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 (NeurIPS 2021)

March 3rd, 2022

About Me

CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY

Radboud University

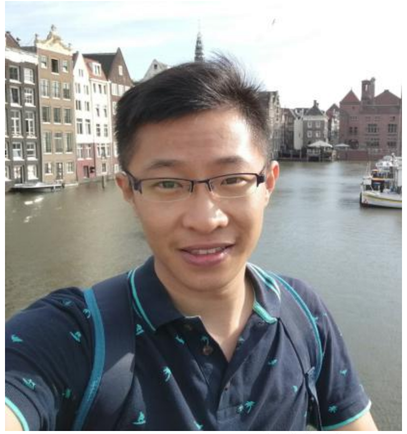
- 个人经历

  2021.12 - present: **Postdoc** @ CISPA (亥姆霍兹信息安全中心), Germany

  2017.09 - 2022.02: **PhD** @ Radboud University, Netherlands

- 研究兴趣

  Security & Privacy in Computer Vision: Adversarial (Image) Examples, Data Poisoning (训练数据投毒), Membership Inference (训练成员推理).



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

Adversarial Examples

Success rate

Transferability

Stealthy

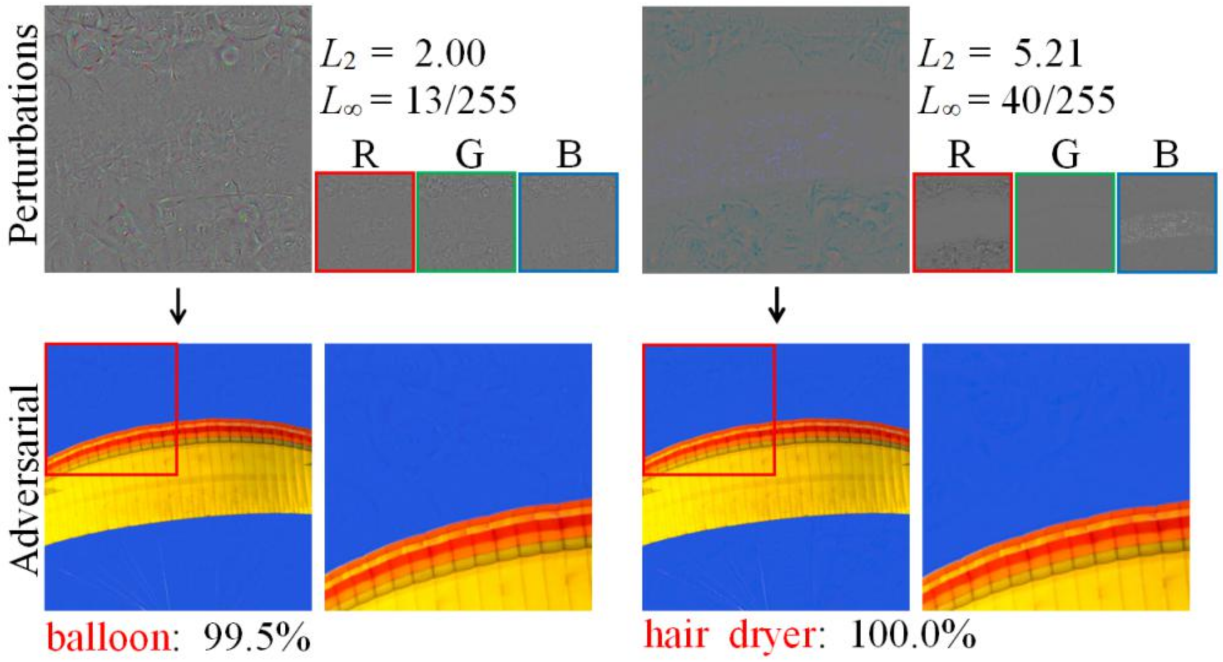
Imperceptible

Perceptible yet natural

Adversarial Examples

Success rate → Transferability

Stealthy → **Imperceptible** | Perceptible yet natural



