About Me

Zhengyu Zhao (赵正宇)

zhengyu.zhao@cispa.de
zhengyuzhao.github.io

Postdoc @ CISPA Helmholtz Center for Information Security, Germany

PhD @ Radboud University, The Netherlands

Research Interests:

Security (e.g. adversarial example and data poisoning) **and Privacy** (e.g. membership inference) **risks of Machine Learning/Computer Vision**.

Vulnerability of Computer Vision to Adversarial Perturbations

16/12/2022

Outline

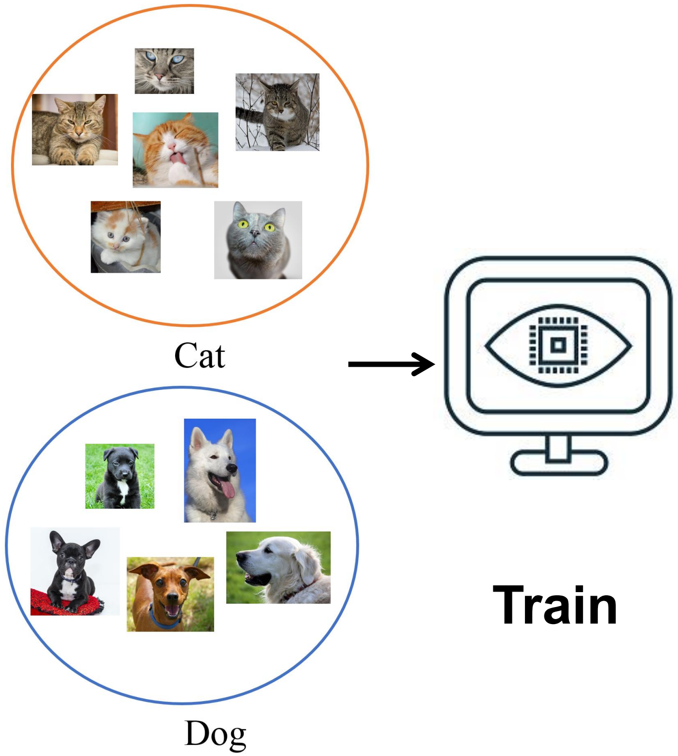
- Background of computer vision (CV) and adversarial images
- Two of our recent projects
- Other related projects

Outline

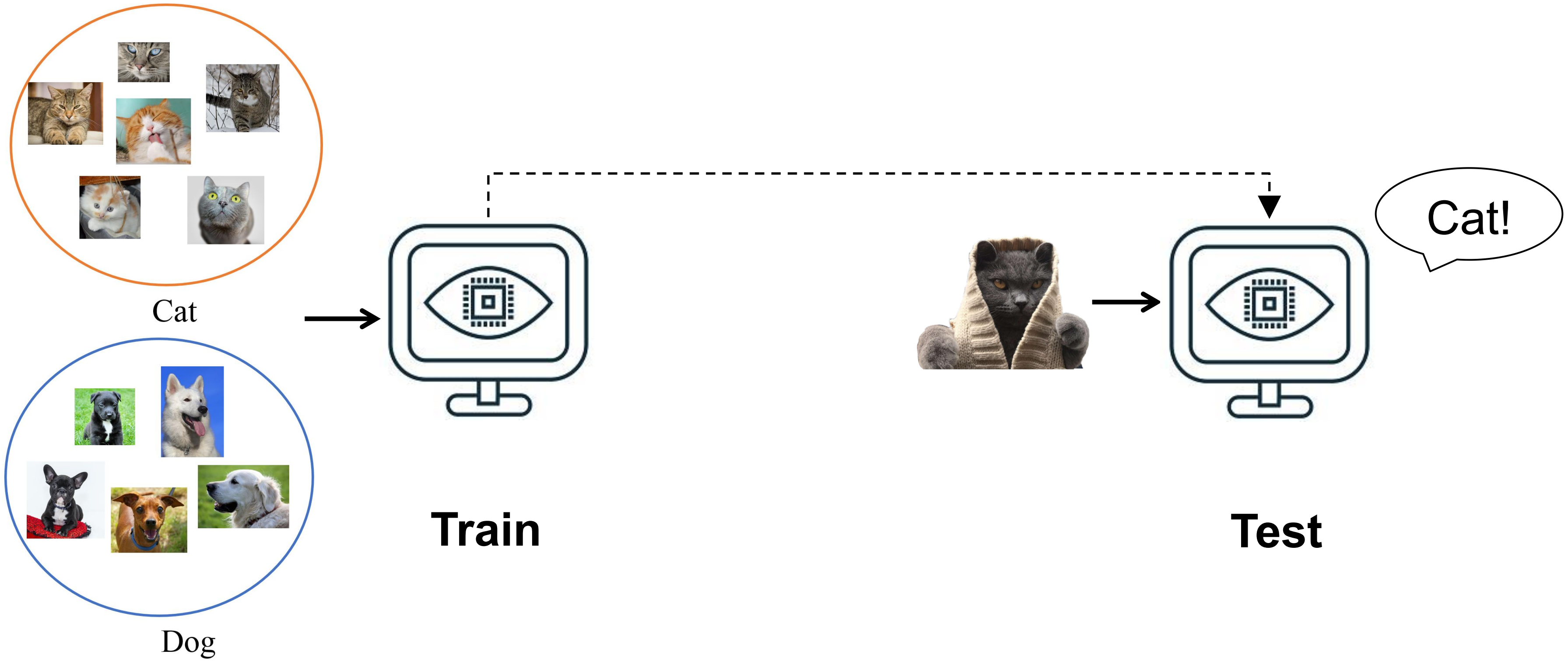
- Background of computer vision (CV) and adversarial images
- Two of our recent projects
- Other related projects

Computer Vision (CV)

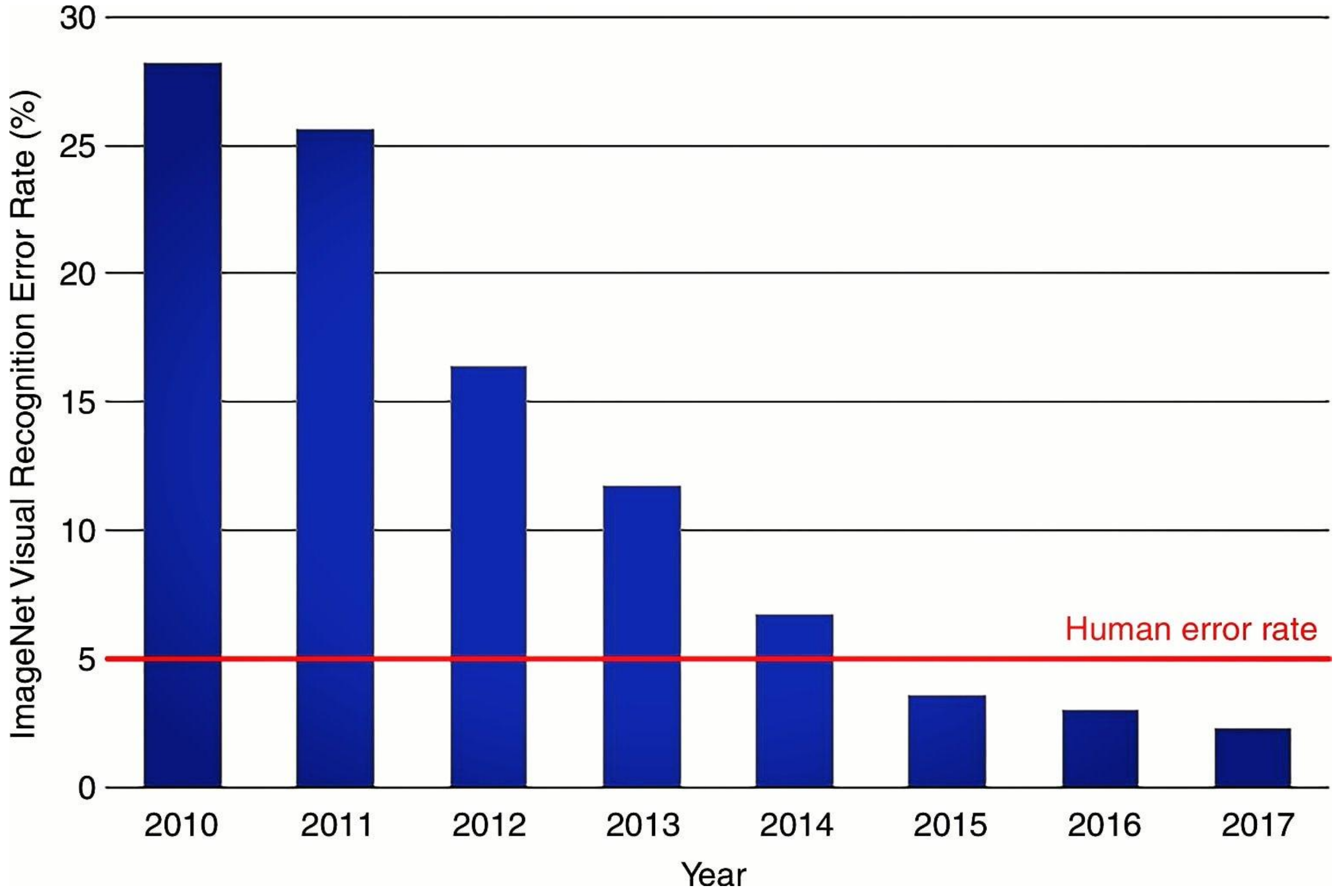
Pipeline of Computer Vision



Pipeline of Computer Vision

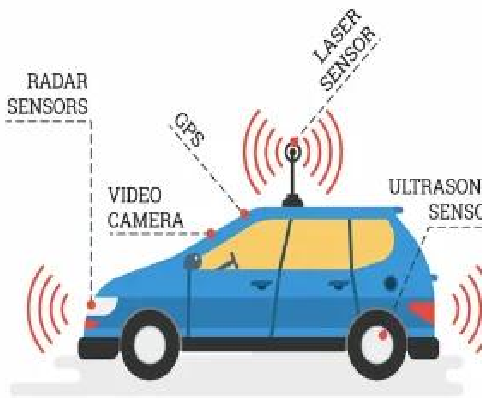


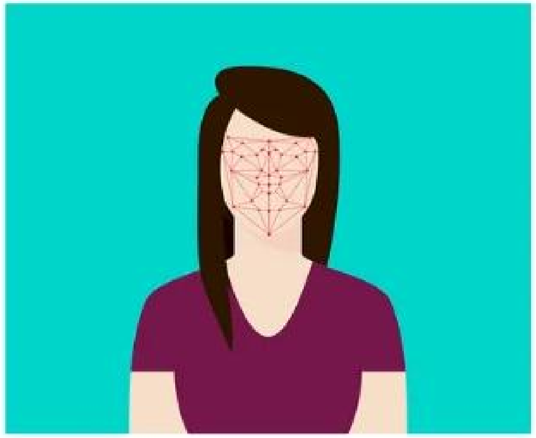
Success of Computer Vision



credit: https://www.synopsys.com/designware-ip/technical-bulletin/computer-vision-lab-life.html

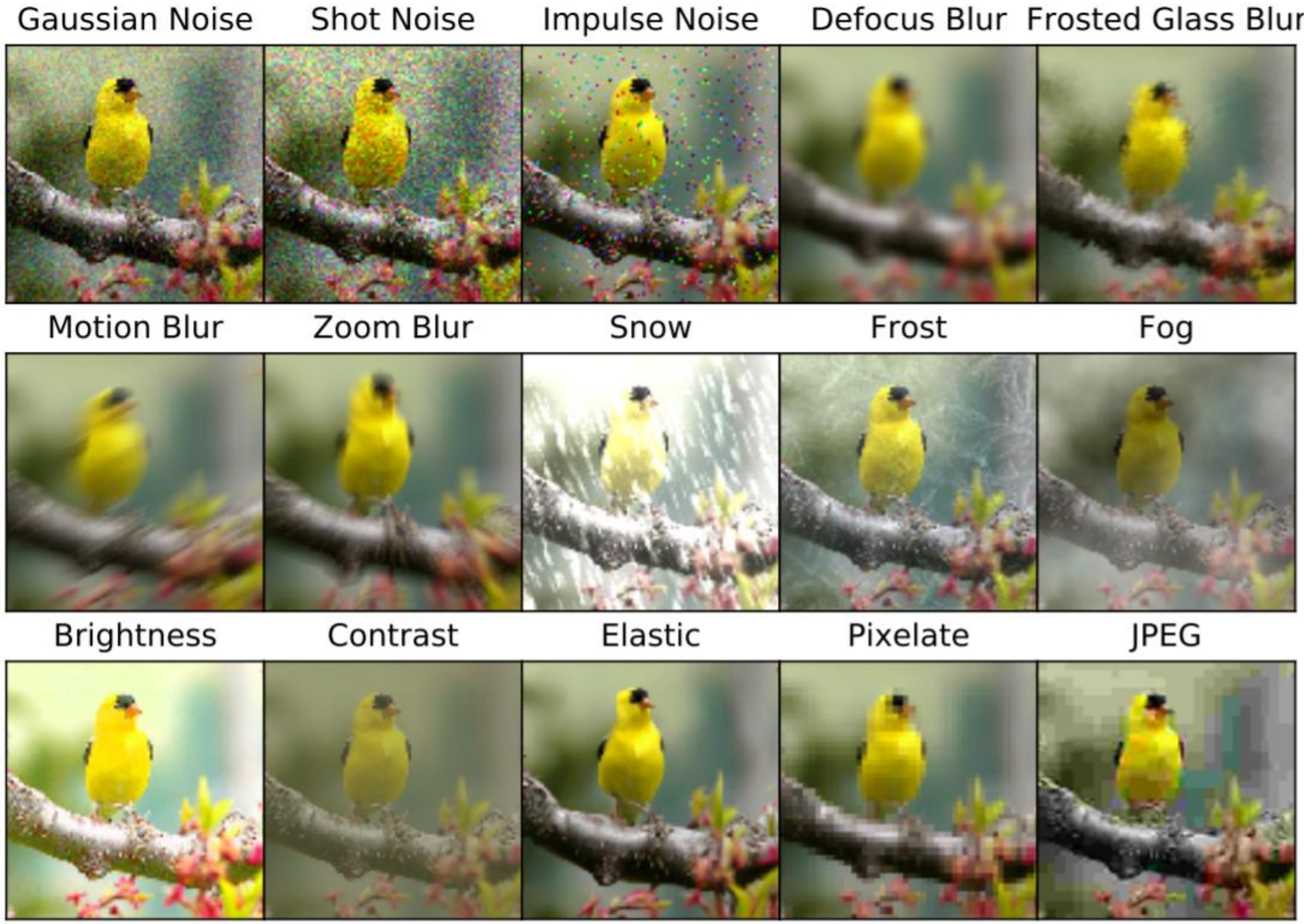
Success of Computer Vision





Google Lens

Vulnerability of Computer Vision



common perturbations

Vulnerability of Computer Vision

face recognition[1]

self-driving cars[2]

[1] https://ipvm.com/reports/face-masks
[2] https://www.theguardian.com/technology/2018/mar/22/video-released-of-uber-self-driving-crash-that-killed-woman-in-arizona

Vulnerability of Computer Vision to Common Perturbations?

Vulnerability of Computer Vision to Adversarial Perturbations!

worst-case

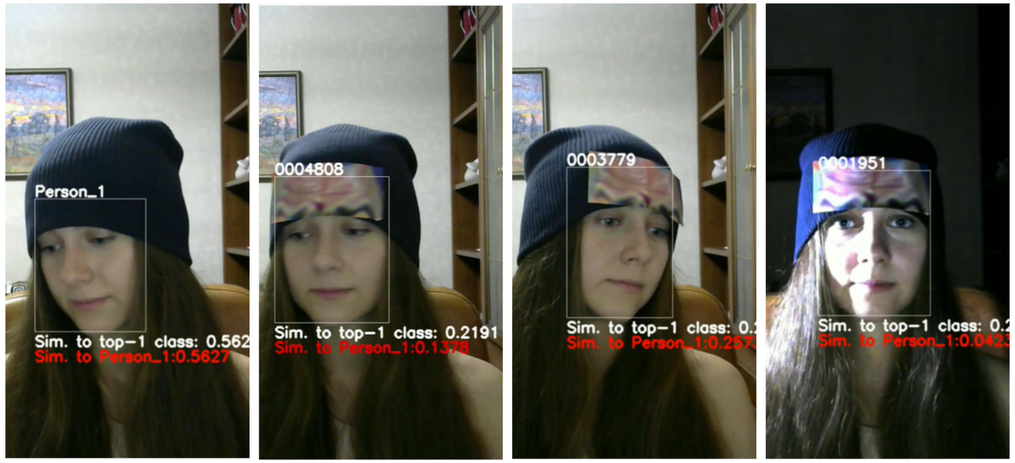
Adversarial Vulnerability of Computer Vision

face recognition[1]

[1] https://ipvm.com/reports/face-masks

Adversarial Vulnerability of Computer Vision

face recognition[1]



adversarial hat[2]

[1] https://ipvm.com/reports/face-masks
[2] Komkov, Stepan, and Aleksandr Petiushko. "Advhat: Real-world adversarial attack on arcface face id system." ICPR 2021.

