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 focus:

Analyzing the vulnerability of deep neural networks to various attacks, e.g., (test-time) adversarial examples and (training-time) data poisons.

Failures of Computer Vision in Adversarial Scenarios

03/03/2023

Outline

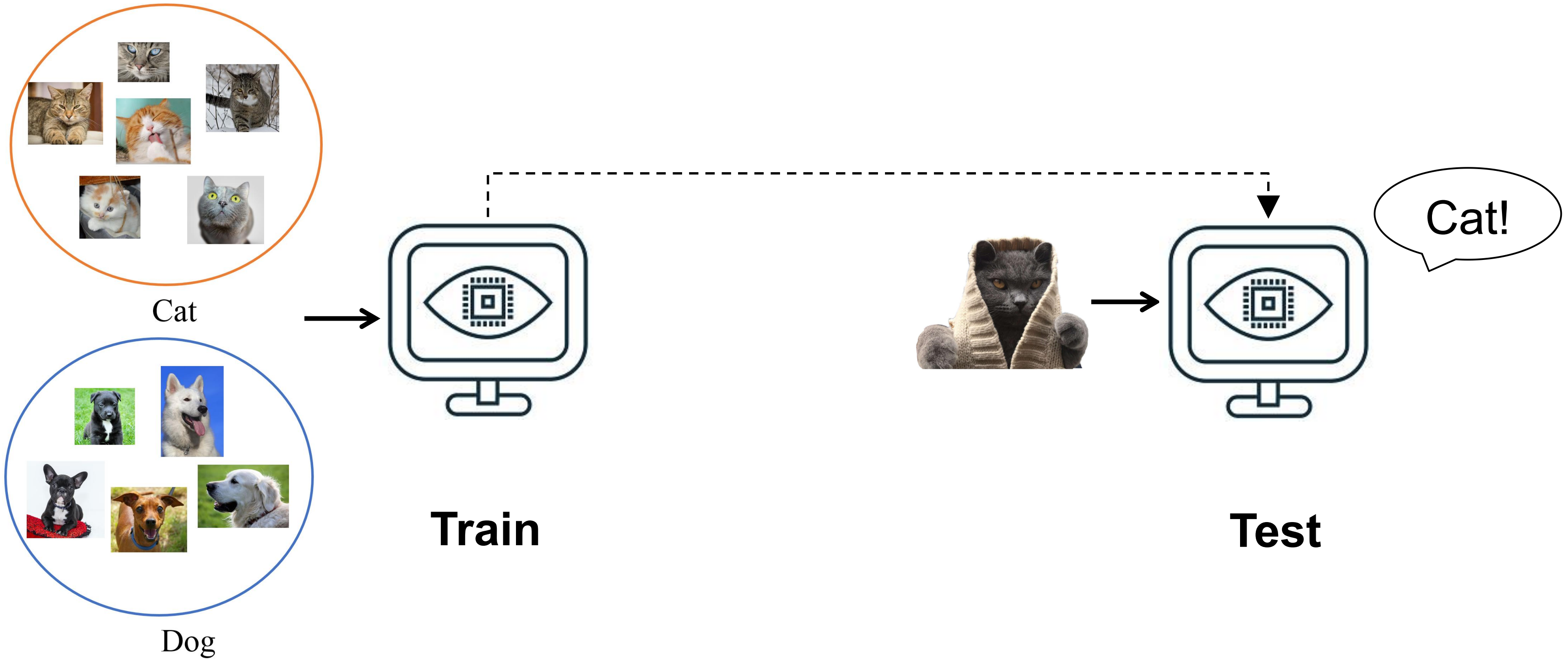
- Overview of adversarial images in computer vision
- Two recent projects
- Other related projects

Outline

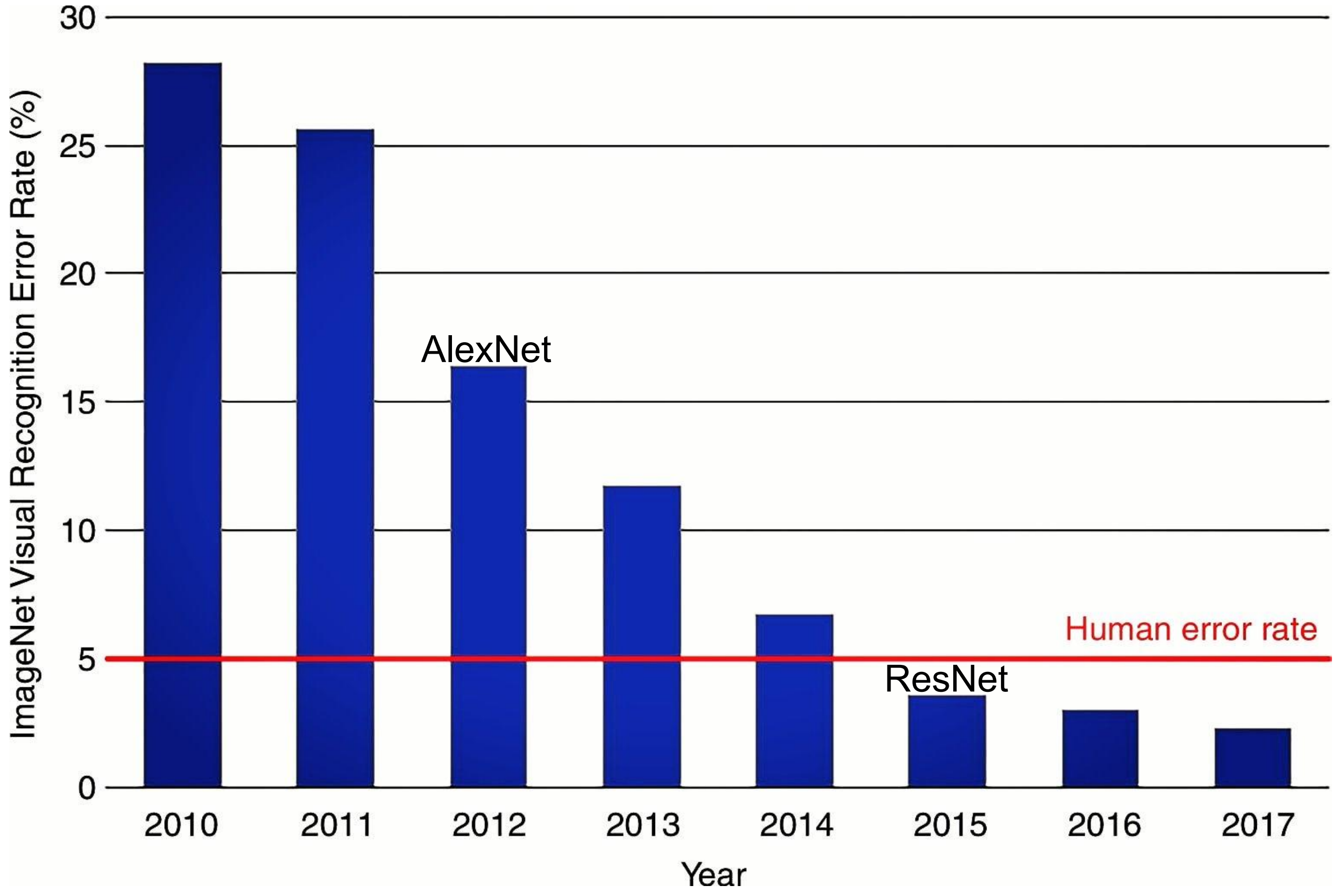
- Overview of adversarial images in computer vision
- Two recent projects
- Other related projects

Computer Vision (CV)

Working pipeline of CV

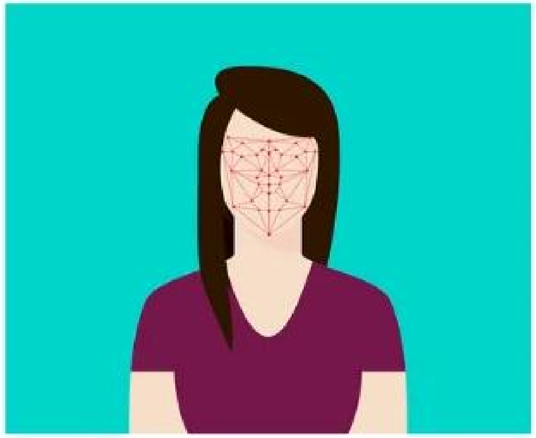


Success of CV



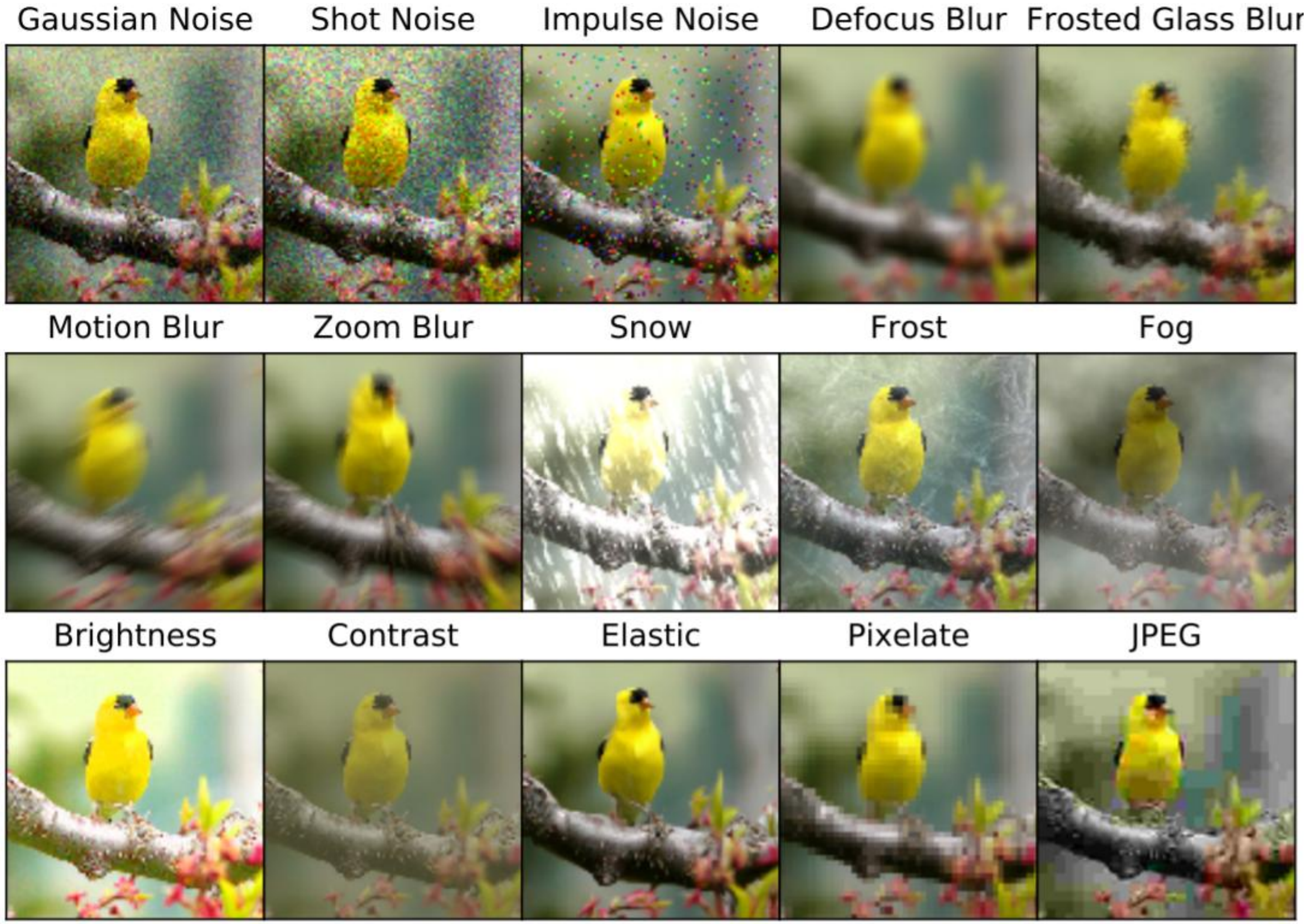
credit: https://www.synopsys.com/designware-ip/technical-bulletin/computer-vision-lab-life.html

Success of CV



Google Lens

Failure of CV (against Real-world Perturbations)



Failure of CV (against Real-world Perturbations)

face recognition[1]

self-driving car[2]

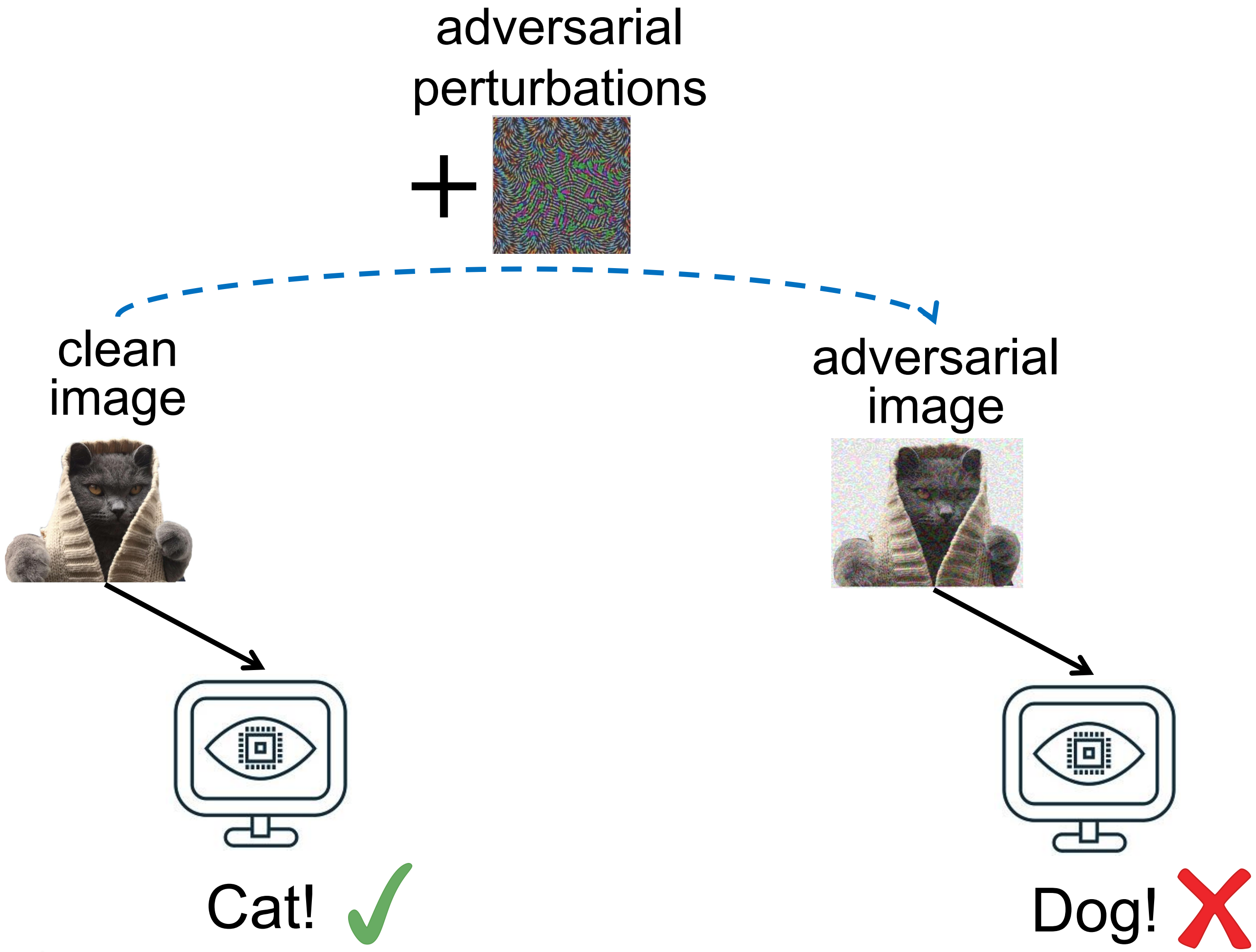
[1] https://ipvm.com/reports/face-masks
[2] https://www.theguardian.com/technology/2018/mar/22/video-released-of-uber-self-driving-crash-that-killed-woman-in-arizona

average-case (real-world) Image perturbations?

