On Success and Simplicity: A Second Look at Transferable Targeted Adversarial Images
(对有目标对抗图像迁移性的反思)

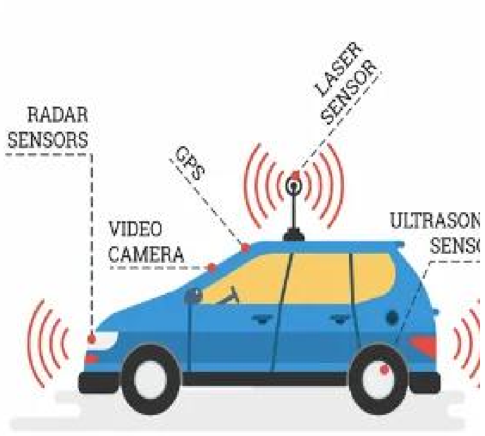
Zhengyu Zhao (赵正宇), Zhuoran Liu, Martha Larson
Radboud University, The Netherlands

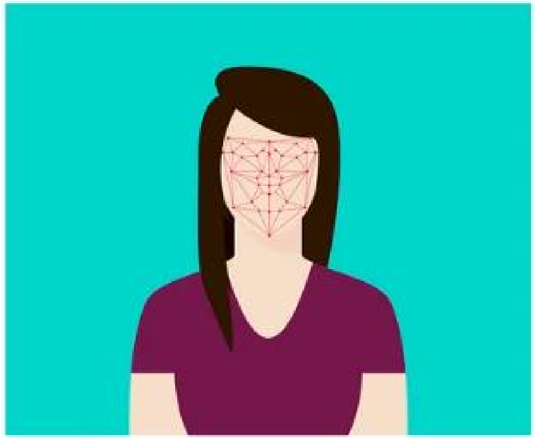
February 17th, 2022

- Background of Computer Vision
- Adversarial Image (对抗图像) and its transferability (迁移性)
- New insights into targeted (有目标) transferability
- Summary & Future work

Background：Computer Vision (计算机视觉)

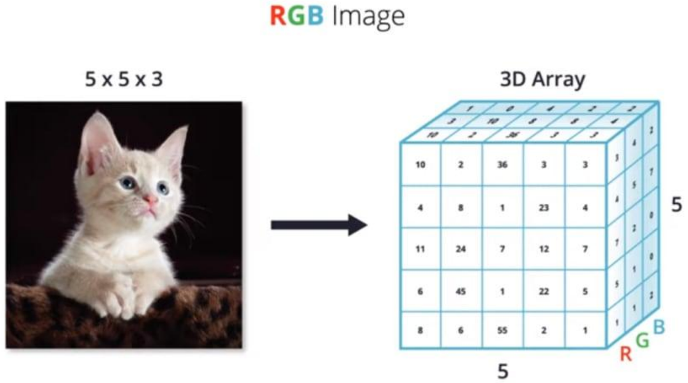
Background： Computer Vision Applications