Adversarial Vulnerability of Computer Vision

self-driving cars[1]

[1] https://www.theguardian.com/technology/2018/mar/22/video-released-of-uber-self-driving-crash-that-killed-woman-in-arizona

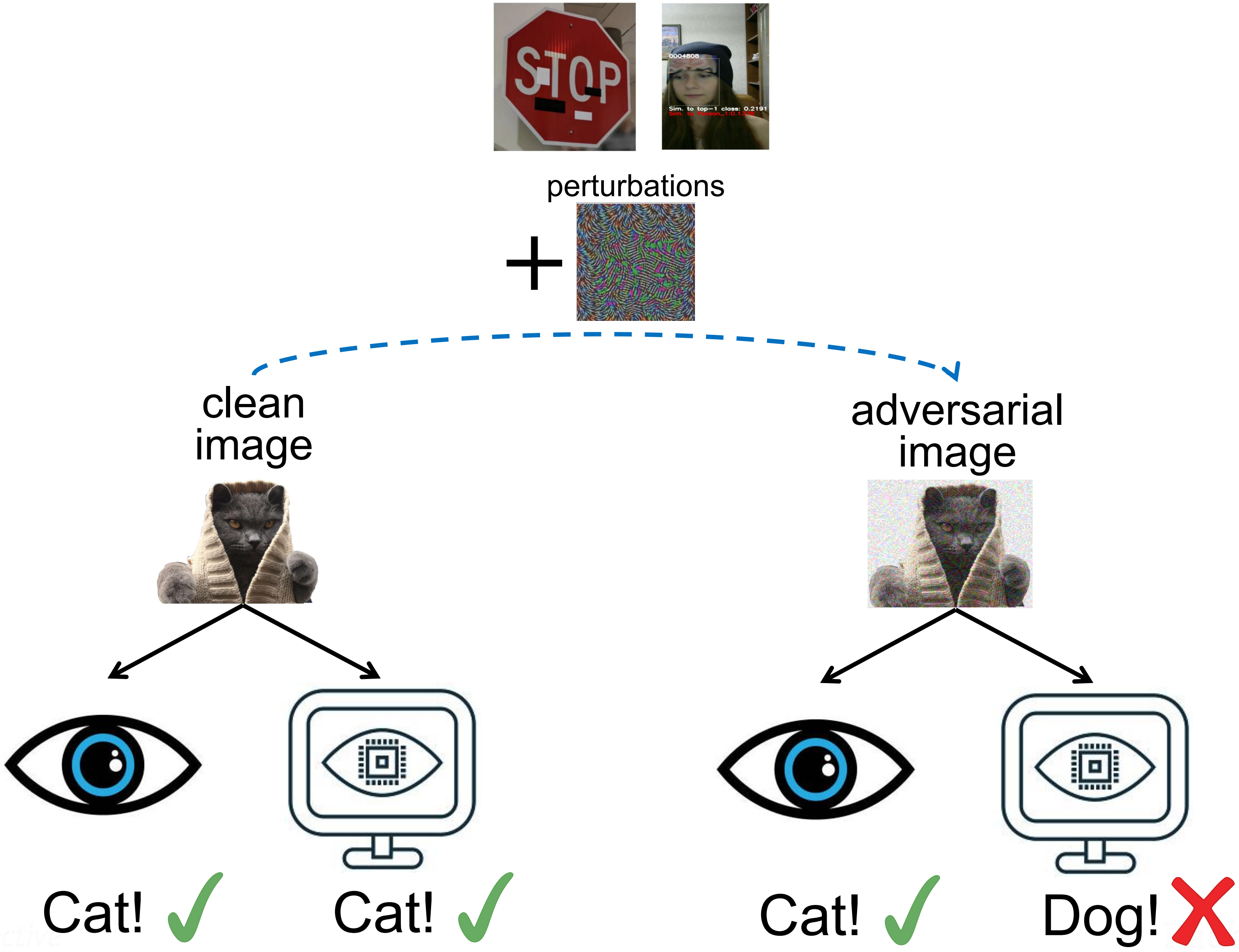
Adversarial Vulnerability of Computer Vision

self-driving cars[1]

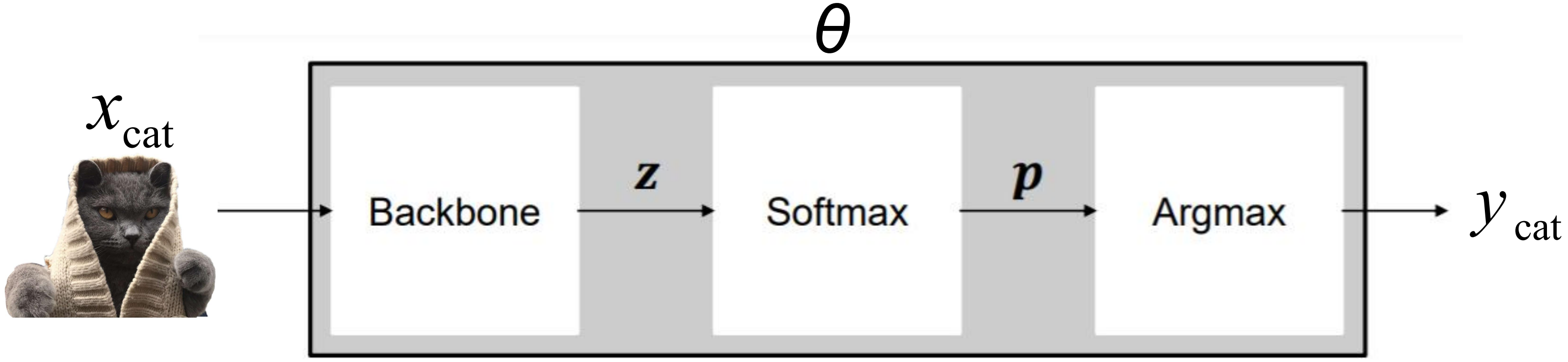
adversarial sticker [2]

[1] https://www.theguardian.com/technology/2018/mar/22/video-released-of-uber-self-driving-crash-that-killed-woman-in-arizona
[2] Eykholt et al. *Robust physical-world attacks on deep learning visual classification*. CVPR 2018.

Formulate Adversarial Images

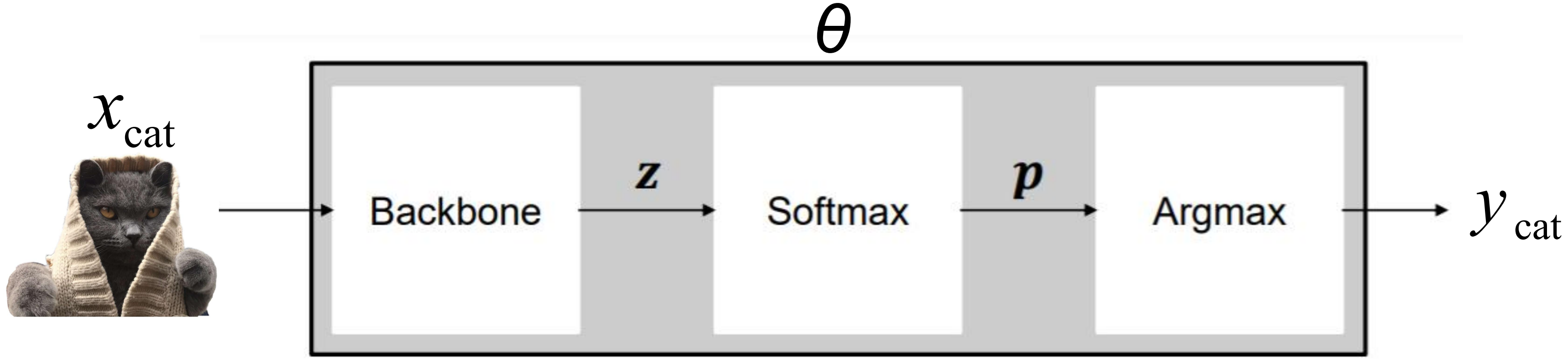


Generate Adversarial Images x’



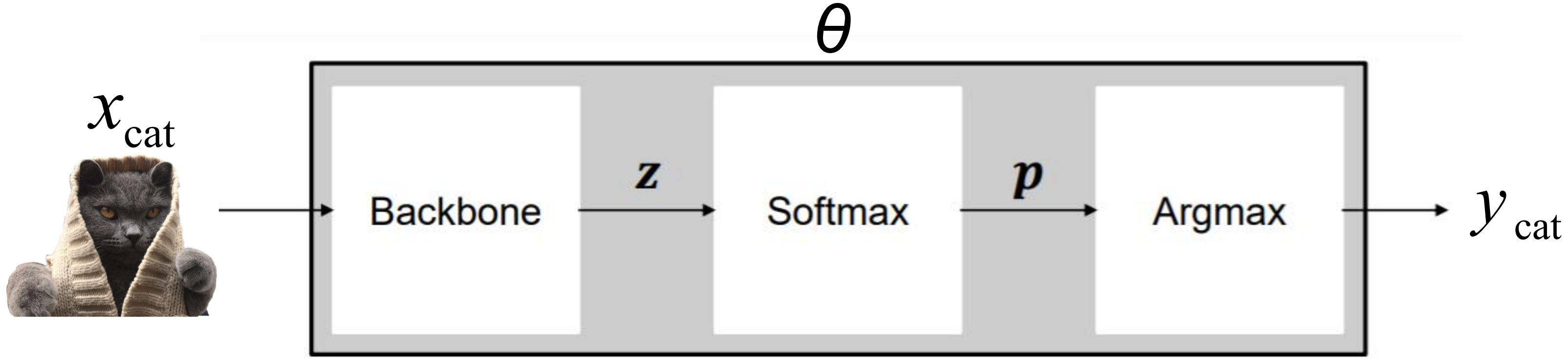
$$\theta' = \arg\min_{\theta} J(\theta, x_{\text{cat}}, y_{\text{cat}})$$

Generate Adversarial Images x’



$$\theta' = \arg\min_{\theta} J(\theta, x_{\text{cat}}, y_{\text{cat}}) \quad \Longleftrightarrow \quad x' = \arg\min_{x} J(\theta_o, x, y_{\text{t}})$$ targeted

Generate Adversarial Images x’



$$\theta' = \arg\min_{\theta} J(\theta, x_{\text{cat}}, y_{\text{cat}}) \quad \Longleftrightarrow$$

$$x' = \arg\min_{x} J(\theta_o, x, y_{\text{t}})$$ targeted

$$x' = \arg\max_{x} J(\theta_o, x, y_{\text{cat}})$$ non-targeted

Generate Adversarial Images x’

θ

x_{cat}

$\rightarrow$ Backbone $\xrightarrow{\boldsymbol{z}}$ Softmax $\xrightarrow{\boldsymbol{p}}$ Argmax $\rightarrow$ y_{cat}

$$\theta' = \arg\min_{\theta} J(\theta, x_{\text{cat}}, y_{\text{cat}}) \longleftrightarrow$$

$$x' = \arg\min_{x} J(\theta_o, x, y_{\text{t}}) \quad \text{targeted}$$

$$x' = \arg\max_{x} J(\theta_o, x, y_{\text{cat}}) \quad \text{non-targeted}$$

x' x_{cat}

$\Big\|$  $-$  $\Big\|_\infty \leq \varepsilon$

Generate Adversarial Images x'

Objective: $x' = \arg\min_{x} J(\theta_o, x, y_t) \quad \text{s.t.} \quad \|x' - x_{\text{cat}}\|_\infty \le \varepsilon$

Optimization: Iterative-Fast Gradient Sign Method (I-FGSM)[1]

$$x'_0 = x_{\text{cat}}, \;\; x'_{i+1} = x'_i - \text{sign}(\nabla_x J(x'_i, y_t))$$

$$x'_{i+1} \leftarrow \text{clip}(x'_{i+1} - x_{\text{cat}}, -\varepsilon, \varepsilon)$$

[1] Kurakin et al. *Adversarial Examples in the Physical World*. ICLR workshop 2017

Recap of Background

Computer vision success

↳ ... vulnerability to common perturbations

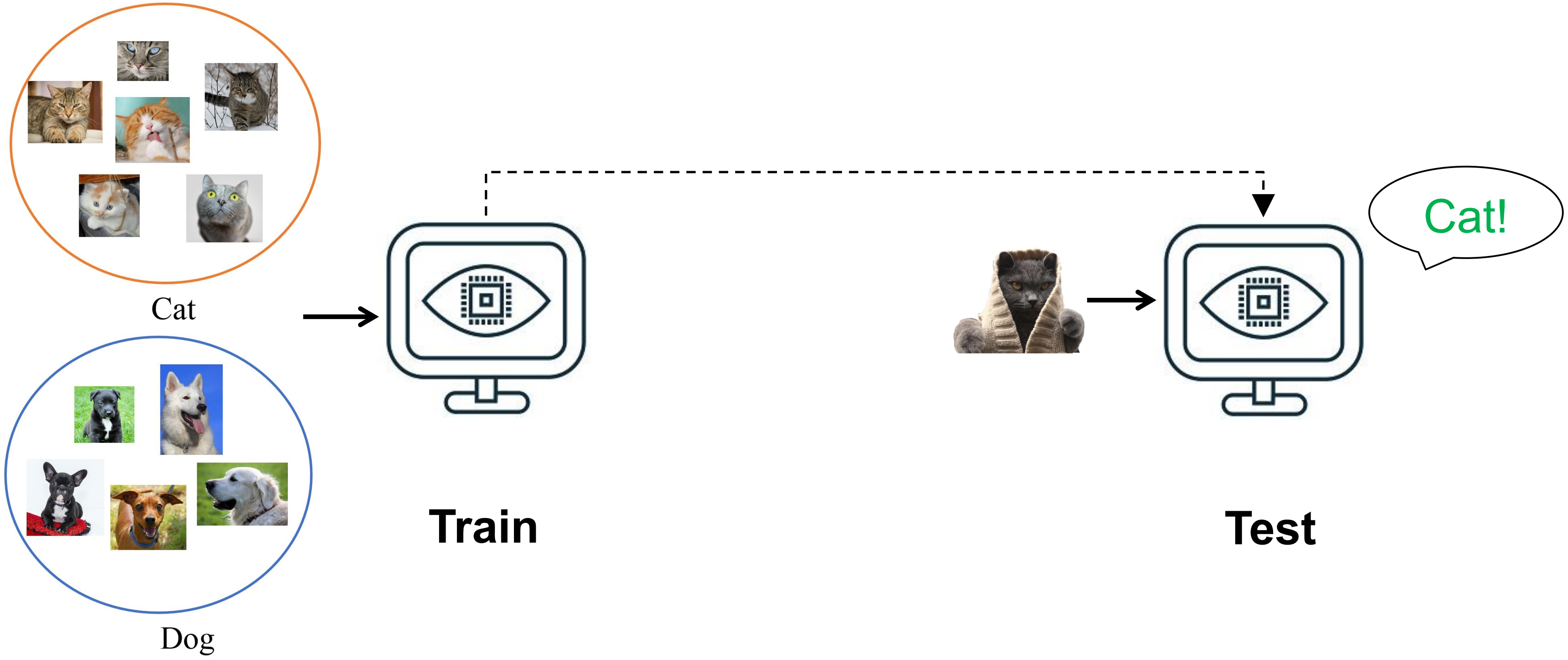
↳ ... adversarial perturbations

↳ generate adversarial images

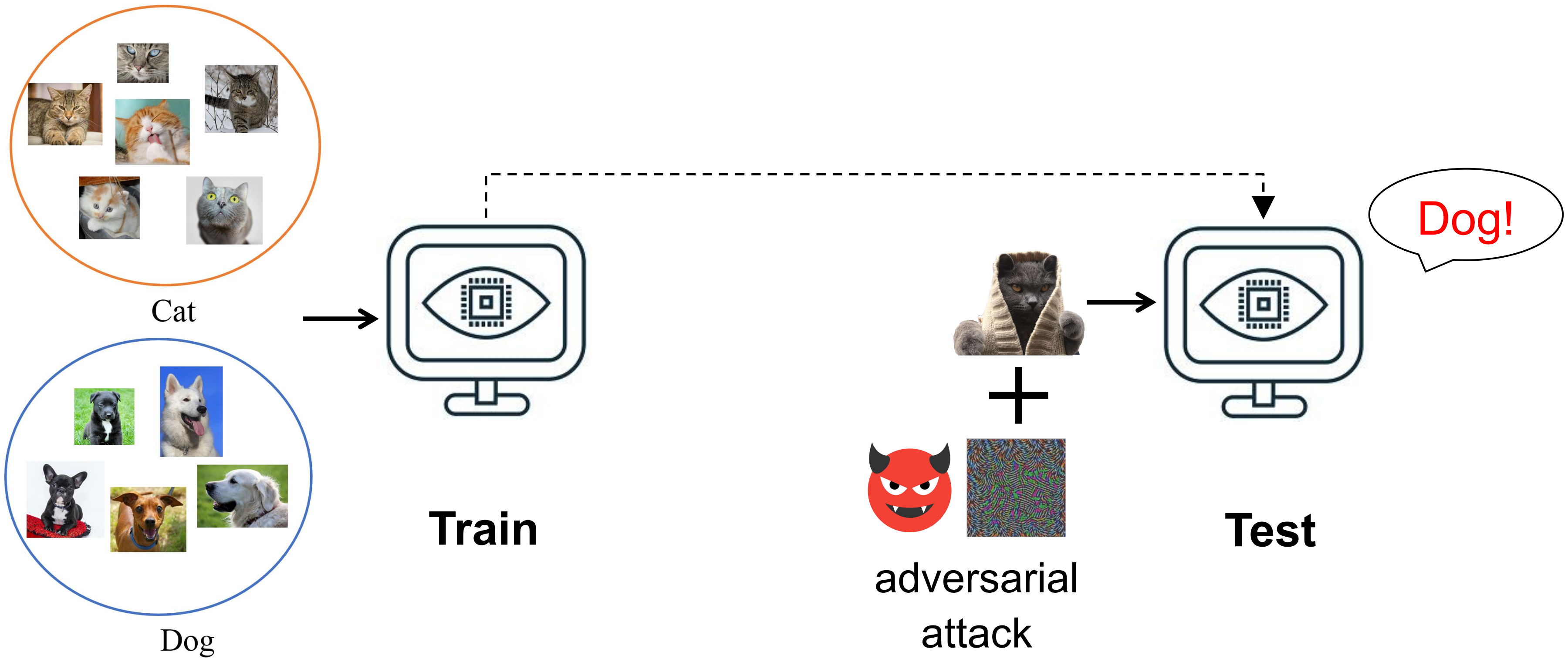
Outline

- Background of computer vision (CV) and adversarial images
- Two of our recent projects
- Other related projects