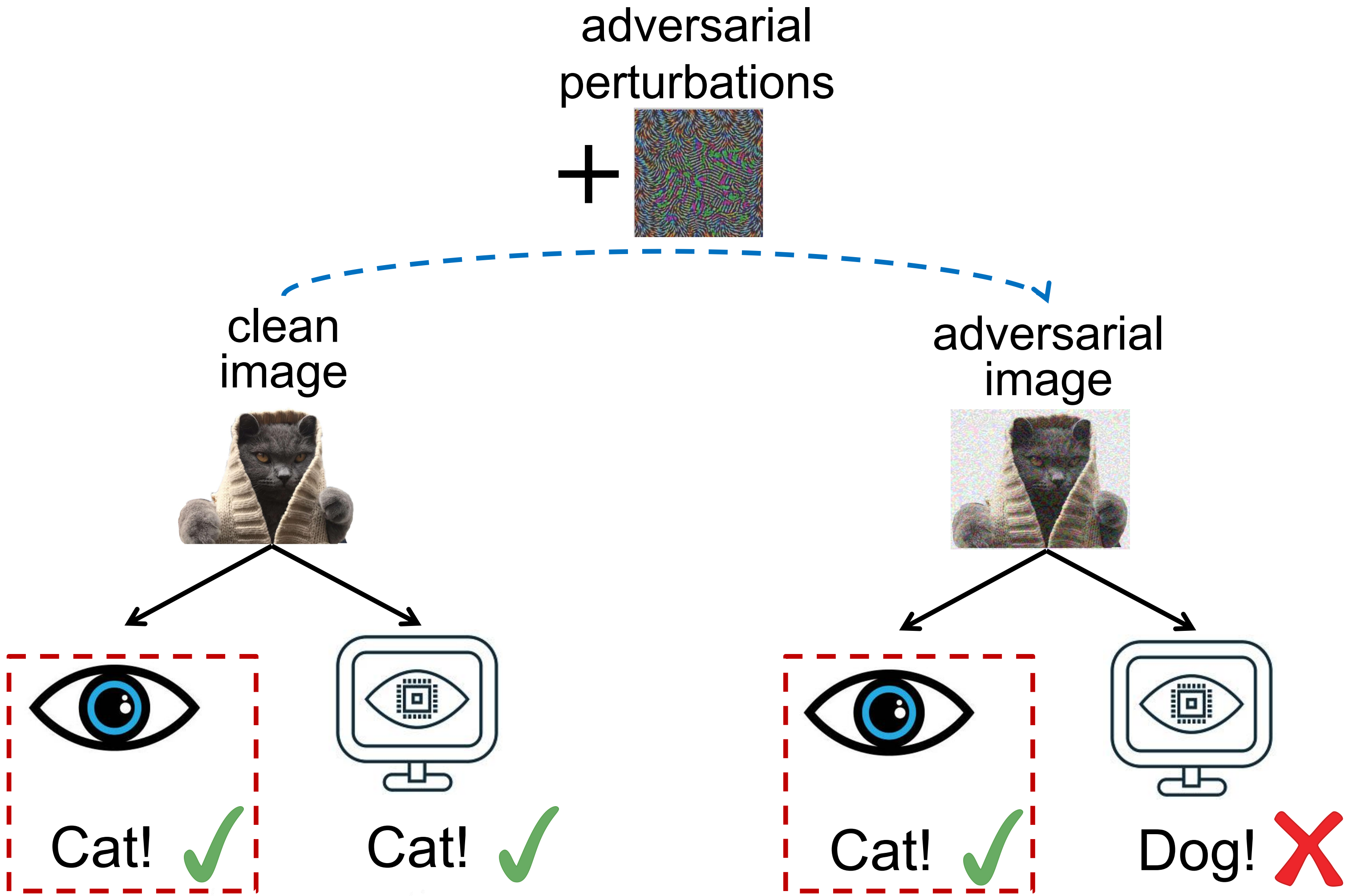
average-case (real-world) Image perturbations?

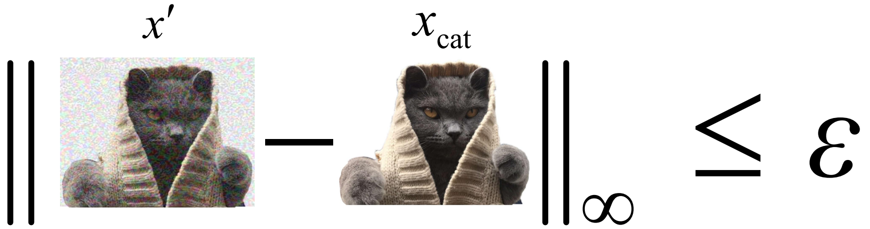
worst-case (adversarial) Image perturbations!

Formalize Adversarial Image Perturbations



Stealthy Attacks with Imperceptible Perturbations





Real-world → Adversarial Image Perturbations

face recognition[1]

adversarial mask[2]

[1] https://ipvm.com/reports/face-masks
[2] https://towardsdatascience.com/fooling-facial-detection-with-fashion-d668ed919eb

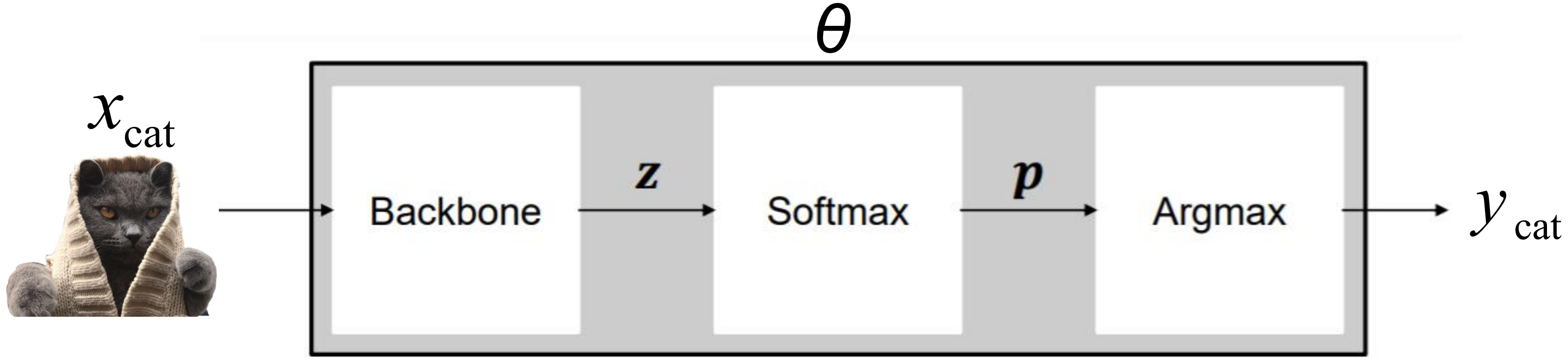
Real-world → Adversarial Image Perturbations

self-driving car[1]

adversarial graffiti[2]

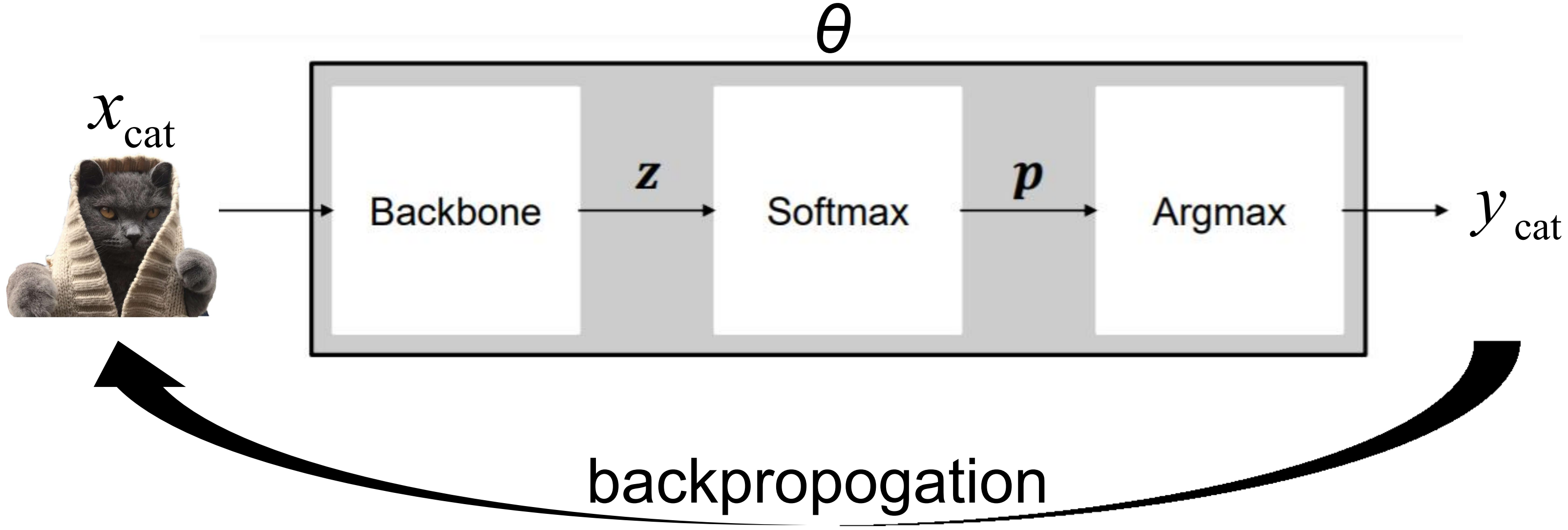
[1] https://www.theguardian.com/technology/2018/mar/22/video-released-of-uber-self-driving-crash-that-killed-woman-in-arizona
[2] Eykholt et al. *Robust physical-world attacks on deep learning visual classification*. CVPR 2018.

Optimize Adversarial Images x’



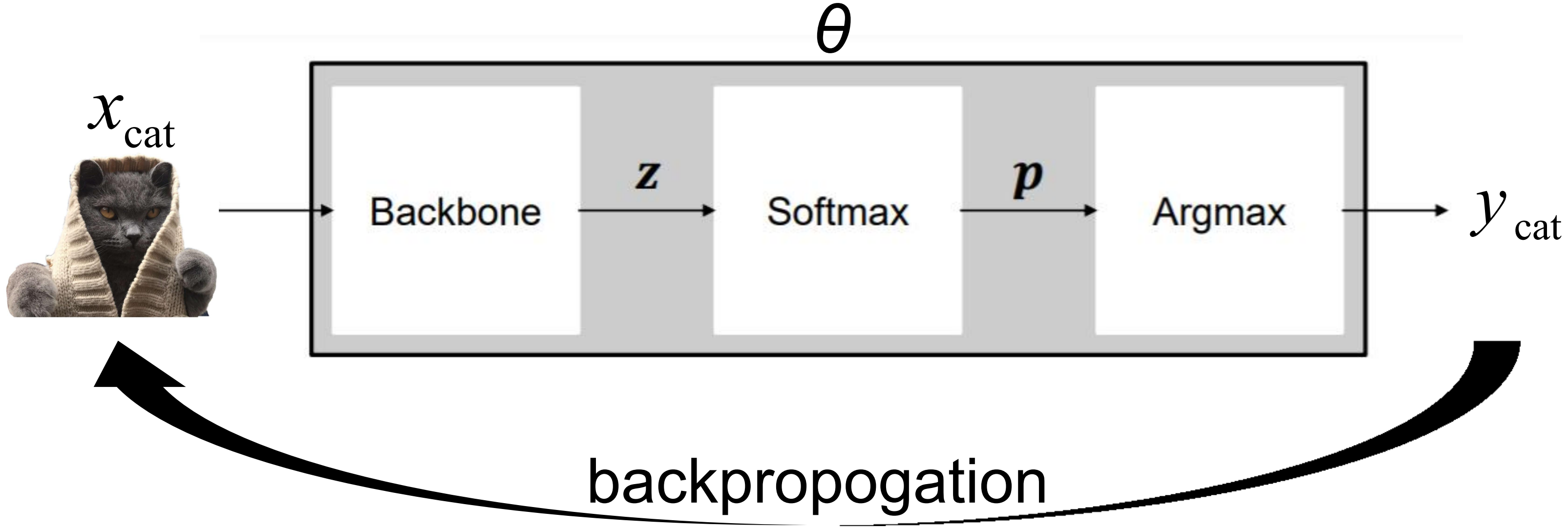
$$\theta' = \arg\min_{\theta} J(\theta, x_{\text{cat}}, y_{\text{cat}})$$

Optimize 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

Optimize 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

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

Optimize 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]

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

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

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

Recap

Success of computer vision

↳ Failures against real-world perturbations

↳ ... adversarial images

↳ optimize adversarial images

Outline

- Overview of adversarial images in computer vision
- Two recent projects
- Other related projects