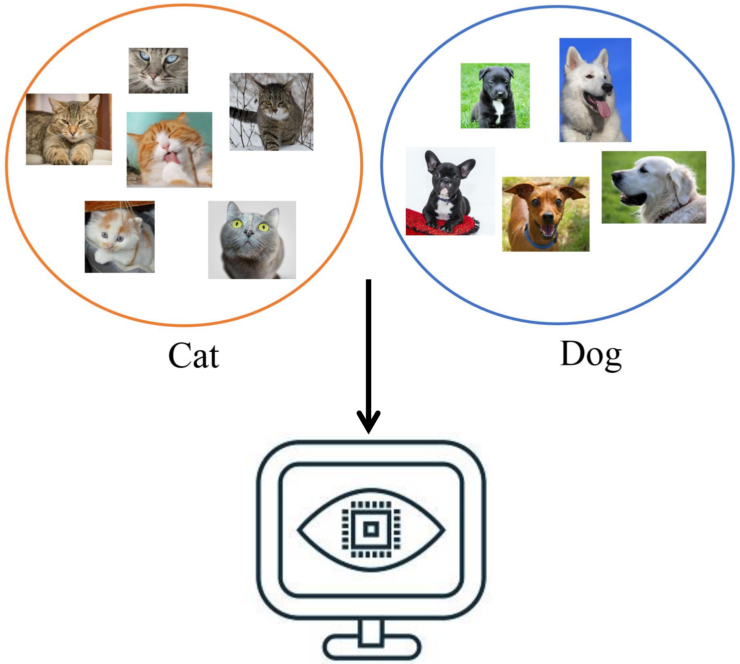


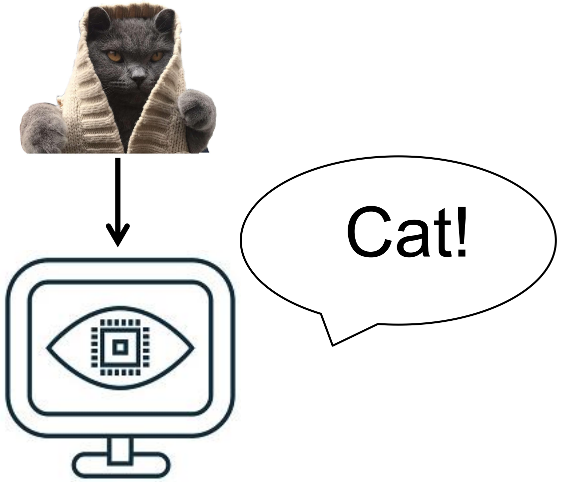

Applications in different areas

Background： Computer Vision Pipeline



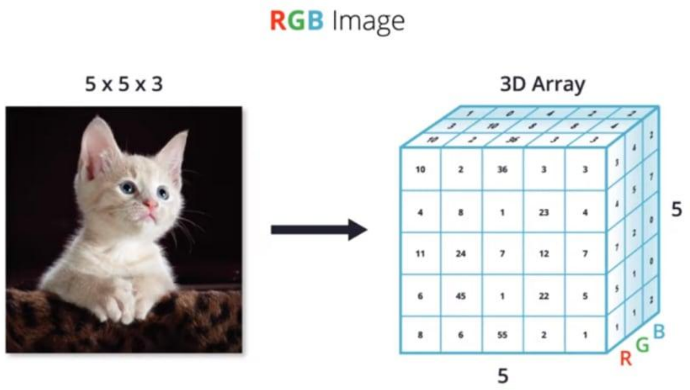


new cat image

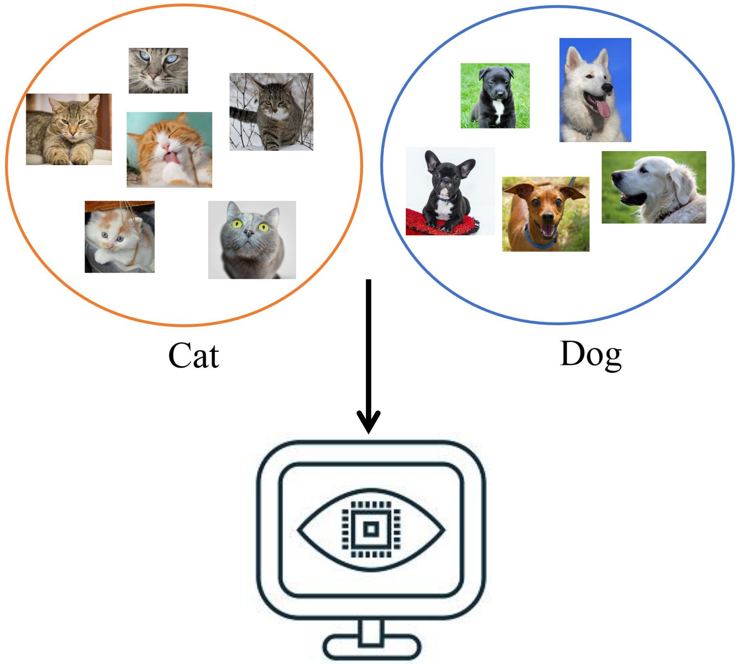


Perceive **Train** **Recognize**

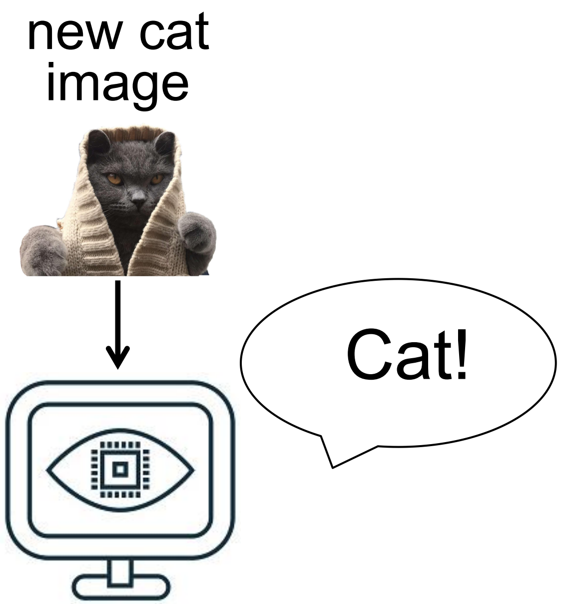
Background: Computer Vision Pipeline



Perceive

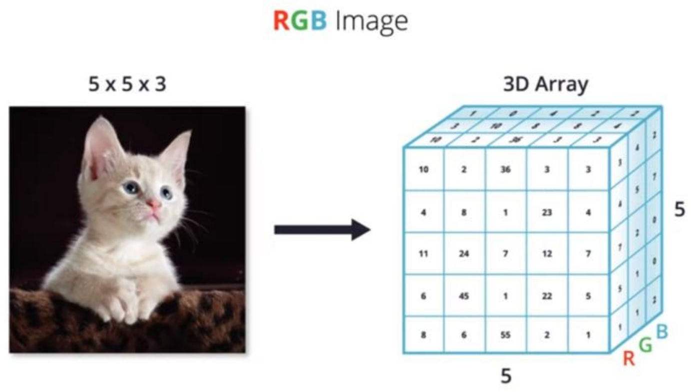