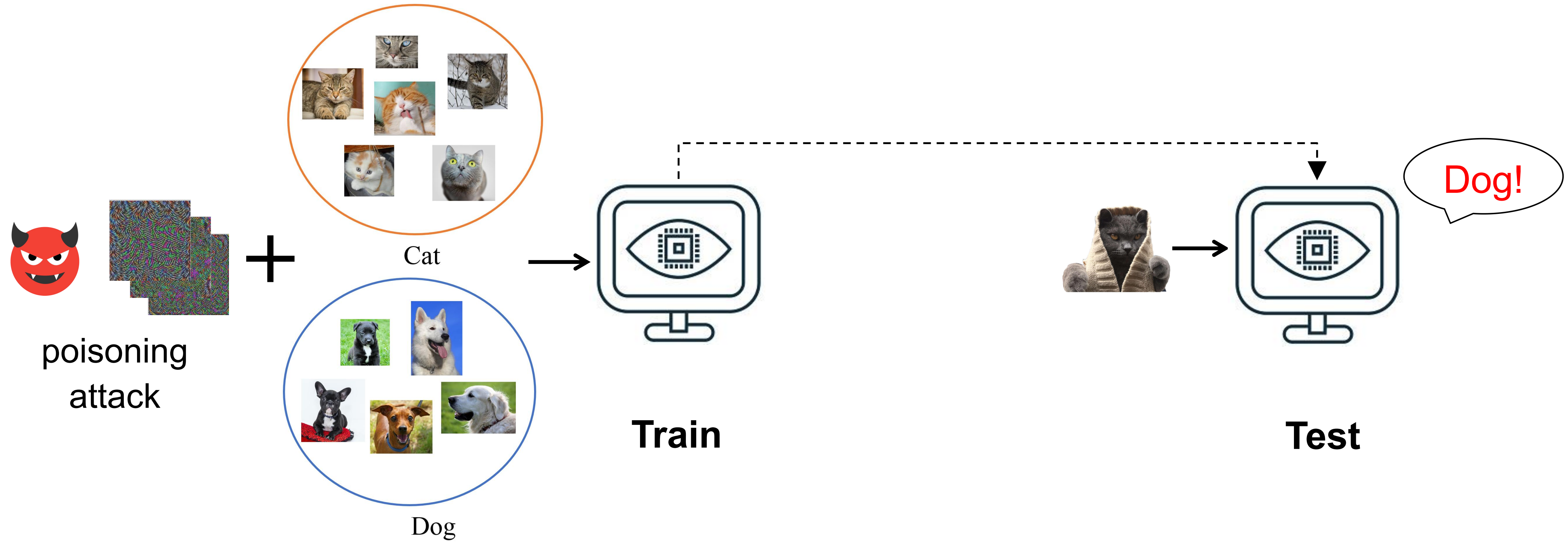
Computer Vision Pipeline



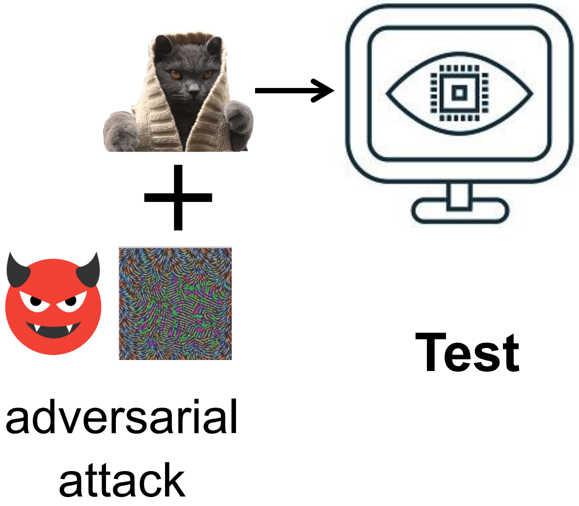
Test-Time Attack

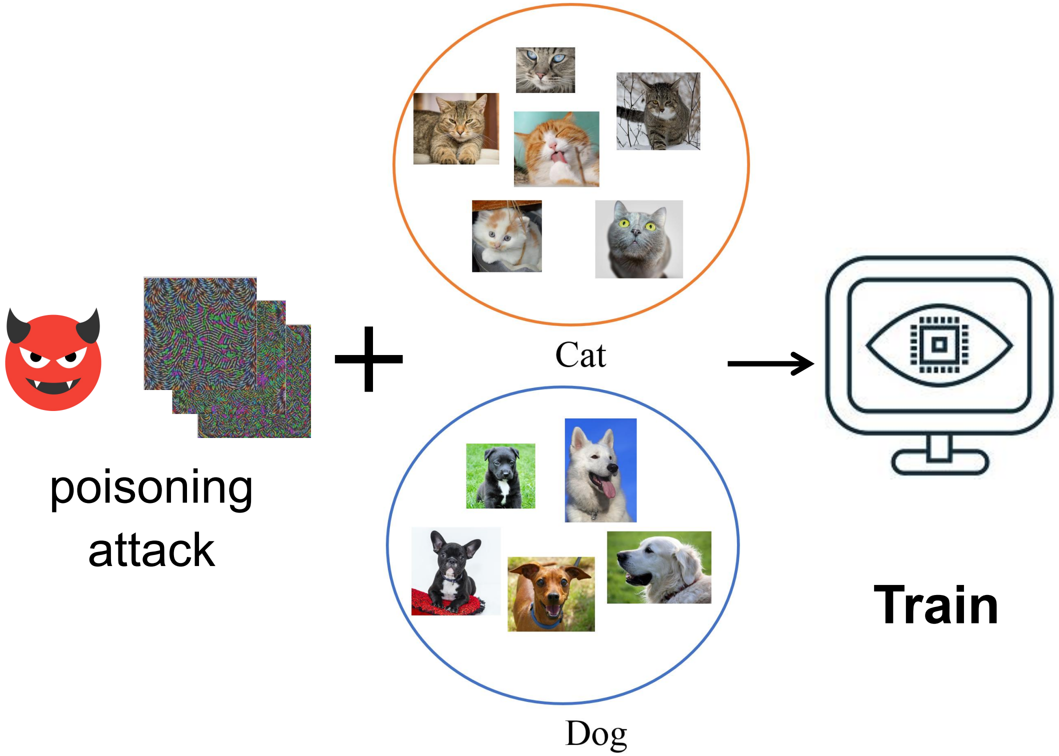


Training-Time Attack



Two projects





On Success and Simplicity: A Second Look at **Transferable Targeted Attacks**. NeurIPS 2021

Data Poisoning against **Adversarial Training**. Under review

Consensus-Challenging Insights

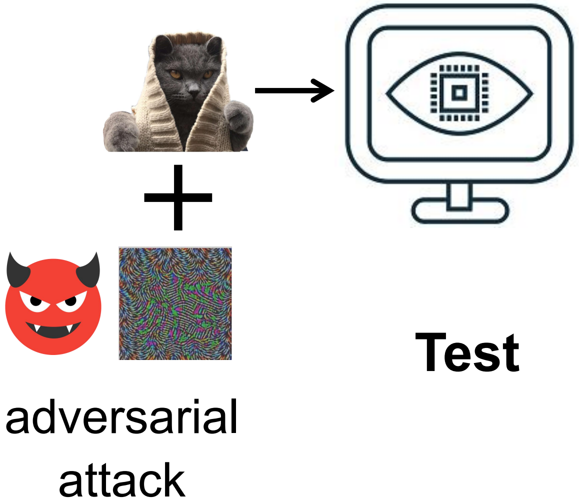
Mission:
Impossible

Existing work[1-N]

Mission:
Possible (and Simple)

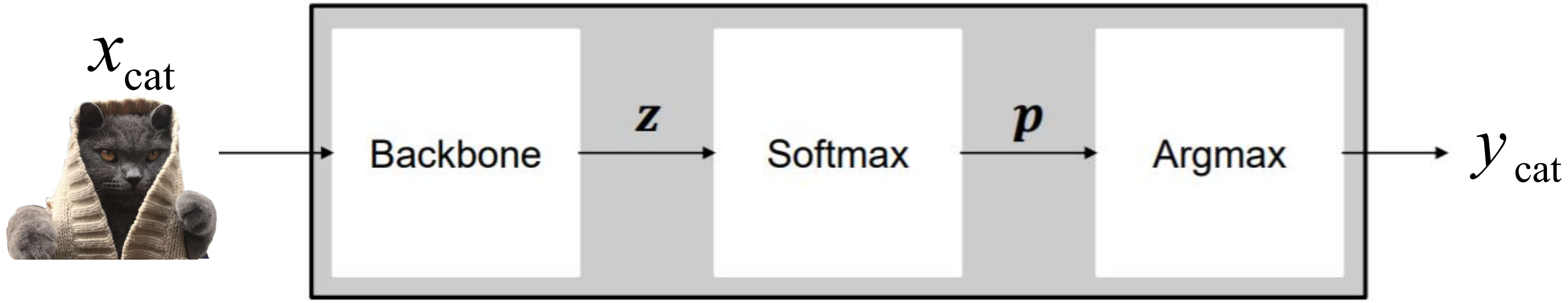
Ours

Project 1. Transferable Targeted Attacks



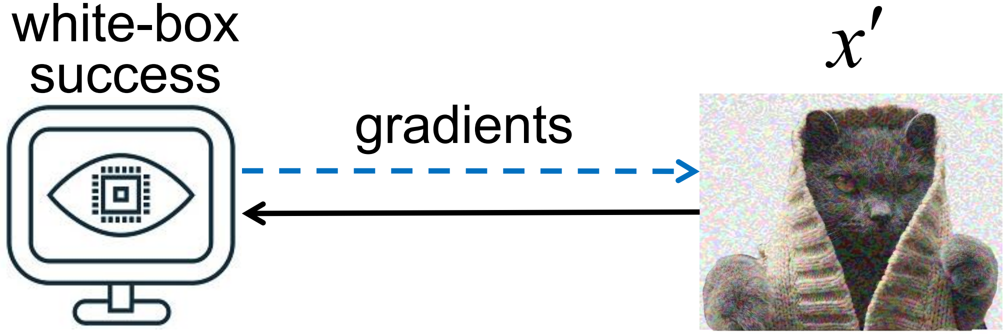
On Success and Simplicity: A Second Look at
Transferable Targeted Attacks. NeurIPS 2021

Transferable Targeted Attacks

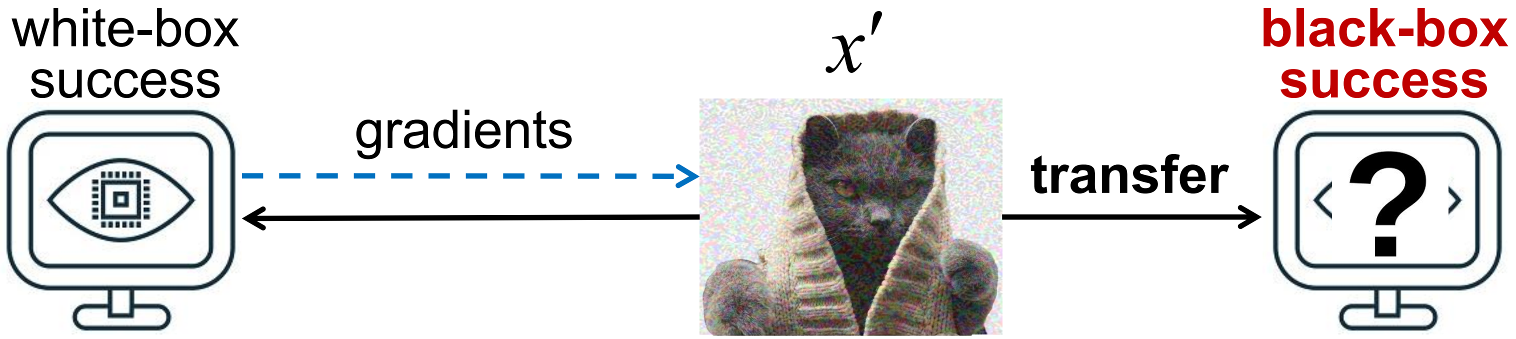


$$\theta' = \arg\min_{\theta} J(x_{cat}, y_{cat}) \quad \Longleftrightarrow \quad \begin{aligned} x' &= \arg\min_{x} J(x, y_t) && \text{targeted} \\ x' &= \arg\max_{x} J(x, y_{cat}) && \text{non-targeted} \end{aligned}$$

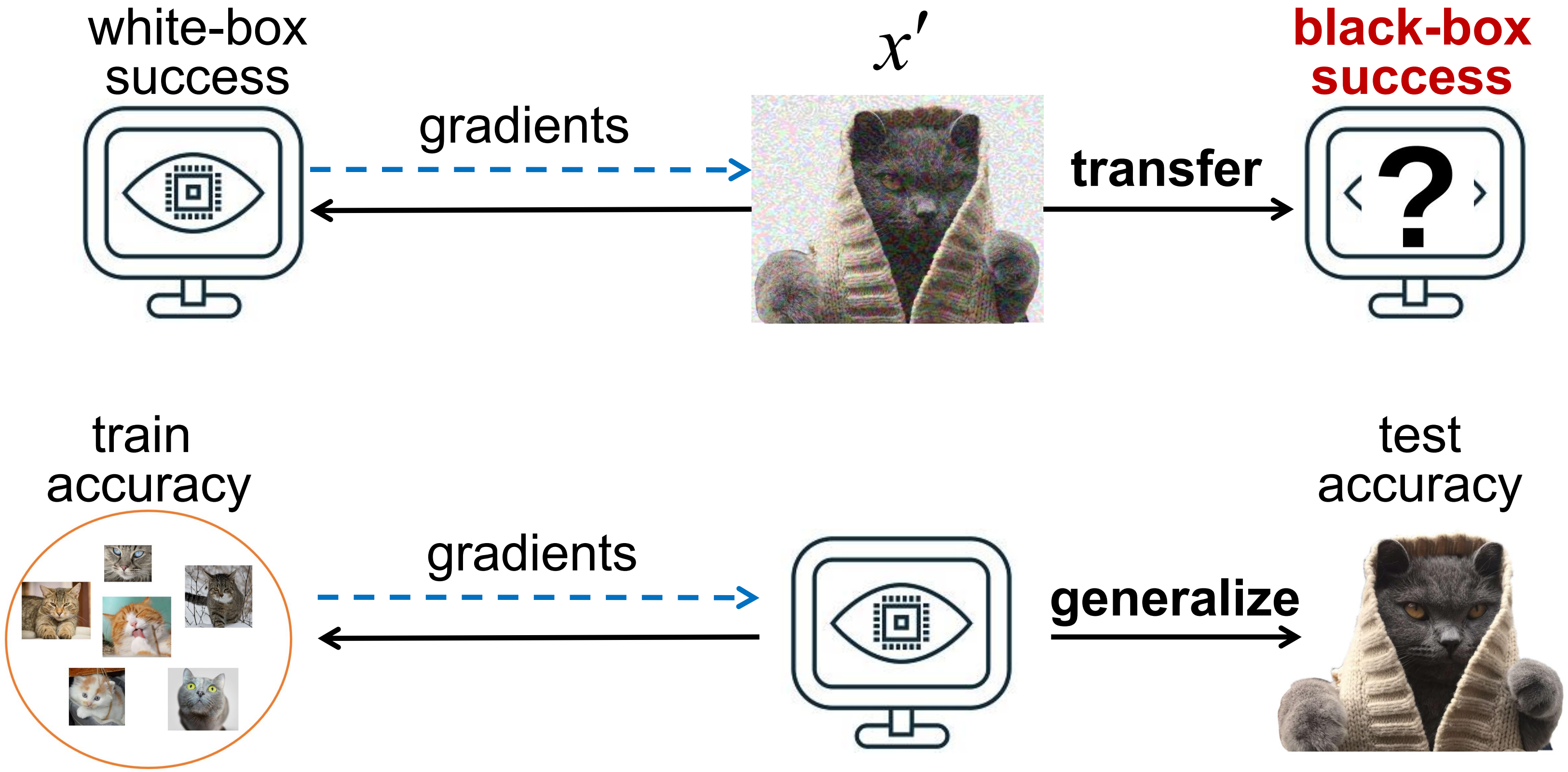
Transferable Targeted Attacks



Transferable Targeted Attacks



Transferable Targeted Attacks



Existing Work for Transferable Attacks

Iterative-Fast Gradient Sign Method (I-FGSM):

$$x'_0 = x_o, \quad x'_{i+1} = x'_i - \alpha \cdot \mathrm{sign}(\nabla_x J(x'_i, y_t))$$

Transfer techniques:

- Gradient stabilization

  e.g., momentum-based (MI-FGSM)[1]:

$$\boldsymbol{g}_{i+1} = \mu \cdot \boldsymbol{g}_i + \frac{\nabla_{\boldsymbol{x}} J(\boldsymbol{x}'_i, y_t)}{\|\nabla_{\boldsymbol{x}} J(\boldsymbol{x}'_i, y_t)\|_1}$$

$$\boldsymbol{x}'_{i+1} = \boldsymbol{x}'_i - \alpha \cdot \mathrm{sign}(\boldsymbol{g}_i)$$

- Input augmentation

  e.g., resizing & padding (DI-FGSM)[2]

  translation (TI-FGSM)[3]:

$$\boldsymbol{x}'_{i+1} = \boldsymbol{x}'_i - \alpha \cdot \mathrm{sign}(\nabla_{\boldsymbol{x}} J(T(\boldsymbol{x}'_i, p), y_t))$$

[1] Dong et al. *Boosting Adversarial Attacks with Momentum*. CVPR 2018.
[2] Xie et al. *Improving Transferability of Adversarial Examples with Input Diversity*. CVPR 2019
[3] Dong et al. *Evading defenses to transferable adversarial examples by translation-invariant attacks*. CVPR 2019.

Consensus-Challenging Insight

Impossible to use I-FGSM for targeted transferability.

Existing work[1-6]

Possible and even SOTA.

Ours

[1] Liu et al. *Delving into transferable adversarial examples and black-box attacks*. ICLR 2017.
[2] Dong et al. *Boosting Adversarial Attacks with Momentum*. CVPR 2018.
[3] Inkawhich et al. *Feature space perturbations yield more transferable adversarial examples*. CVPR 2019.
[4] Inkawhich et al. *Transferable perturbations of deep feature distributions*. ICLR 2020.
[5] Inkawhich et al. *Perturbing across the feature hierarchy to improve standard and strict blackbox attack transferability*. NeurIPS 2020.
[6] Naseer et al. *On generating transferable targeted perturbations*. ICCV 2021.

Revive I-FGSM: Step 1. Ensemble (0% →15%)

ResNet50 → DenseNet121 (Iter. =10)

I-FGSM: ~0%
MI-FGSM: ~0.5%
TI-FGSM: ~0.5%
DI-FGSM: ~5%
MTDI-FGSM: ~15%

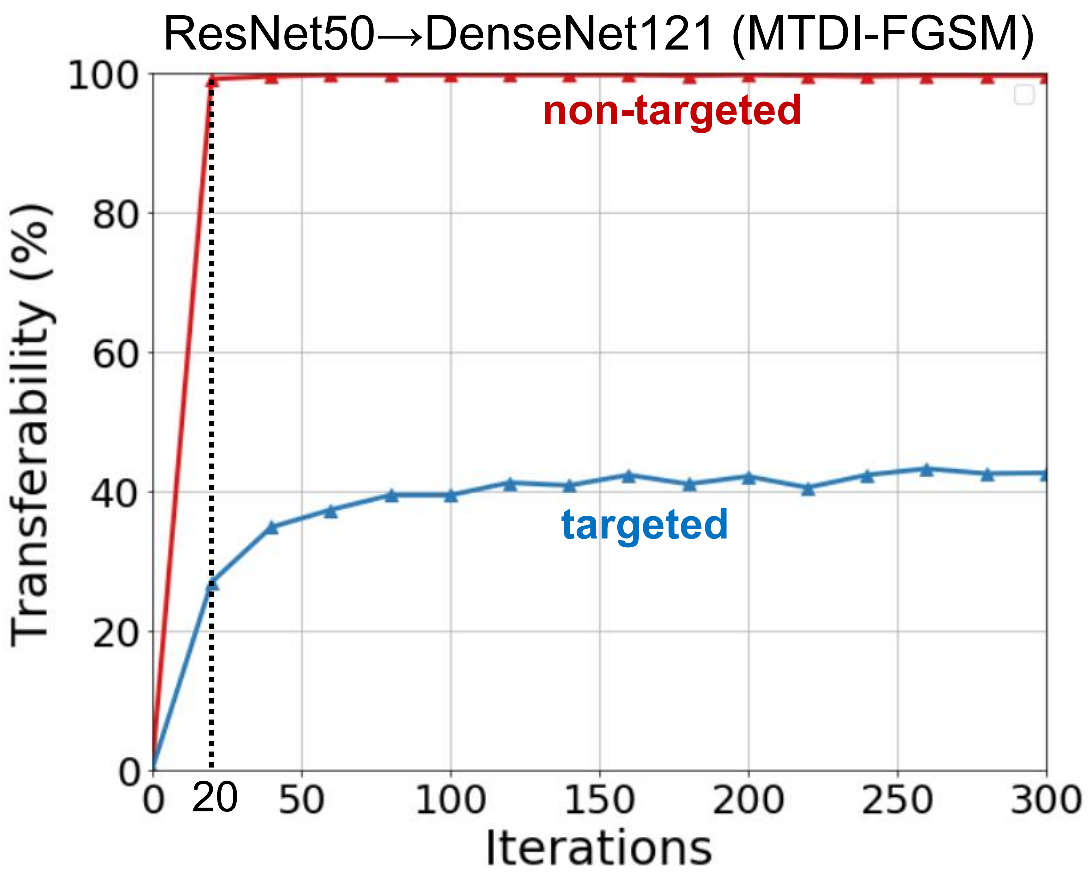
Revive I-FGSM: Step 1. Ensemble (0% →15%)