Consensus-Challenging Insights

Mission:
(almost) impossible

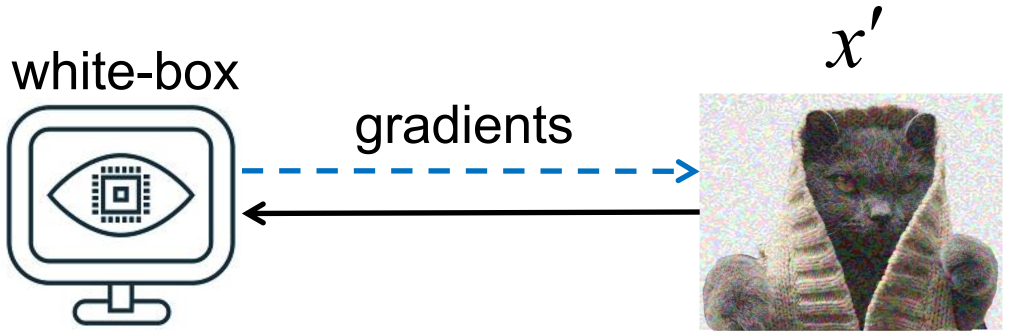
Existing work[1-N]

Mission:
possible (and not hard)

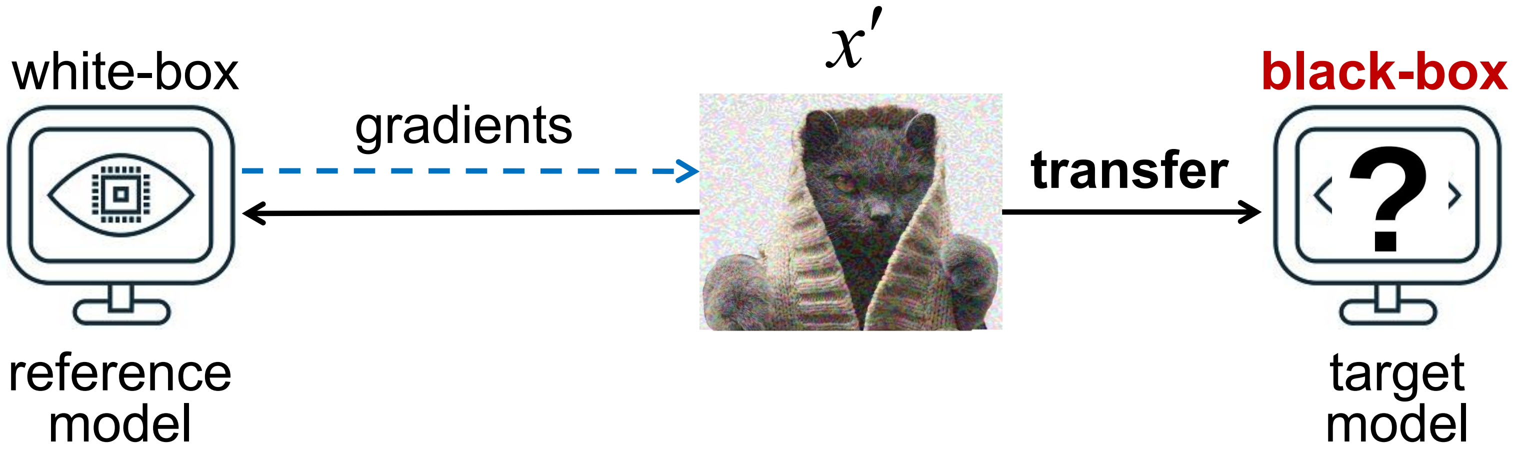
Ours

Project 1. Transferable Targeted Attacks

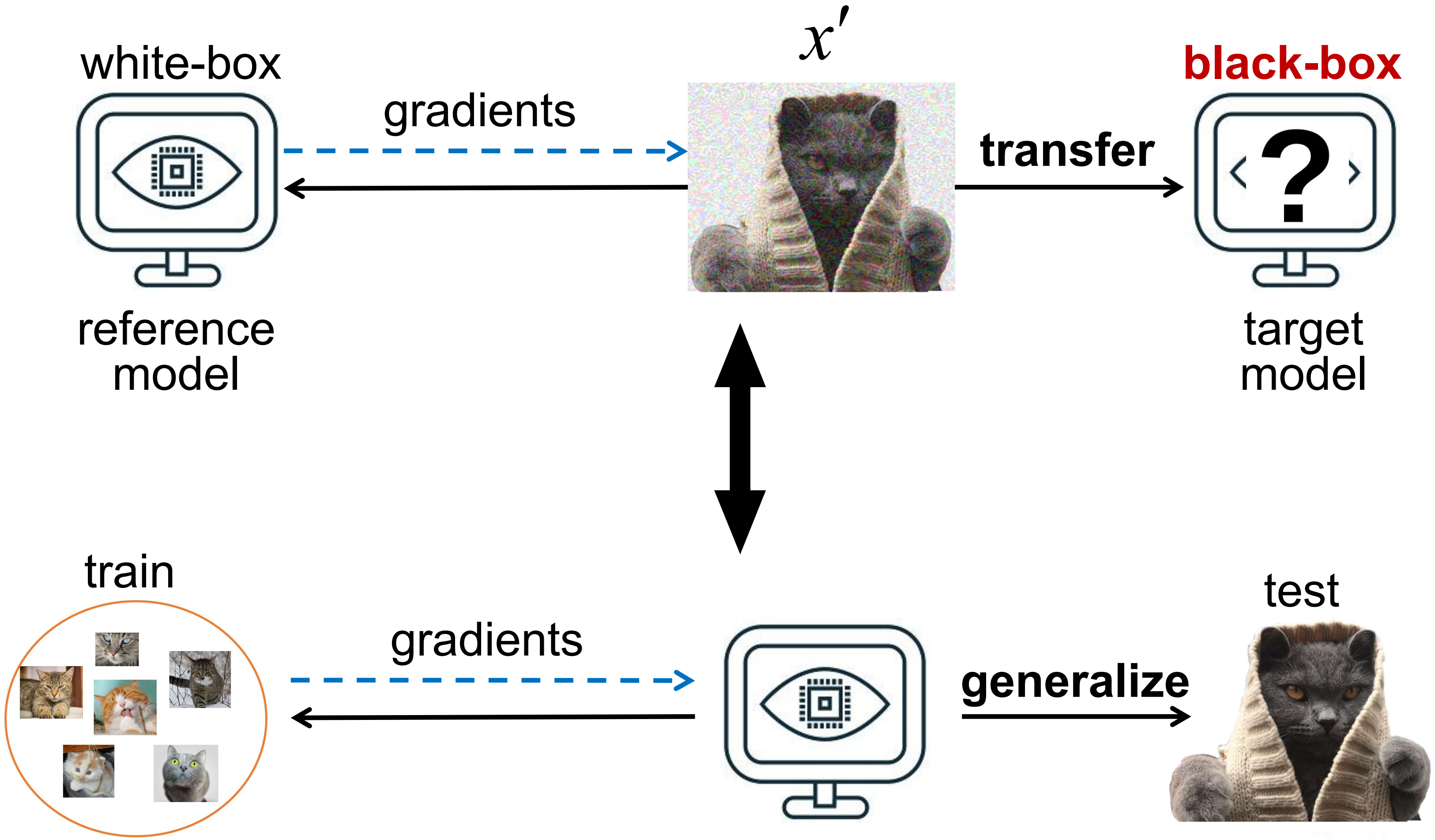
Transferable Targeted Attacks



Transferable Targeted Attacks



Transferable Targeted Attacks



Transfer Techniques

- Gradient stabilization

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

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

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

- Data augmentation

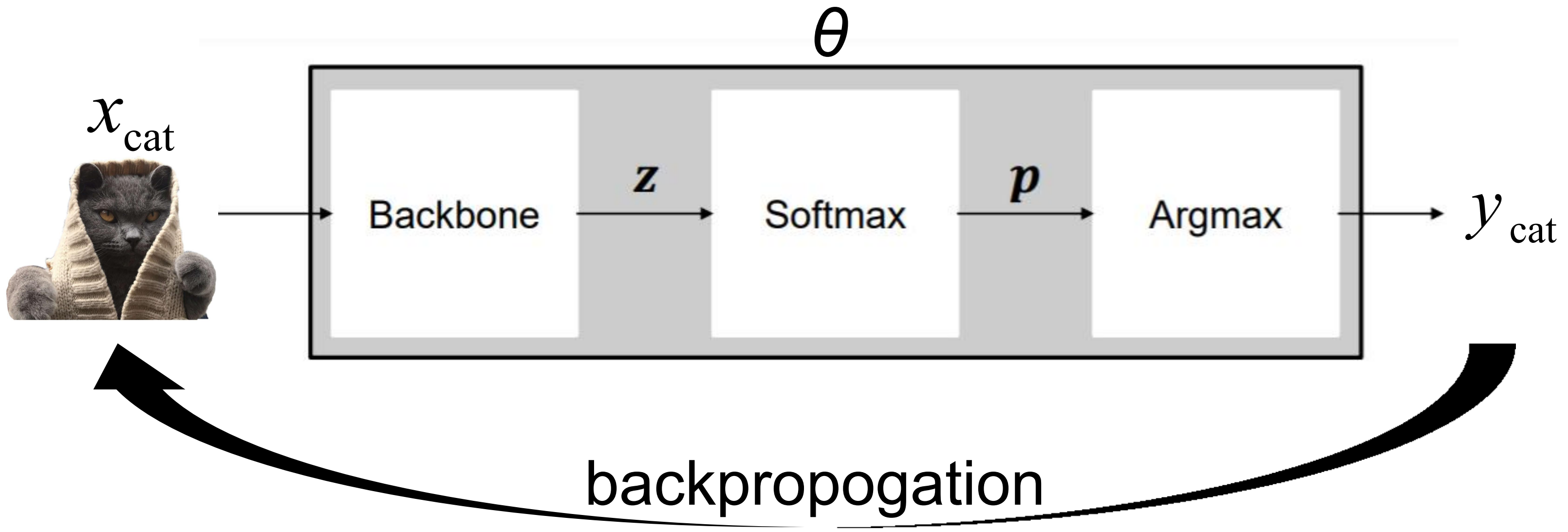
  e.g., resizing & padding (DI-FGSM)[2]

  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.

Transferable **Targeted** Attacks



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

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

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

Consensus-Challenging Insight

Impossible for I-FGSM to achieve 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.

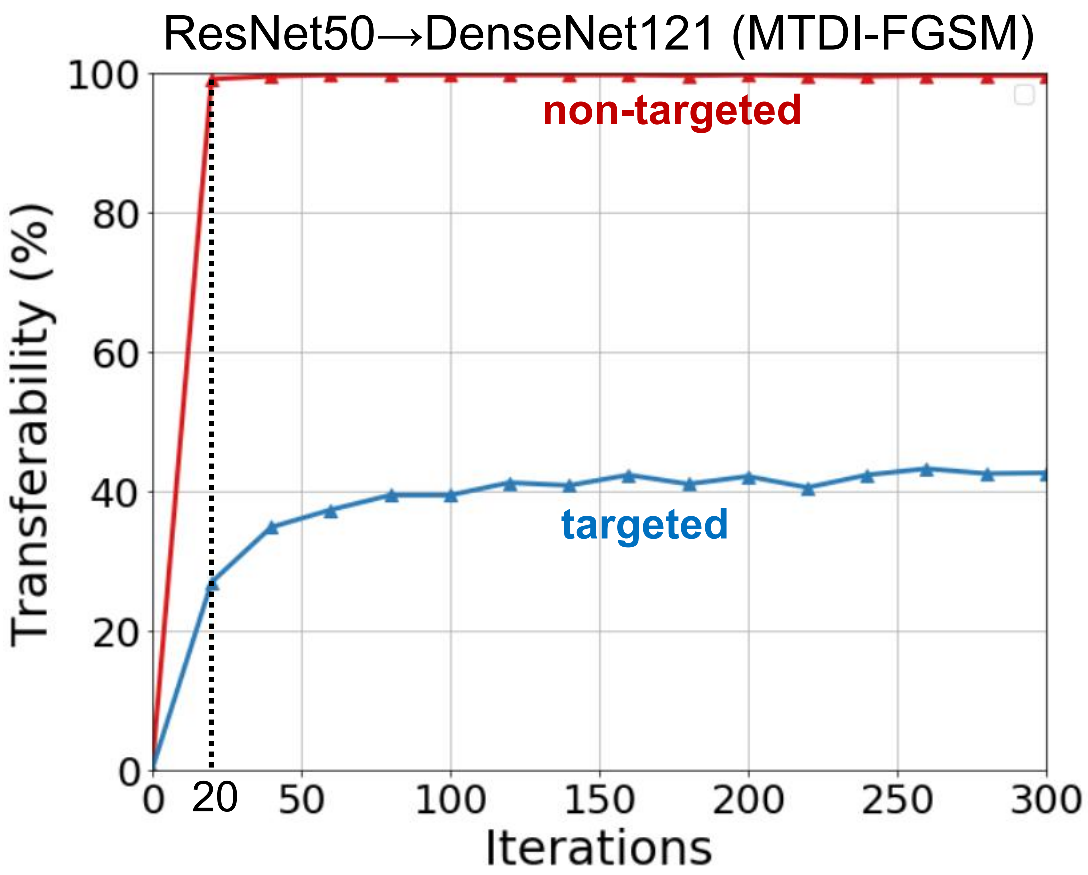
Fix I-FGSM: Step 1. Ensemble (0% →15%)

ResNet50 → DenseNet121 (Iter. =10)

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

single technique in existing work

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



<20 iterations in existing work:

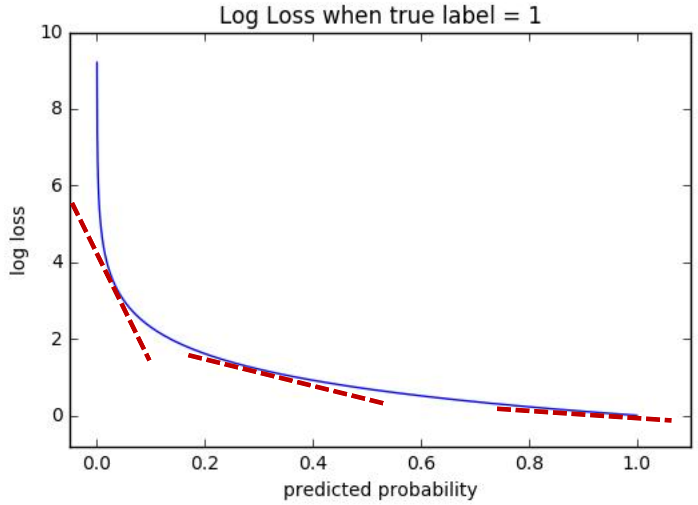
- fail to converge
- efficiency is not important