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

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

Adversarial Examples

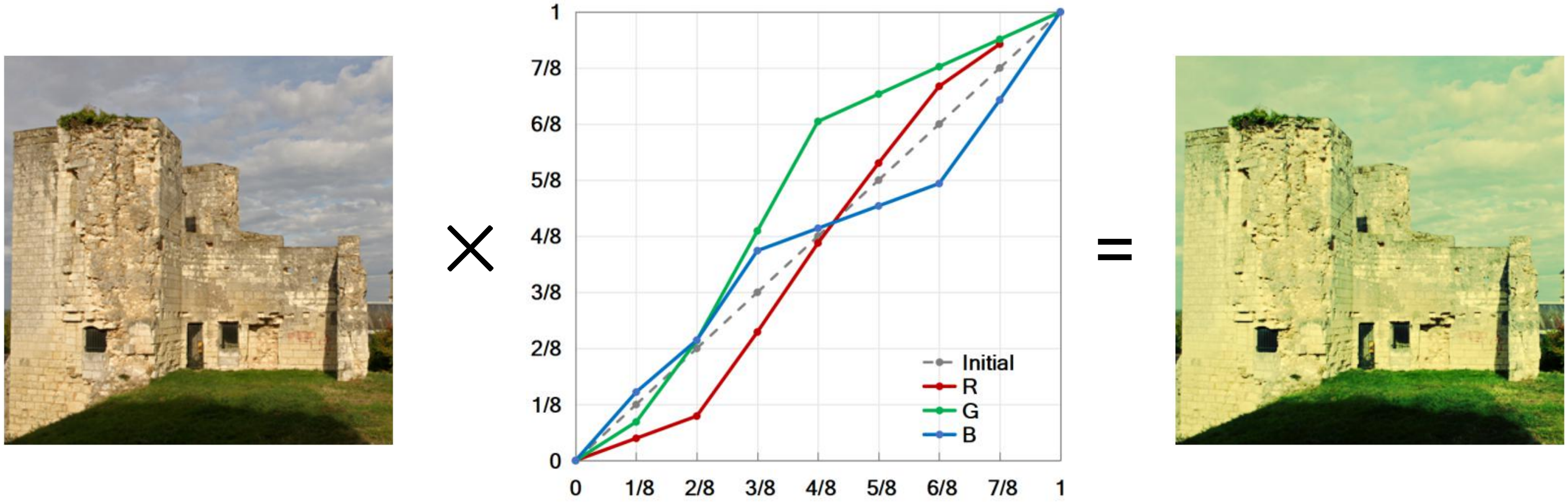
Success rate

Stealthy

Transferability

Imperceptible

Perceptible yet Natural



Zhengyu Zhao, Zhuoran Liu, Martha Larson, "*Adversarial Robustness Against Human-Interpretable Image Color Transformations within Principled Color Filter Space*", Under Revision, IEEE T-PAMI. Preliminary version published at BMVC 2020.

Adversarial Examples

Success rate

Transferability

Stealthy

Imperceptible

Perceptible yet natural

On Success and Simplicity: A Second Look at Transferable Targeted Adversarial Images

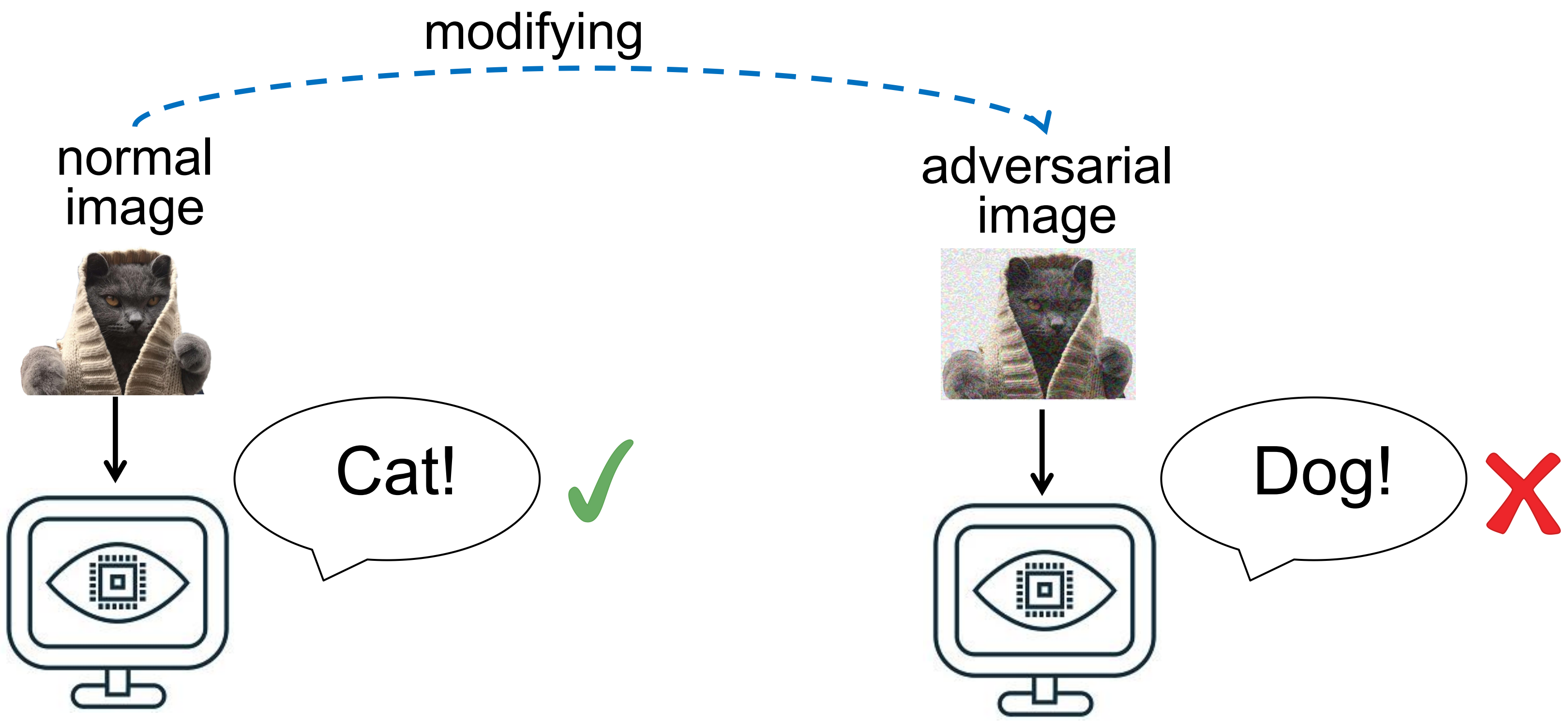
Zhengyu Zhao (赵正宇), Zhuoran Liu, Martha Larson.