ResNet50 → DenseNet121 (Iter. =10)

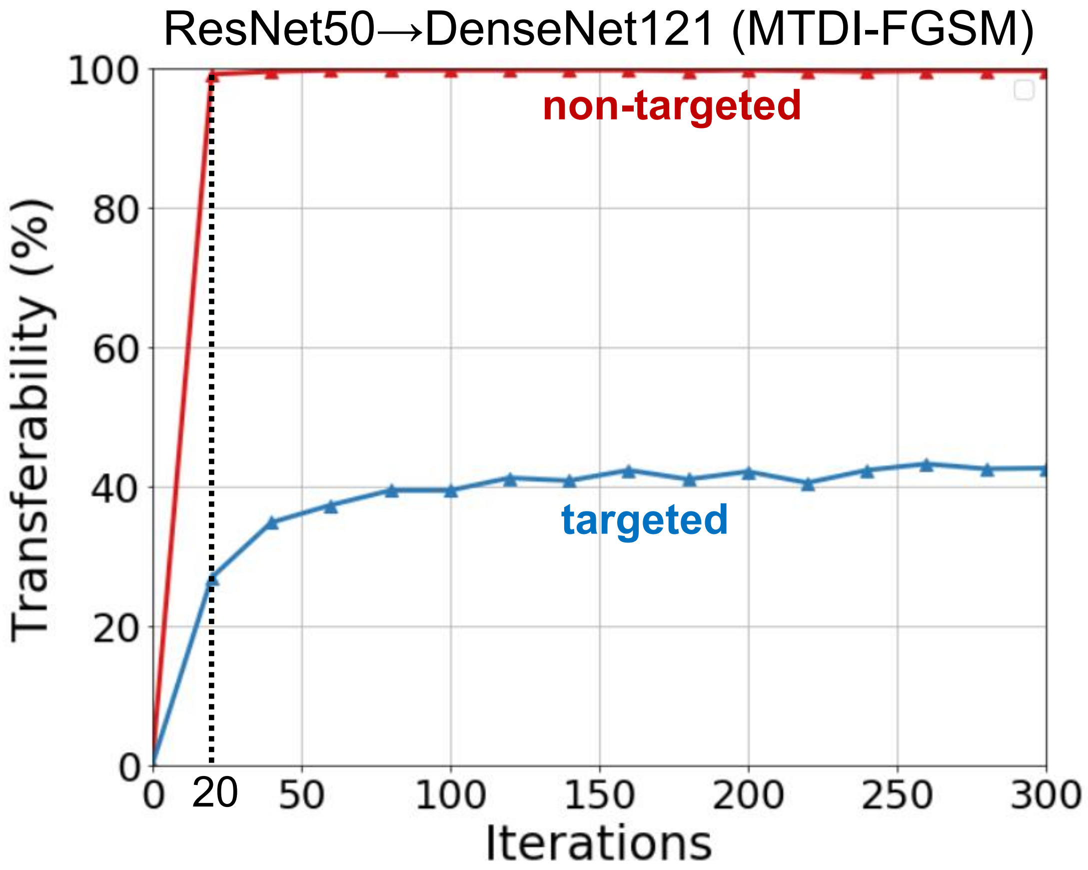
I-FGSM: ~0%
MI-FGSM: ~0.5%
TI-FGSM: ~0.5%
DI-FGSM: ~5%
MTDI-FGSM: ~15%

Mostly MI-FGSM in existing work

Revive I-FGSM: Step 2. More Iterations (15% →42%)



Revive I-FGSM: Step 2. More Iterations (15% →42%)



<20 iterations in existing work:

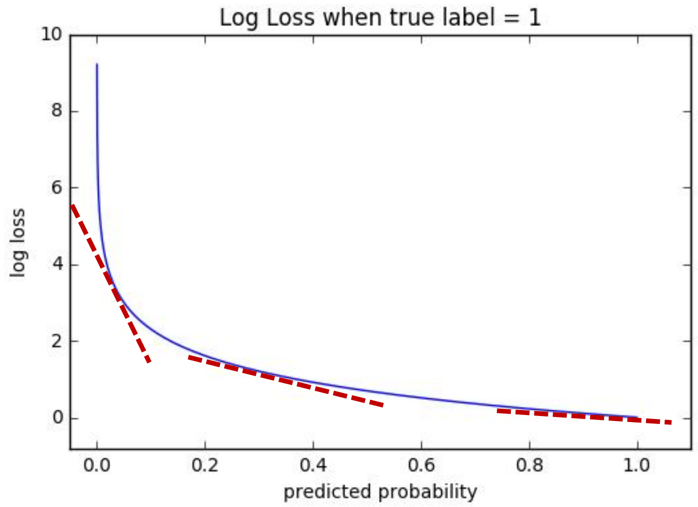
- fail to converge
- unnecessary constraint

Revive I-FGSM: Step 3. Better Loss

Cross-Entropy Loss (L_{CE}) causes **decreasing gradient** problem:

$$L_{CE} = -1 \cdot \log(p_t) = -\log(\frac{e^{z_t}}{\sum e^{z_j}}) = -z_t + \log(\sum e^{z_j}),$$

$$\frac{\partial L_{CE}}{\partial z_t} = -1 + \frac{\partial \log(\sum e^{z_j})}{\partial e^{z_t}} \cdot \frac{\partial e^{z_t}}{\partial z_t} = -1 + \frac{e^{z_t}}{\sum e^{z_j}} = \underline{-1 + p_t}.$$

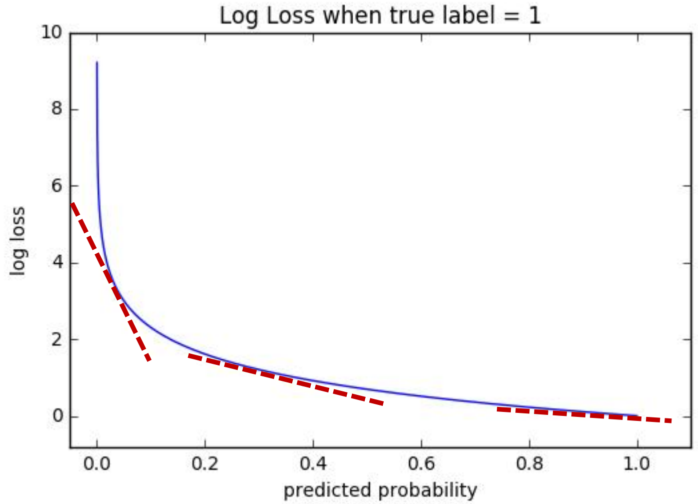


Revive I-FGSM: Step 3. Better Loss

Cross-Entropy Loss (L_{CE}) causes **decreasing gradient** problem:

$$L_{CE} = -1 \cdot \log(p_t) = -\log(\frac{e^{z_t}}{\sum e^{z_j}}) = \underline{-z_t} + \log(\sum e^{z_j}),$$

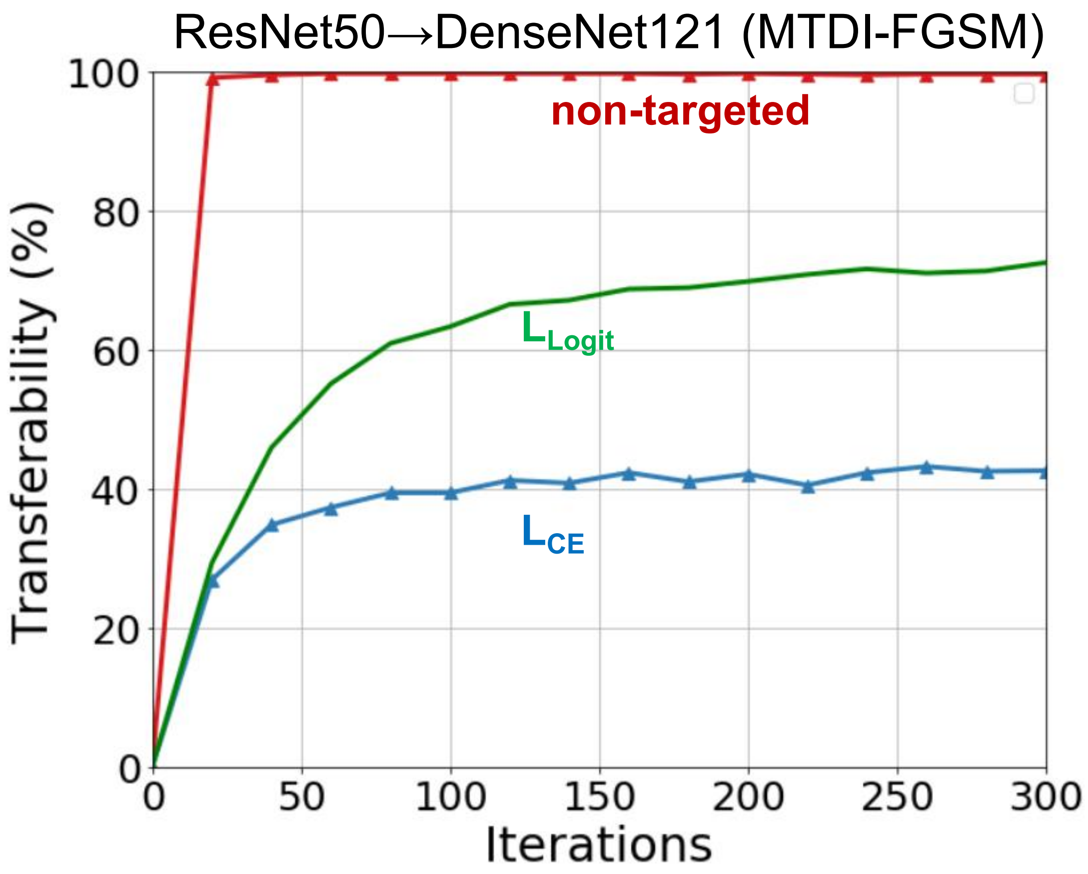
$$\frac{\partial L_{CE}}{\partial z_t} = -1 + \frac{\partial \log(\sum e^{z_j})}{\partial e^{z_t}} \cdot \frac{\partial e^{z_t}}{\partial z_t} = -1 + \frac{e^{z_t}}{\sum e^{z_j}} = -1 + p_t.$$



Logit Loss (L_{Logit}) is better:

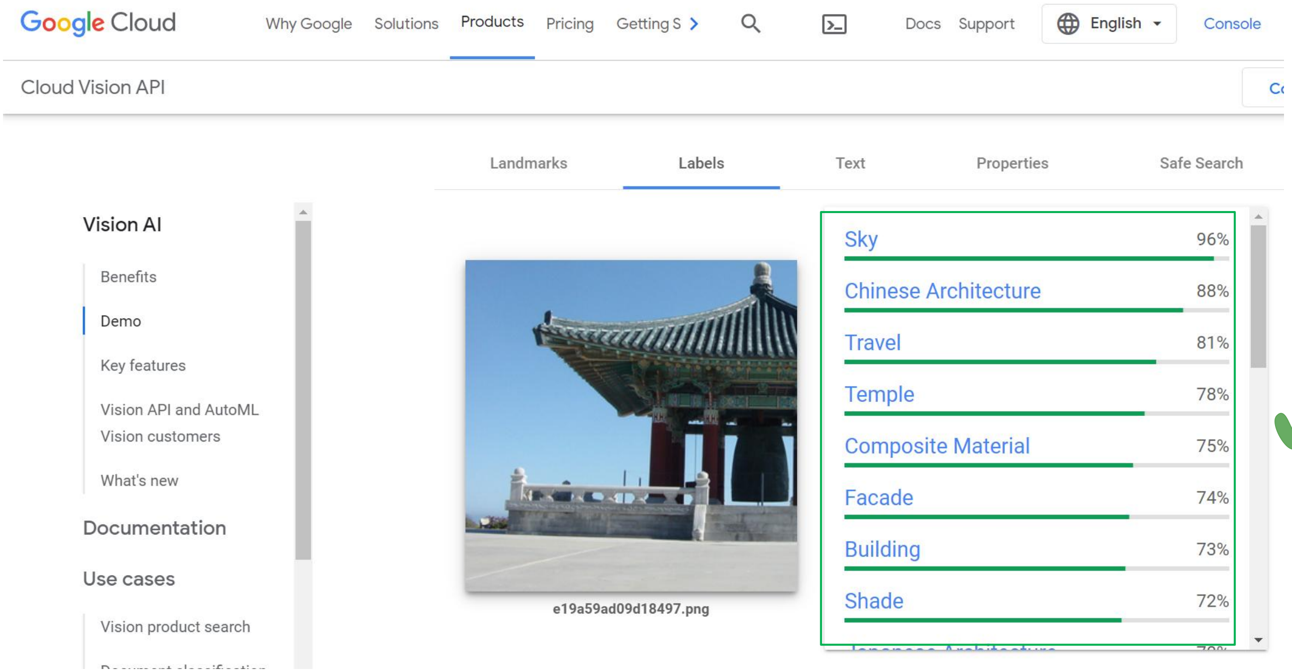
$$L_{Logit} = \underline{-z_t}, \; \frac{\partial L_{Logit}}{\partial z_t} = -1.$$

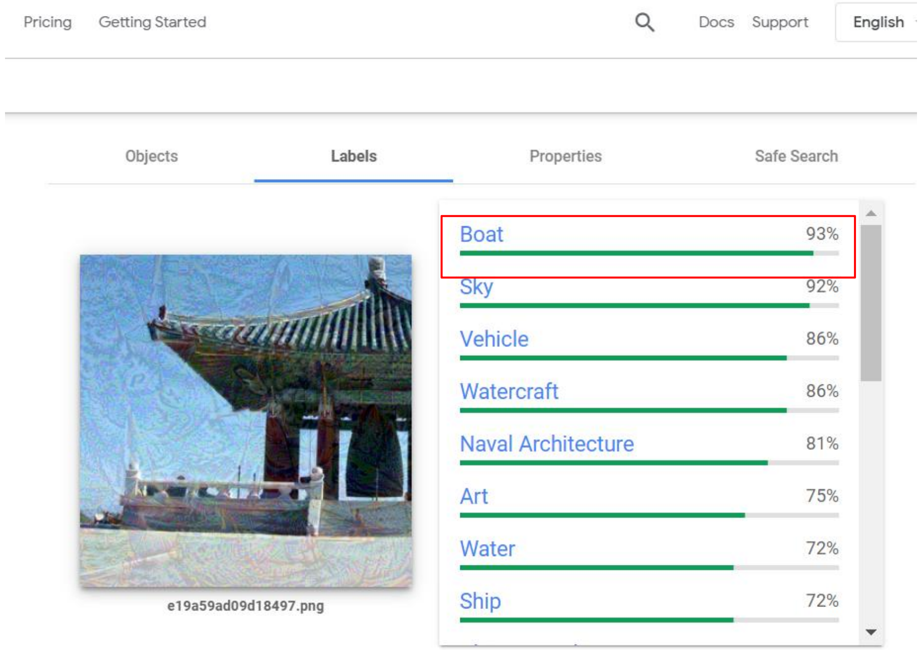
Revive I-FGSM: Step 3. Better Loss (42% →72%)



Other Analyses: Real-World Attacks

Services	Evaluation	Ori	CE	Po+Trip	Logit
Object localization	non-targeted	31.50	53.00	51.75	**62.50**
	targeted	0	9.00	8.50	**19.25**
Label detection	non-targeted	9.75	34.00	22.50	**35.00**
	targeted	0	4.50	2.25	**6.25**



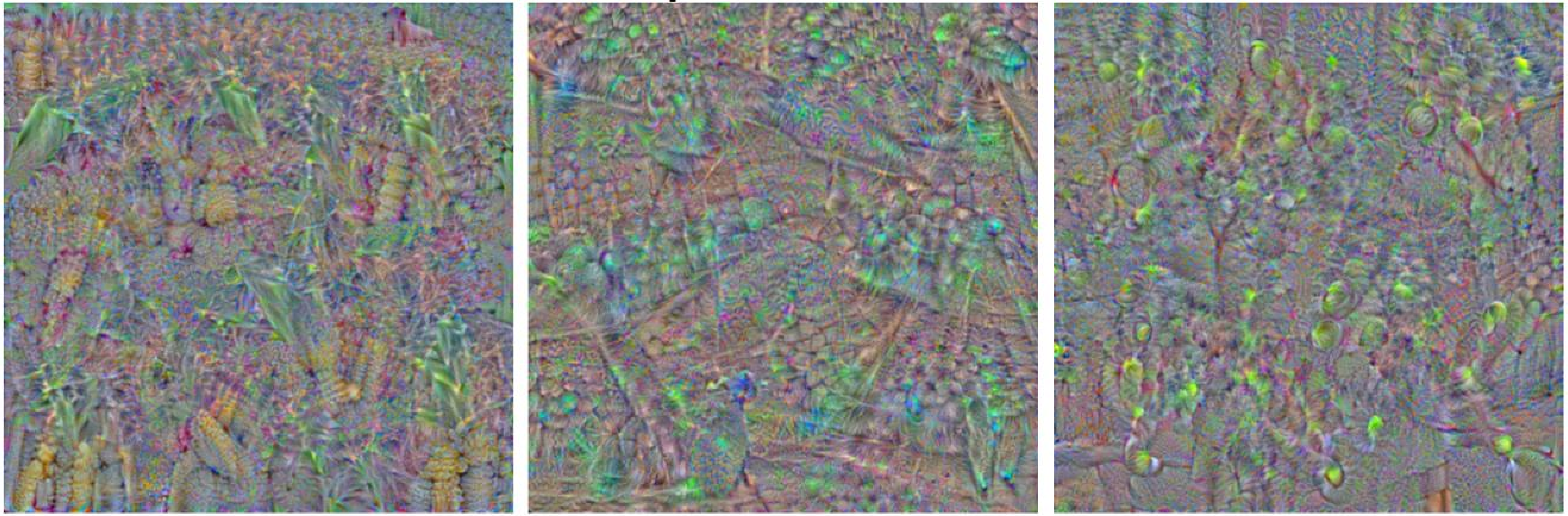


y_t = “yawl” (a type of boat)

[8] Zhao et al. *The Importance of Image Interpretation: Patterns of Semantic Misclassification in Real-World Adversarial Images*. MMM 2023.