Fix I-FGSM: Step 3. Suitable Loss

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

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

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

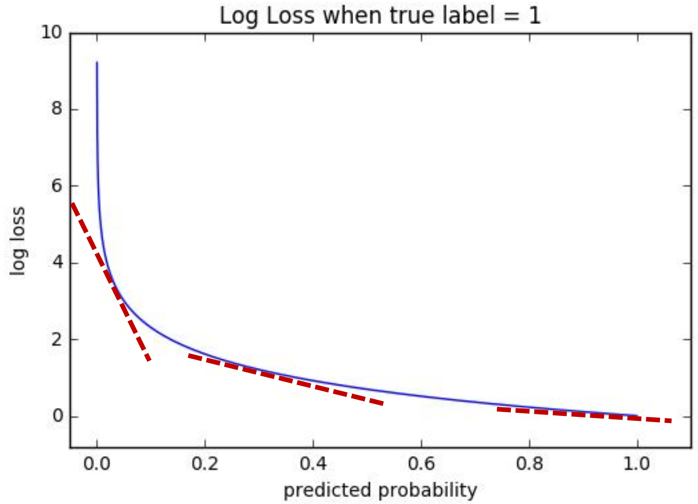


Fix I-FGSM: Step 3. Suitable Loss

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

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

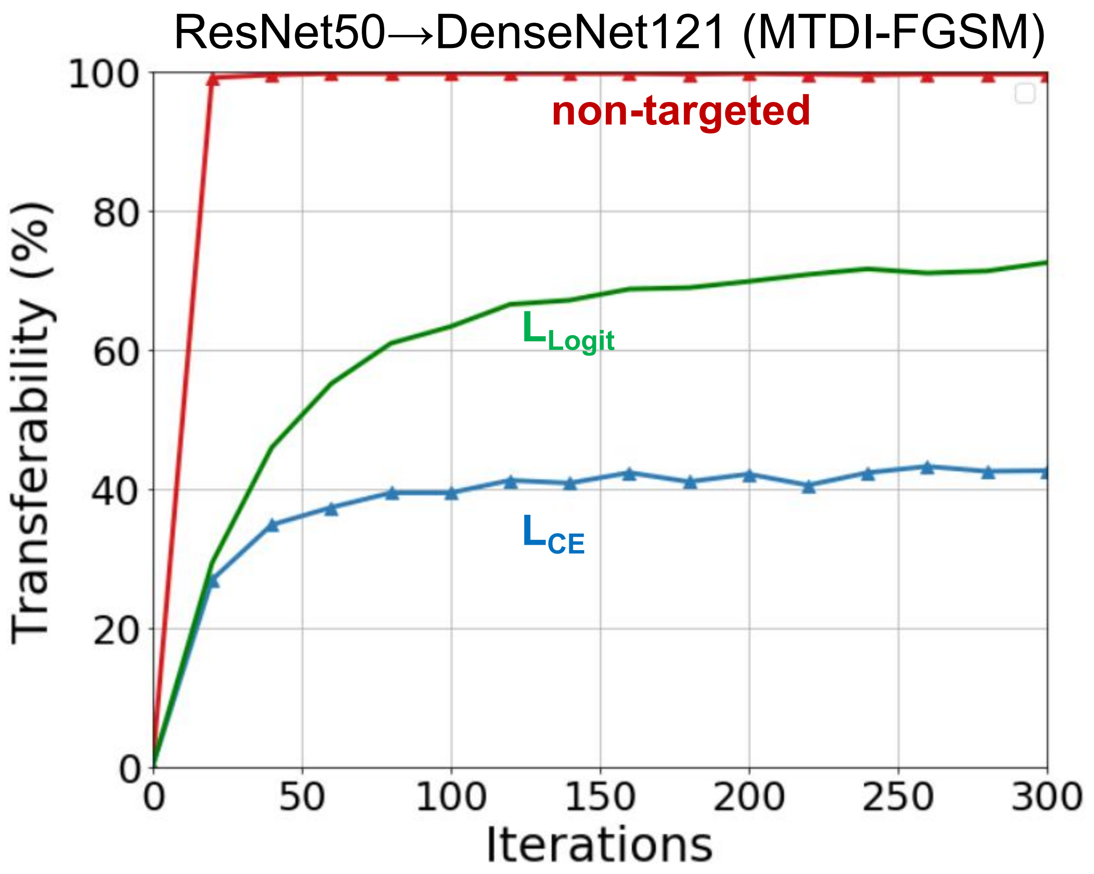
$$\frac{\partial L_{CE}}{\partial z_t} = -1 + \frac{\partial \log(\sum e^{z_j})}{\partial e^{z_t}} \cdot \frac{\partial e^{z_t}}{\partial z_t} = -1 + \frac{e^{z_t}}{\sum e^{z_j}} = -1 + p_t.$$



Logit Loss (L_{Logit}):

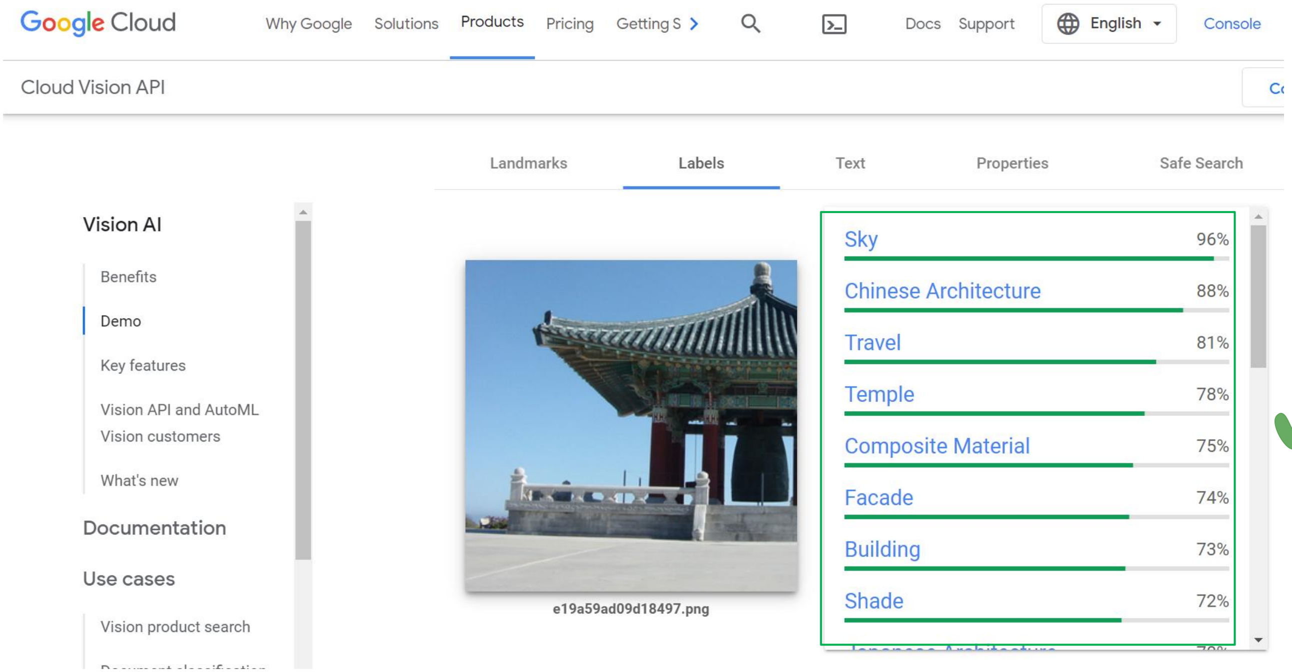
$$L_{Logit} = \underline{-z_t}, \ \frac{\partial L_{Logit}}{\partial z_t} = -1.$$

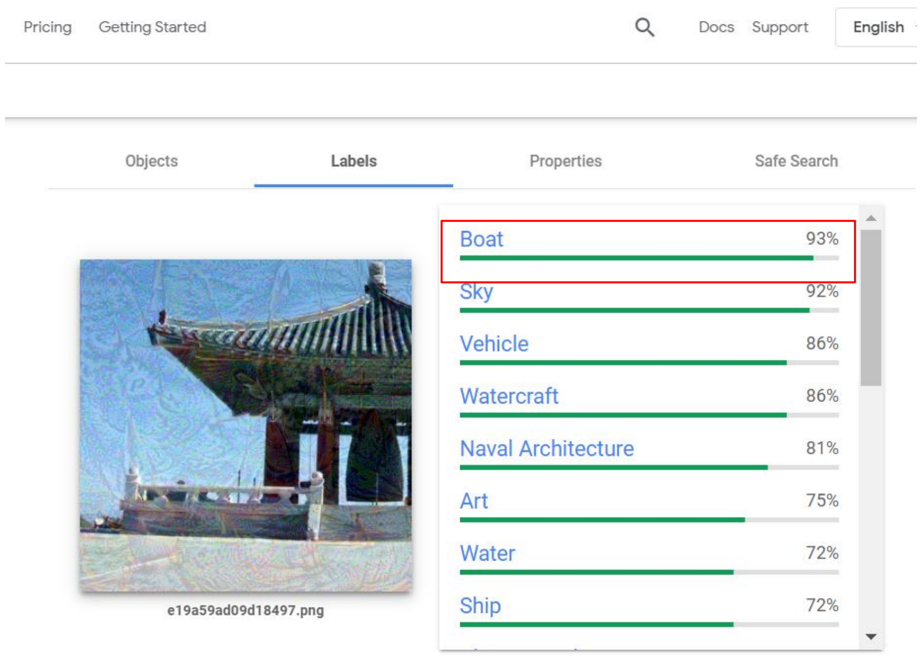
Fix I-FGSM: Step 3. Suitable Loss (42% →72%)



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**



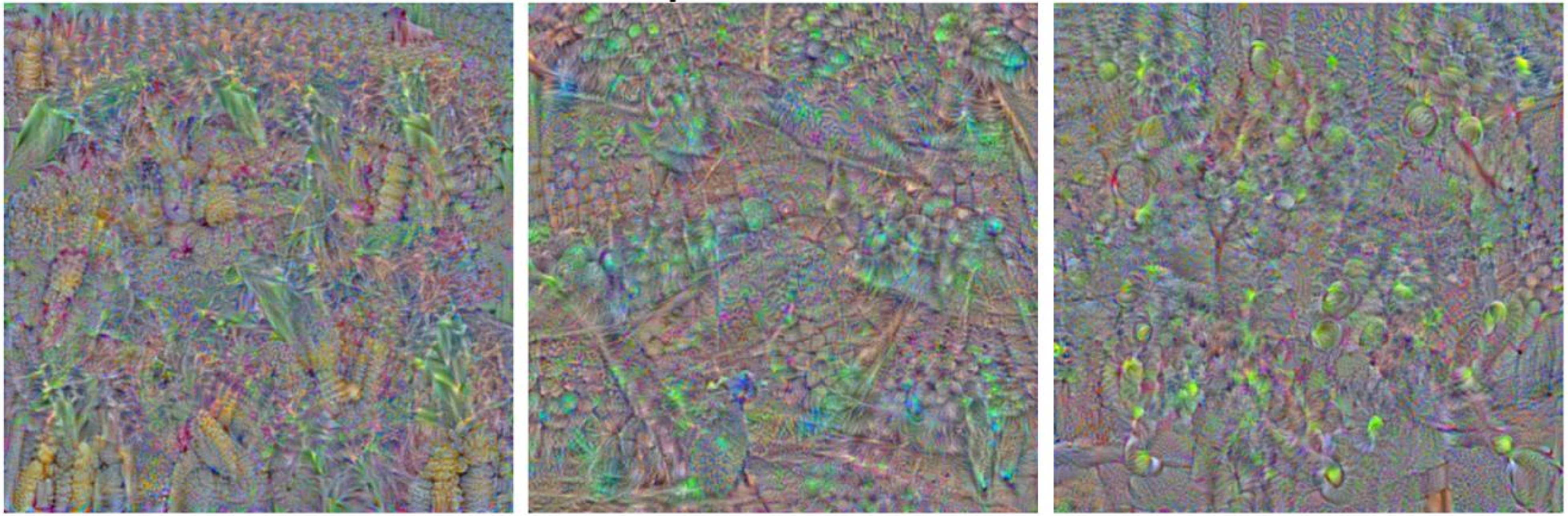


y_t = “yawl” (a type of boat)

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

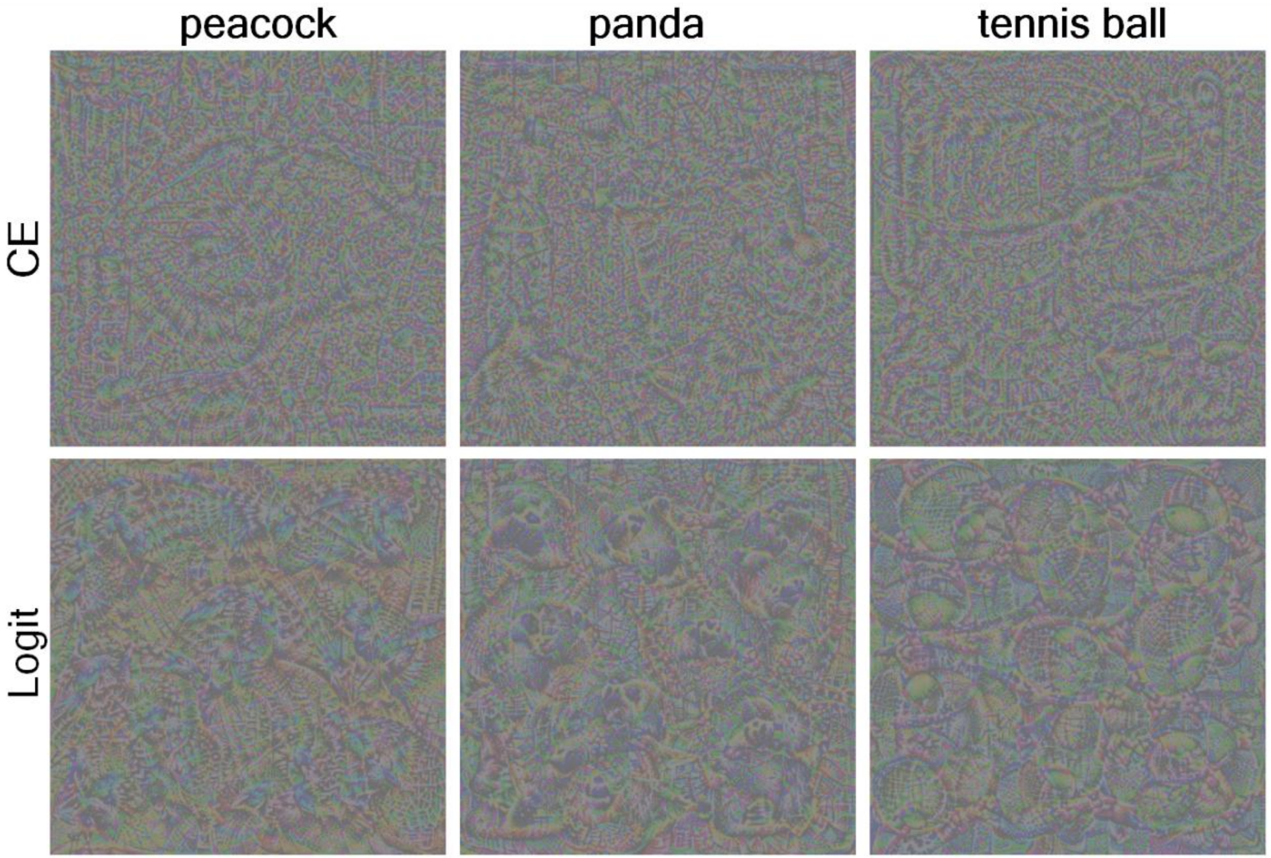
Other Analyses: Perturbation Semantics

target label: “corn” “peacock” “tennis ball”