Train

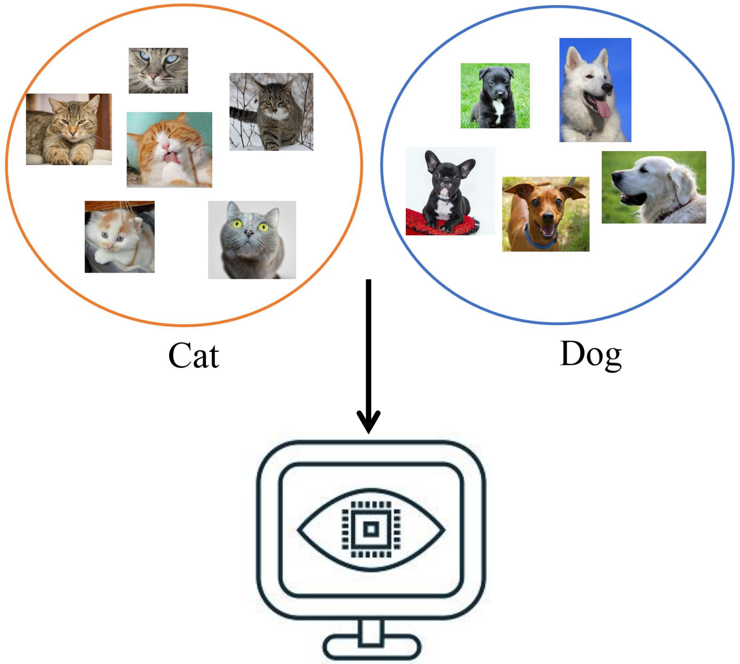


Recognize

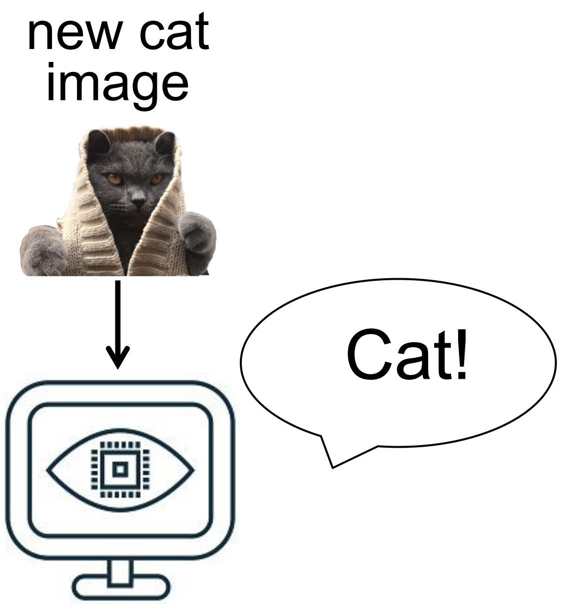
Background： Computer Vision Pipeline



Perceive

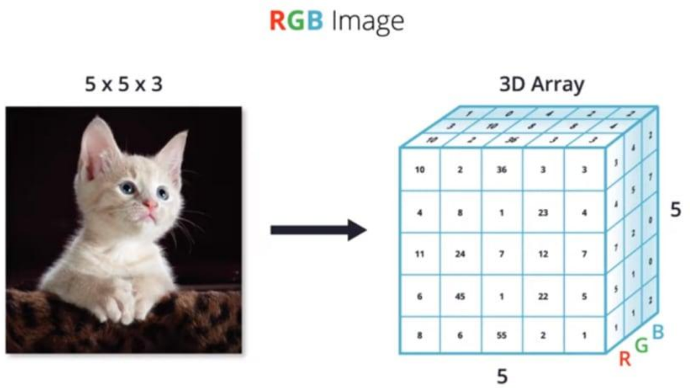


Train

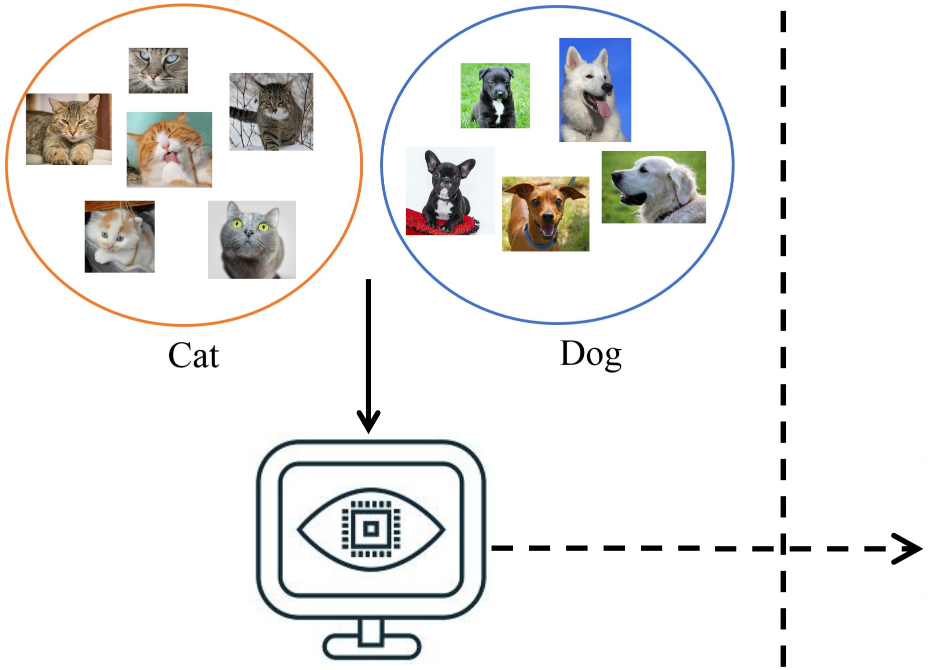


Recognize

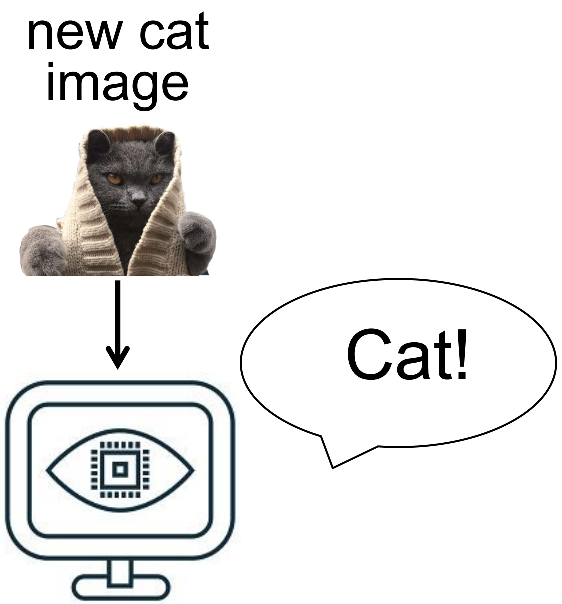
Background: Computer Vision Pipeline



Perceive

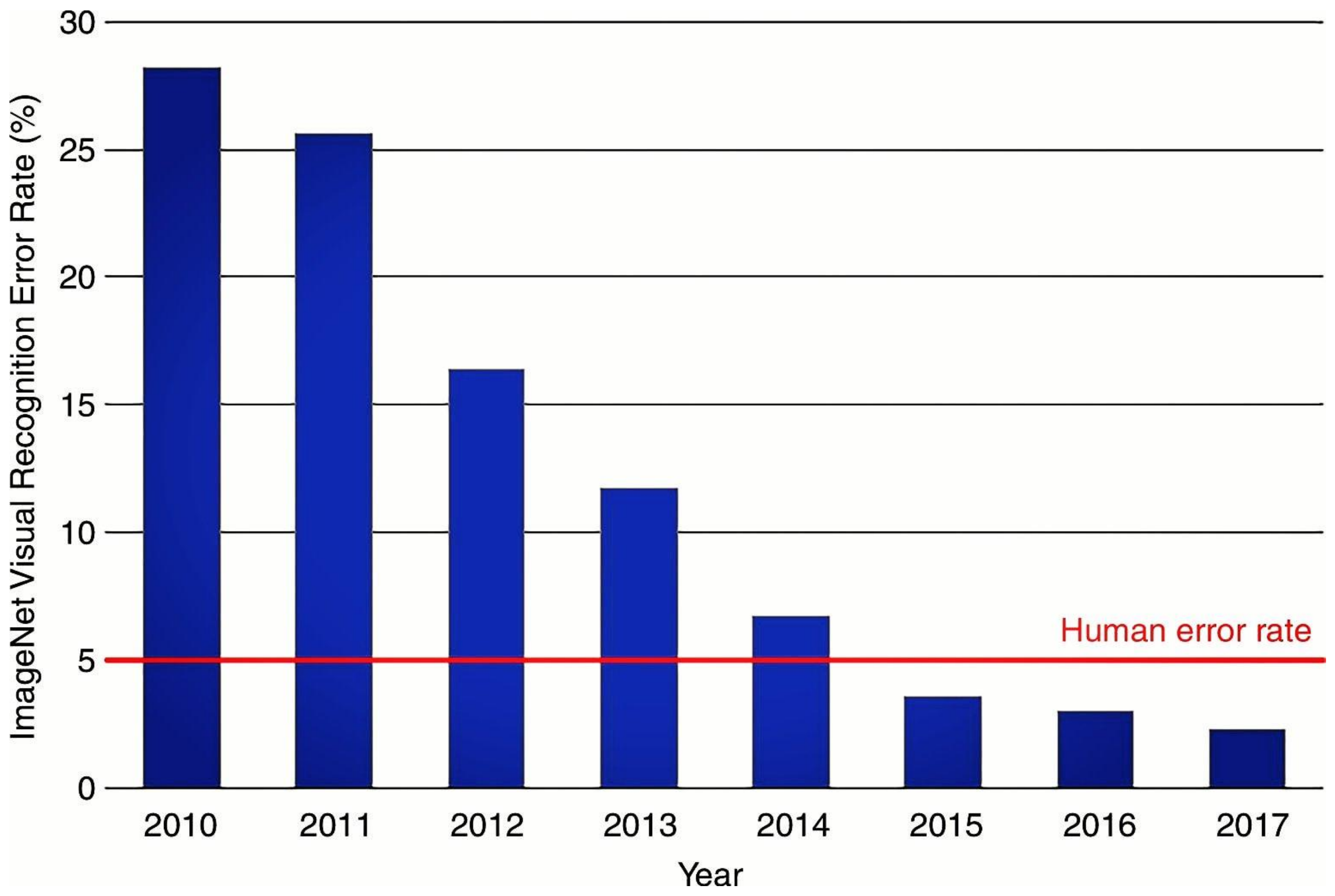


Train



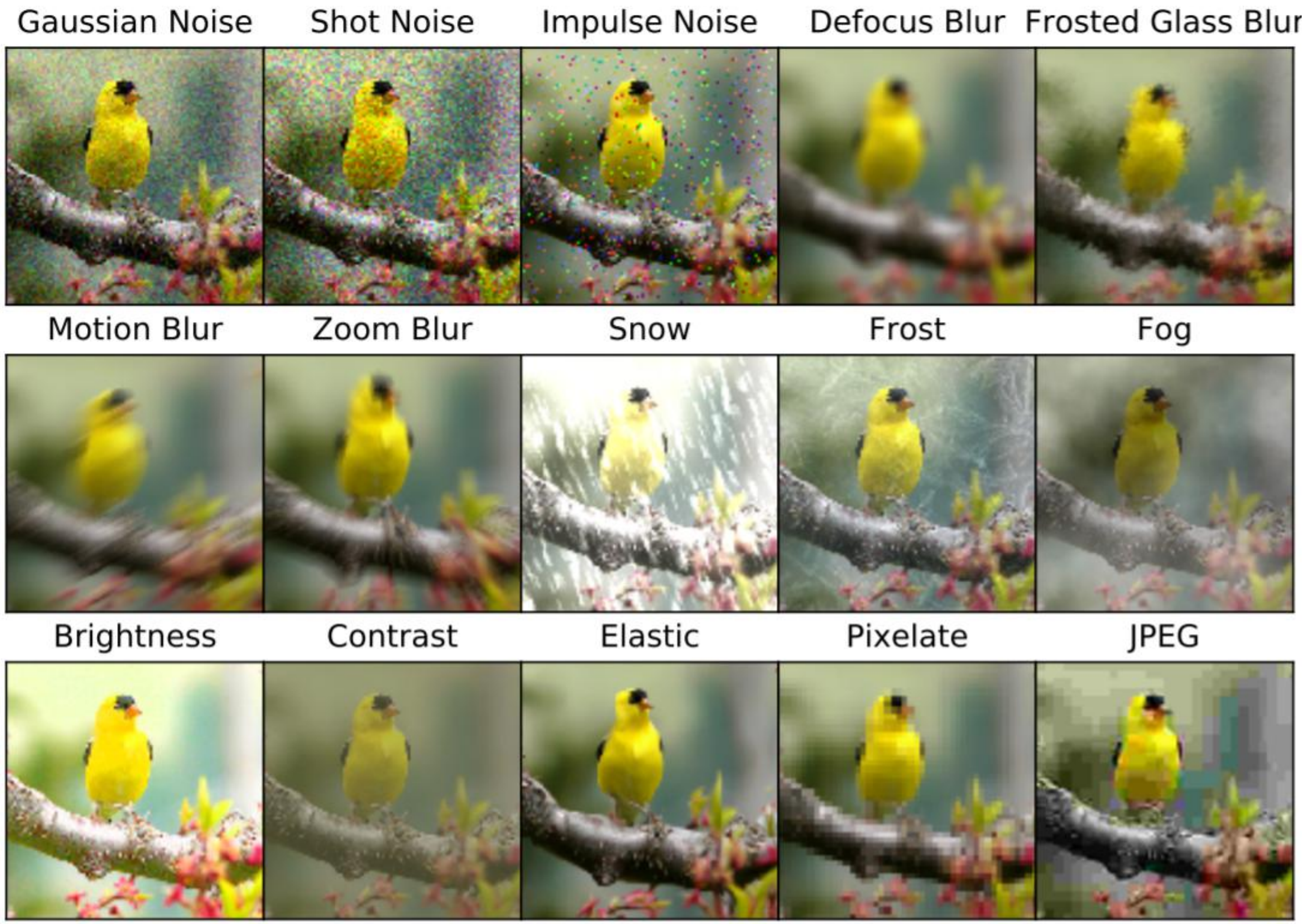