NeurIPS 2021

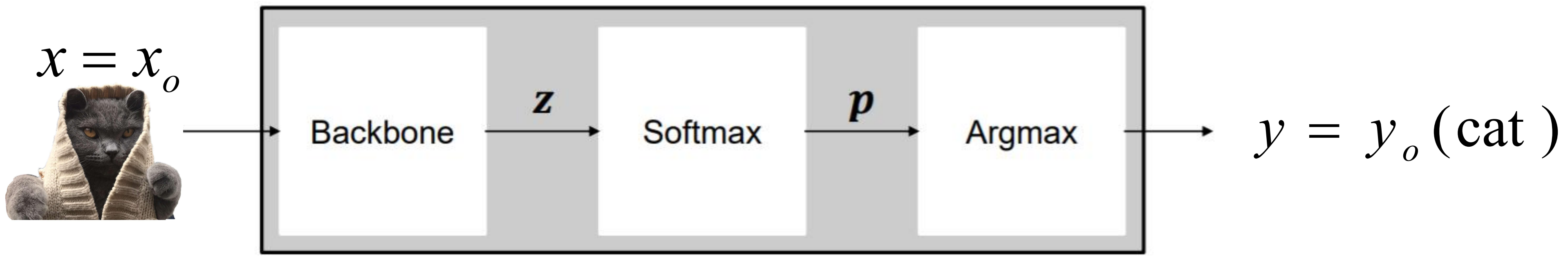
Outline

- Recap adversarial image and targeted transferability
- Our work on revisiting targeted transferability
- Summary & future challenges

Adversarial Images: Definition



Adversarial Images: Formulation



Loss function

$$\theta^* = \arg\min_{\theta} J(x_o, y_o) \qquad x' = \arg\min_{x} J(x, \underline{y_t})$$

Learned parameters

Ground-truth label of $\boldsymbol{X}$

Target class

x' $\qquad$ x_o

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

Adversarial Images: Iterative Optimization

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

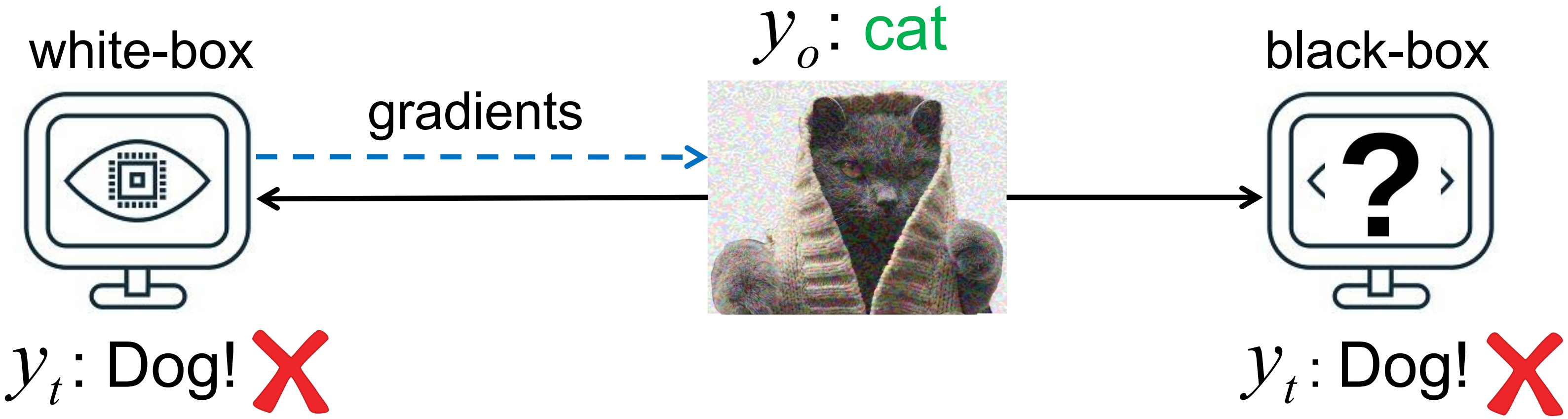
Optimization: Projected Gradient Descent (PGD)

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

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

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

Adversarial Images: (Targeted) Transferability



Targeted Transferability via Iterative Approach

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

Improve transferability

- Gradient stabilization[1,2]

  e.g. momentum-based[1]:

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

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

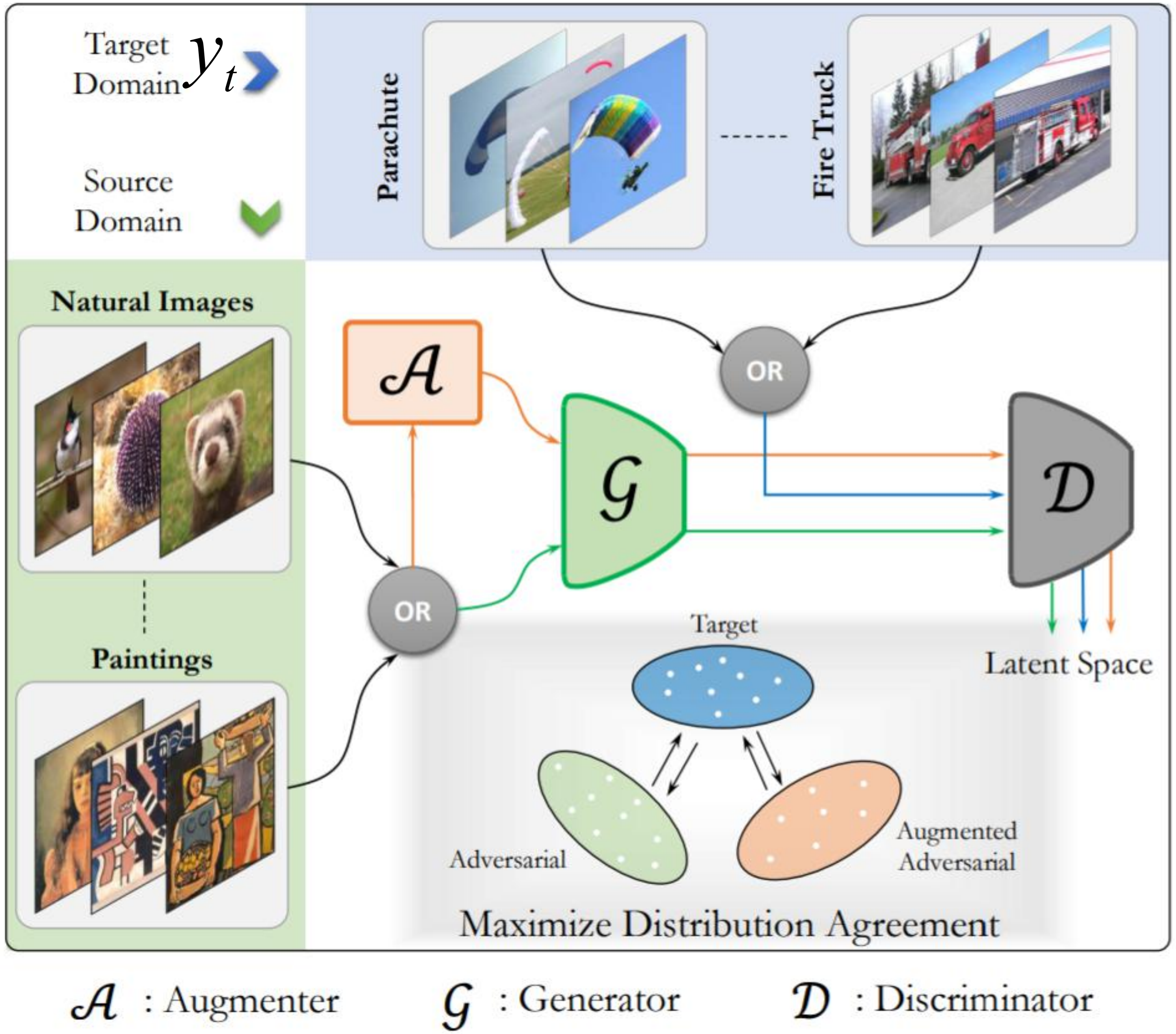
- Input augmentation[3,4,5]

  e.g. random resizing & padding[4]:

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

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

Targeted Transferability via Generative Approach



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

Iterative vs. Generative Approaches

	Iterative	VS	**Generative**
• Data:	Single Input image		Massive training data
• Model:	$1 \times$ target-agnostic classifier		$1000 \times$ target-specific generators