without ϵ

Other Analyses: Targeted Universal Perturbations[1]



with ϵ=16

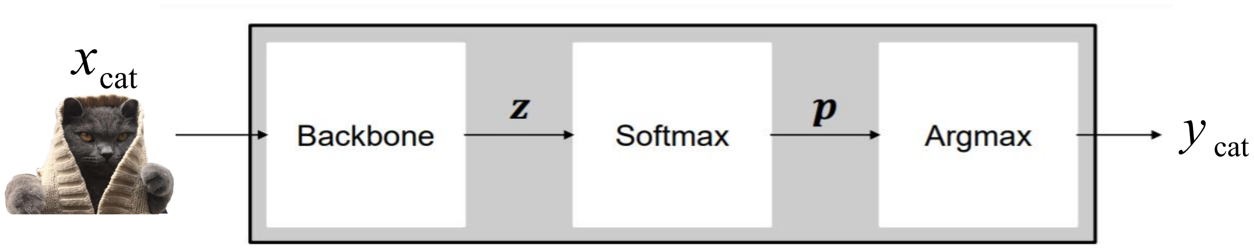
Success rates (%)

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

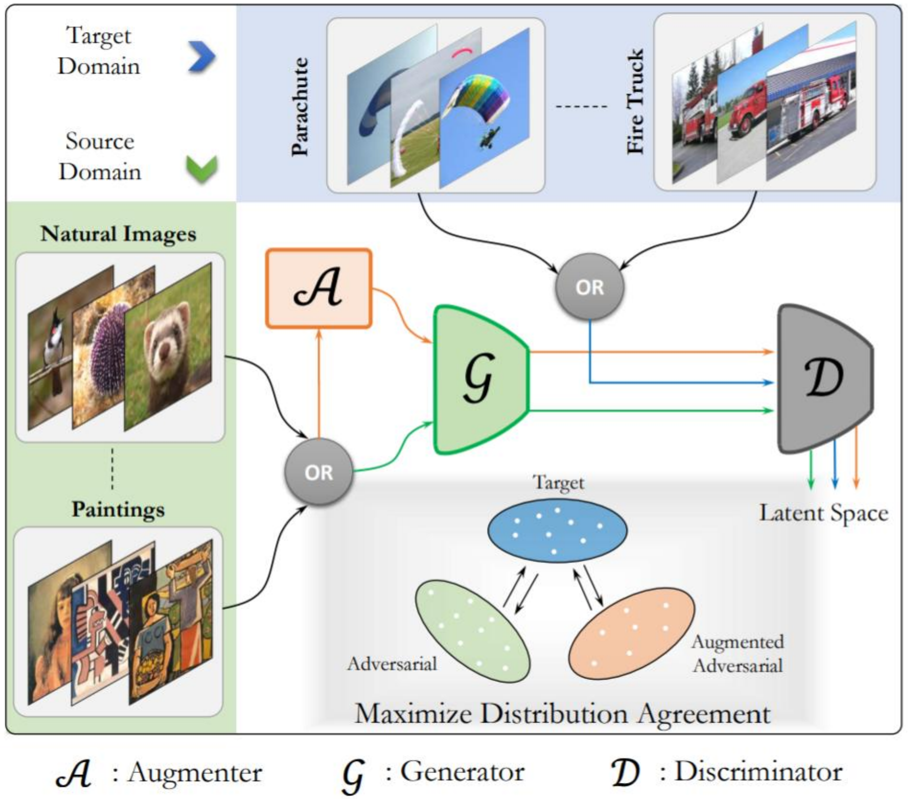
[1] Moosavi-Dezfooli et al. *Universal Adversarial Perturbations*. CVPR 2017.

Other Analyses: I-FGSM (ours) vs. Generative (SOTA)

Ours



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

Other Analyses: I-FGSM (ours) vs. Generative (SOTA)

Targeted Transferability (%)

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

Summary of Project 1

- 3 steps to fix I-FGSM
  - ensemble
  - more iterations
  - suitable (logit) loss

- Other Analyses
  - real-world attacks
  - universal perturbations
  - I-FGSM (data/training-free) vs. generative

Summary of Project 1

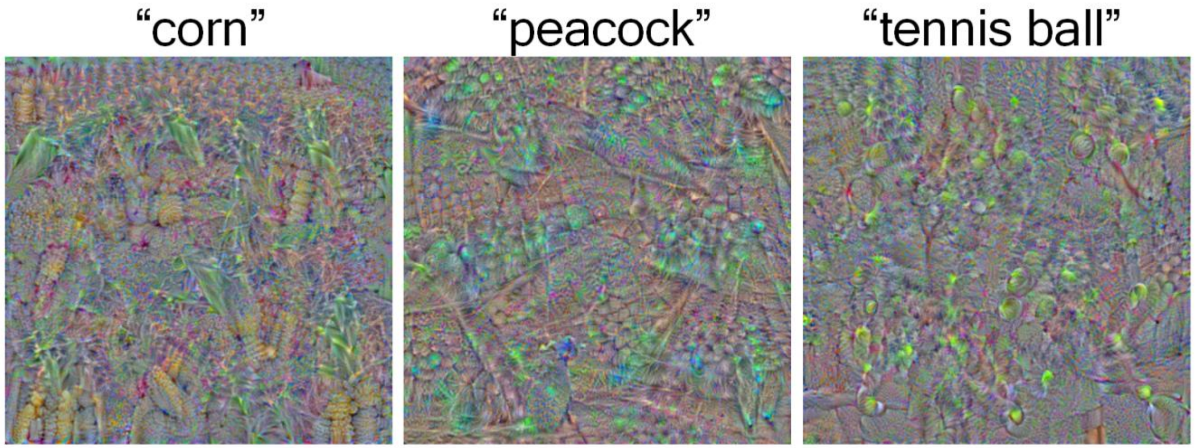
- 3 steps to revive I-FGSM
  - ensemble
  - more iterations
  - suitable (logit) loss
- Other Analyses
  - real-world attacks
  - universal perturbations
  - I-FGSM (data/training-free) vs. generative

"God is in the details"

Future Work

- Explaining transferability

feature similarity or model similarity



Res50 → Dense121: ~70% 😄

Res50 → Incv3: ~10% 😭

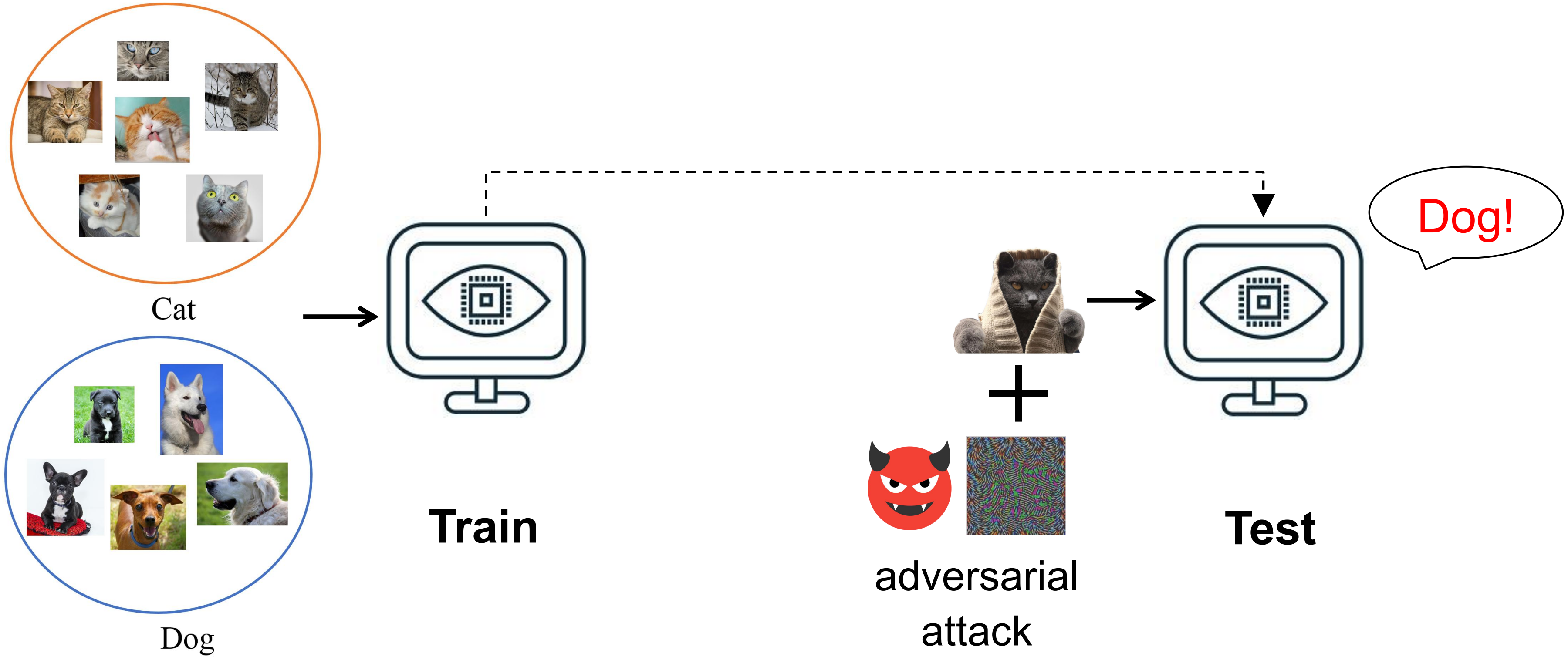
- Benchmarking transferability

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

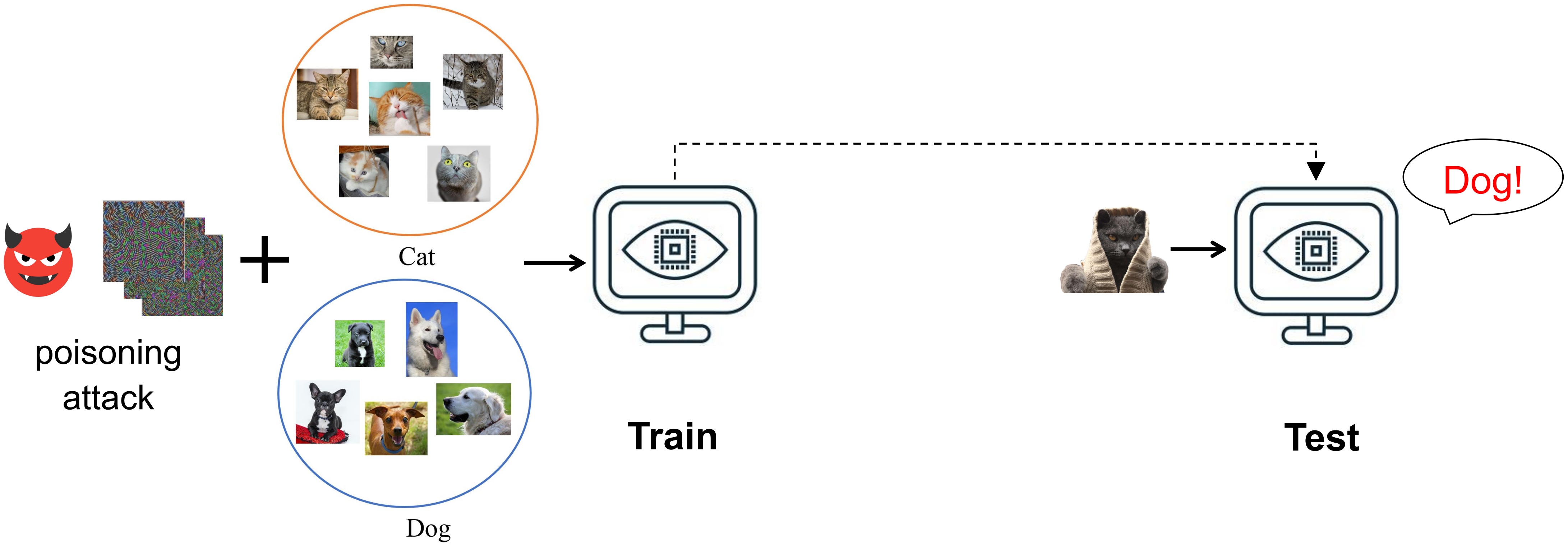
https://github.com/ZhengyuZhao/TransferAttackEval

- Systematic categorization of 40+ transfer attacks
- 23 representative attacks against 9 representative defenses on ImageNet
- Consensus-challenging insights

