Robust Physical-World Attacks on Deep Learning Visual Classification

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, and Dawn Song

SYSSEC ChoManGi

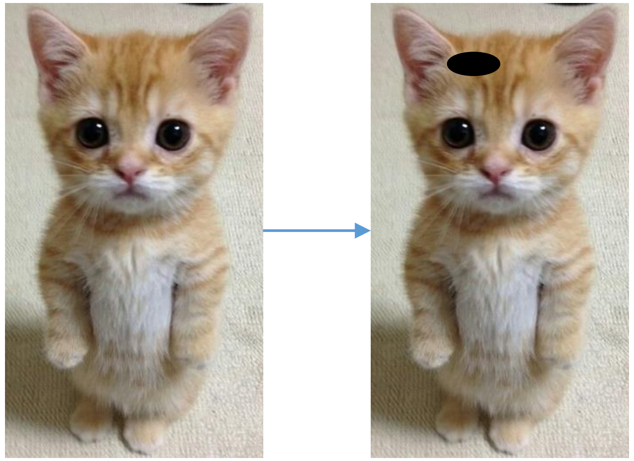
What is Adversarial Example?

Cat

Cat

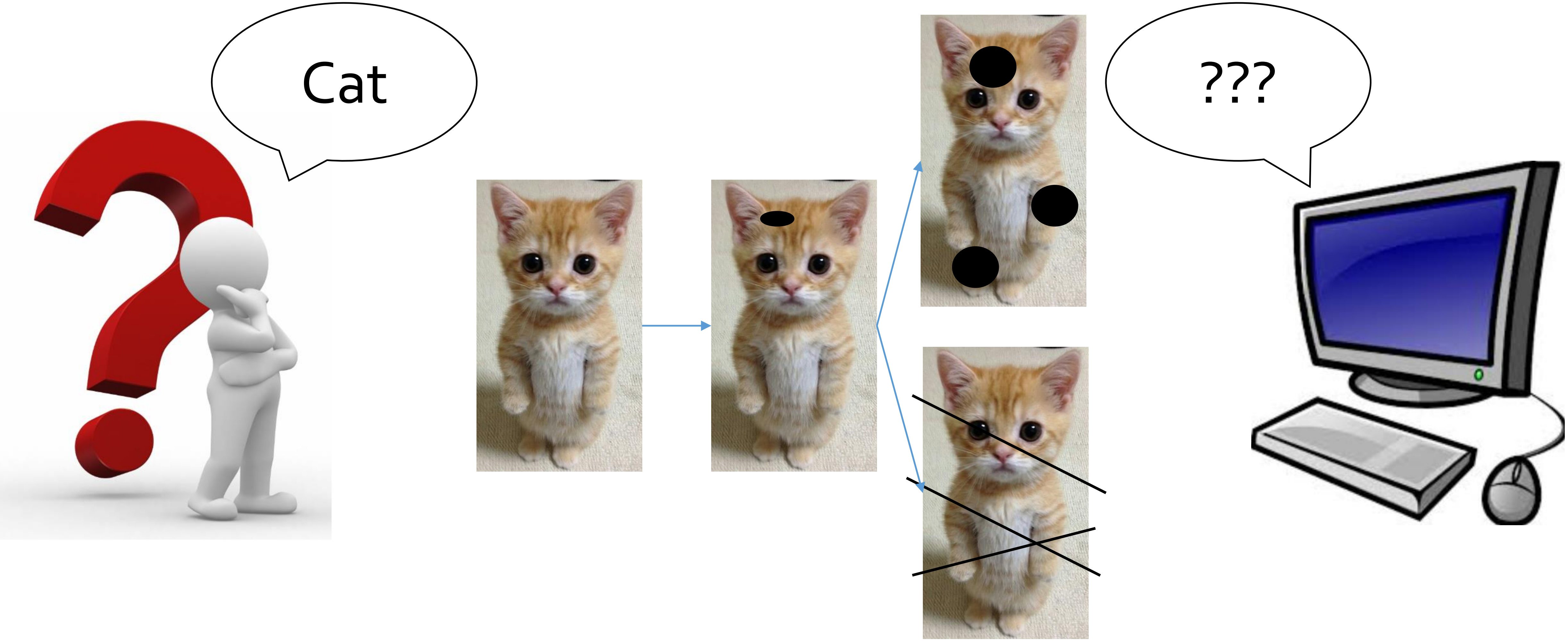
What is Adversarial Example?