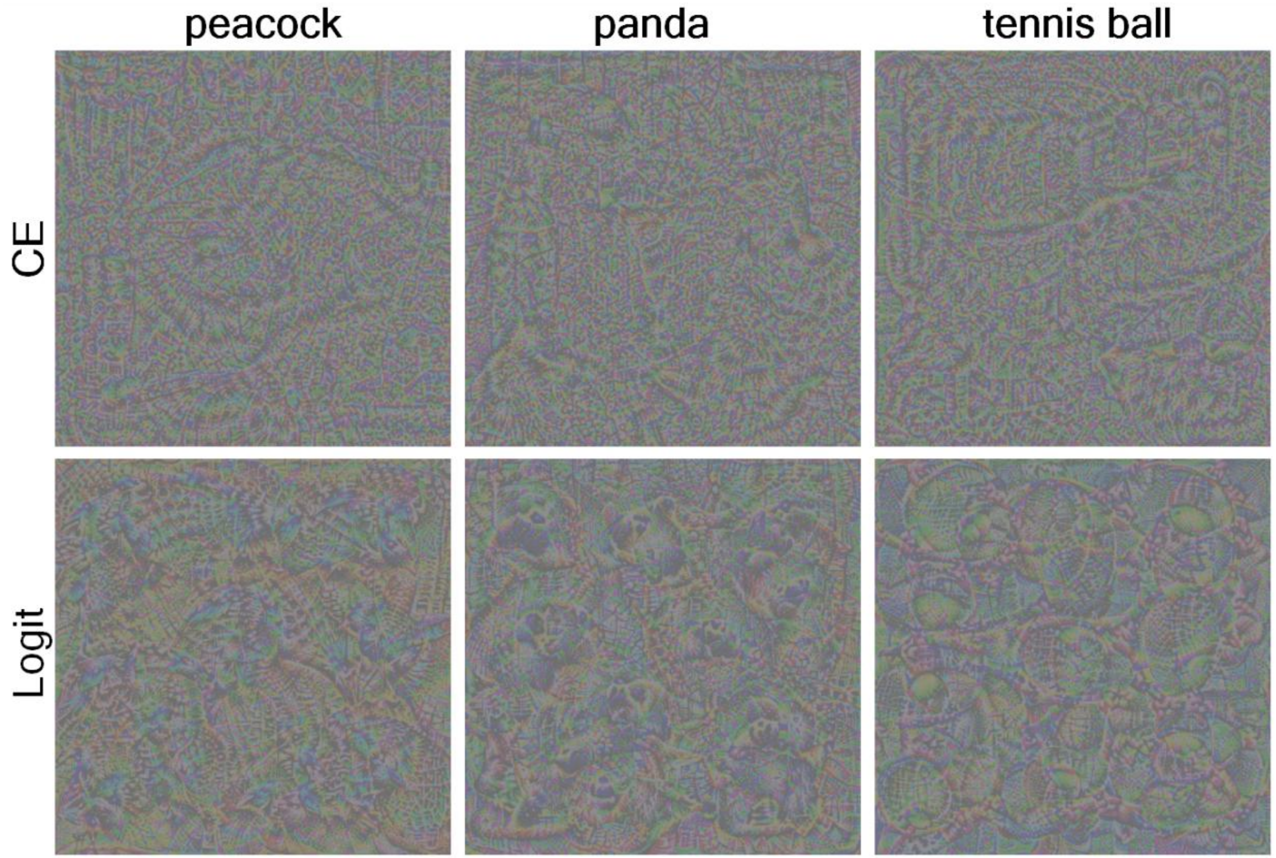
Other Analyses: Perturbation Semantics

“corn” “peacock” “tennis ball”



without ϵ

Targeted Universal Adversarial Perturbations (UAPs)[1]



with ϵ=16

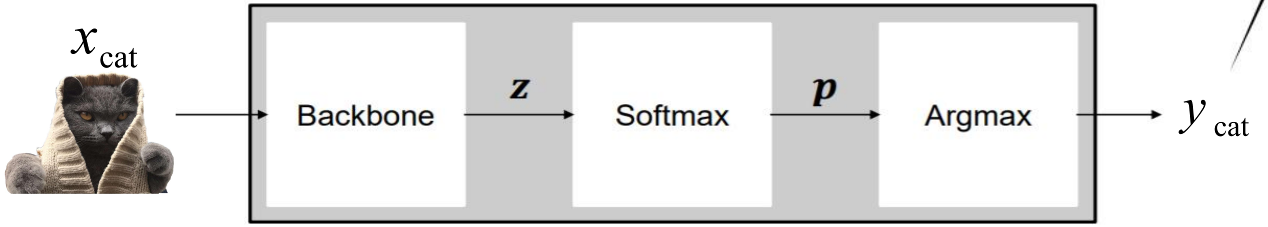
Success rates (%) of Targeted UAPs (ϵ=16)

Attack	Inc-v3	Res50	Dense121	VGG16
CE	2.6	9.2	8.7	20.1
Logit	**4.7**	**22.8**	**21.8**	**65.9**

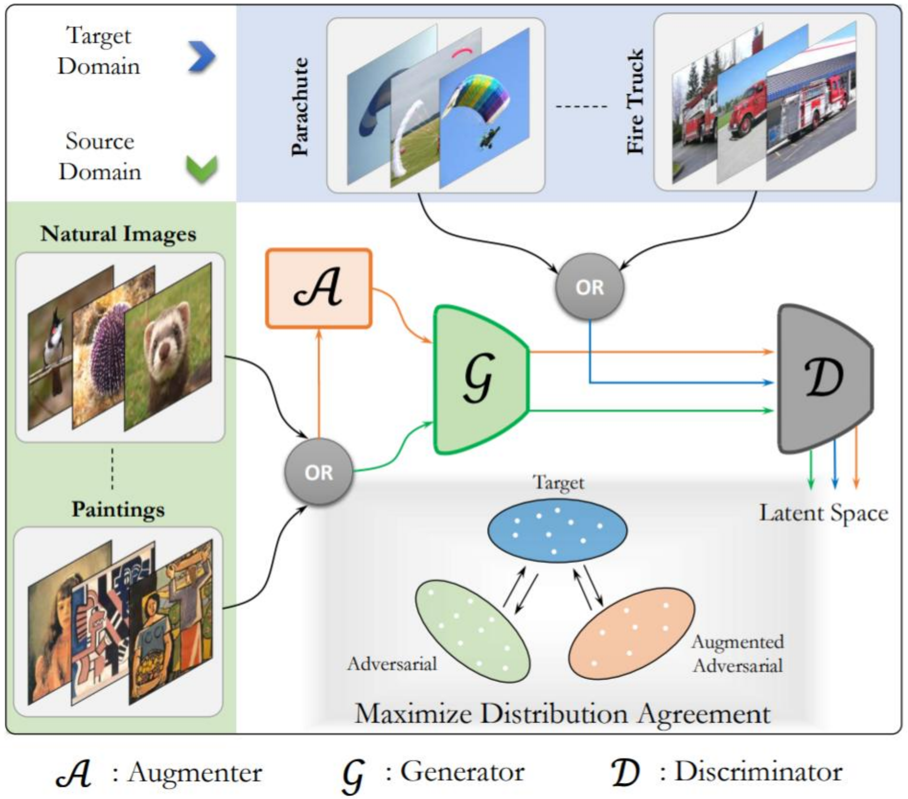
[1] Moosavi-Dezfooli et al. *Universal Adversarial Perturbations*. CVPR 2017.

Iterative (I-FGSM) vs. Generative

Iterative



Generative[1]



- Data: Single Input image | Massive training data
- Model: 1 × surrogate classifier | 1000 × target-specific generators

[1] Naseer et al. *On Generating Transferable Targeted Perturbation*. ICCV 2021

Iterative (I-FGSM) vs. Generative

x' $\quad$ x_{cat}

$\Big\|$  $-$  $\Big\|_{\infty} \leq \varepsilon$

Targeted Transferability (%)

Bound	Attack	D121	V16	D121-ens	V16-ens
$\epsilon = 16$	TTP [8]	**79.6**	**78.6**	92.9	89.6
	ours	75.9	72.5	**99.4**	**97.7**
$\epsilon = 8$	TTP [8]	37.5	46.7	63.2	66.2
	ours	**44.5**	**46.8**	**92.6**	**87.0**

[8] Naseer et al. *On Generating Transferable Targeted Perturbation*. ICCV 2021

Summary of Project 1

- 3 steps to revive I-FGSM
  - ensemble
  - more iterations
  - better (logit) loss

- Other Analyses
  - real-world attacks
  - targeted UAPs
  - iterative (I-FGSM) vs. generative

Summary of Project 1

- 3 steps to revive I-FGSM
  - ensemble
  - more iterations
  - better (logit) loss

- Other Analyses
  - real-world attacks
  - targeted UAPs
  - iterative (I-FGSM) vs. generative

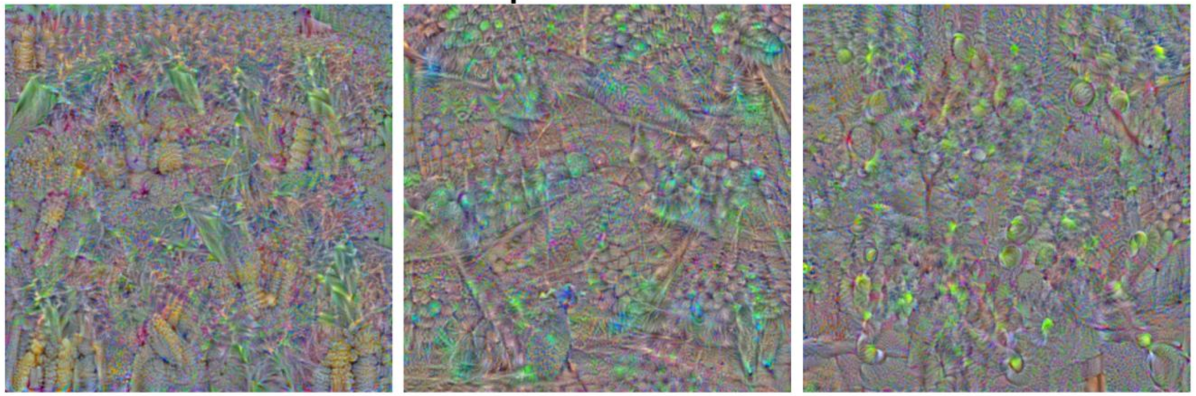
"God is in the details"

Future Work

Why transferable?

Semantic similarity

“corn” “peacock” “tennis ball”



and/or

Model similarity

Res50 → Dense121: ~70% 😄

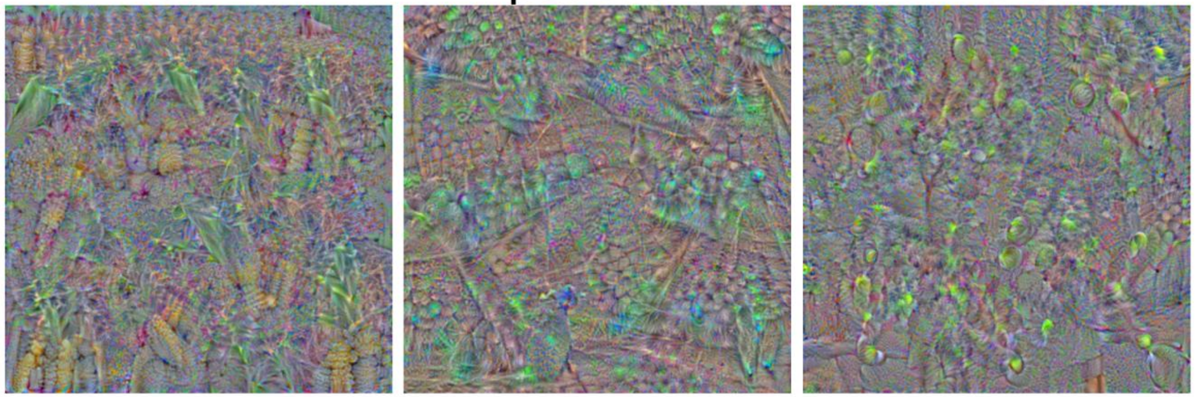
Res50 → Incv3: ~10% 😭

Future Work

Why transferable?

Semantic similarity and/or Model similarity

“corn” “peacock” “tennis ball”



Res50 → Dense121: ~70% 😄

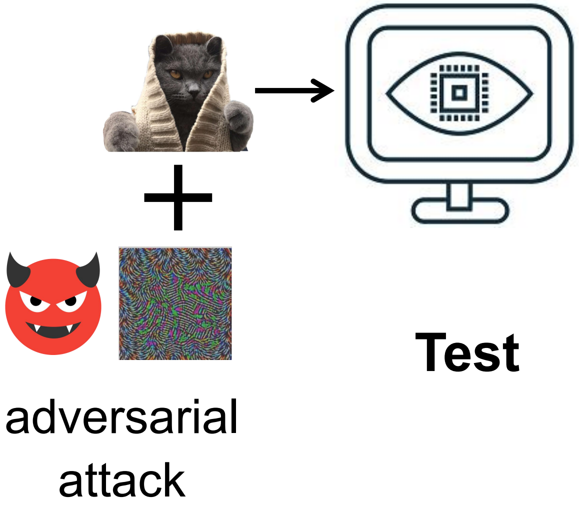
Res50 → Incv3: ~10% 😭

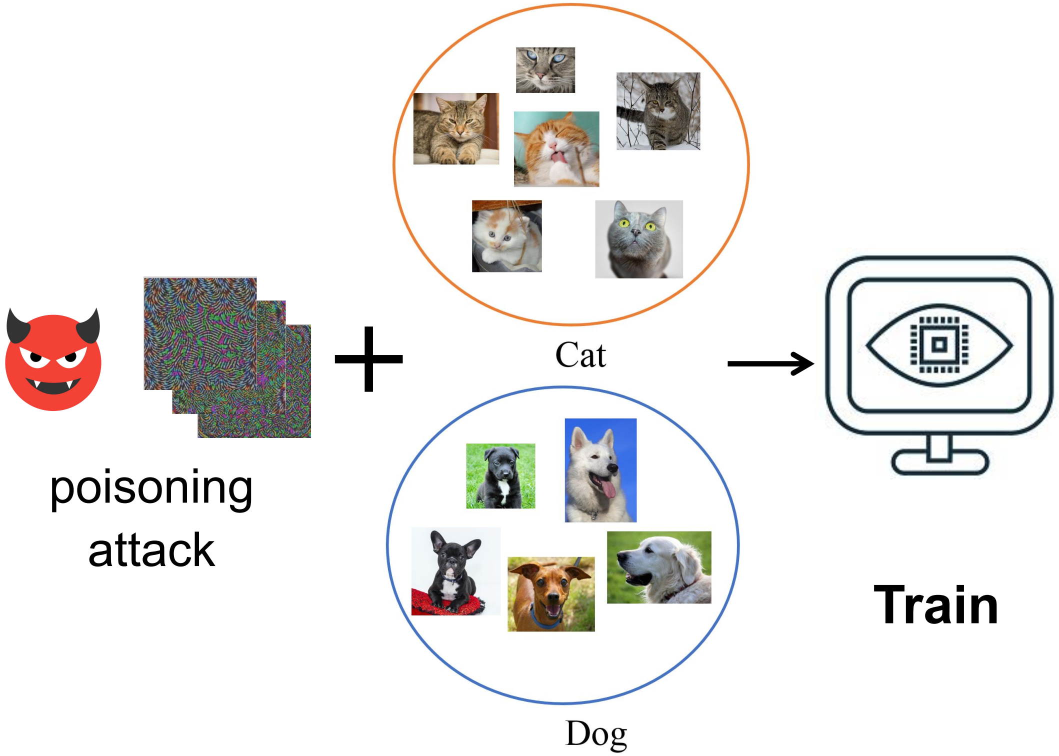
Zhao et al. *Towards Good Practices in Evaluating Transfer Adversarial Attacks*. arXiv 2022

https://github.com/ZhengyuZhao/TransferAttackEval

“We design good practices in evaluating transfer adversarial attacks. We systematically categorize 40+ recent attacks and comprehensively evaluate 23 representative ones against 9 defenses on ImageNet.”

Two projects

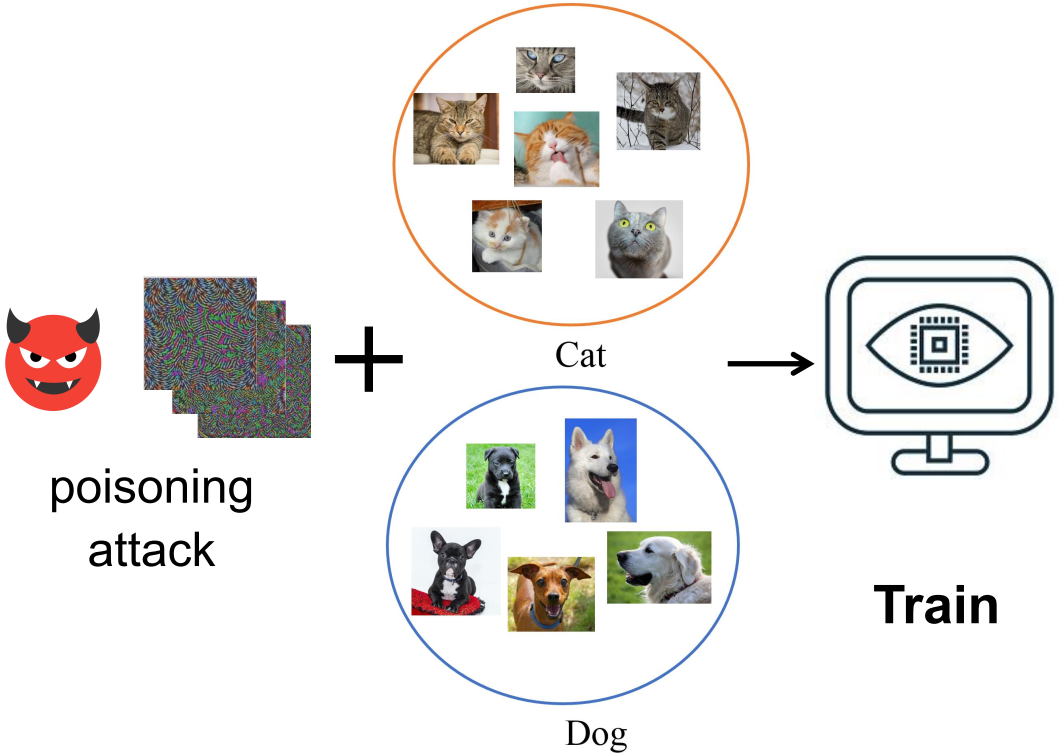




On Success and Simplicity: A Second Look at **Transferable Targeted Attacks**. NeurIPS 2021

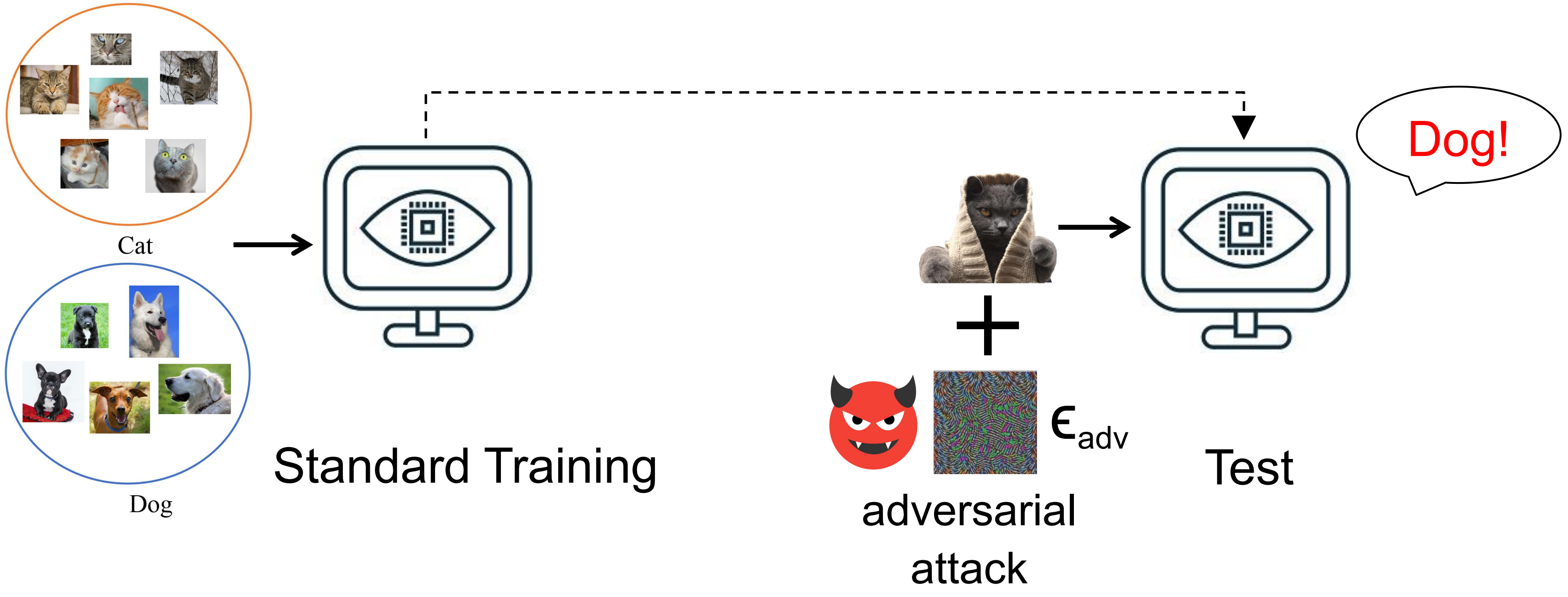
Data Poisoning against **Adversarial Training**. Under review