Targeted Transferability: Iterative approach **<<** Generative approach

Revisiting Targeted Transferability: Key Message

Targeted Transferability: Iterative approach $\cancel{<}$ Generative approach

$>$

$\Big\|$ 

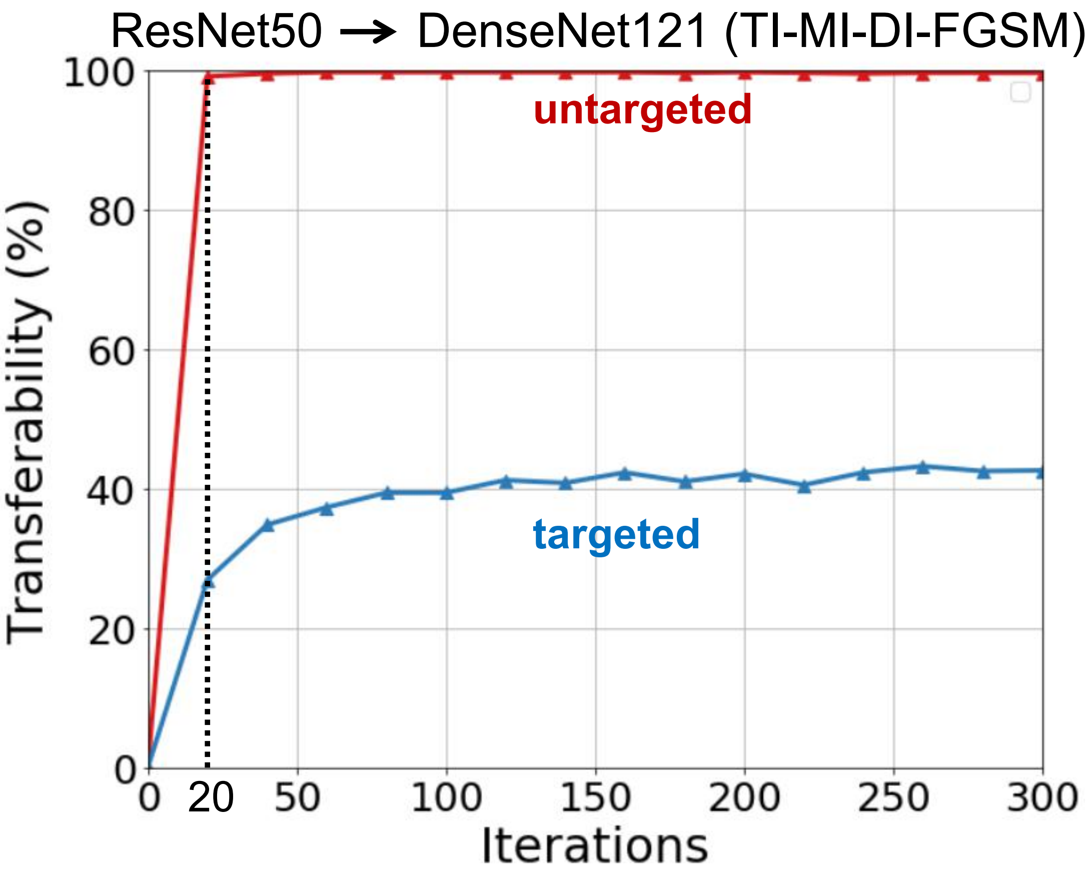
$-$ 

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

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**

Revisiting Targeted Transferability: More Iterations



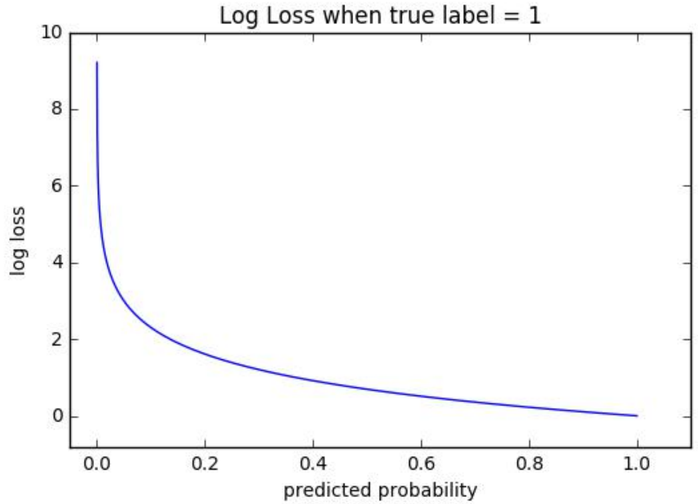
Few (≤ 20) iterations in the literature:

- not converge to optimal performance
- unrealistic iteration budget

Revisiting Targeted Transferability: Better Loss

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

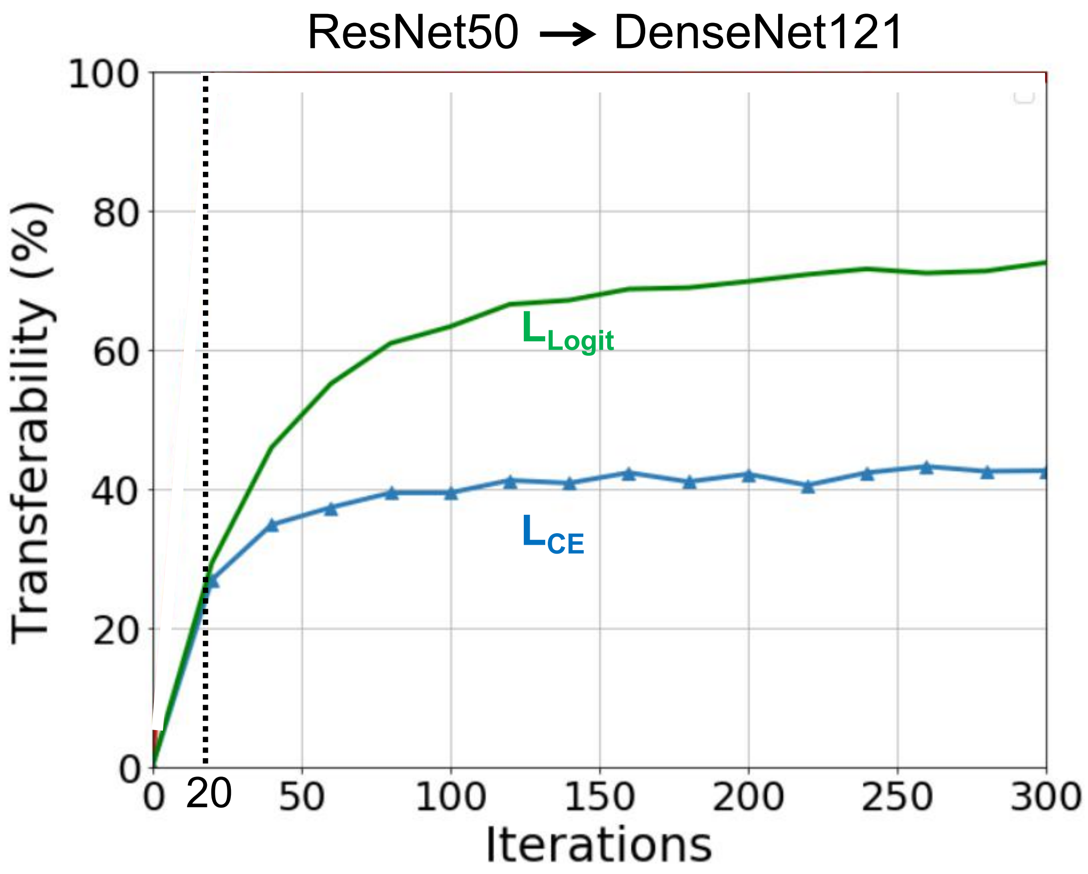
$$L_{CE} = -1 \cdot \log(p_t) = -\log(\frac{e^{z_t}}{\sum e^{z_j}}) = -z_t + \log(\sum e^{z_j}),$$
$$\frac{\partial L_{CE}}{\partial z_t} = -1 + \frac{\partial \log(\sum e^{z_j})}{\partial e^{z_t}} \cdot \frac{\partial e^{z_t}}{\partial z_t} = -1 + \frac{e^{z_t}}{\sum e^{z_j}} = -1 + p_t.$$