Cat



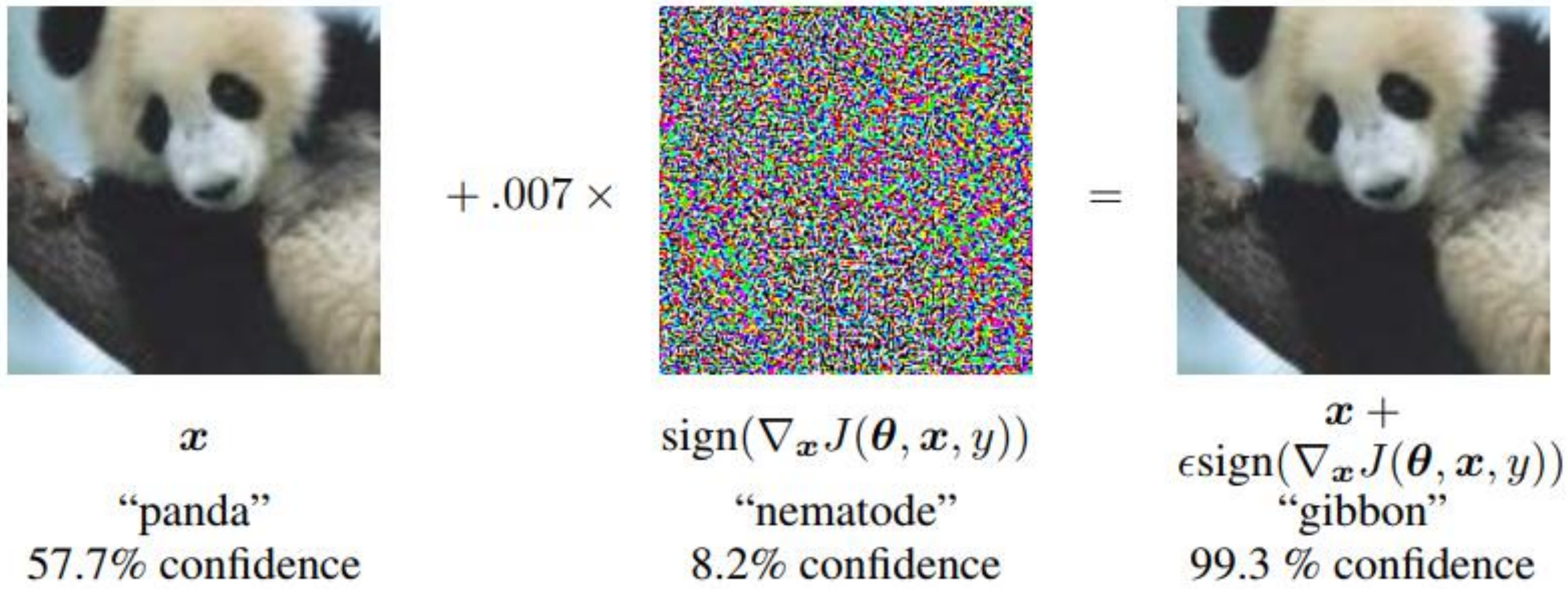
Cat?

What is Adversarial Example?

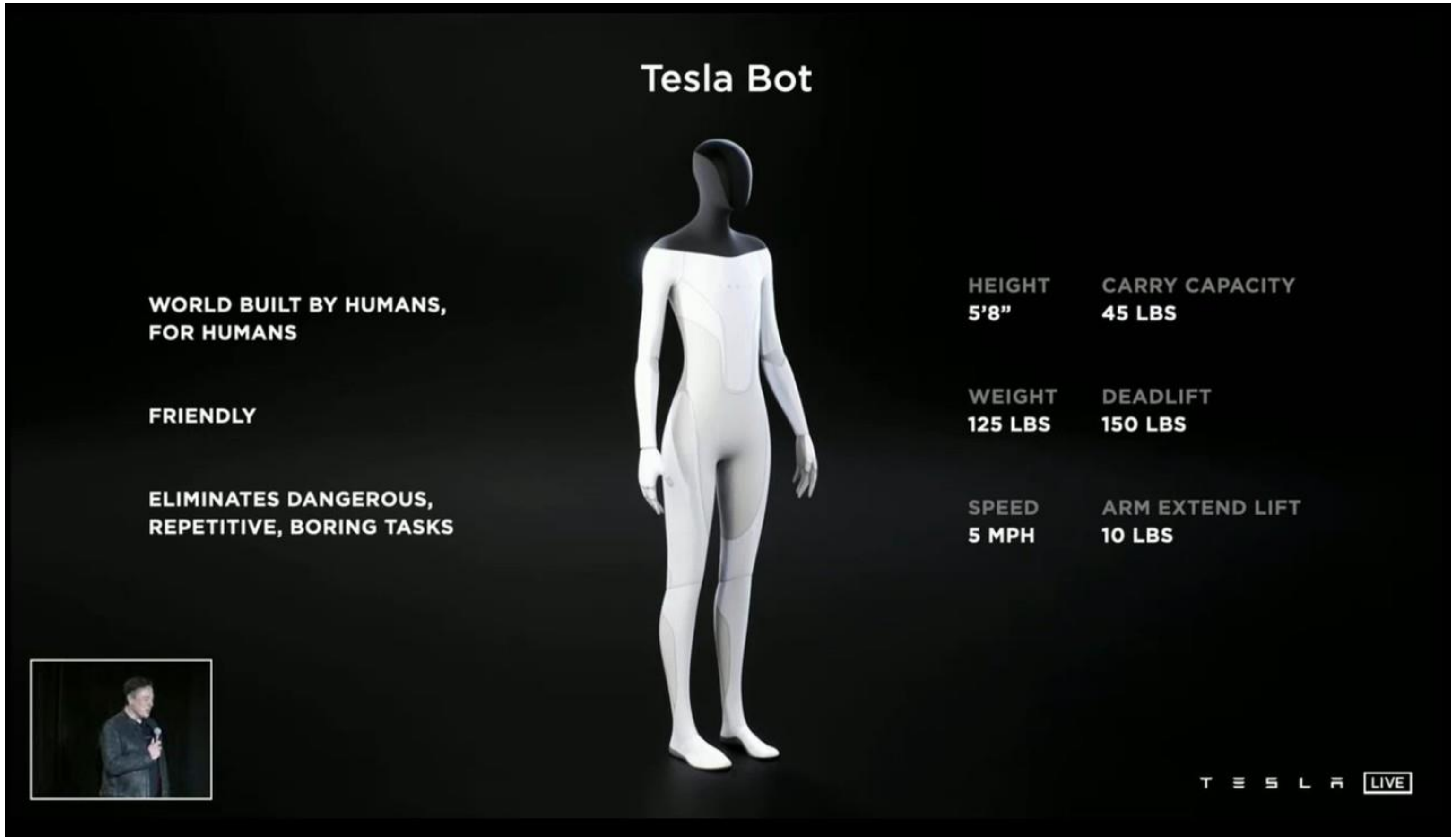


What is adversarial example?

