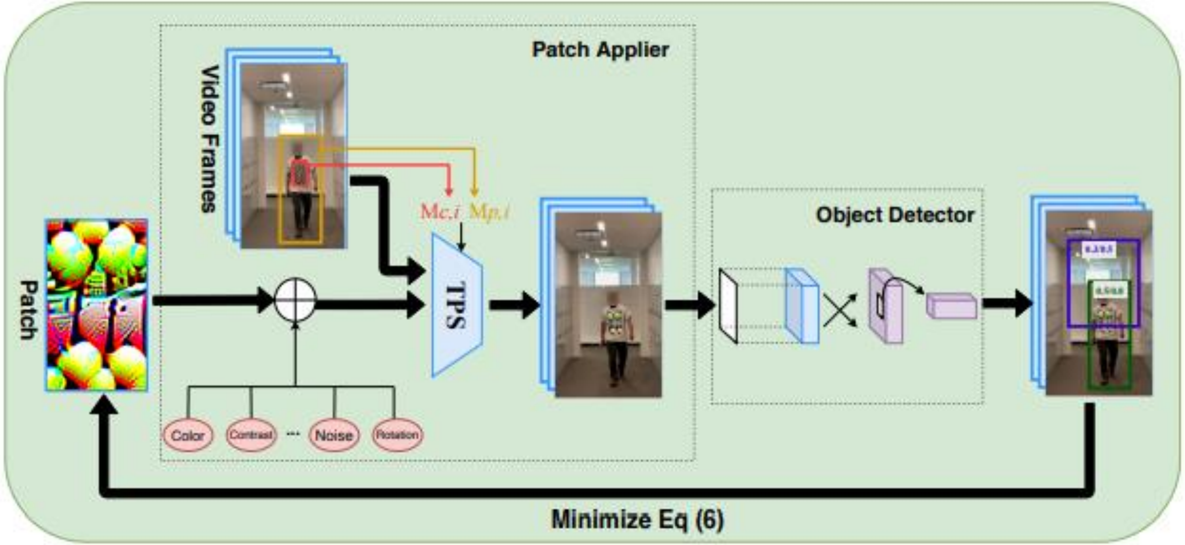
Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., ... & Song, D. (2018). Robust physical-world attacks on deep learning visual classification. CVPR

Physical Adversarial Attacks

Distance/Angle	Subtle Poster	Subtle Poster Right Turn	Camouflage Graffiti	Camouflage Art (LISA-CNN)	Camouflage Art (GTSRB-CNN)
5′ 0°					
5′ 15°					
10′ 0°					
10′ 30°					
40′ 0°					
Targeted-Attack Success	100%	73.33%	66.67%	100%	80%

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., ... & Song, D. (2018). Robust physical-world attacks on deep learning visual classification. CVPR





Xu, K., Zhang, G., Liu, S., Fan, Q., Sun, M., Chen, H., ... & Lin, X. (2020). Adversarial t-shirt! evading person detectors in a physical world. ECCV

Physical Adversarial Attacks

How would you create a physical adversarial attack?

Xu, K., Zhang, G., Liu, S., Fan, Q., Sun, M., Chen, H., ... & Lin, X. (2020). Adversarial t-shirt! evading person detectors in a physical world. ECCV

Physical Adversarial Attacks

Other examples:

Wei, H., Tang, H., Jia, X., Yu, H., Li, Z., Wang, Z., ... & Wang, Z. (2022). Physical Adversarial Attack meets Computer Vision: A Decade Survey. arXiv preprint arXiv:2209.15179.

A Brief Intro to Adversarial Examples

Questions?