Logit Loss (L_{Logit}) is better:

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

Revisiting Targeted Transferability: Better Loss



Revisiting Targeted Transferability: Better Evaluation

More challenging&realistic scenarios:

1. Model diversity 2. Target class diversity

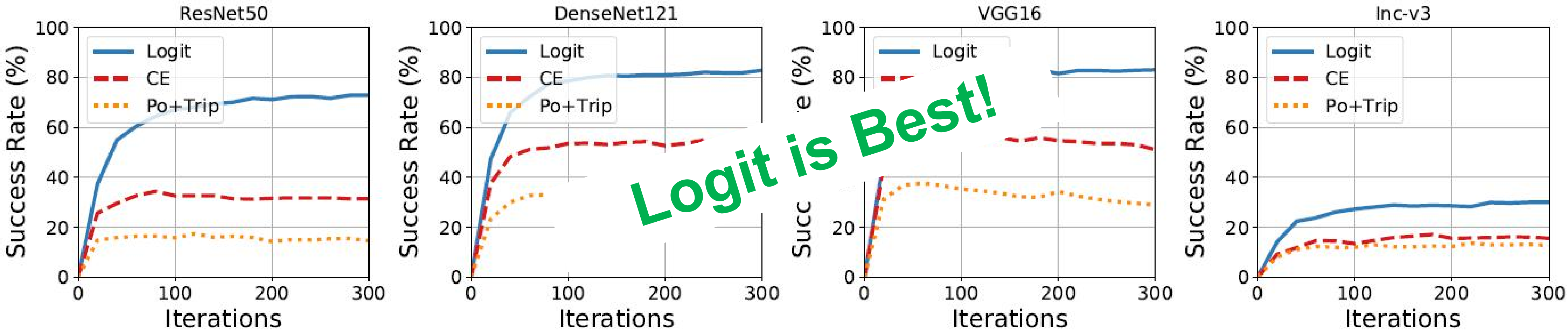
Revisiting Targeted Transferability: Better Evaluation

1. Model diversity　　2. Target class diversity

Attack	-Inc-v3	-Inc-v4	-IncRes-v2	-R[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!

↓

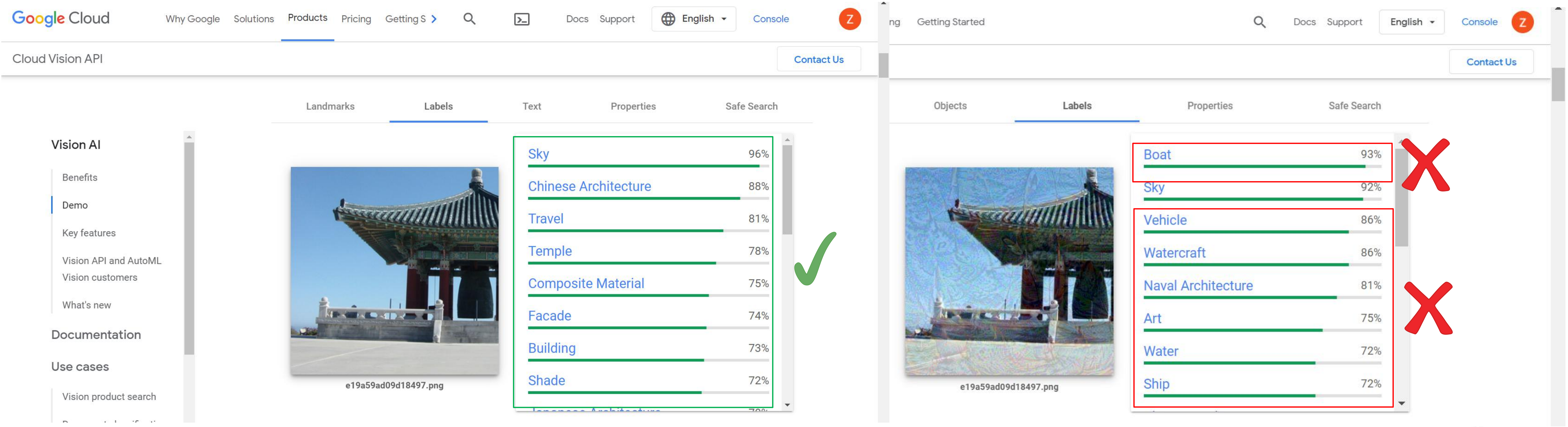


Revisiting Targeted Transferability: Better Evaluation

1. Model diversity   2. Target class diversity

Results on Google Cloud Vision API (100 images)