- Samples that cause a machine learning model to make a false prediction by using the difference between human and AI
  - It uses the different classification standards between human and AI
  - Human and machines make different decision about adversarial samples.



Importance of A.E.?

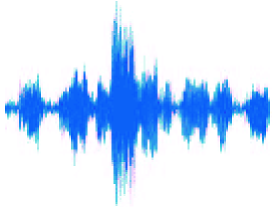


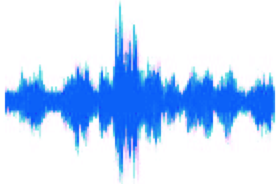

Original: Temple


Adversarial: Ostrich


Original: Duck


Adversarial: Horse


Original:
"How are you?"


Adversarial:
"Open the door"

Principles of adversarial example

❖ target AI: Classifier

– $Y = F(X)$ is model process. X: model input, Y: classification result

❖ Basic idea

– Add minimal perturbation to input x so that it can be determined that it is not X.

– When $F(X) = Y$, solving optimization problem that

$$argmin_{\|\delta X\|}\ s.t.\ F(X + \delta X) = Y^* \neq Y$$

❖ Category

– Untargeted: Classify X as not Y.

- $minimize_{\delta X}\ J(X + \delta X, Y^* \neq Y)$, J: Cost function

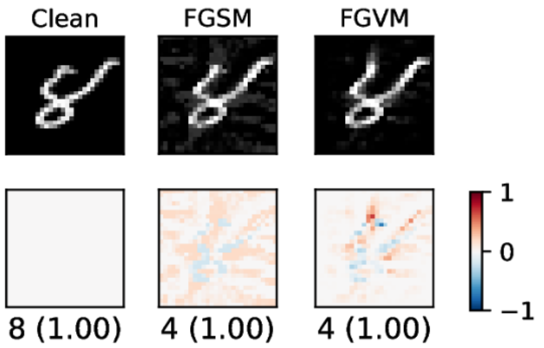
– Targeted: Classify X as targeted class Y^*

- $minimize_{\delta X}\ J\left(X + \delta X, Y^* = Y_{target} \neq Y\right)$

Previous Works on Digital images

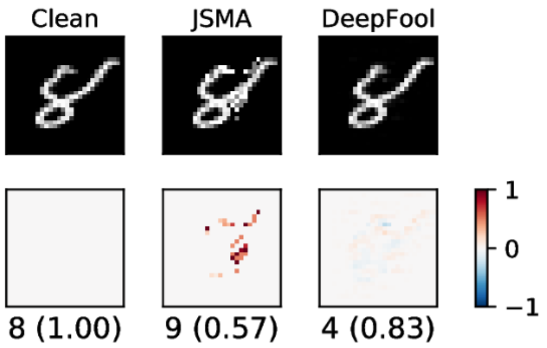
- **Fast/Iterative Gradient Sign/Value Method (FGSM/FGVM, IGSM)**
  - First suggested adversarial example method – Goodfellow et al. 2014.
  - FGSM: $\boldsymbol{X}_{adv} = \boldsymbol{X} + \epsilon \cdot \mathrm{sign}(\nabla_{\boldsymbol{X}} J(\boldsymbol{X}, y_{true}))$, FGVM: $\boldsymbol{X}_{adv} = \boldsymbol{X} + \epsilon \cdot \nabla_{\boldsymbol{X}} J(\boldsymbol{X}, y_{true})$
  - IGSM: $\boldsymbol{X}_{adv,0} = \boldsymbol{X},$

$$\boldsymbol{X}_{adv,N+1} = Clip_{\boldsymbol{X},\epsilon}\{\boldsymbol{X}_{adv,N} + \alpha \cdot \mathrm{sign}(\nabla_{\boldsymbol{X}} J(\boldsymbol{X}_N, y_{true}))\}.$$

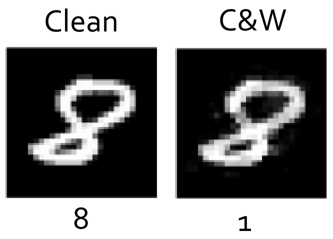


- **DeepFool - 2015**
  - Calculate a minimum of L2 perturbation through model structure approximation
- **Jacobian-based Saliency Map Attack (JSMA) - 2016**
  - Modulates only pixels that have the greatest impact on classification results

  ➔ Minimize the number of pixels modulated(minimize L_0 distance)



- **Carlini & Wagner (C&W) - 2017**
  - Defending the adversarial defense method, Defensive Distillation.
  - Considered to be one of the most powerful attack method
  - Attacks on L_0, L_2, L_∞ metrics are possible



Why Physical World?

- ❖ After the Carlini's paper, adversarial example research on digital images is not active. – Why?

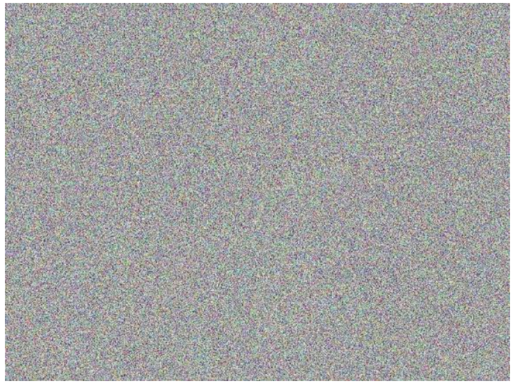
- ❖ Previous techniques for digital images presented were all to add calculated minimum noise to the whole picture, including the background.

- ❖ Verification about practicality is essential to actual use of the proposed techniques.