Recognize

Background: Successful Computer Vision



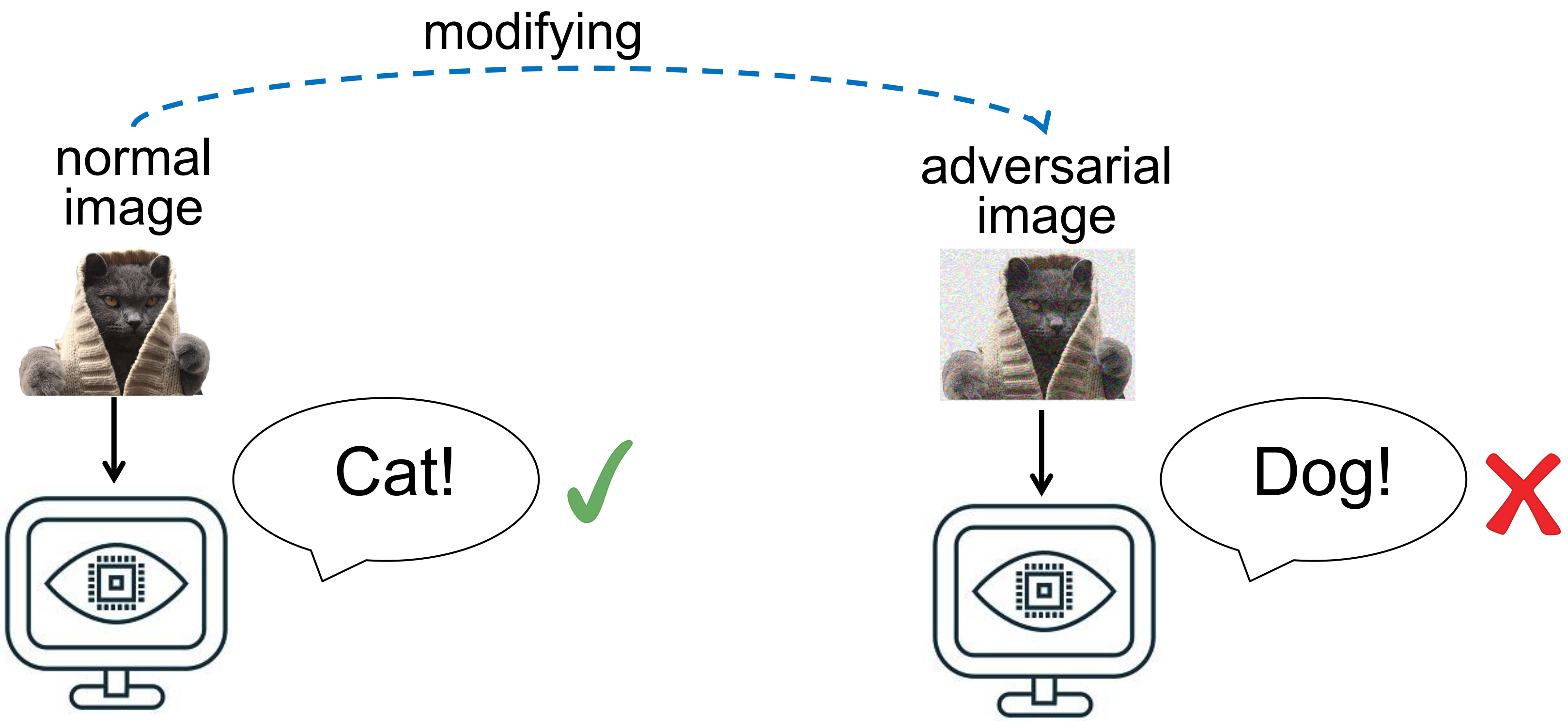
credit: https://www.synopsys.com/designware-ip/technical-bulletin/computer-vision-lab-life.html

Background: Failed Computer Vision



Abnormal images

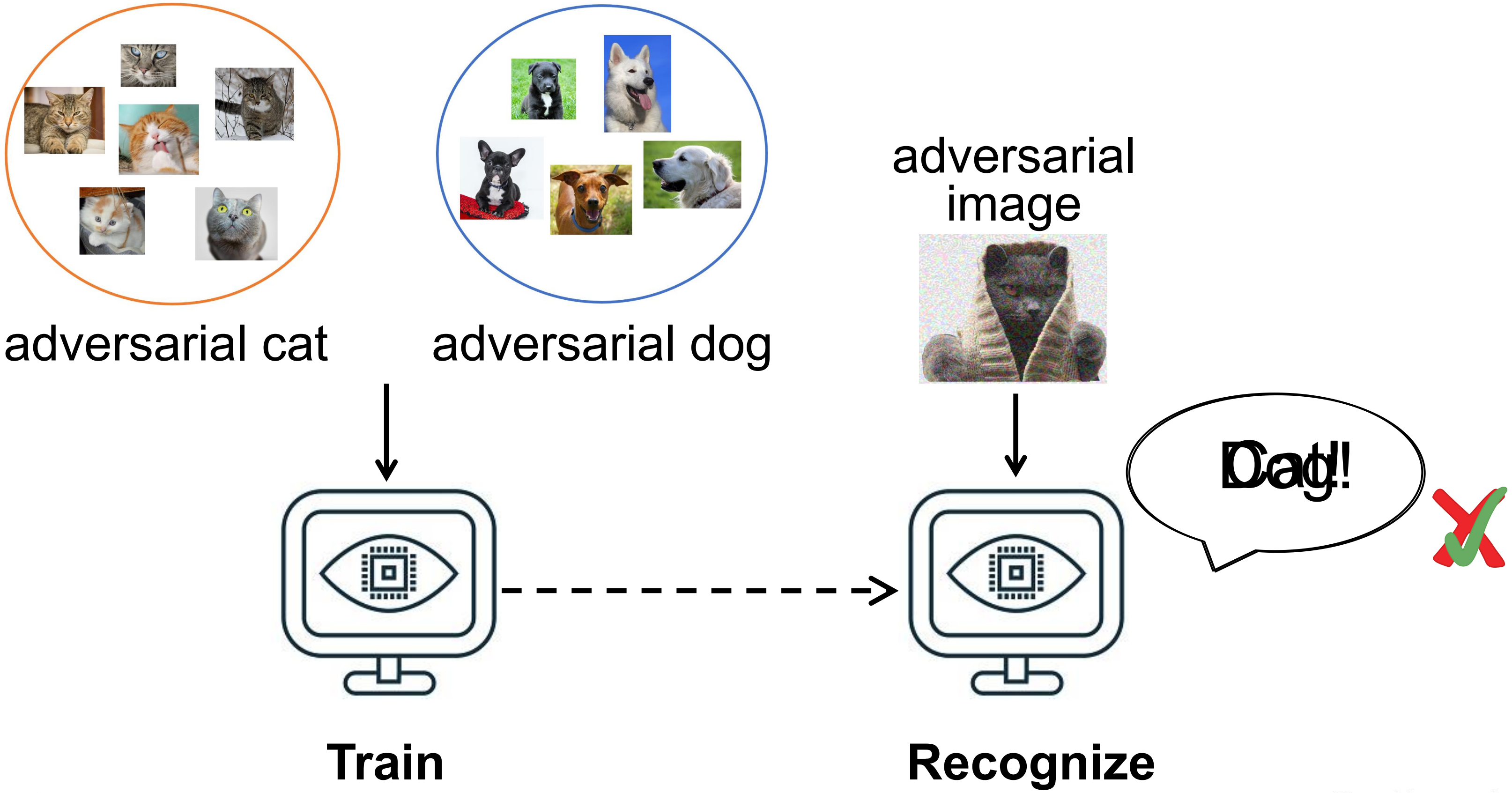
Failed Computer Vision: Adversarial Images (对抗图像)



Adversarial Images: Motivations

Improve **good** computer vision:

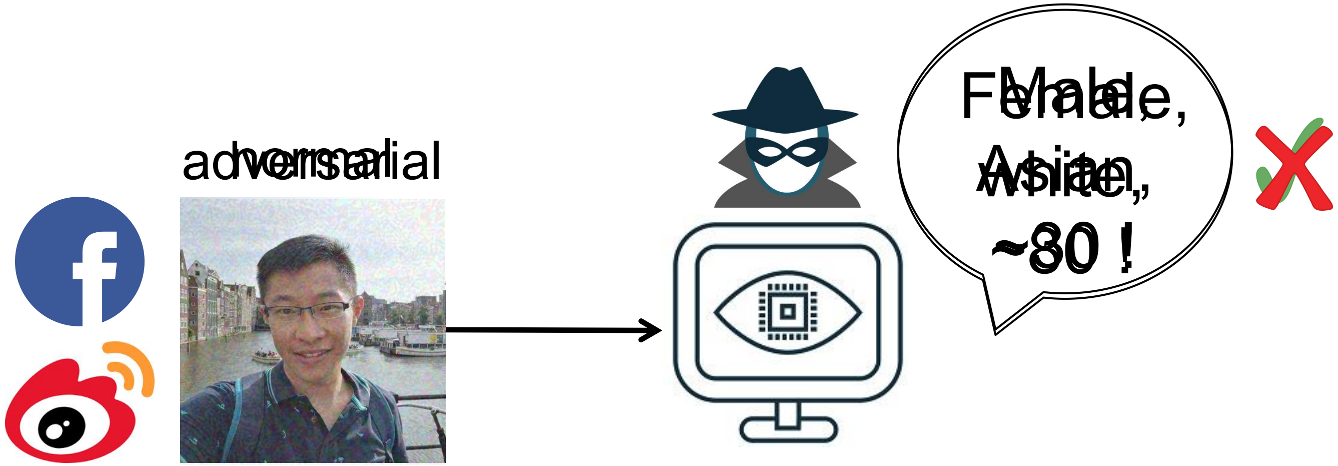
Weaken **bad** computer vision:



Adversarial Images: Motivations

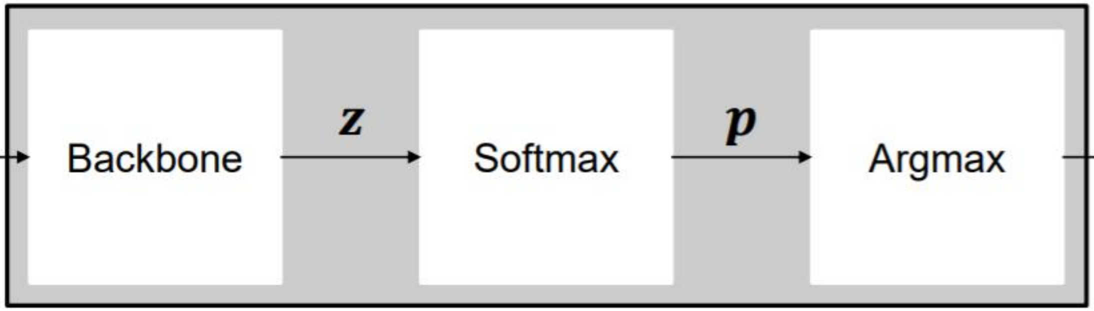
Improve **good** computer vision:

Weaken **bad** computer vision:



Adversarial Images: How to generate?

$x = x_o$



$y = y_o\,(\text{cat})$

Loss function

$$\theta^* = \arg\min_{\theta} J(x, y)$$

Learned parameters

Ground-truth label of ***X***

$$x' = \arg\max_{x} J(x, y_o)$$

Untargeted (无目标): any class other than y_o

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

Targeted (有目标): one specific class y_t

x' $\quad$ x_o

$$\| x' - x_o \|_\infty \le \varepsilon$$

(Targeted) Adversarial Images: Optimization

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

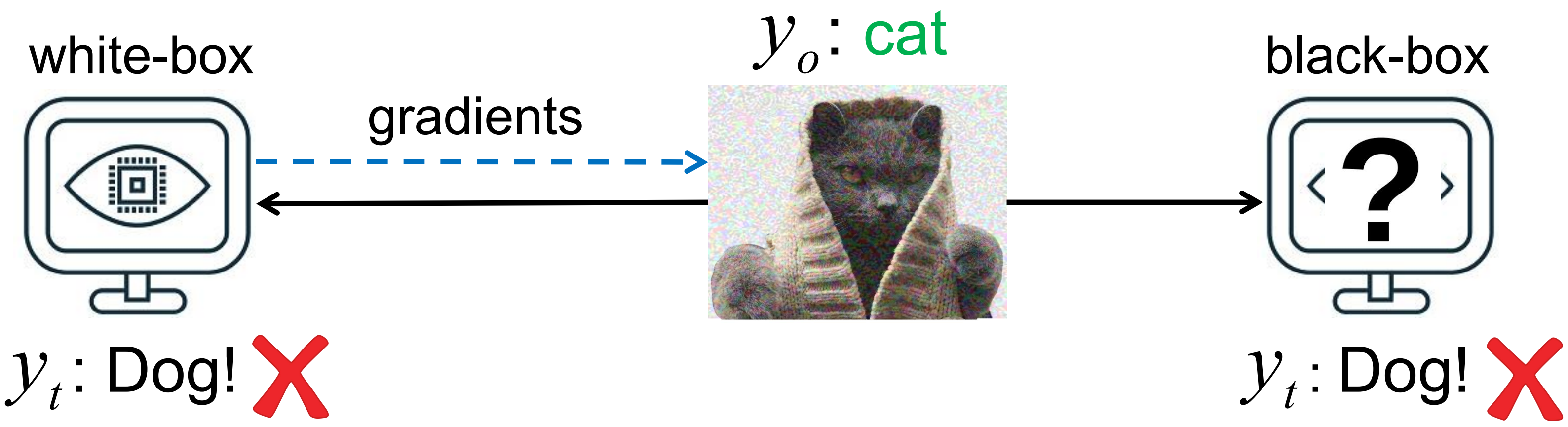
Optimization: Iterative-Fast Gradient Sign Method (I-FGSM)[1]

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

$$x'_{i+1} \leftarrow \mathrm{clip}(x'_{i+1} - x_o, -\varepsilon, \varepsilon)$$

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

(Targeted) Adversarial Images: Transferability



Targeted Transferability via Iterative Methods

Iterative-Fast Gradient Sign Method (I-FGSM)[1]: $x'_0 = x_o, \;\; x'_{i+1} = x'_i - \alpha \cdot \mathrm{sign}(\nabla_x J(x'_i, y_t))$

Improve transferability

- Gradient stabilization[2,3]

  e.g. momentum-based[2]:

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

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

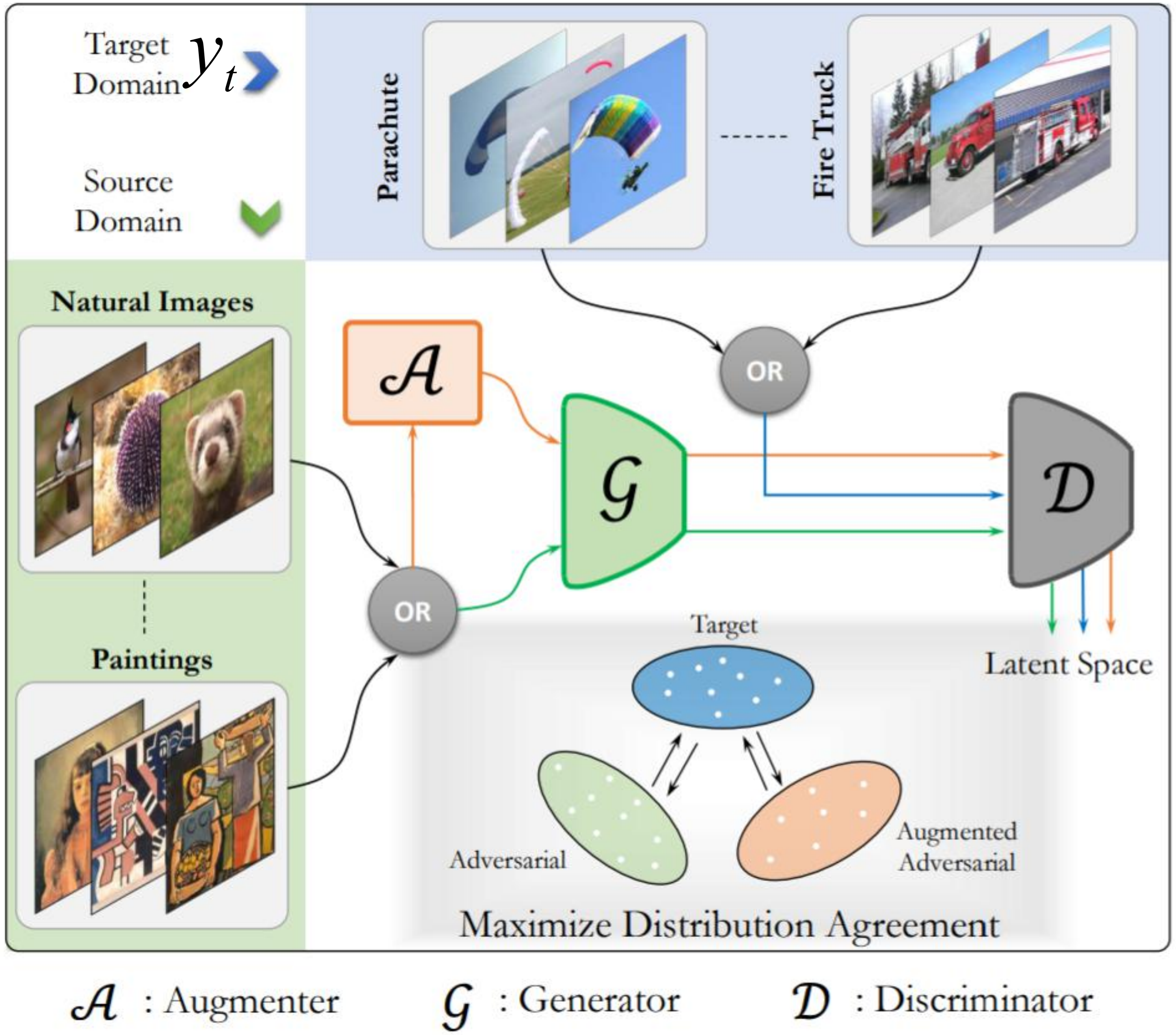
- Input augmentation[4,5,6]

  e.g. random transformations[5]:

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

1. Kurakin et al. *Adversarial Examples in the Physical World*. ICLR workshop 2017
2. Dong et al. *Boosting Adversarial Attacks with Momentum*. CVPR 2018.
3. Lin et al. *Nesterov Accelerated Gradient and Scale Invariance for Adversarial Attacks*. ICLR 2020
4. Dong et al. *Evading Defenses to Transferable Adversarial Examples by Translation-Invariant Attacks*. CVPR 2019
5. Xie et al. *Improving Transferability of Adversarial Examples with Input Diversity*. CVPR 2019
6. Wang et al. *Admix: Enhancing the transferability of adversarial attacks.* ICCV, 2021.

Targeted Transferability via Generative Methods



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

Iterative vs. Generative Methods

	Iterative methods	vs.	Generative methods
• Data:	Single Input image		Massive additional Data
• Model:	Single target-agnostic model		Multiple target-specific GANs

- (Targeted) Transferability: Iterative methods **<<** Generative methods

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

New Insights into Iterative Methods: Conclusion

- (Targeted) Transferability: Iterative methods $\cancel{<}$ Generative methods

$>$

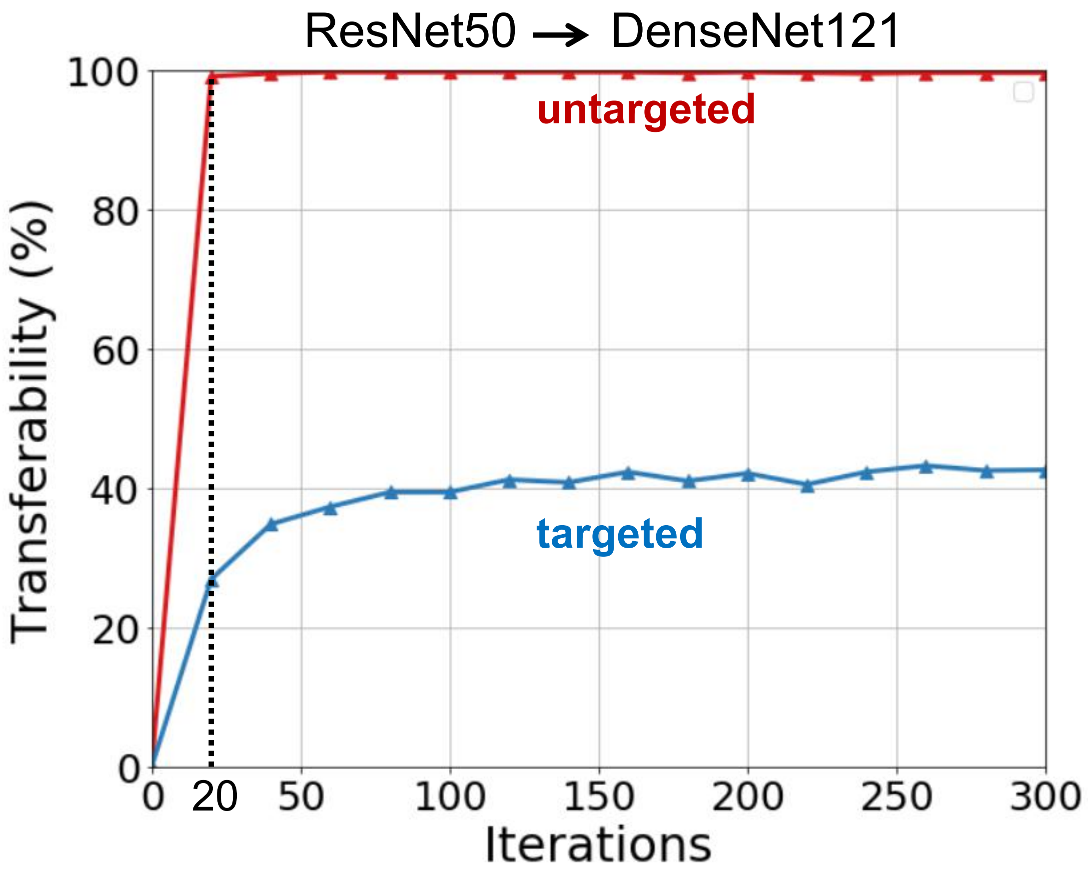
$$\left\| x' - x_o \right\|_\infty \leq \varepsilon$$

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**

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

New Insights into Iterative Methods: More Iterations



Few (≤ 20) iterations in the literature:

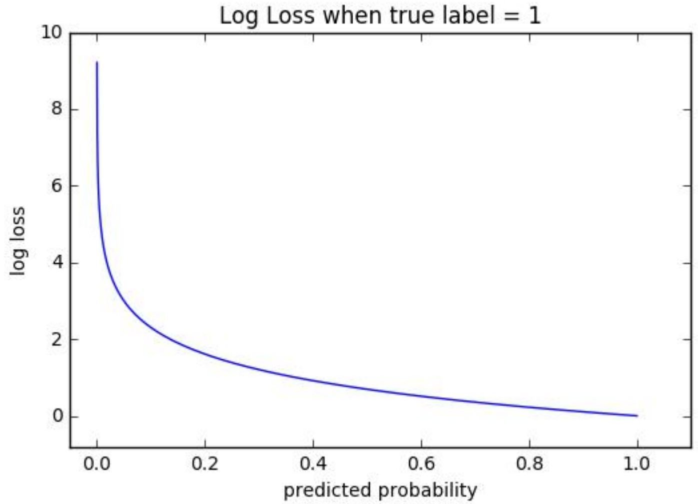
- not converge to optimal
- unrealistic iteration budget.