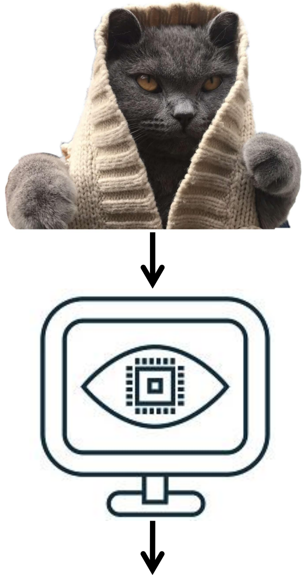
	CE	Po+Trip	Logit
Transferability (%)	7	8	**18**



y_t is “yawl” (a type of boat)

Revisiting Targeted Transferability: Better Evaluation

1. Model diversity   2. Target class diversity



y_t

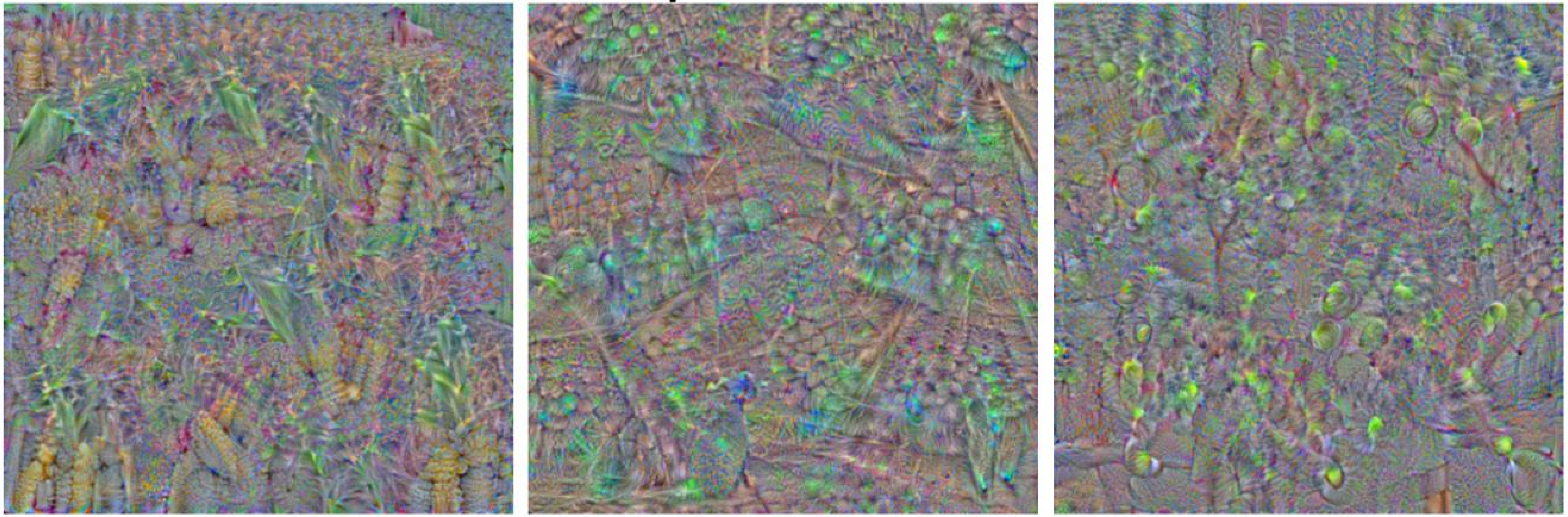
1st: tabby cat 66%
2nd: tiger cat 28%
3rd: Egyptian cat 5%
...
1000th: airplane 0.1%

Targeted transferability (%) from best to worst case

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**

Perturbation Semantics

“corn” “peacock” “tennis ball”