Why Physical World?

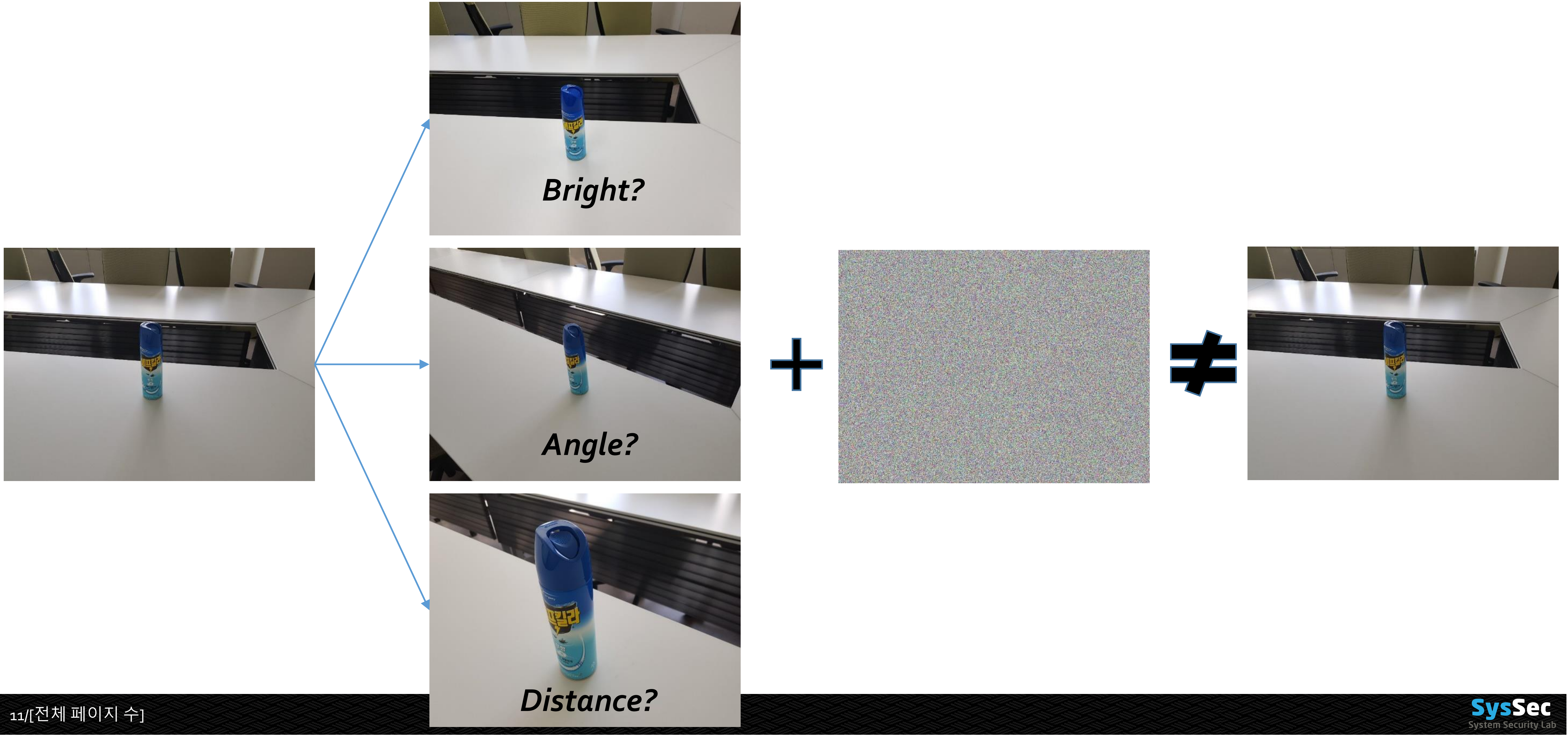
Possible??

+



=

Why Physical World?



Purpose

❖Propose an white-box adversarial example attack on road signs in physical world conditions.

❖Why they focused on road signs?

- Simple!
- Various angles, distances!
- Impoartance!
- Effective!