New Insights into Iterative Methods: Better Loss

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

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

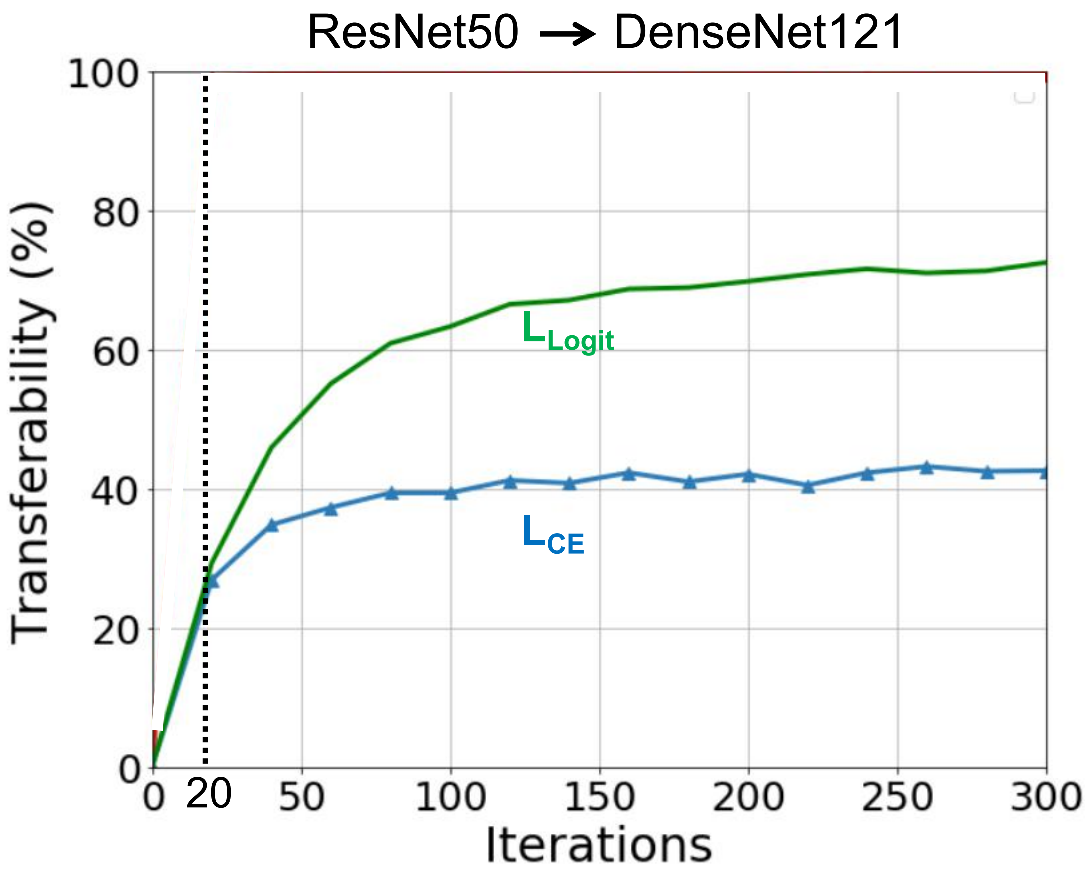
$$\frac{\partial L_{CE}}{\partial z_t} = -1 + \frac{\partial \log(\sum e^{z_j})}{\partial e^{z_t}} \cdot \frac{\partial e^{z_t}}{\partial z_t} = -1 + \frac{e^{z_t}}{\sum e^{z_j}} = -1 + p_t.$$



Logit loss (L_{Logit}) is better:

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

New Insights into Iterative Methods: Better Loss



New Insights into Iterative Methods: Better Evaluation

More challenging&realistic scenarios:

1. Model diversity

2. Target class diversity

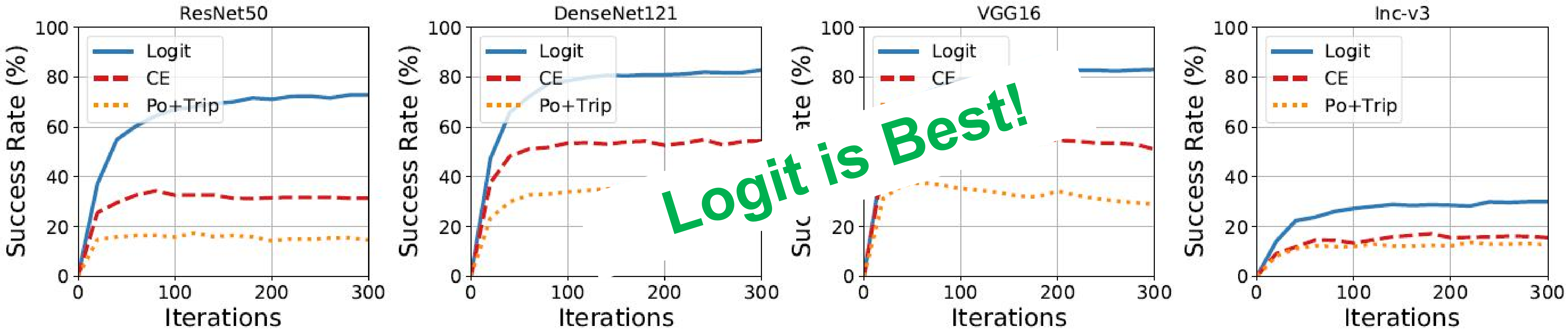
New Insights into Iterative Methods: Better Evaluation

1. Model diversity

Attack	-Inc-v3	-Inc-v4	-IncRes-v2	[illegible]	-Res101	-Res152	Average
CE	85.3	83.3	[illegible]	[illegible]	**93.2**	90.7	87.7
Po+Trip	84.4	82.4	[illegible]	85.0	87.9	85.7	84.4
Logit	**85.5**	**85.8**	[illegible]	**90.0**	91.4	**90.8**	**88.1**

Saturation!