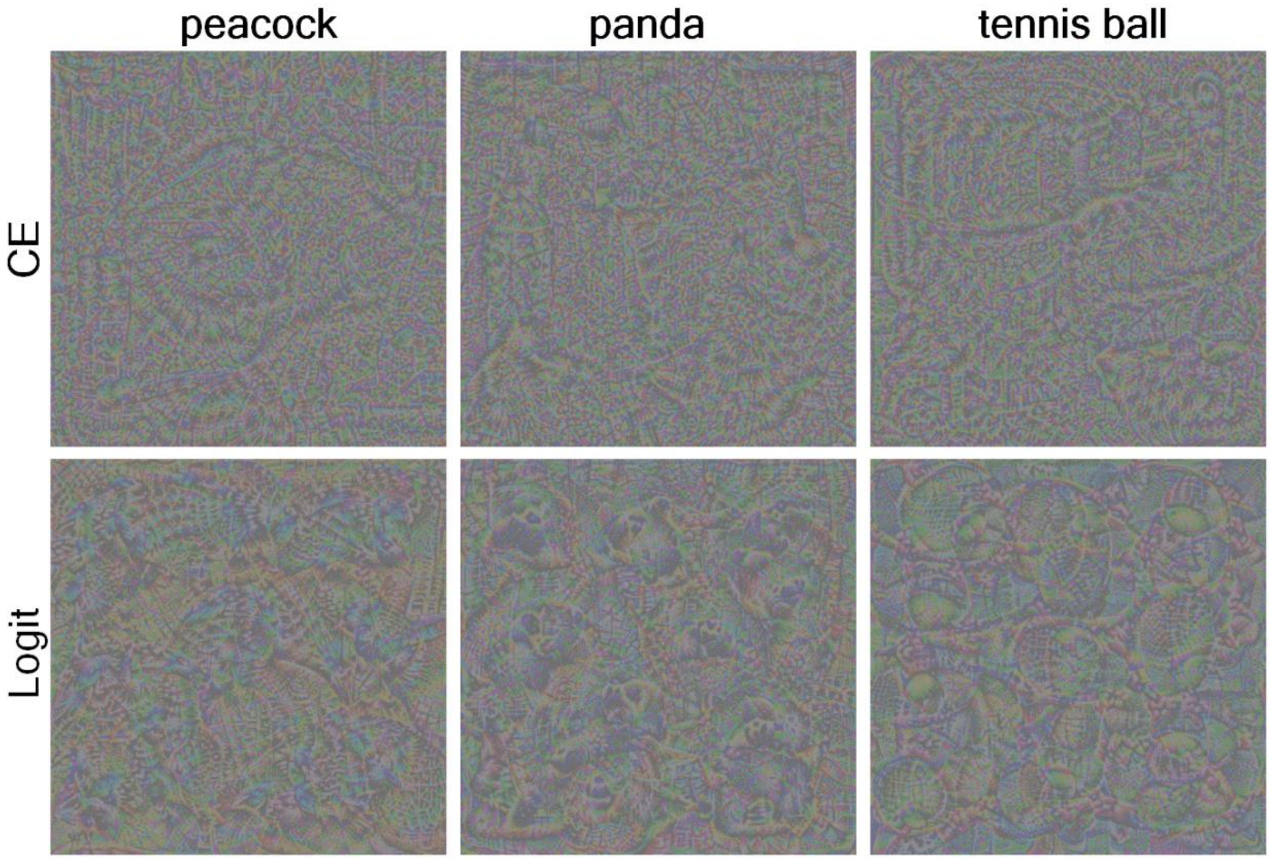


Unbounded Adversarial perturbations

Data/Training-free Targeted UAPs

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**



Summary

- Targeted Transferability: Iterative approach ~~<~~ > Generative approach
  - More iterations
  - Better (Logit) loss
- Better evaluation with more challenging scenarios
  - Model diversity
  - Target class diversity
- Semantic perturbations for UAPs

Follow-ups using our iterative baseline

1. Slightly robust source model; Other target models (e.g. transformer, CLIP)

Springer et al. *A Little Robustness Goes a Long Way: Leveraging Robust Features for Targeted Transfer Attacks*. NeurIPS 2021.

2. Data/Training-free targeted UAPs with a better initialization

Li et al. Learning Universal Adversarial Perturbation by Adversarial Example. AAAI 2022

3. Seeking local worst-case perturbations

Qin et al. Boosting the Transferability of Adversarial Attacks with Reverse Adversarial Perturbation. OpenReview 2022

...

Future Challenge 1

Iterative: lightweight (data-free and model-free) but slow (many iterations).
😄 ☹️

Generative: heavy (additional data and models) but fast (single forward pass).
☹️ 😄

Iterative + Generative: Better source classifier with fewer iterations

1. Wu et al. *Skip Connections Matter: On the Transferability of Adversarial Examples Generated with ResNets*. ICLR 2020.
2. Guo et al. *Backpropagating Linearly Improves Transferability of Adversarial Examples*. NeurIPS 2020.
3. Zhang et al. *Backpropagating Smoothly Improves Transferability of Adversarial Examples*. CVPRW 2021
4. Zhu et al. *Rethinking Adversarial Transferability from a Data Distribution Perspective*. ICLR 2022.

Future Challenge 2

Interpreting & improving targeted transferability on specific architectures:

Attack	Source Model: Res50 →Dense121	→VGG16	→Inc-v3	Source Model: Dense121 →Res50	→VGG16	→Inc-v3
Logit	**29.3/63.3/72.5**	**24.0/55.7/62.7**	**3.0/7.2/9.4**	**17.2/39.7/43.7**	**13.5/35.3/38.7**	**2.7/6.9/7.6**
Attack	Source Model: VGG16 →Res50	→Dense121	→Inc-v3	Source Model: Inc-v3 →Res50	→Dense121	→VGG16
Logit	**3.3/8.7/11.2**	**3.6/11.7/13.2**	**0.2/0.7/0.9**	**0.8**/1.6/**2.9**	**1.2**/2.8/**5.3**	**0.7/2.2/3.7**