SysSec
System Security Lab

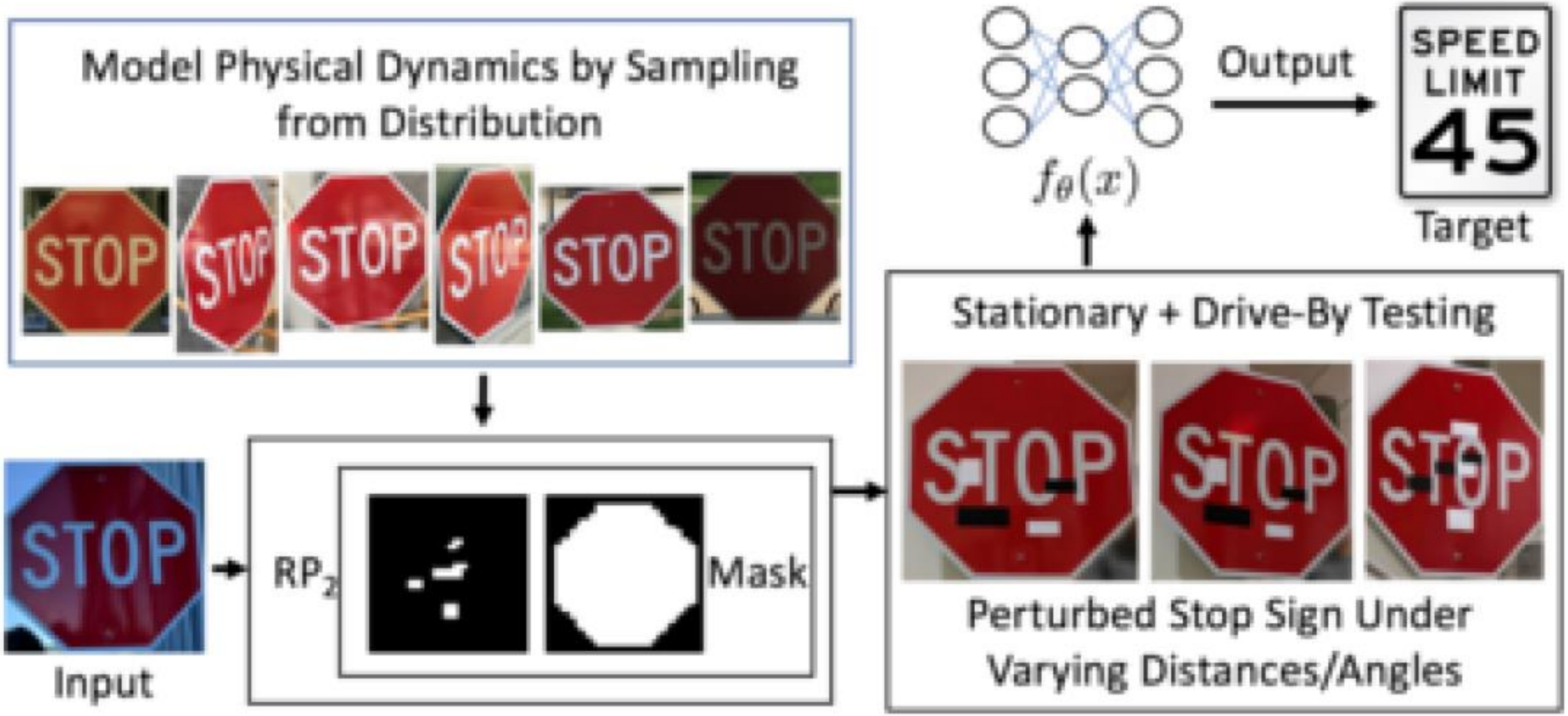
Contribution(Work)

- ❖Propose RP2(Robust Physical Perturbation) algorithm to generate physical perturbations that can consistently cause misclassification under various physical conditions.

- ❖Evaluate algorithm with famous road sign dataset, LISA and GTSRB.

- ❖To show the generality of algorithm, they tested their attacks on Inception-v3 classifier to misclassify the microwave as "phone".

Attack Pipeline



SysSec
System Security Lab

Algorithm - basic

1) Untargeted : arg $min_{\delta}\lambda\|\delta\|_{p} - J(f_{\theta}(x+\delta), y)$
2) Targeted : arg $min_{\delta}\lambda\|\delta\|_{p} + J(f_{\theta}(x+\delta), y^{*})$

δ : perturbation, λ: regularization coefficient, x : input,

y : authentic class, y* : target class, $\|.\|_{p}$: 2D p-norm($\sum_{i,j}(\left|\delta_{(i,j)}\right|^{p})^{1/p}$), J : cross entropy, θ : hyper parameter

SysSec
System Security Lab

Algorithm – (1) various conditions

Consider various distances, angles, brightness for loss calculations.

1) Untargeted : $\arg min_\delta \lambda \|\delta\|_p - J(f_\theta(x+\delta), y)$
2) Targeted : $\arg min_\delta \lambda \|\delta\|_p + J(f_\theta(x+\delta), y^*)$

1) Untargeted : $\arg min_\delta \lambda \|\delta\|_p - \frac{1}{k}\sum_{i=1}^{k} J(f_\theta(x+\delta), y)$
2) Targeted : $\arg min_\delta \lambda \|\delta\|_p + \frac{1}{k}\sum_{i=1}^{k} J(f_\theta(x+\delta), y^*)$

Average for distance, angle, brightness!

Algorithm – (2) use mask

Use mask matrix to modify specific areas only.

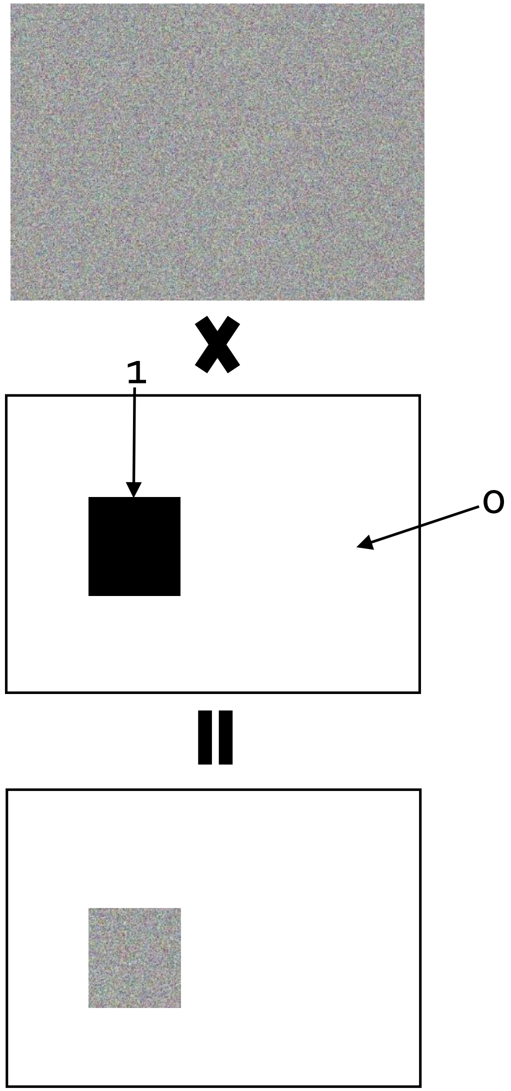
1) Untargeted : $\arg min_{\delta} \lambda \|\delta\|_p - \frac{1}{k}\sum_{i=1}^{k} J(f_\theta(x+\delta), y)$