↓

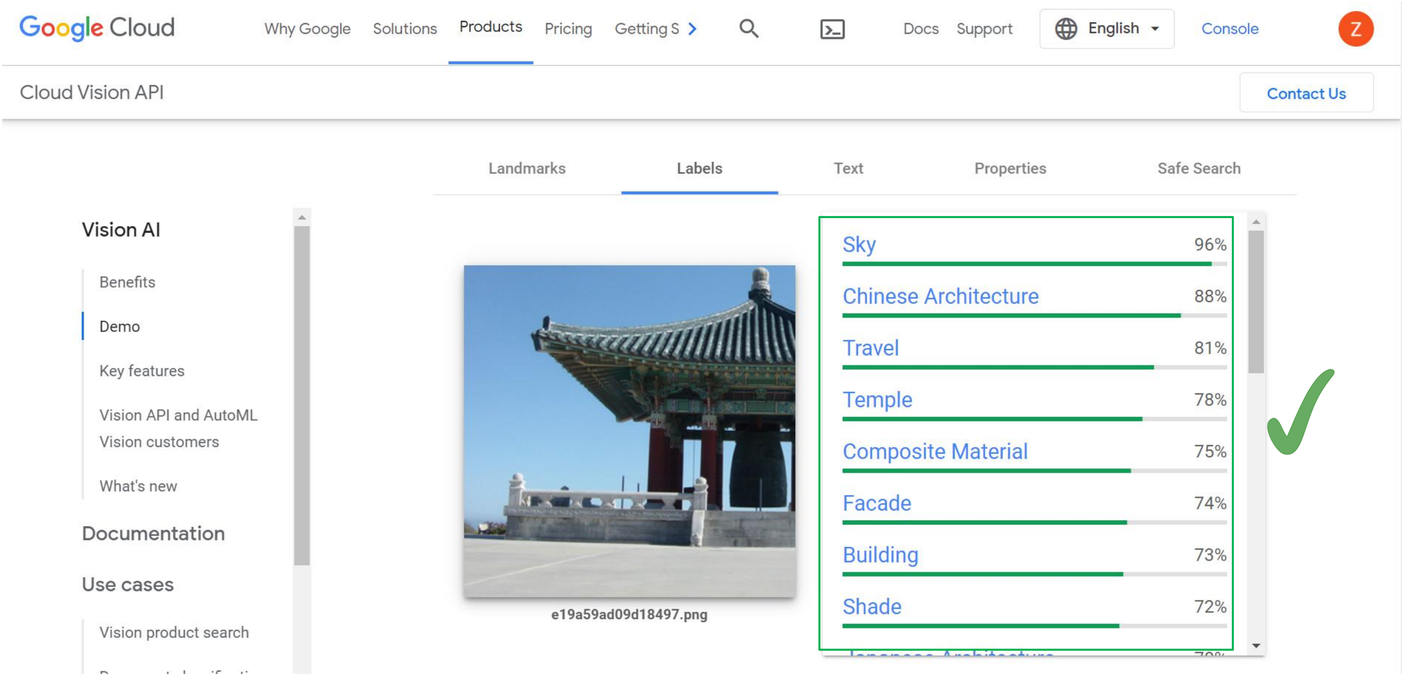


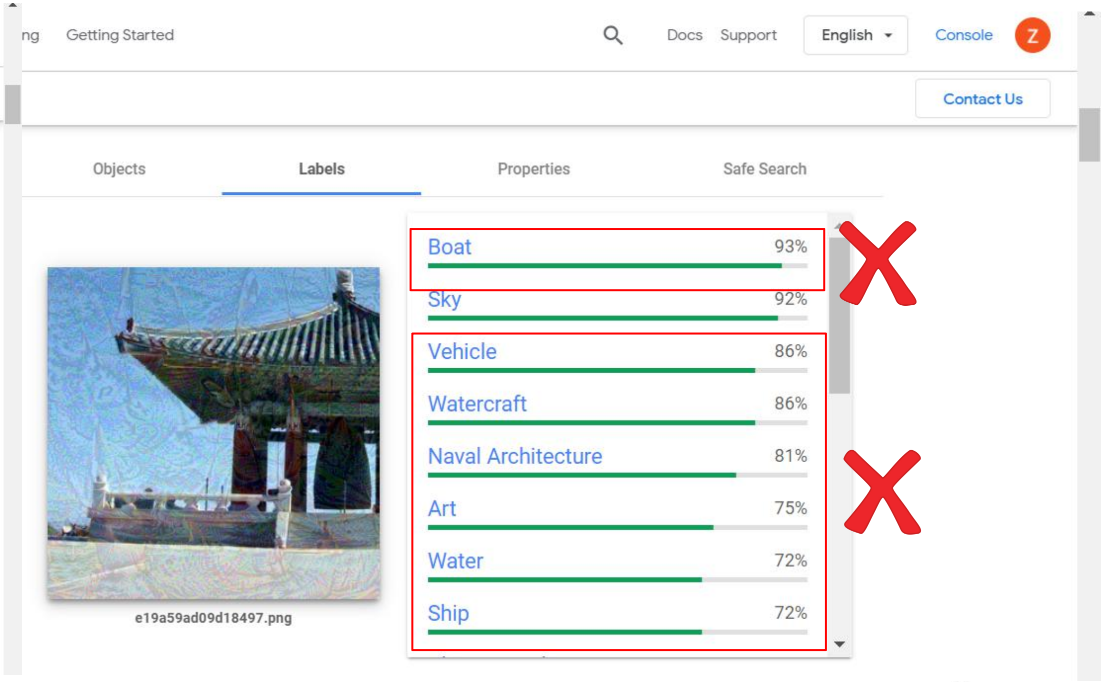
New Insights into Iterative Methods: Better Evaluation

1. Model diversity

Results on Google Cloud Vision API

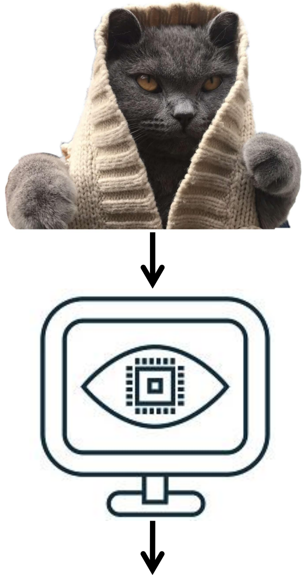
	CE	Po+Trip	Logit
Transferability (%)	7	8	**18**





New Insights into Iterative Methods: Better Evaluation

2. Target class diversity



y_t

1st: tabby cat 66%
2nd: tiger cat 28%
3rd: Egyptian cat 5%
...
1000th: airplane 0.1%

Transferability (%) when varying the target class y_t.

Attack	2nd	10th	200th	500th	800th	1000th
CE	**89.9**	76.7	49.7	43.1	37.0	25.1
Po+Trip	82.6	77.6	58.4	53.6	49.1	38.2
Logit	83.8	**81.3**	**75.0**	**71.0**	**65.1**	**52.8**

The further the target is, the more difficult it is to transfer.

Logit is best.

Summary

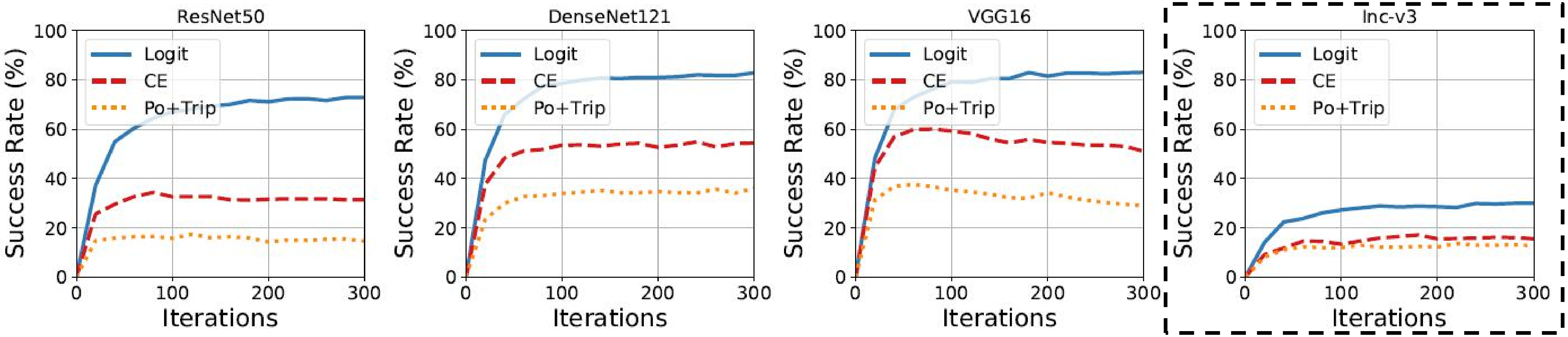
- (Targeted) Transferability: Iterative methods ~~<<~~ Generative methods

  >

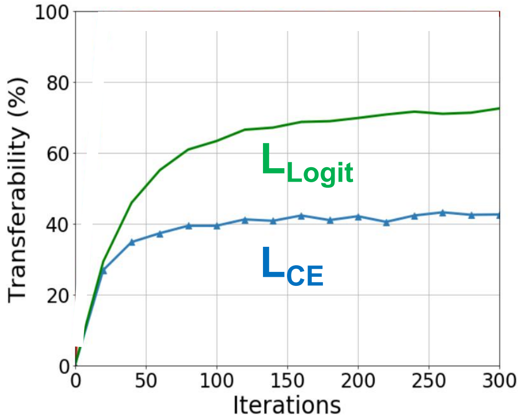
  - More iterations
  - Better loss: Logit
- Better evaluation (More challenging&realistic scenarios)
  - Model diversity
  - Target class diversity

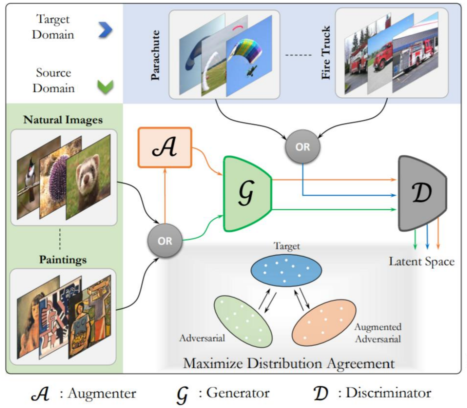
Future Work

1. Improve targeted transferability on specific models (Inception).



2. Speed up iterative methods with generative priors.





论文: https://arxiv.org/abs/2012.11207
代码: https://github.com/ZhengyuZhao/Targeted-Tansfer
个人主页: https://zhengyuzhao.github.io/