Project 2. Poisoning against Adversarial Training

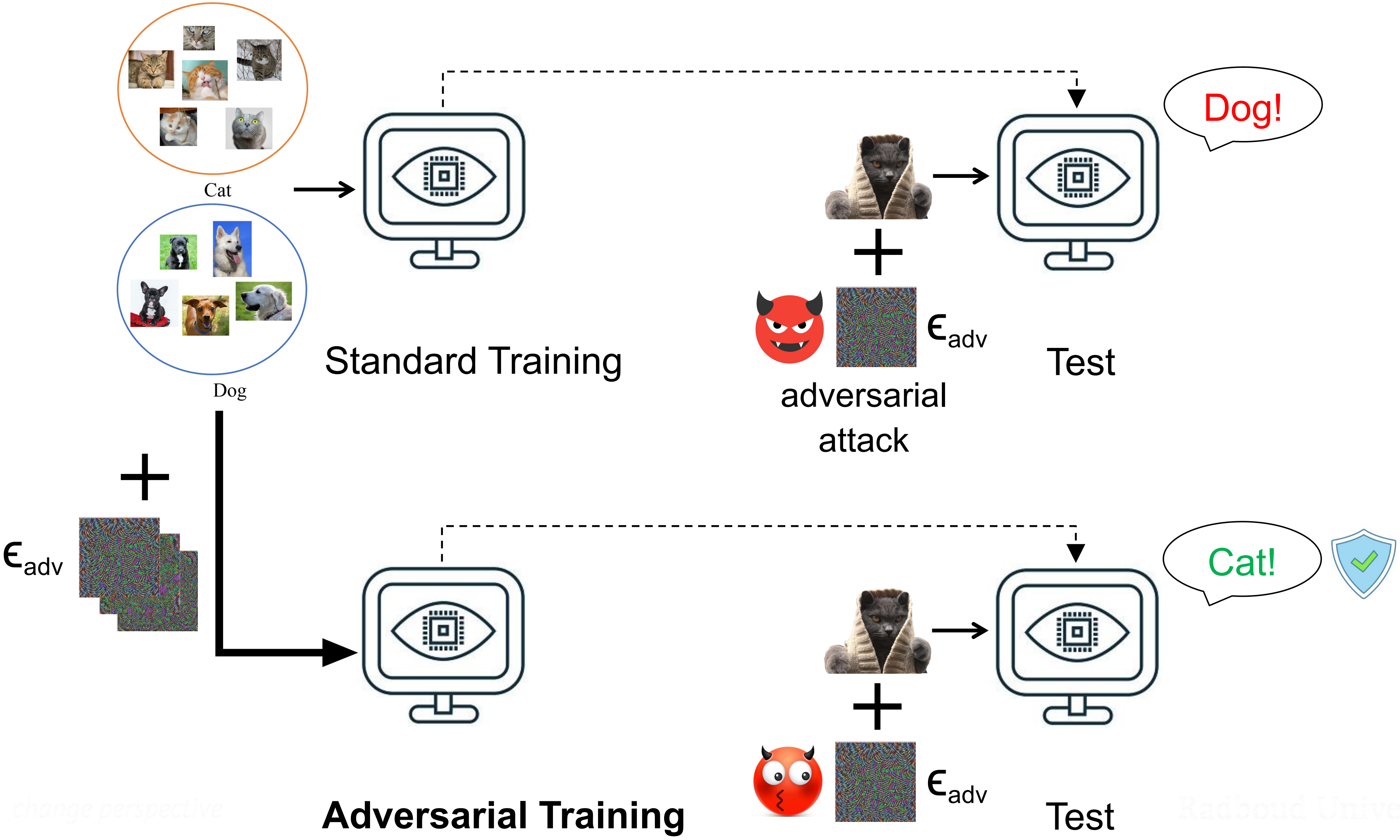


Data Poisoning against **Adversarial Training**.
Under review

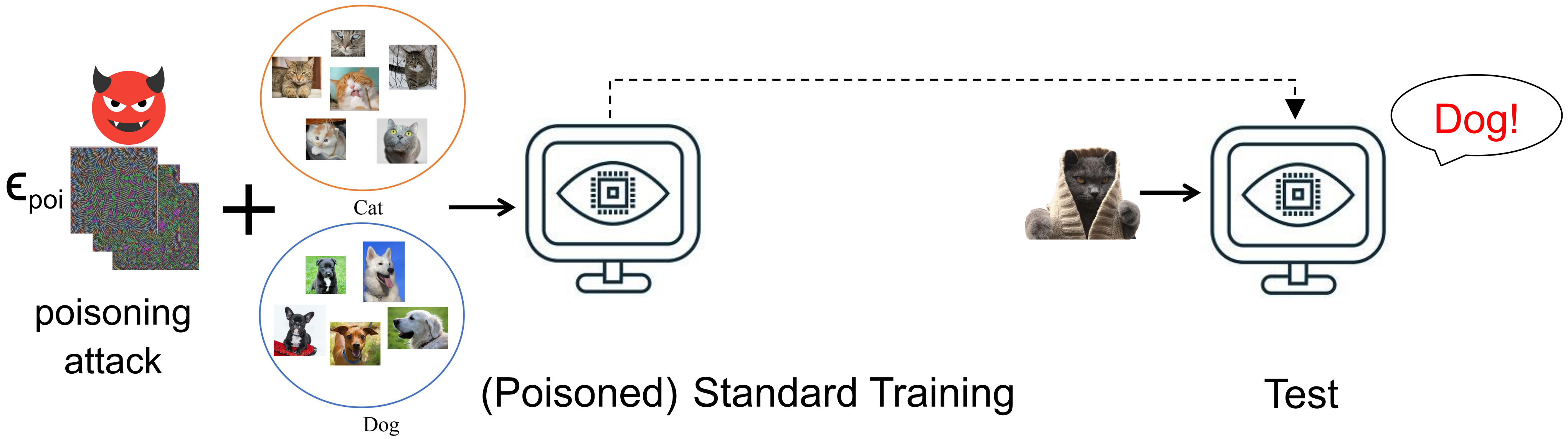
Adversarial Training for Adversarial attacks



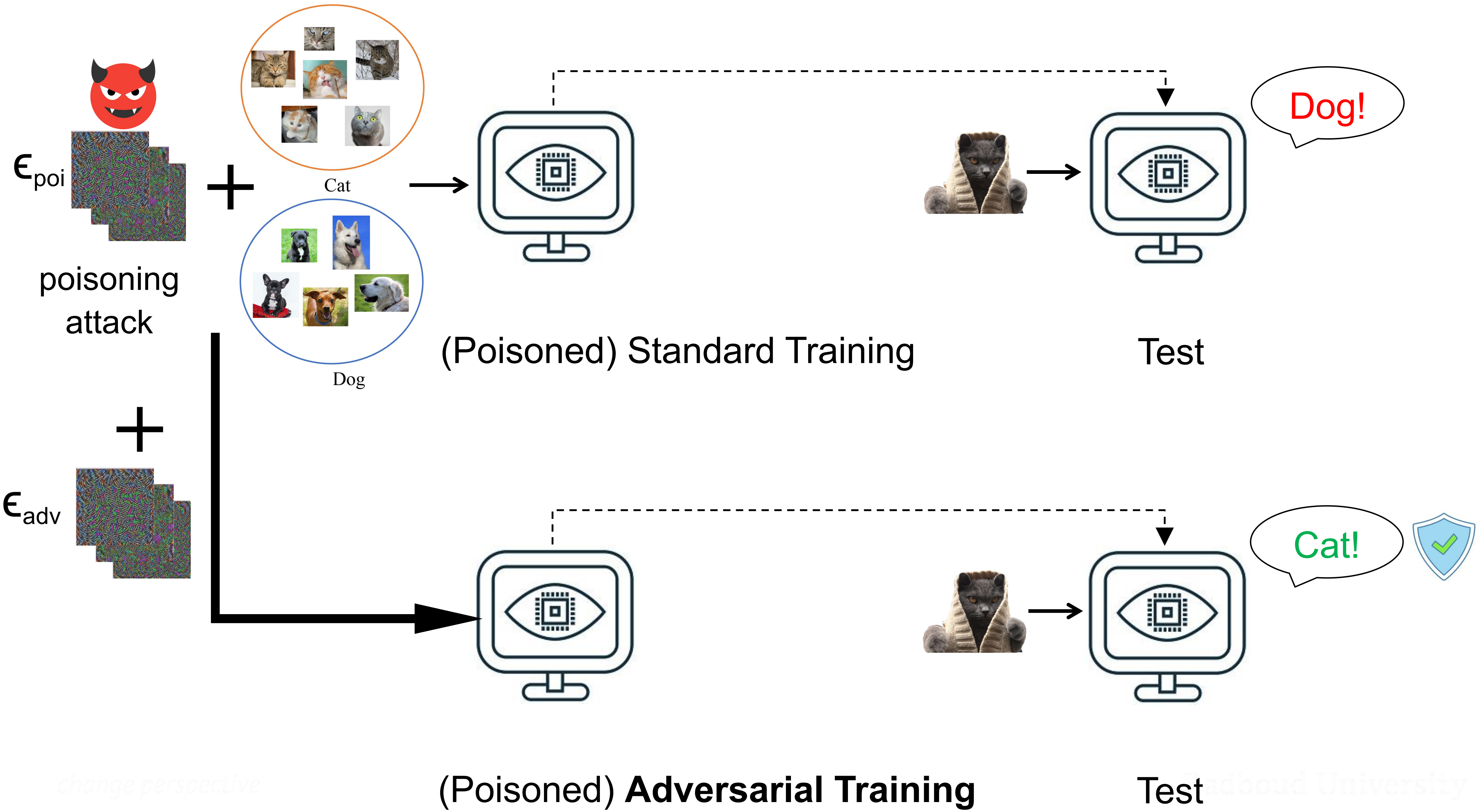
Adversarial Training for Adversarial attacks



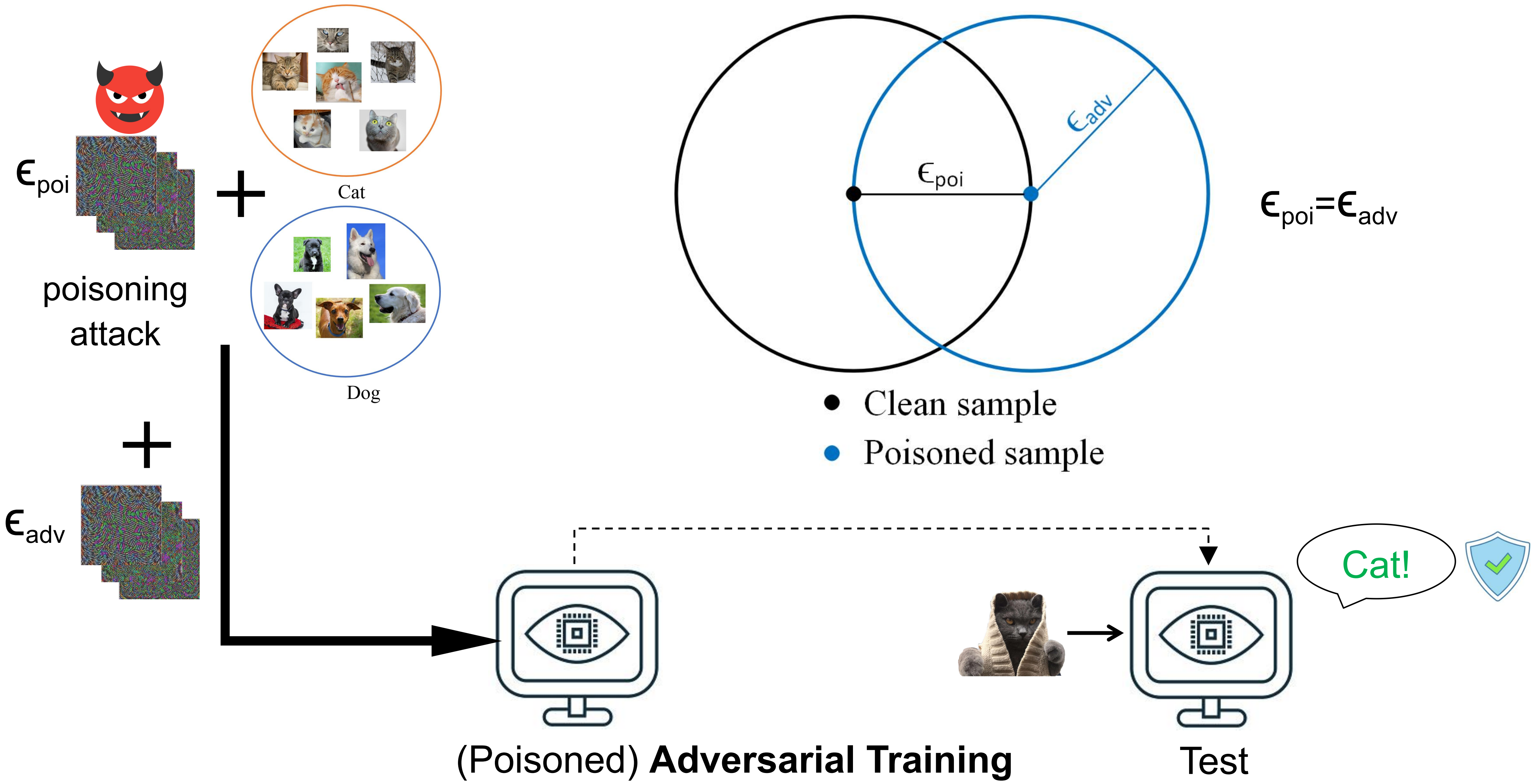
Adversarial Training for Poisoning Attacks



Adversarial Training for Poisoning Attacks



Adversarial Training for Poisoning Attacks



[1] Tao et al. *Better Safe Than Sorry: Preventing Delusive Adversaries with Adversarial Training*. NeurIPS 2021.

Consensus-Challenging Insights

Impossible to poison adversarially-trained models

Existing work[1-6]

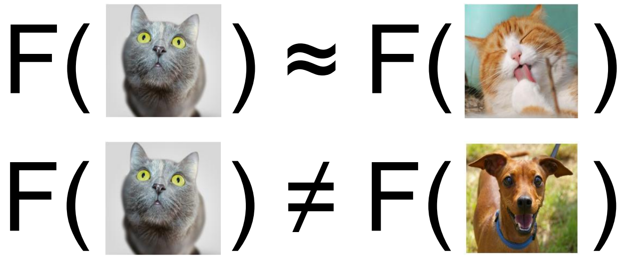
Possible (from a new attack perspective)

Ours

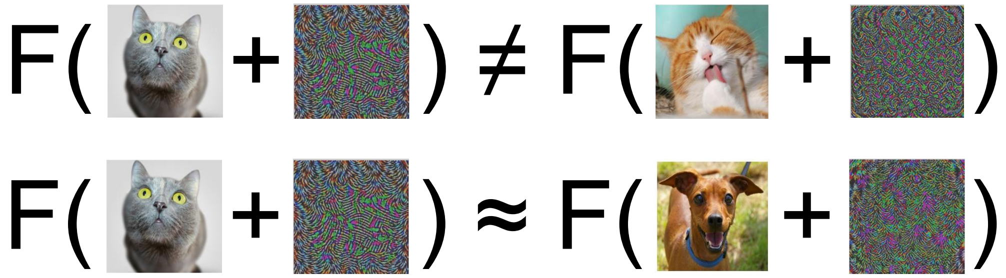
[1] Fowl et al. *Adversarial Examples Make Strong Poisons*. NeurIPS 2021.
[2] Huang et al. *Unlearnable Examples: Making Personal Data Unexploitable*. ICLR 2021.
[3] Tao et al. *Better Safe Than Sorry: Preventing Delusive Adversaries with Adversarial Training*. NeurIPS 2021.
[4] Wang et al. *Fooling Adversarial Training with Inducing Noise*. arXiv 2021.
[5] Fu et al. *Robust Unlearnable Examples: Protecting Data Against Adversarial Learning*. ICLR 2022.
[6] Tao et al. *Can Adversarial Training Be Manipulated By Non-Robust Features*? NeurIPS 2022.