2) Targeted : $\arg min_{\delta} \lambda \|\delta\|_p + \frac{1}{k}\sum_{i=1}^{k} J(f_\theta(x+\delta), y^*)$

3) Untargeted : $\arg min_{\delta} \lambda \|M_x\delta\|_p - \frac{1}{k}\sum_{i=1}^{k} J(f_\theta(x+M_x\delta), y)$

4) Targeted : $\arg min_{\delta} \lambda \|M_x\delta\|_p + \frac{1}{k}\sum_{i=1}^{k} J(f_\theta(x+M_x\delta), y^*)$

Perturbate only matrix area!



Algorithm – (3) available colors

NPS(p) = $\prod_{\hat{p} \in P} |p - \hat{p}|$, P = set of printable colors, p = color of each pixels

1) Untargeted : arg $min_{\delta} \lambda ||M_x \delta||_p - \frac{1}{k} \sum_{i=1}^{k} J(f_\theta(x + M_x \delta), y)$

2) Targeted : arg $min_{\delta} \lambda ||M_x \delta||_p + \frac{1}{k} \sum_{i=1}^{k} J(f_\theta(x + M_x \delta), y^*)$

3) Untargeted : arg $min_{\delta} \lambda ||M_x \delta||_p + \underline{NPS(M_x \delta)} - \frac{1}{k} \sum_{i=1}^{k} J(f_\theta(x + M_x \delta), y)$

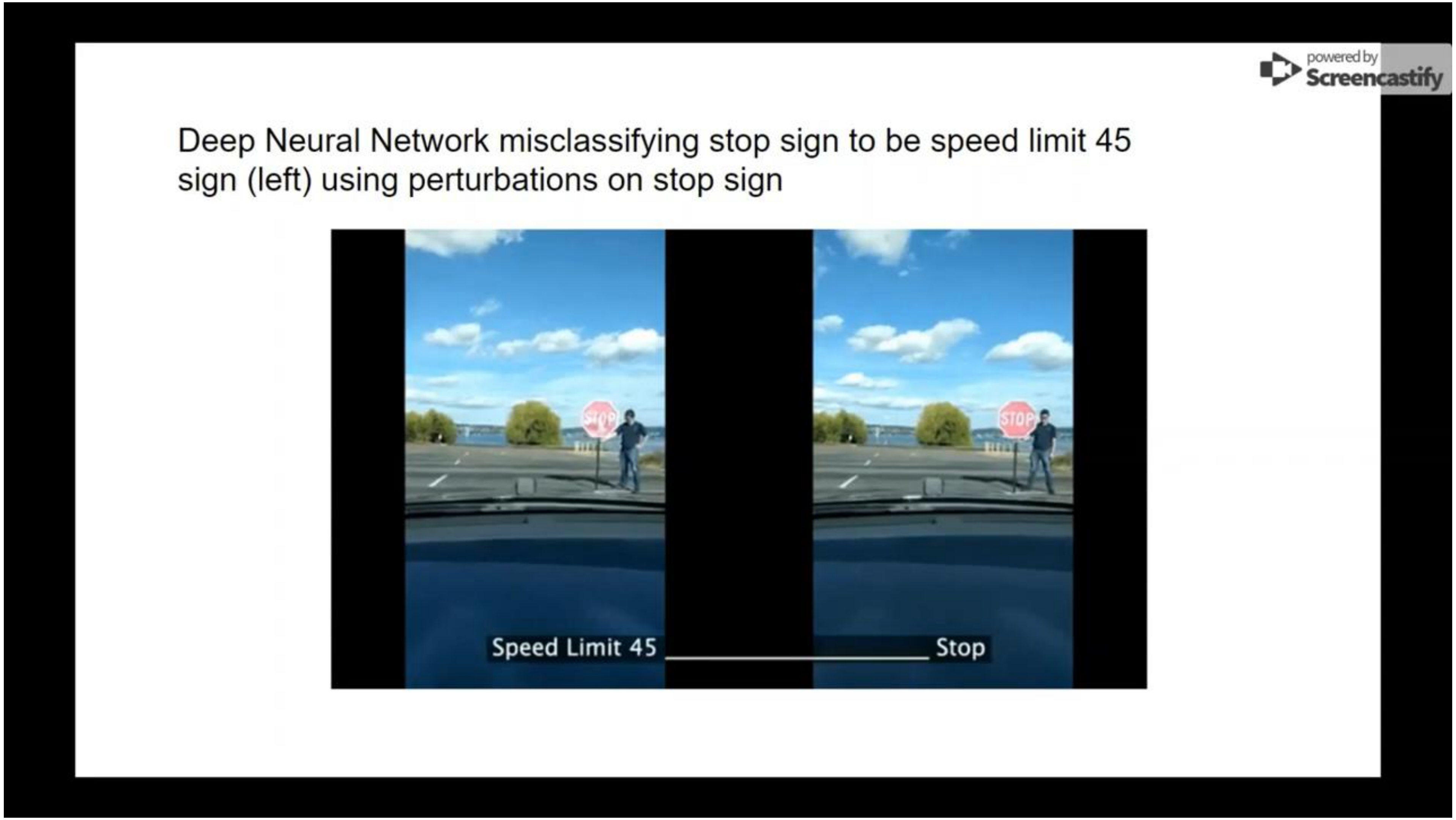
4) Targeted : arg $min_{\delta} \lambda ||M_x \delta||_p + \underline{NPS(M_x \delta)} + \frac{1}{k} \sum_{i=1}^{k} J(f_\theta(x + M_x \delta), y^*)$

Don’t use non-printable color!