Testing-Stage Attack

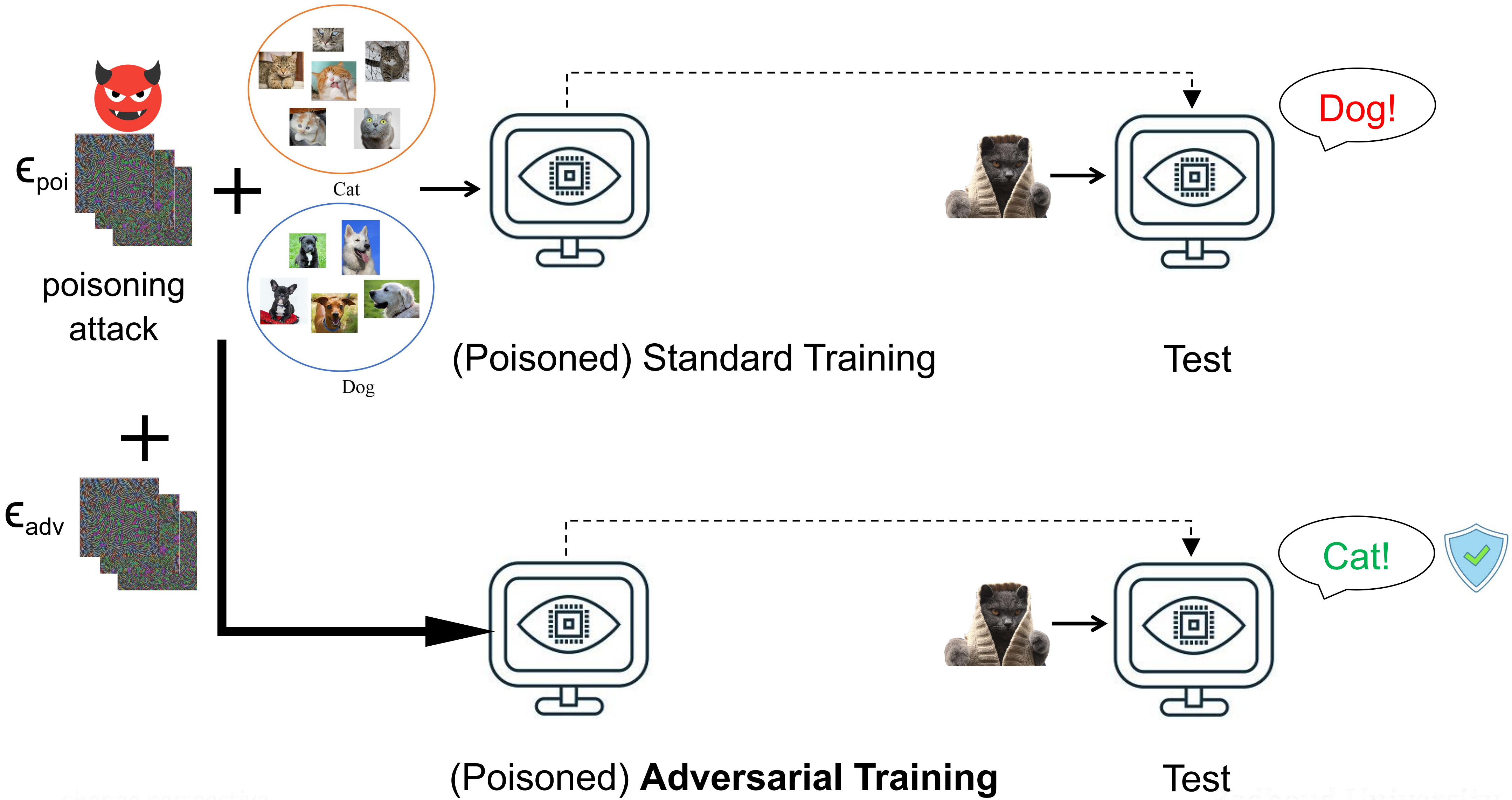


Training-Stage Attack

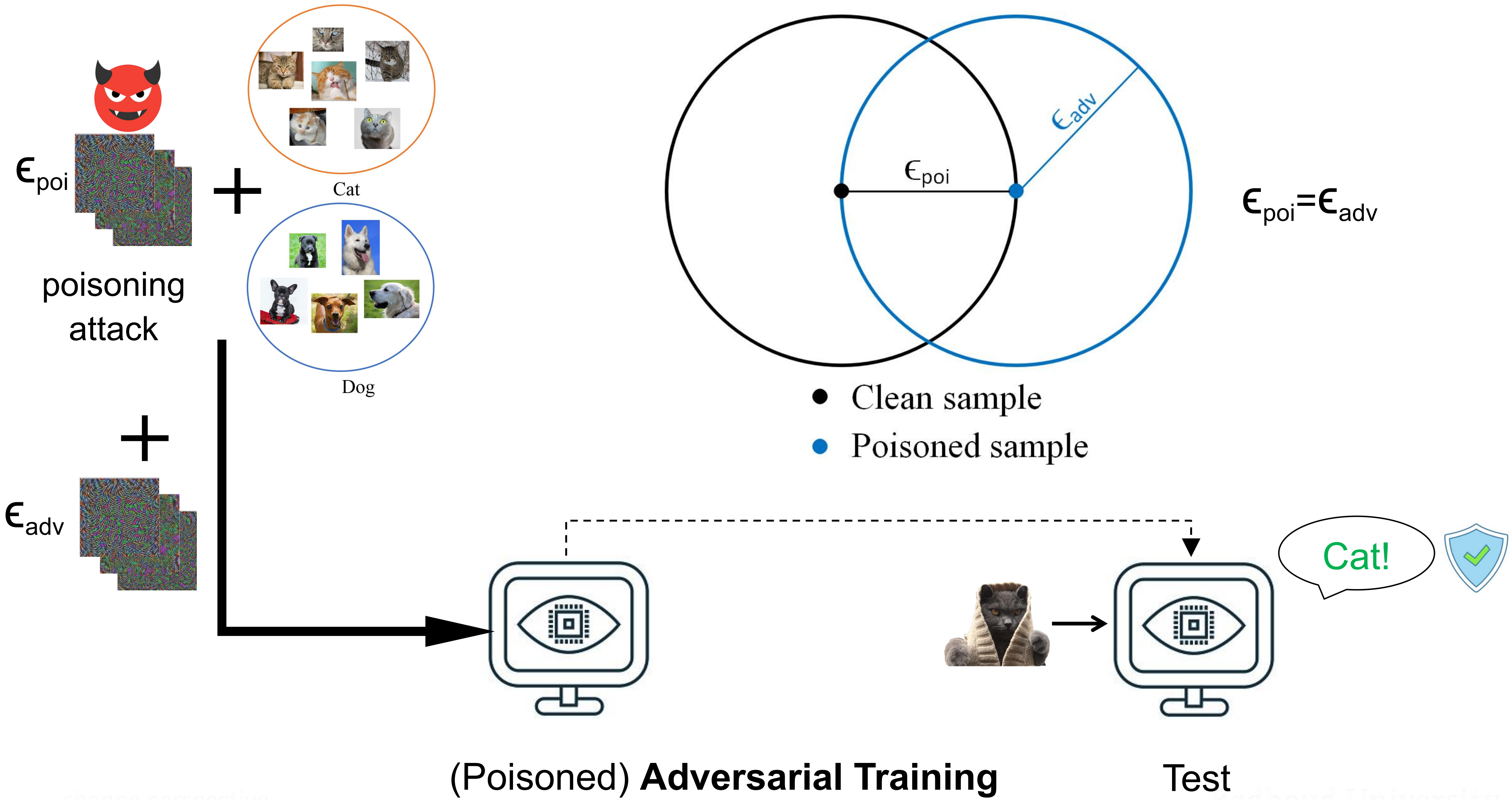


Project 2. Poisoning Against Adversarial Training

Adversarial Training-based Defense



Adversarial Training-based Defense



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

Consensus-Challenging Insight

Impossible to poison
AT models

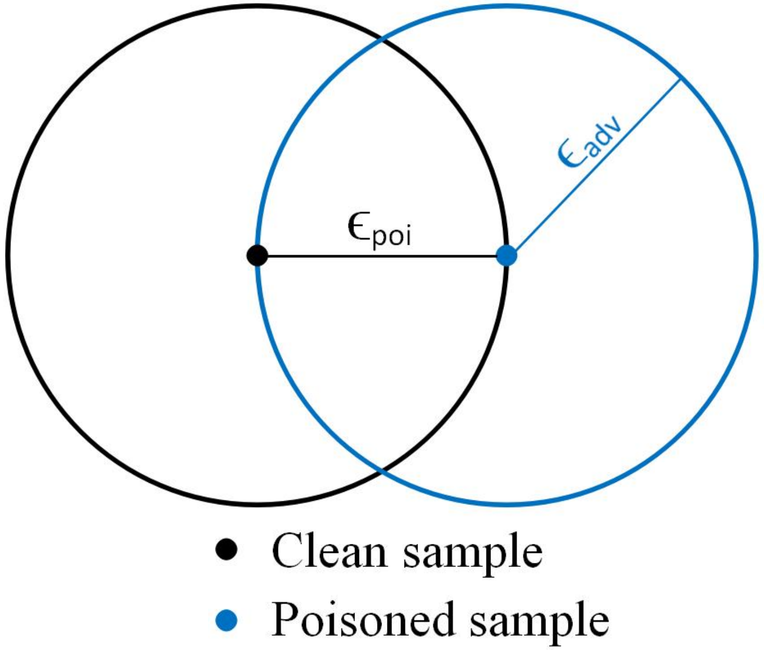
Existing work[1-6]

Possible (with a new
attack strategy)

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.