New Attack Perspective

(Clean) adversarial/standard training



Inter-class entanglement (ours)

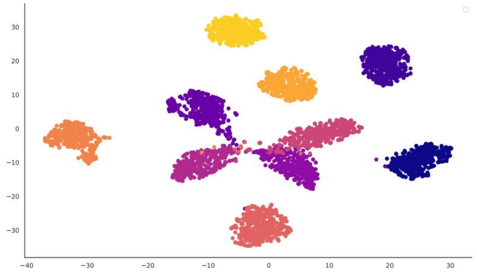


New Attack Perspective

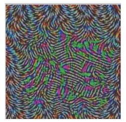
(Clean) adversarial/standard training

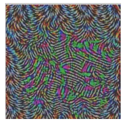
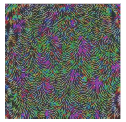
F() ≈ F()

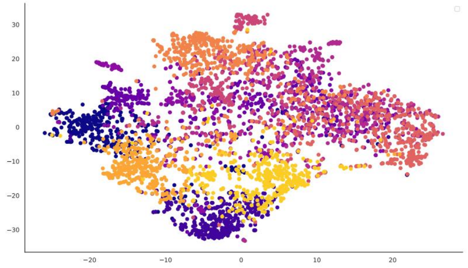
F() ≠ F()



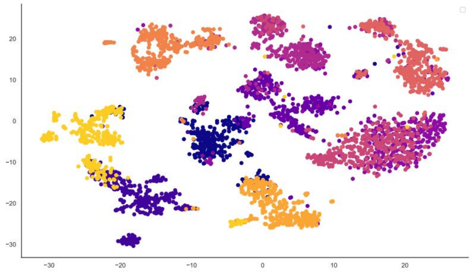
Inter-class entanglement (ours)

F( + ) ≠ F( + ) (a)

F( + ) ≈ F( + ) (b)



(a)



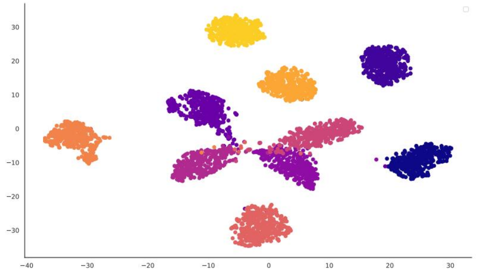
(b)

New Attack Perspective

(Clean) adversarial/standard training

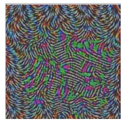
F() ≈ F()

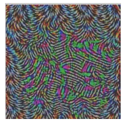
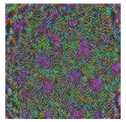
F() ≠ F()

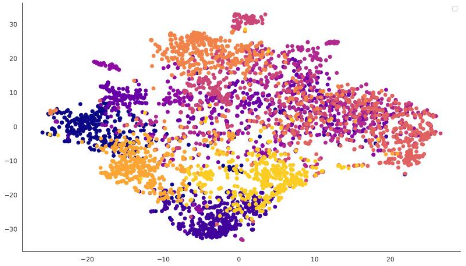


Test Acc: 84.88%

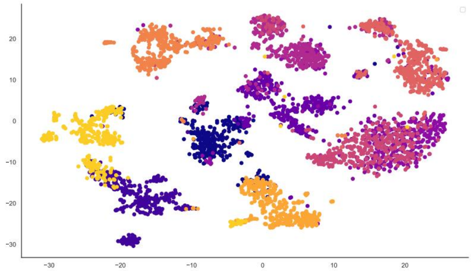
Inter-class entanglement (ours)

F( + ) ≠ F( + ) (a)

F( + ) ≈ F( + ) (b)



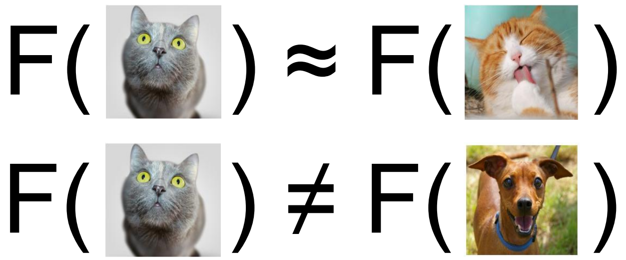
(a) Test Acc: 71.57% 😄

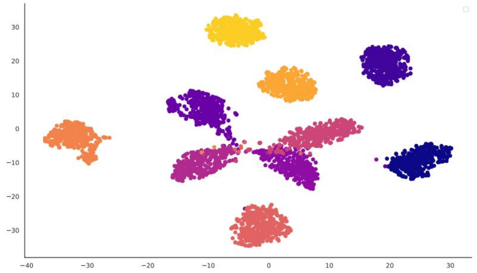


(b)Test Acc: 72.99% 😄

New Attack Perspective

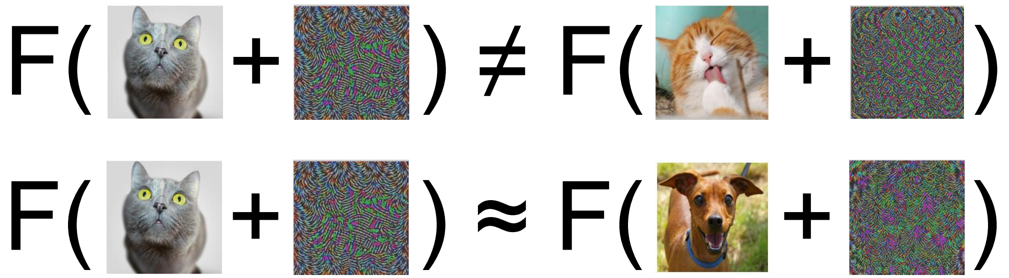
(Clean) adversarial/standard training

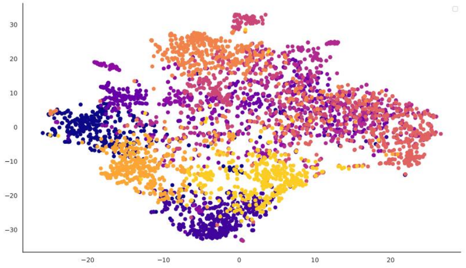




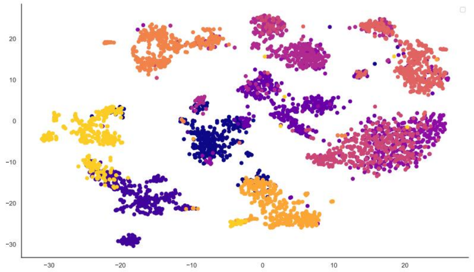
Test Acc: 84.88%

Inter-class entanglement (ours)





(a) Test Acc: 71.57% 😄



(b)Test Acc: 72.99% 😄

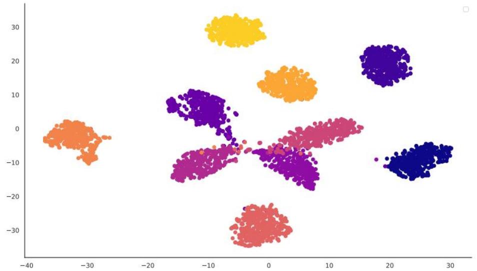
≈ discarding 83% training data!

New Attack Perspective

(Clean) adversarial/standard training

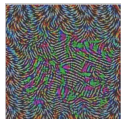
F() ≈ F()

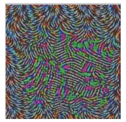
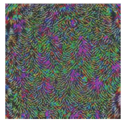
F() ≠ F()

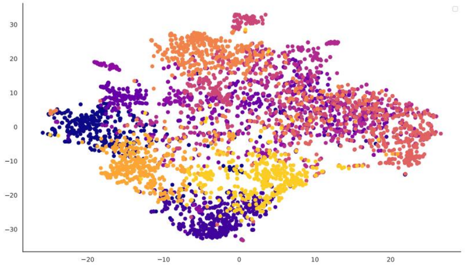


Test Acc: 84.88%

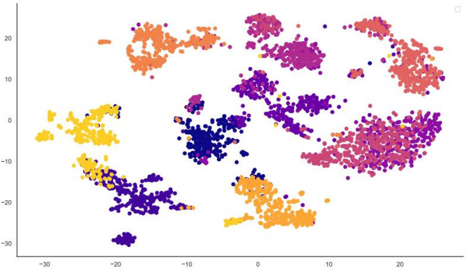
Inter-class entanglement (ours)

F( + ) ≠ F( + ) (a)

F( + ) ≈ F( + ) (b)

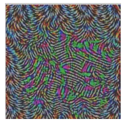


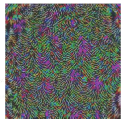
(a) Test Acc: 71.57% 😄



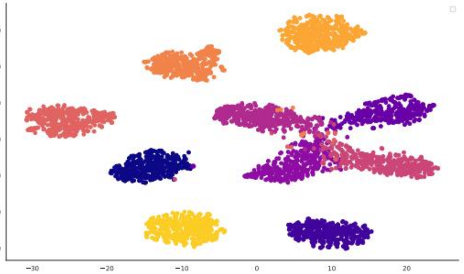
(b)Test Acc: 72.99% 😄

Whole-class swap (existing)

F( + ) ≈ F()

F( + ) ≈ F()

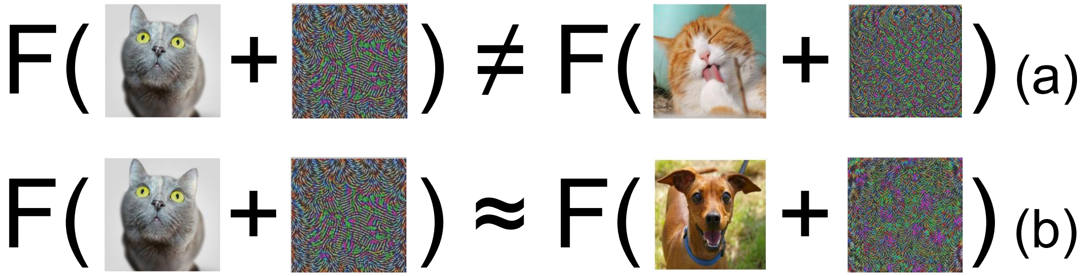
$$x' = \arg\min_{x} J(x, y_t)$$

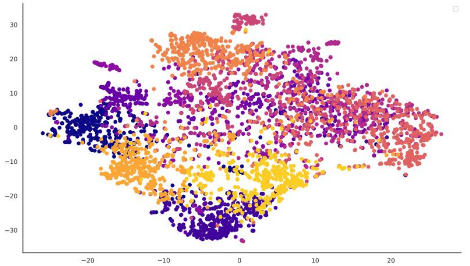


Test Acc: 83.11% 😭

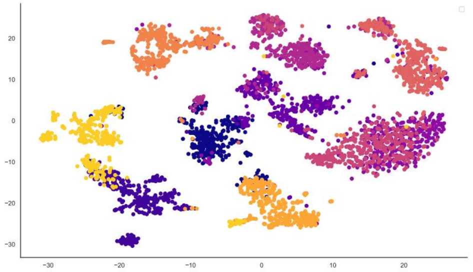
New Attack Perspective

Inter-class entanglement (ours)





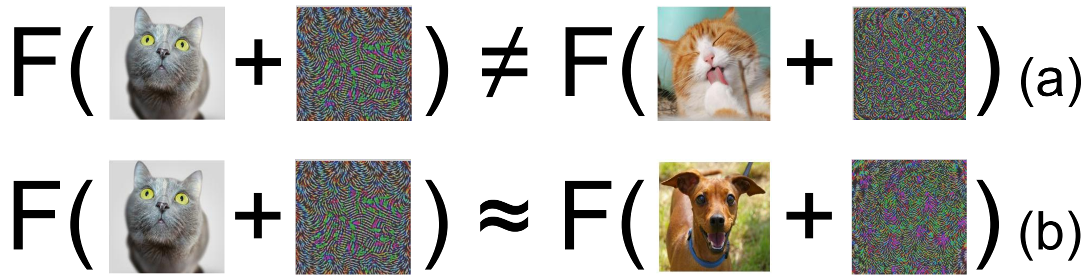
(a) Test Acc: 71.57%

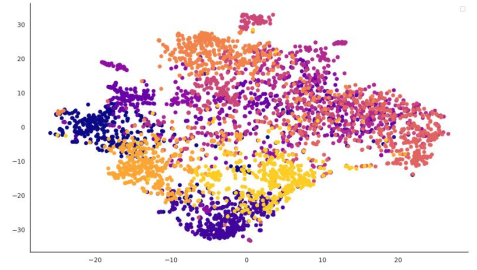


(b)Test Acc: 72.99%

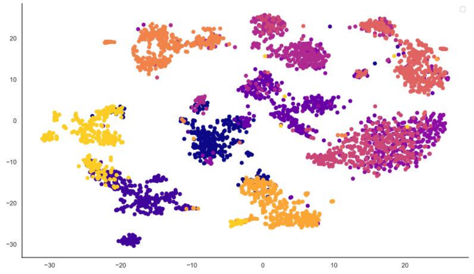
New Attack Perspective

Inter-class entanglement (ours)





(a) Test Acc: 71.57%



(b)Test Acc: 72.99%

$$\boldsymbol{\mu} = \frac{1}{|\mathcal{X}|} \sum_{\boldsymbol{x} \in \mathcal{X}} F^*_{L-1}(\boldsymbol{x})$$

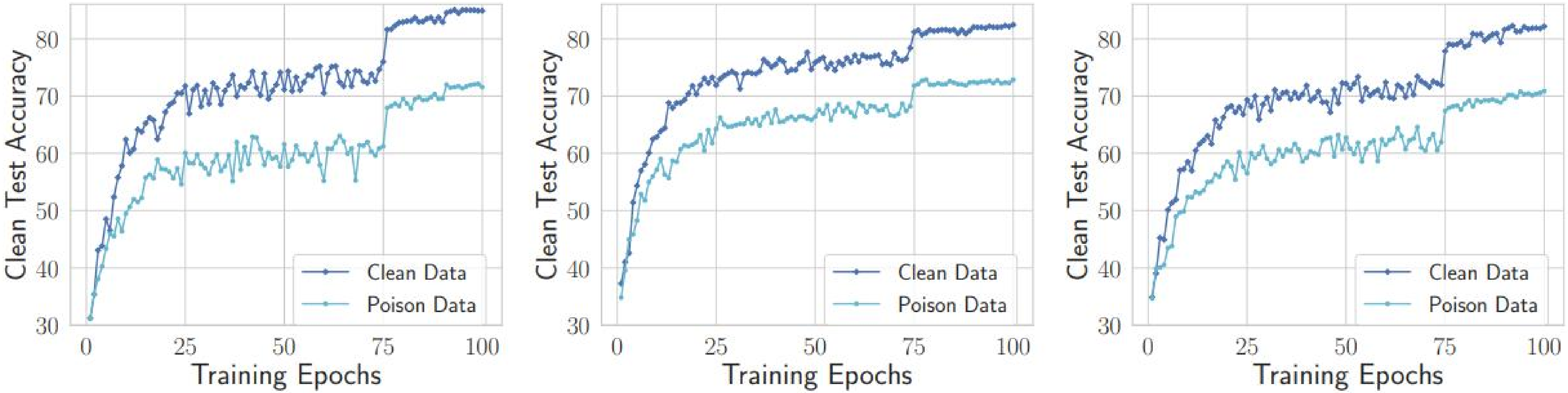
$$\mathcal{L}_{\text{push}} = \max_{\boldsymbol{\delta}^{\text{poi}}} \| F^*_{L-1}(\boldsymbol{x} + \boldsymbol{\delta}^{\text{poi}}) - \boldsymbol{\mu}_y \|_2 \quad \text{(a)}$$

$$\mathcal{L}_{\text{pull}} = \min_{\boldsymbol{\delta}^{\text{poi}}} \| F^*_{L-1}(\boldsymbol{x} + \boldsymbol{\delta}^{\text{poi}}) - \boldsymbol{\mu}_{y'} \|_2 \quad \text{(b)}$$

Results

Table 2: Evaluating INF on different datasets.

POISON METHOD \ DATASET	CIFAR-10	CIFAR-100	TINYIMAGENET
NONE (CLEAN)	84.88	59.50	51.95
INF (OURS)	71.57	47.29	41.32



(a) Madry (Madry et al., 2018) (b) TRADES (Zhang et al., 2019) (c) MART (Wang et al., 2020)

Figure 2: Evaluating INF against three different well-known adversarial training frameworks.

Results

Table 5: Transferability of INF poisons from ResNet-18 to other model architectures.

POISON METHOD \ TARGET	RESNET-18	RESNET-34	VGG-19	DENSENET-121	MOBILENETV2
NONE (CLEAN)	84.88	86.58	75.99	87.22	80.11
INF	71.57	73.05	64.66	74.35	67.21

Table 6: Evaluating INF against defenses that apply both data augmentations and AT.

DEFENSE	CLEAN TEST ACCURACY (%)
NONE (CLEAN)	84.88
ADVERSARIAL TRAINING	71.57
+RANDOM NOISE	71.88
+JPEG COMPRESSION	70.40
+MIXUP (ZHANG ET AL., 2018)	71.84
+CUTOUT (DEVRIES AND TAYLOR, 2017)	69.81
+CUTMIX (YUN ET AL., 2019)	68.85
+GRAYSCALE (LIU ET AL., 2021)	68.67

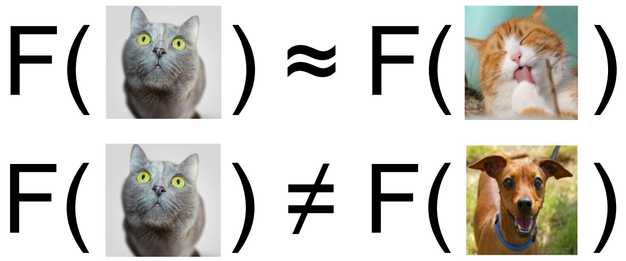
Other Results

- Poison only partial training data
- Adaptive defense to our attack strategy/algorithm
- Adaptive defense with adapted adversarial training

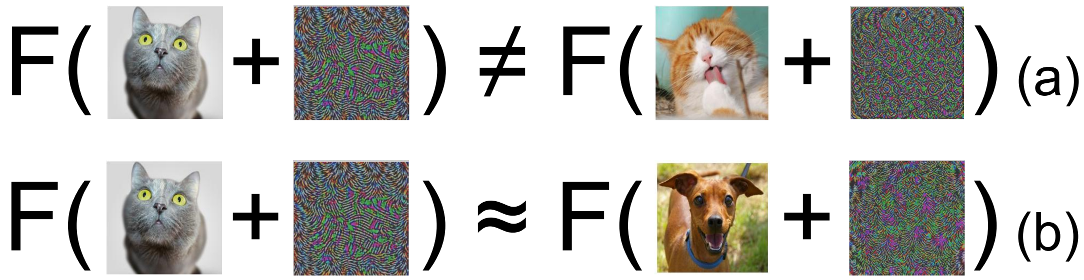
...