Evaluation – (1) road sign

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%

Evaluation – (1) road sign



Evaluation – (1) road sign

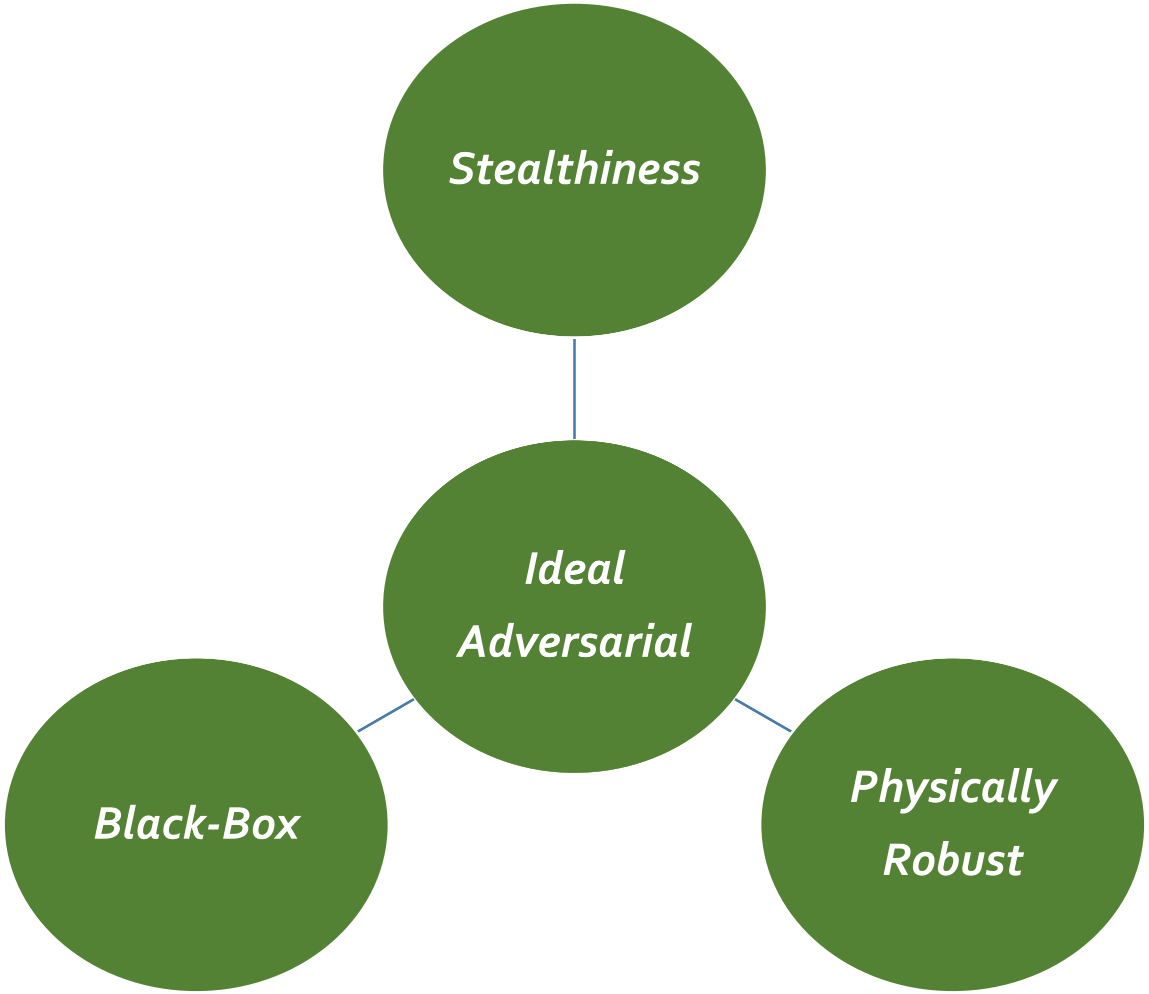
Perturbation	Attack Success	A Subset of Sampled Frames $k = 10$
Subtle poster	100%	
Camouflage abstract art	84.8%	