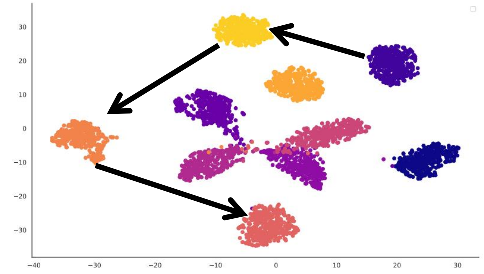
Consensus-Challenging Insight



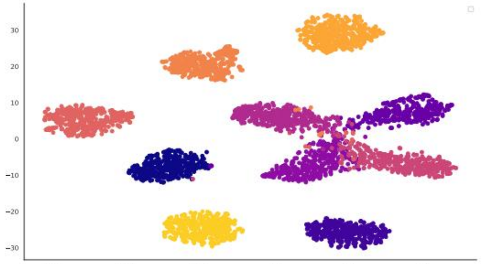
$\epsilon_{poi}=\epsilon_{adv}$

Existing Poisoning

existing poisoning

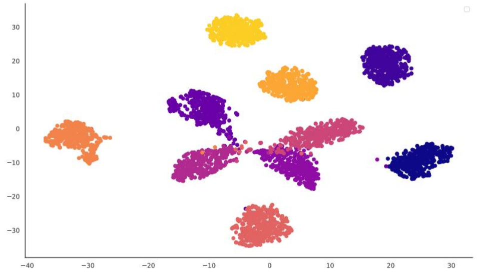


$$x' = \arg\min_{x} J(x, y_t)$$



Test Acc: 83.11% 😲

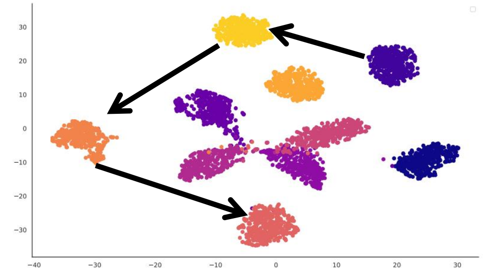
clean training



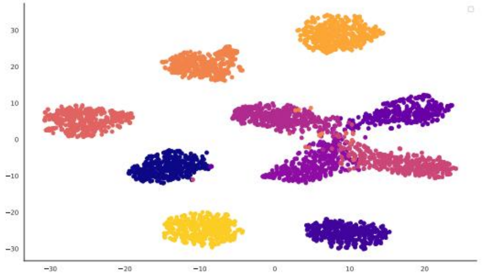
Test Acc: 84.88%

Our Poisoning

existing poisoning

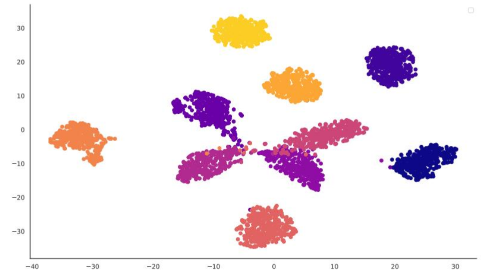


$$x' = \arg\min_{x} J(x, y_t)$$

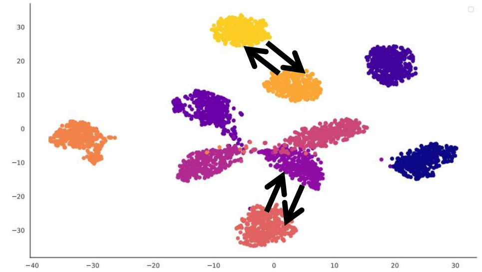


Test Acc: 83.11% 😵

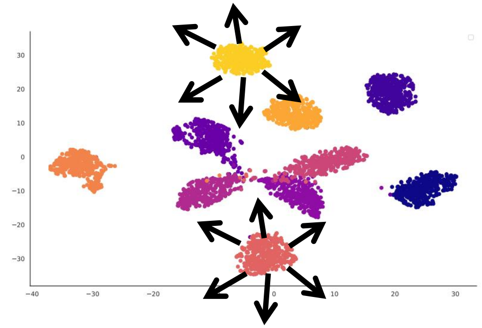
clean training



Test Acc: 84.88%

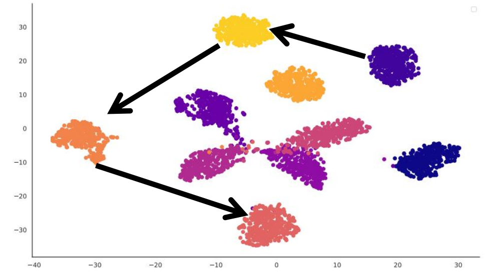


our poisoning

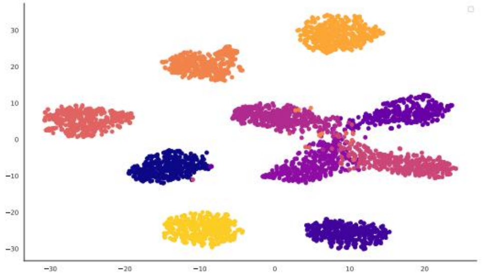


Our Poisoning

existing poisoning

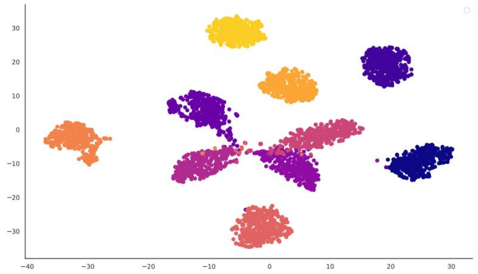


$$x' = \arg\min_{x} J(x, y_t)$$

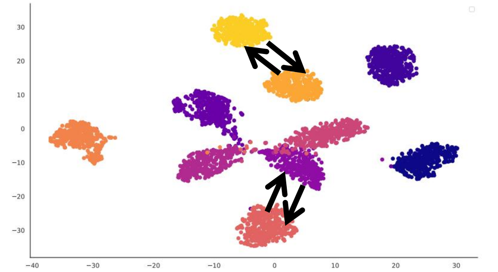


Test Acc: 83.11% 😵

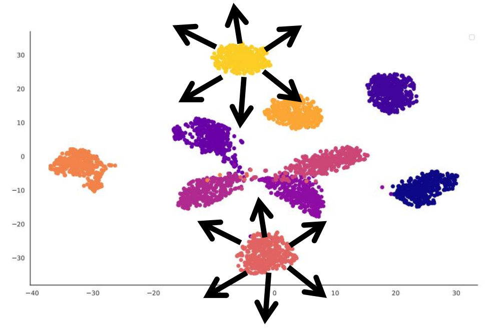
clean training

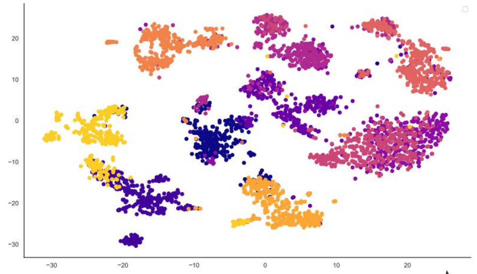


Test Acc: 84.88%

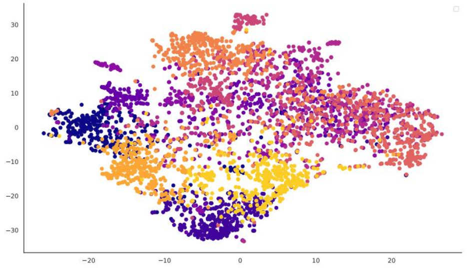


our poisoning





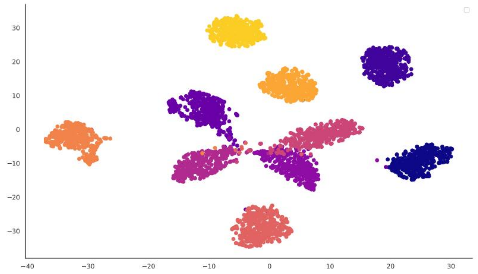
Test Acc: 72.99% 😈



Test Acc: 71.57% 😈

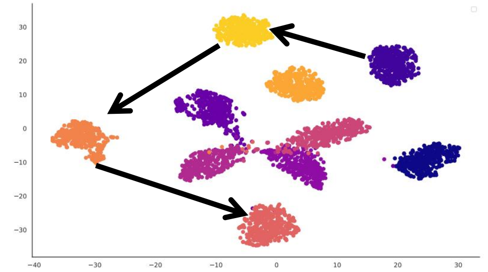
Our Poisoning

clean training

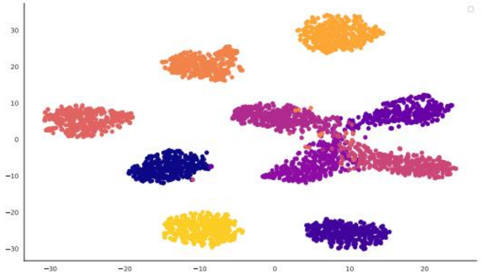


Test Acc: 84.88%

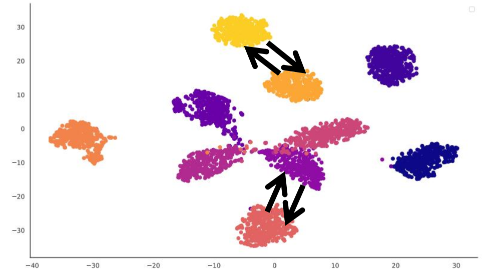
existing poisoning



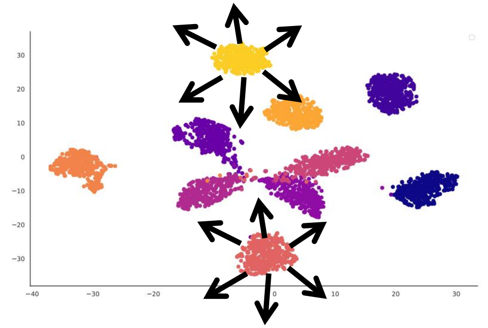
$$x' = \arg\min_{x} J(x, y_{\mathrm{t}})$$



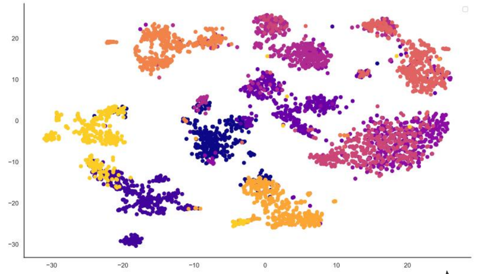
Test Acc: 83.11%

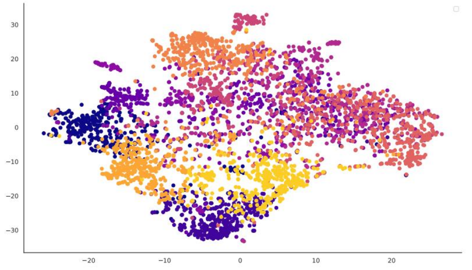


our poisoning



Test Acc: 72.99%



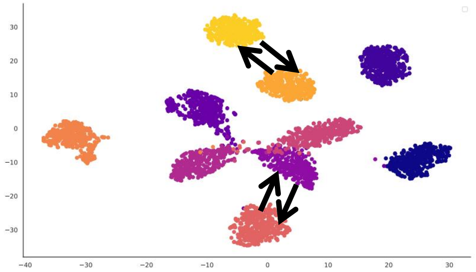


Test Acc: 71.57%

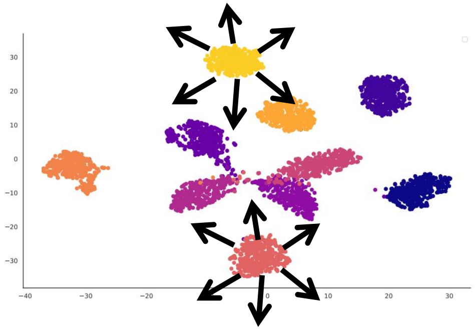
equal to discarding
83% training data!

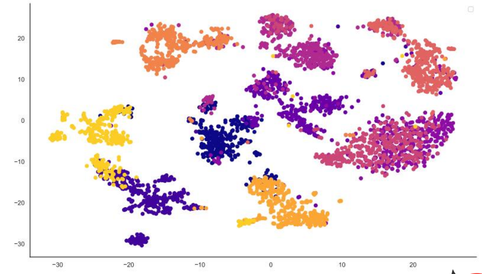
Our Poisoning

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

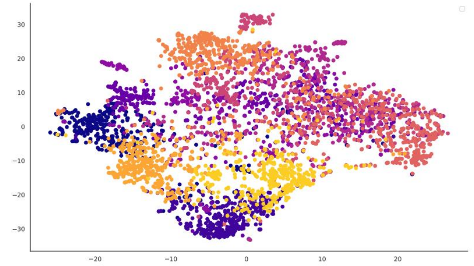


our poisoning





Test Acc: 72.99%



Test Acc: 71.57%

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

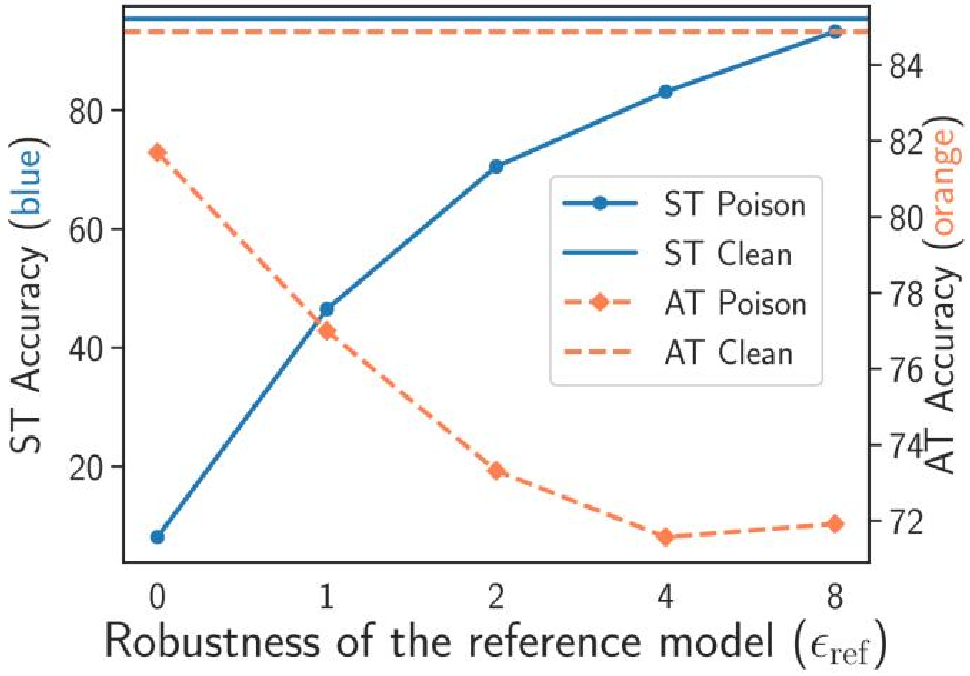
$$\mathcal{L}_{\text{push}} = \max_{\boldsymbol{\delta}^{\text{poi}}} \|F^*_{L-1}(\boldsymbol{x} + \boldsymbol{\delta}^{\text{poi}}) - \boldsymbol{\mu}_{y}\|_2$$

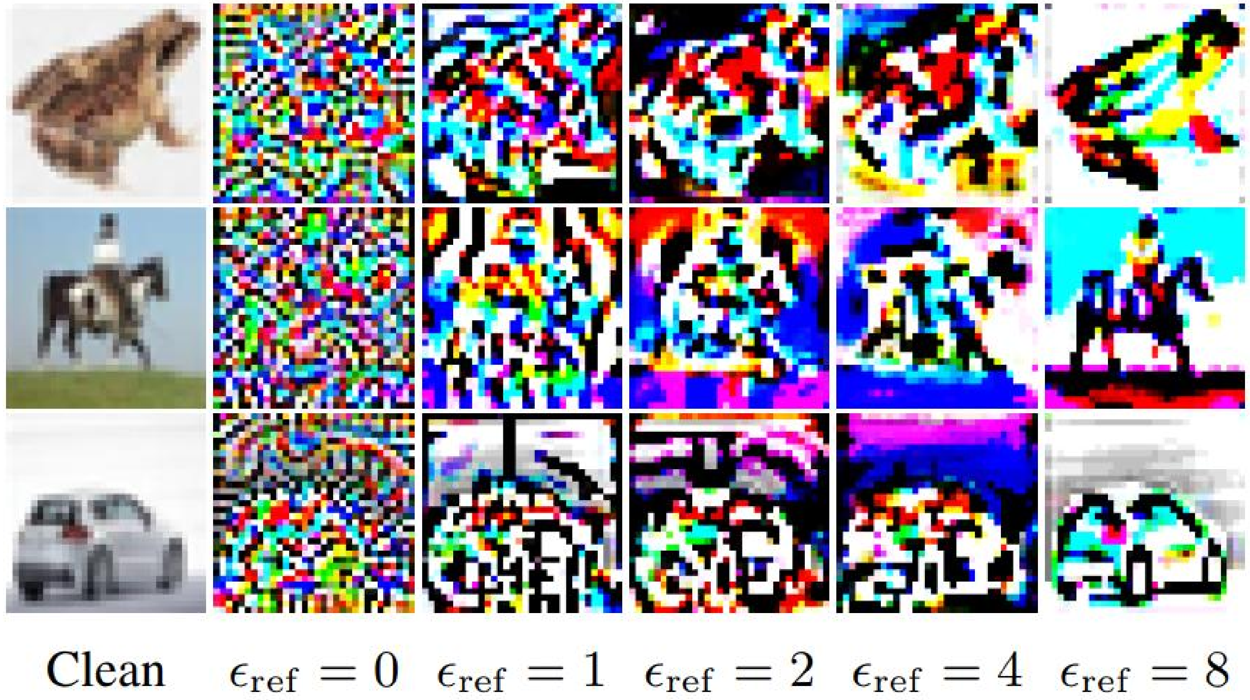
Results

- Different datasets
- Different AT frameworks
- Transferability
- Partial data Poisoning training data
- Ensemble defenses
- Adaptive defenses

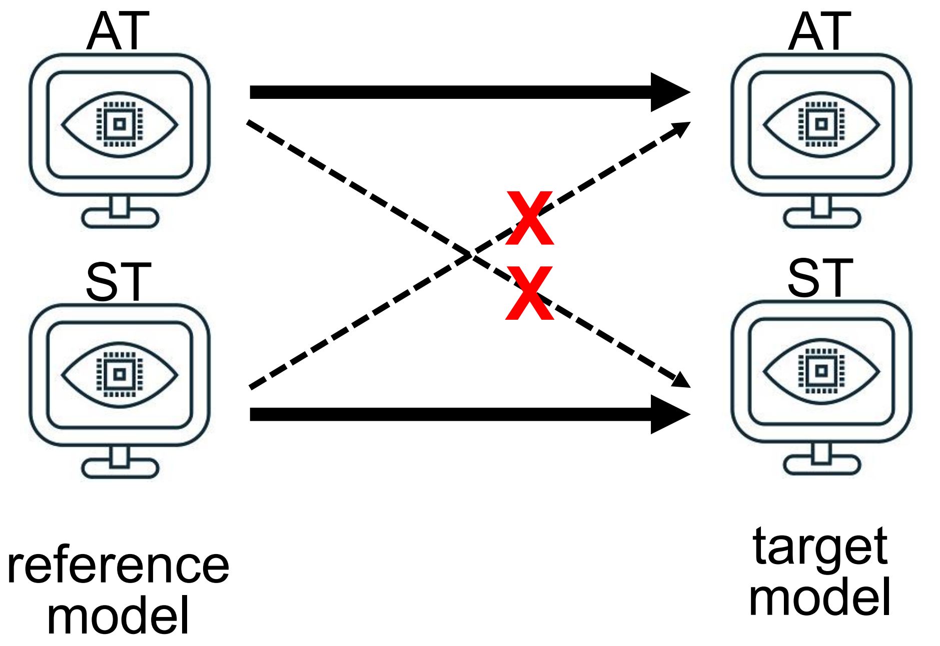
...