Standard Training (ST) vs. Adversarial Training (AT)

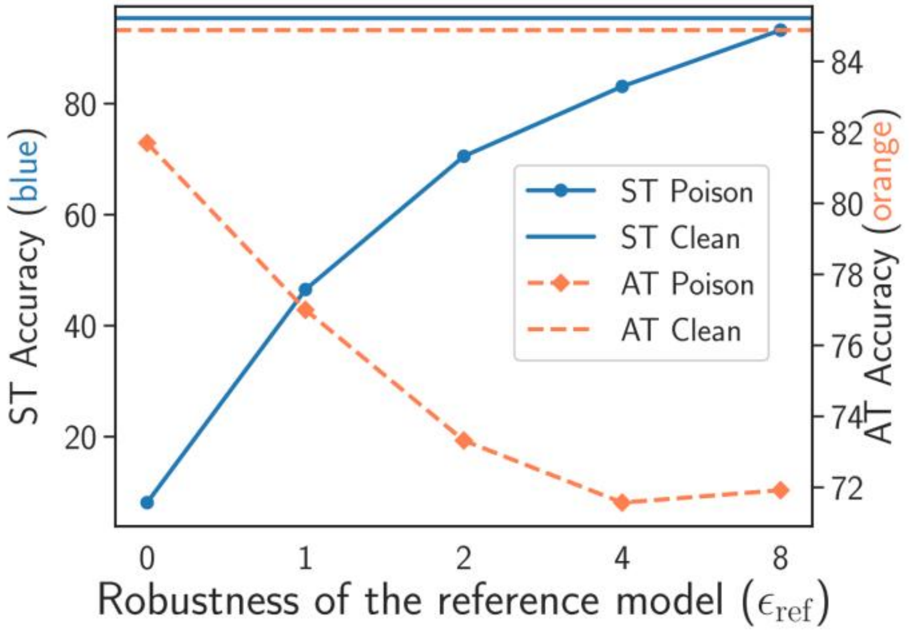
(Clean) **Adversarial/Standard** training



Inter-class entanglement (ours)

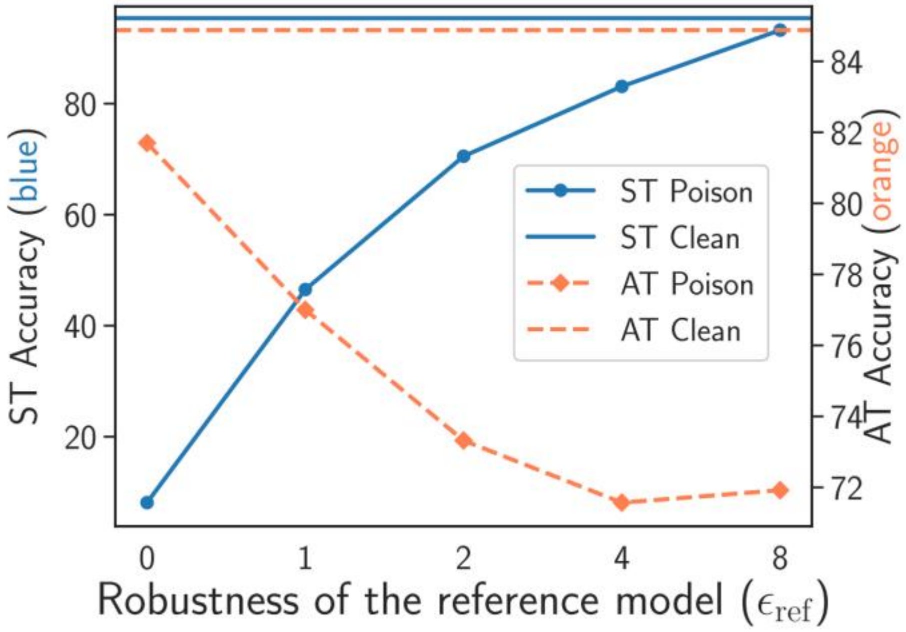


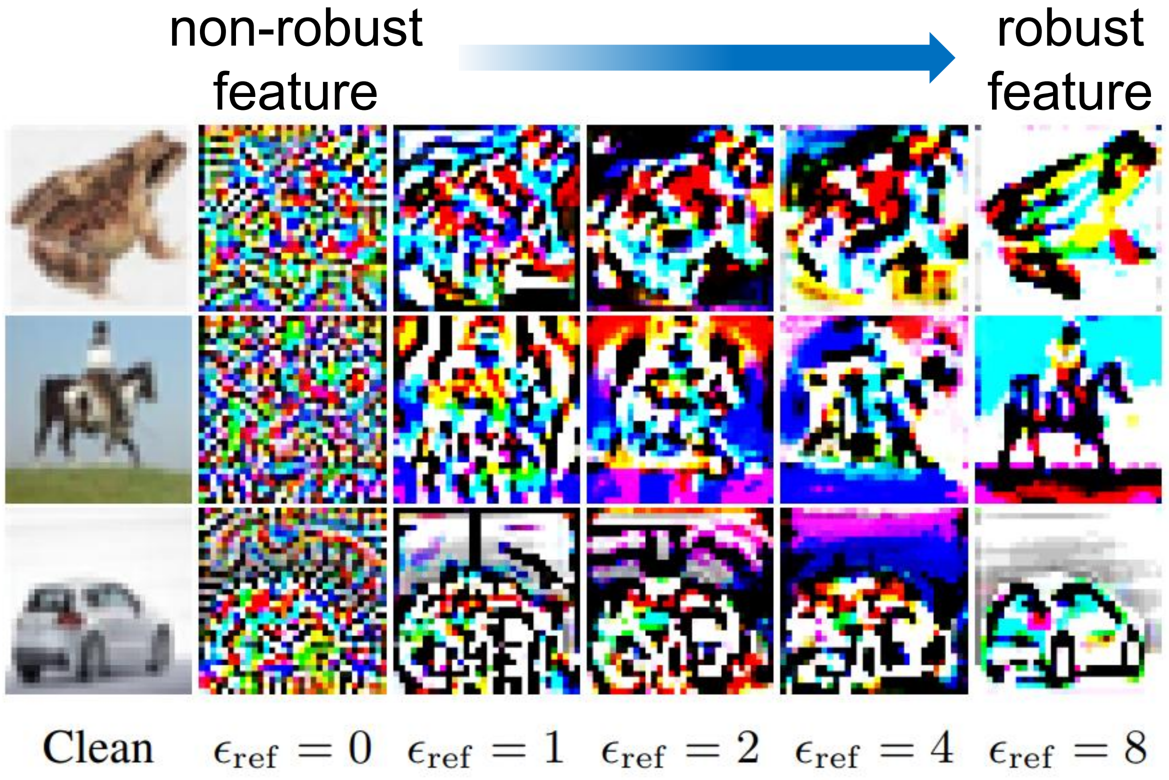
Standard Training (ST) vs. Adversarial Training (AT)



$$\mathcal{L}_{\text{push}} = \max_{\boldsymbol{\delta}^{\text{poi}}} \| F^*_{L-1}(\boldsymbol{x} + \boldsymbol{\delta}^{\text{poi}}) - \boldsymbol{\mu}_y \|_2$$

Standard Training (ST) vs. Adversarial Training (AT)





$$\mathcal{L}_{\text{push}} = \max_{\boldsymbol{\delta}^{\text{poi}}} \| F^*_{L-1}(\boldsymbol{x} + \boldsymbol{\delta}^{\text{poi}}) - \boldsymbol{\mu}_y \|_2$$

Hybrid Attack against Unknown Defense

$$\mathcal{L}_{\text{push}} = \max_{\boldsymbol{\delta}^{\text{poi}}} \|F^*_{L-1}(\boldsymbol{x} + \boldsymbol{\delta}^{\text{poi}}) - \boldsymbol{\mu}_y\|_2$$

$$\downarrow$$

$$\mathcal{L}_{\text{hybrid}} = \max_{\boldsymbol{\delta}^{\text{poi}}} \|F^*_{L-1,\text{ST}}(\boldsymbol{x} + \boldsymbol{\delta}^{\text{poi}}) - \boldsymbol{\mu}_{y,\text{ST}}\|_2 + \lambda\|F^*_{L-1,\text{AT}}(\boldsymbol{x} + \boldsymbol{\delta}^{\text{poi}}) - \boldsymbol{\mu}_{y,\text{AT}}\|_2$$

Hybrid Attack against Unknown Defense

$$\mathcal{L}_{\text{push}} = \max_{\boldsymbol{\delta}^{\text{poi}}} \|F^*_{L-1}(\boldsymbol{x} + \boldsymbol{\delta}^{\text{poi}}) - \boldsymbol{\mu}_y\|_2$$

$$\downarrow$$

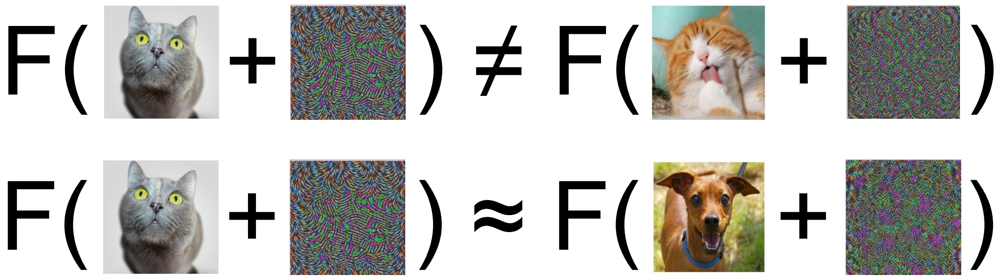
$$\mathcal{L}_{\text{hybrid}} = \max_{\boldsymbol{\delta}^{\text{poi}}} \|F^*_{L-1,\text{ST}}(\boldsymbol{x} + \boldsymbol{\delta}^{\text{poi}}) - \boldsymbol{\mu}_{y,\text{ST}}\|_2 + \lambda \|F^*_{L-1,\text{AT}}(\boldsymbol{x} + \boldsymbol{\delta}^{\text{poi}}) - \boldsymbol{\mu}_{y,\text{AT}}\|_2$$

METHOD ($\epsilon_{\text{poi}} = 8/255$) \ ϵ_{adv}	0/255	4/255	8/255	16/255	OPTIMAL TEST ACC.
NONE (CLEAN)	94.59	90.31	84.88	73.78	94.59
ADVPOISON	**9.91**	88.98	83.11	71.31	88.98
REM	25.59	**46.57**	84.21	85.76	85.76
ADVIN	77.31	90.08	86.76	72.16	90.08
UNLEARNABLE	25.69	90.47	84.91	79.81	90.47
HYPOCRITICAL	74.06	91.18	84.96	73.33	91.18
HYPOCRITICAL+	75.22	84.82	86.56	82.26	86.56
OURS	83.10	75.39	**71.51**	**63.73**	83.10
OURS (HYBRID)	12.93	76.55	74.30	65.75	**76.55**

Summary of Project 2

- Poisoning AT is possible based on a new attack perspective

Inter-class entanglement



- Robust features for poisoning AT, non-robust for ST
- Hybrid attack

Future Directions

- Possible defenses against our new attack
  - general: training techniques for entangled/noisy data?
  - specific: detecting/pre-filtering our attack?

- Better hybrid attack than $\mathcal{L}_{\text{hybrid}} = \max_{\boldsymbol{\delta}^{\text{poi}}} \|F^*_{L-1,\text{ST}}(\boldsymbol{x} + \boldsymbol{\delta}^{\text{poi}}) - \boldsymbol{\mu}_{y,\text{ST}}\|_2 + \lambda \|F^*_{L-1,\text{AT}}(\boldsymbol{x} + \boldsymbol{\delta}^{\text{poi}}) - \boldsymbol{\mu}_{y,\text{AT}}\|_2$
  - more effective
  - more efficient

Paper and code will be released in January!

Outline

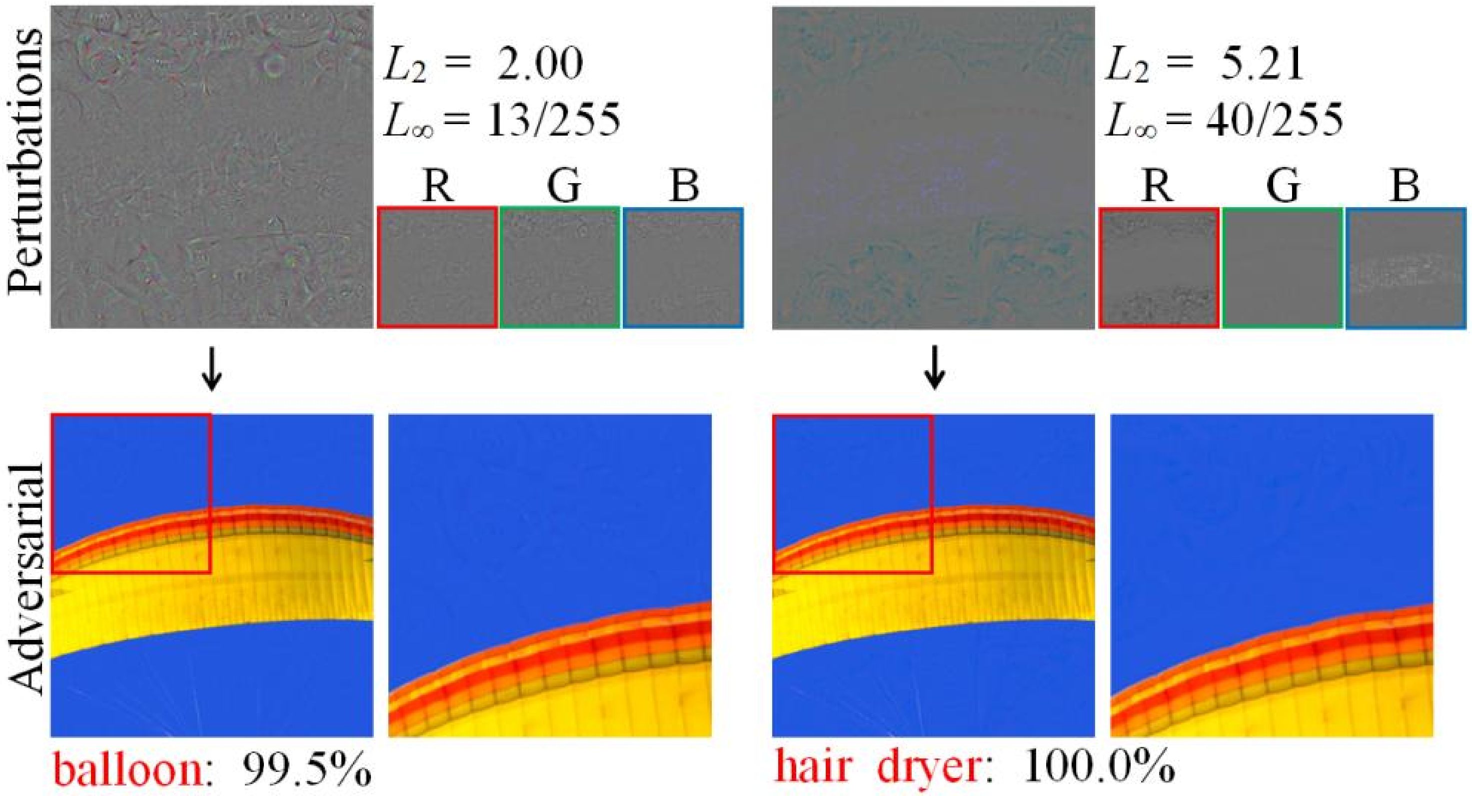
- Background of computer vision (CV) and adversarial images
- Two of our recent projects
- Other related projects

Imperceptible Perturbations

x'  $-$ x_{cat} 

$$\| x' - x_{\text{cat}} \|_\infty \leq \varepsilon \quad \longrightarrow \quad \| x' - x_{\text{cat}} \|_{\text{perc\_dist}} \leq \varepsilon$$

x'  $-$ x_{cat} 



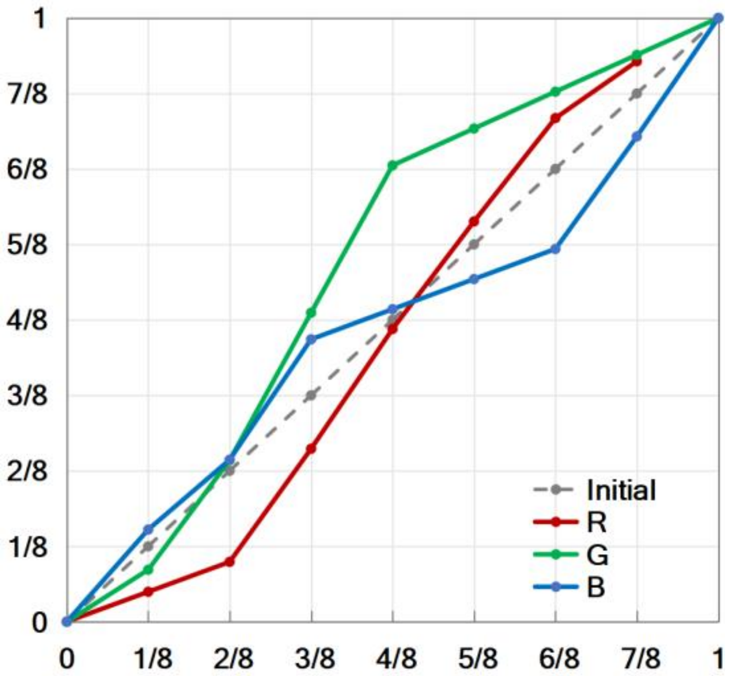
(a) C&W (b) PerC-C&W (ours)

Zhao et al. *Towards Large yet Imperceptible Adversarial Image Perturbations with Perceptual Color Distance*. CVPR 2020.

Perceptible yet Stealthy Perturbations

x

$\times$



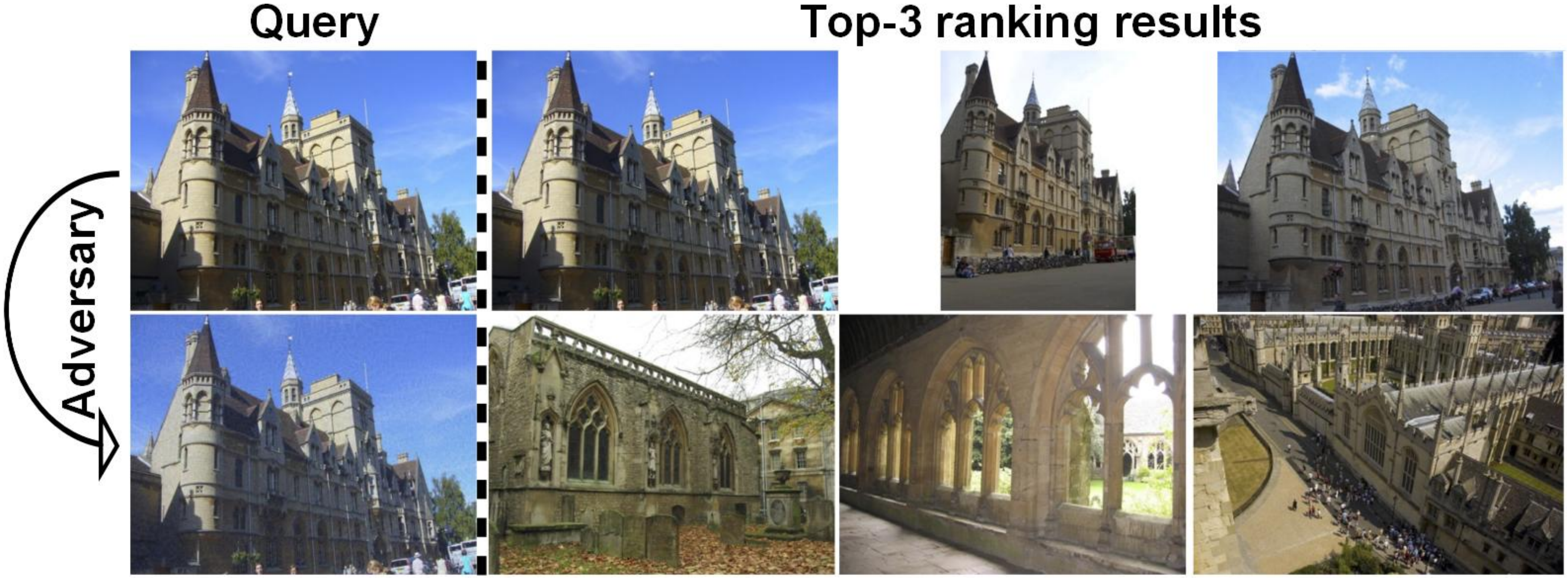
color curve adjust

$=$

x'

Zhao et al. *Adversarial Image Color Transformations in Explicit Color Filter Space*. Under review by IEEE TIFS. Preliminary version at BMVC 2020.

Adversarial attacks on Image Retrieval



Liu et al. *Who's Afraid of Adversarial Queries? The Impact of Image Modifications on Content-based Image Retrieval*. ICMR 2019

Thank you!

Zhengyu Zhao (赵正宇)

zhengyuzhao.github.io zhengyu.zhao@cispa.de

CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY

Radboud University

Research Interests:

Security (e.g. adversarial example and data poisoning) **and Privacy** (e.g. membership inference) **risks of Machine Learning/Computer Vision**.