Evaluation – (2) Microwave

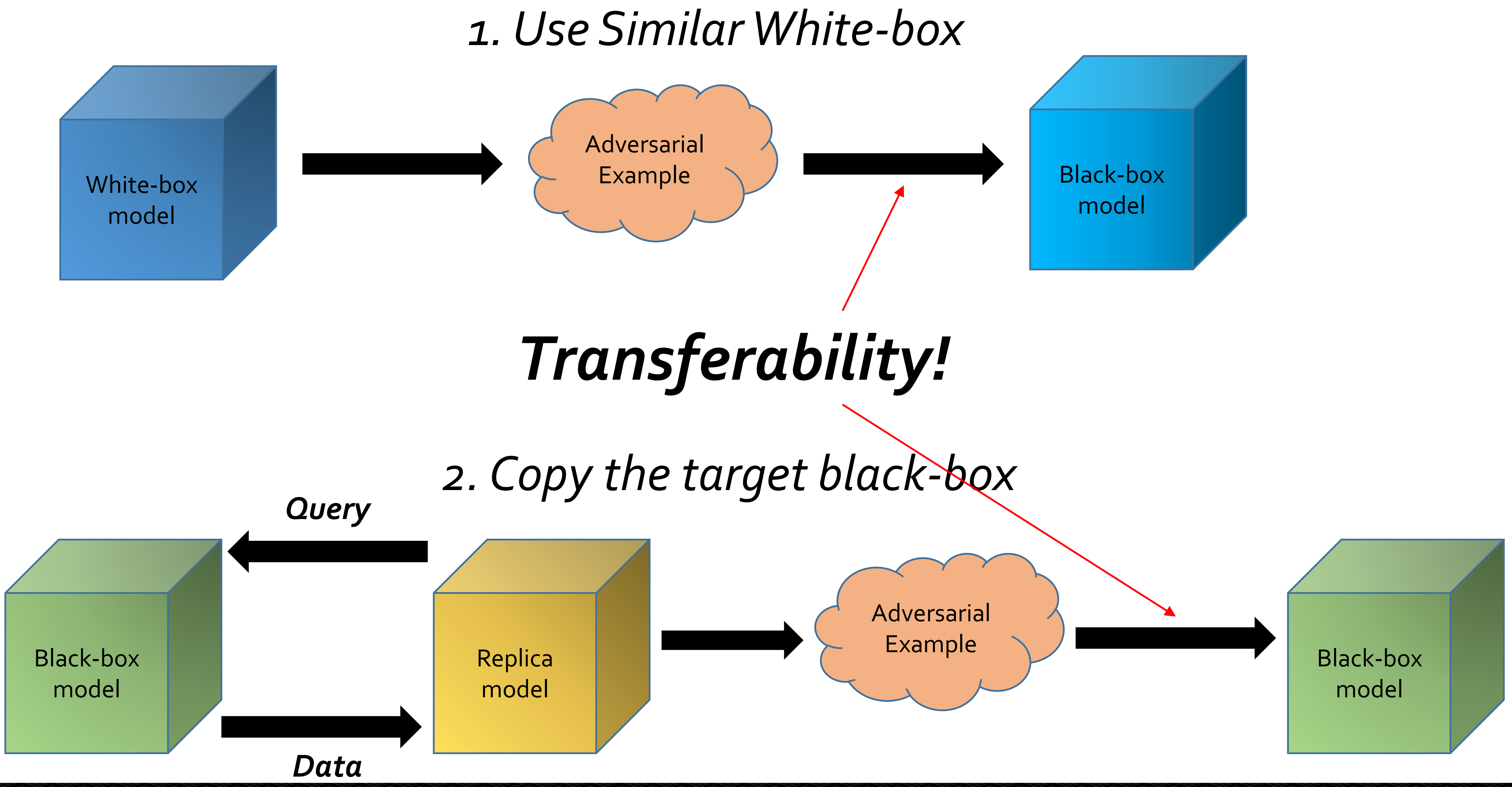
Distance/Angle	Image	Distance/Angle	Image
2′ 0°		2′ 15°	
5′ 0°		5′ 15°	
7′ 0°		7′ 15°	
10′ 0°		10′ 15°	
15′ 0°		20′ 0°	

Distance & Angle	Top Class (Confid.)	Second Class (Confid.)
2’ 0°	Phone (0.78)	Microwave (0.03)
2’ 15°	Phone (0.60)	Microwave (0.11)
5’ 0°	Phone (0.71)	Microwave (0.07)
5’ 15°	Phone (0.53)	Microwave (0.25)
7’ 0°	Phone (0.47)	Microwave (0.26)
7’ 15°	Phone (0.59)	Microwave (0.18)
10’ 0°	Phone (0.70)	Microwave (0.09)
10’ 15°	Phone (0.43)	Microwave (0.28)
15’ 0°	Microwave (0.36)	Phone (0.20)
20’ 0°	Phone (0.31)	Microwave (0.10)

Future Work


Stealthiness
Ideal
Adversarial
Black-Box
Physically
Robust

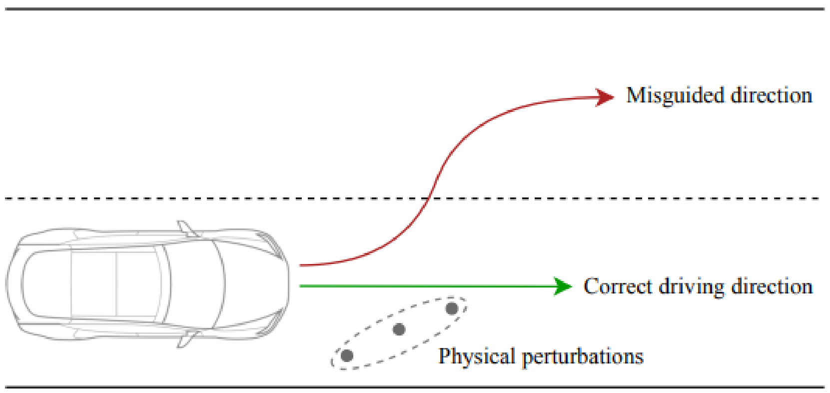
How to Attack Black-box?



Follow-up Work

❖ Too Good to Be Safe: Tricking Lane Detection in Autonomous Driving with Crafted Perturbations (Usenix 2021)

- A.E attack on lane detection system of Tesla
- physical perturbations to misguide the direction of Tesla



SysSec
System Security Lab

Questions to the presenter

- ◆They consider various distances, angles, brightness for loss calculations. I wonder if this attack is possible regardless of the weather, such as sunny, rainy, cloudy, or snowy days (김한나) – Best question
  - It seems very difficult to consider weather conditions.
- ◆In this paper, there is no detailed comment about attack mitigations or defenses. What is the general or proper mitigation targeting DNNs? (이용화)
  - Adversarial training can be the general mitigation against A.E.
- ◆ Are there any advanced ml techniques to learn find unknown vulnerability besides adversarial perturbations? (김경태)
  - The main idea of A.E is to threat the perception differences between human and AI. Just finding the ml vulnerabilities can be found with input mutations(fuzzing).

END