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

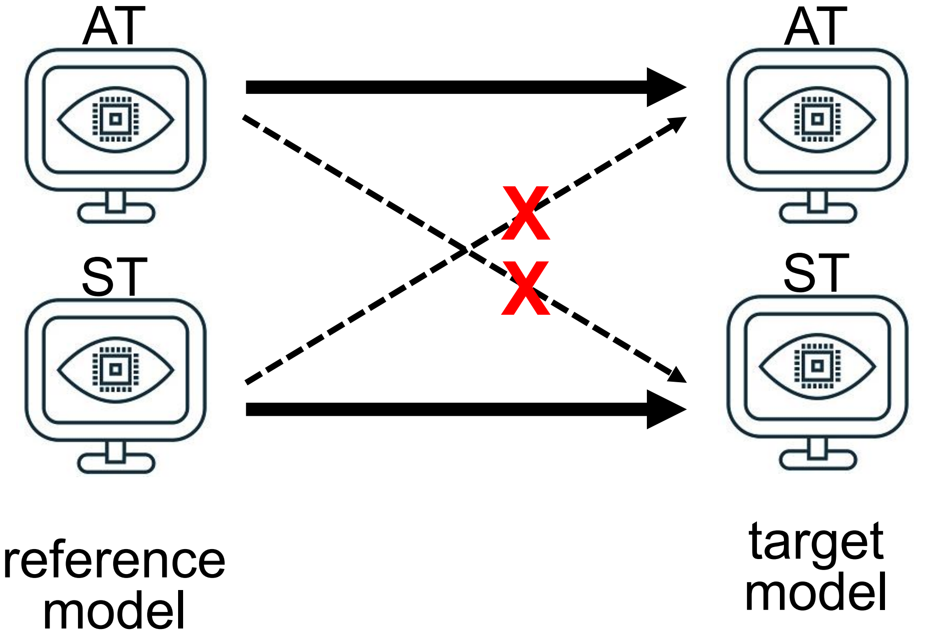




Hybrid Attack



Hybrid Attack



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

$$\downarrow$$

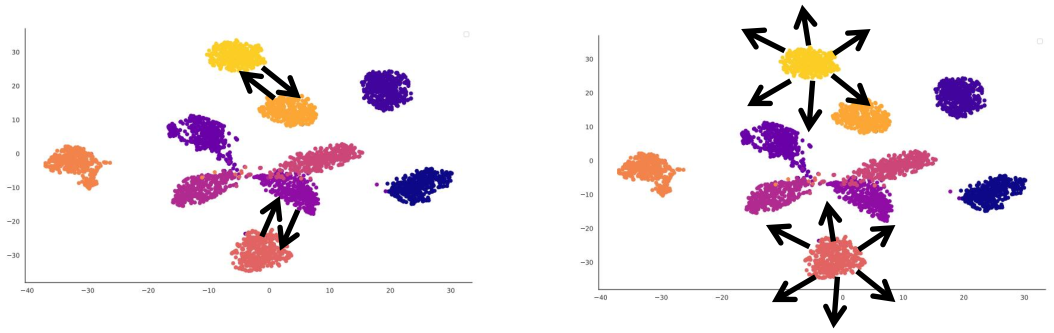
$$\mathcal{L}_{\text{hybrid}} = \max_{\boldsymbol{\delta}^{\text{poi}}} \|F^*_{L-1,\text{ST}}(\boldsymbol{x} + \boldsymbol{\delta}^{\text{poi}}) - \boldsymbol{\mu}_{y,\boxed{\text{ST}}}\|_2 + \lambda\|F^*_{L-1,\text{AT}}(\boldsymbol{x} + \boldsymbol{\delta}^{\text{poi}}) - \boldsymbol{\mu}_{y,\boxed{\text{AT}}}\|_2$$

Hybrid Attack

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 strategy



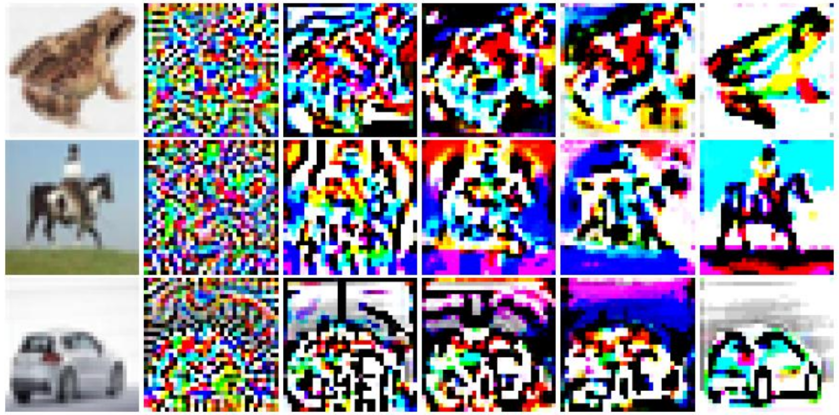
- Poisoning AT vs. ST
- Hybrid attack

Future Directions

- Possible defenses against our new attack
  - generic: training techniques for noisy labels?
  - specific: detecting/pre-filtering our attack?

- More efficient hybrid attack than

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



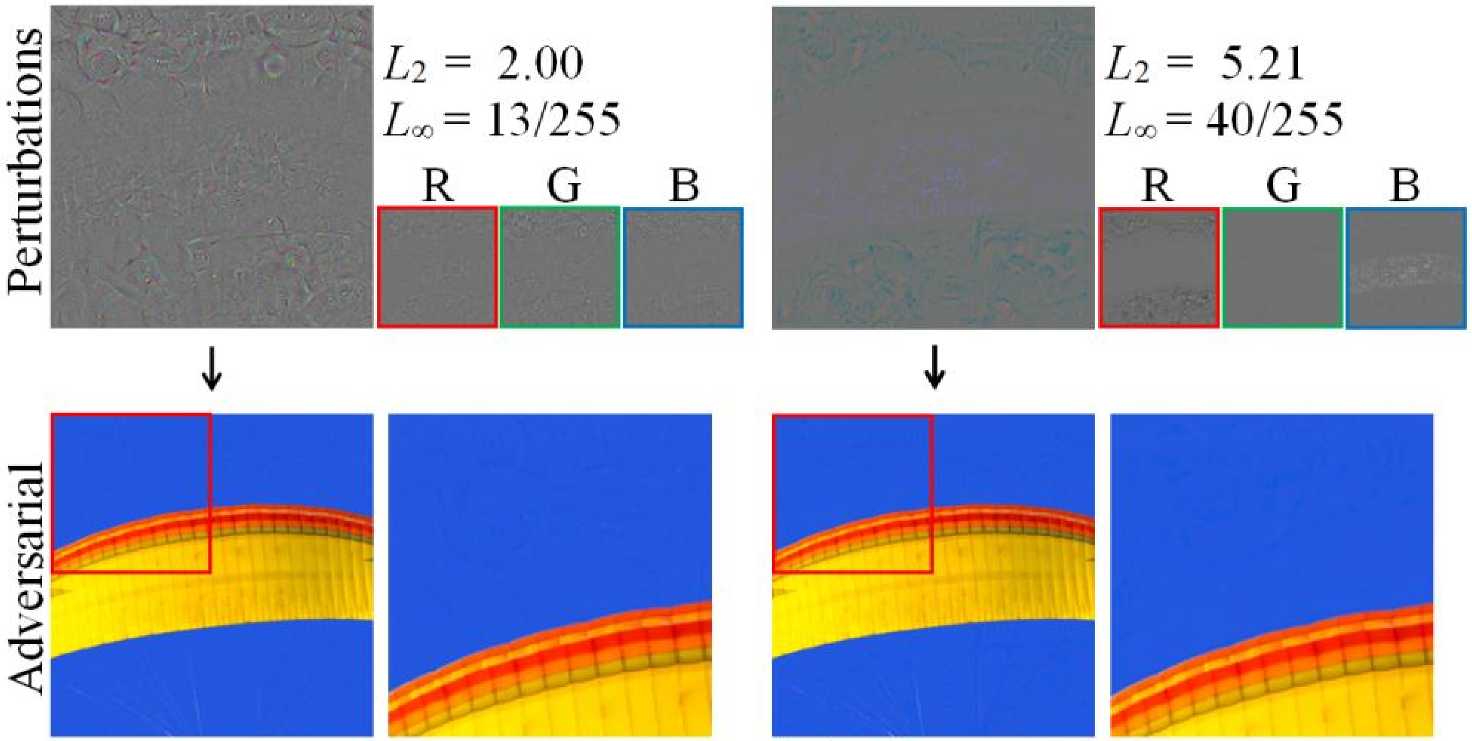
Clean $\epsilon_{\text{ref}} = 0$ $\epsilon_{\text{ref}} = 1$ $\epsilon_{\text{ref}} = 2$ $\epsilon_{\text{ref}} = 4$ $\epsilon_{\text{ref}} = 8$

Outline

- Overview of adversarial images in computer vision
- Two recent projects
- Other related projects

Imperceptible Perturbations


$$\left\| x' - x_{\text{cat}} \right\|_\infty \leq \varepsilon \quad \longrightarrow \quad \left\| x' - x_{\text{cat}} \right\|_{\text{CIEDE2000}} \leq \varepsilon$$

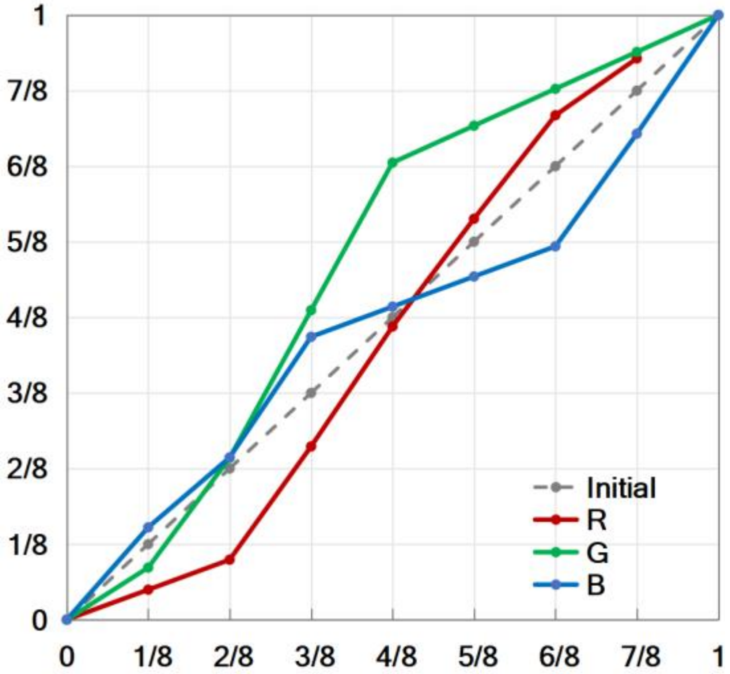
(a) C&W

(b) PerC-C&W (ours)

Perceptible yet Stealthy Attacks

x

$\times$

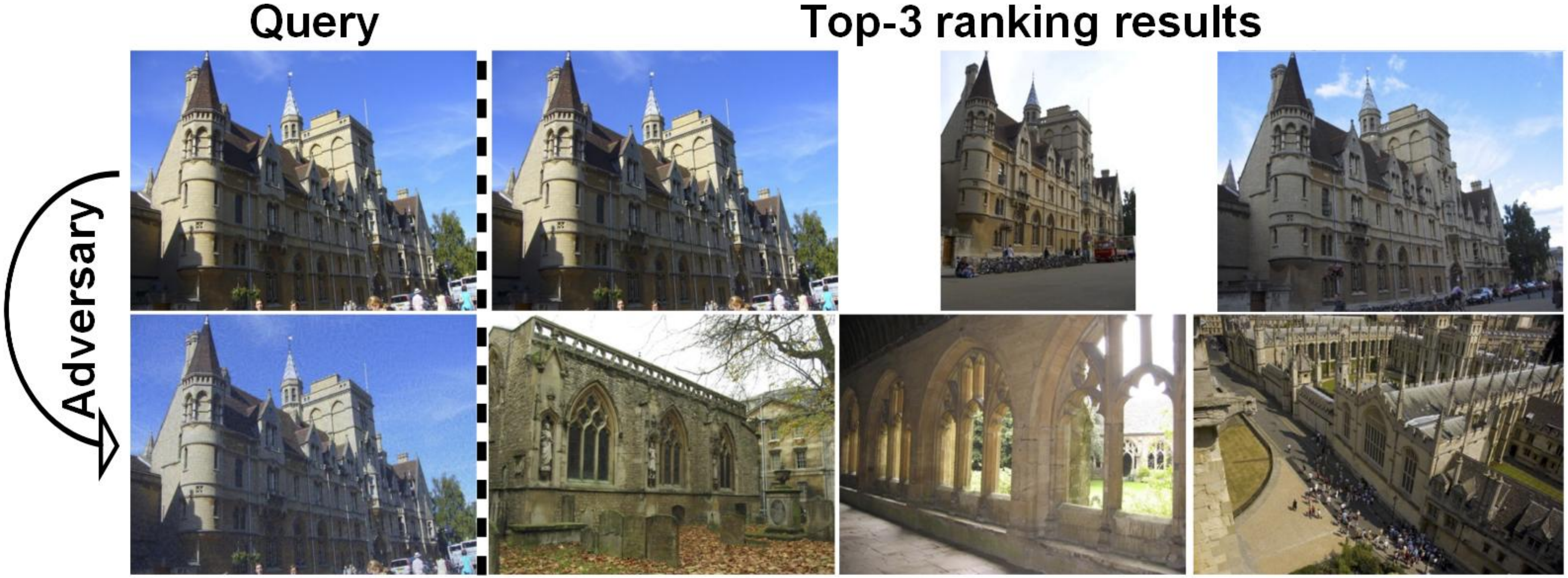


color transformation

$=$

x'

Adversarial attacks on Image Retrieval



CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

Radboud University

- **On Success and Simplicity: A Second Look at Transferable Targeted Attacks (Project 1)**
  Zhengyu Zhao, Zhuoran Liu, Martha Larson. NeurIPS 2021.

- **Is Adversarial Training Really a Silver Bullet for Mitigating Data Poisoning? (Project 2)**
  Rui Wen, Zhengyu Zhao, Zhuoran Liu, Michael Backes, Tianhao Wang, Yang Zhang. ICLR 2023.

- **Towards Good Practices in Evaluating Transfer Adversarial Attacks**
  Zhengyu Zhao*, Hanwei Zhang*, Renjue Li*, Ronan Sicre, Laurent Amsaleg, Michael Backes. arXiv 2022.

- **Towards Large yet Imperceptible Adversarial Image Perturbations with Perceptual Color Distance**
  Zhengyu Zhao, Zhuoran Liu, Martha Larson. CVPR 2020.

- **Adversarial Image Color Transformations in Explicit Color Filter Space**
  Zhengyu Zhao, Zhuoran Liu, Martha Larson. BMVC 2020.

- **Who's Afraid of Adversarial Queries? The Impact of Image Modifications on Content-based Image Retrieval**
  Zhuoran Liu, Zhengyu Zhao, Martha Larson. ICMR 2019.

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

Thank you!

Q&A