Simple Attacks on Deep Networks. Transfer between Networks and Tasks Simple Defenses.

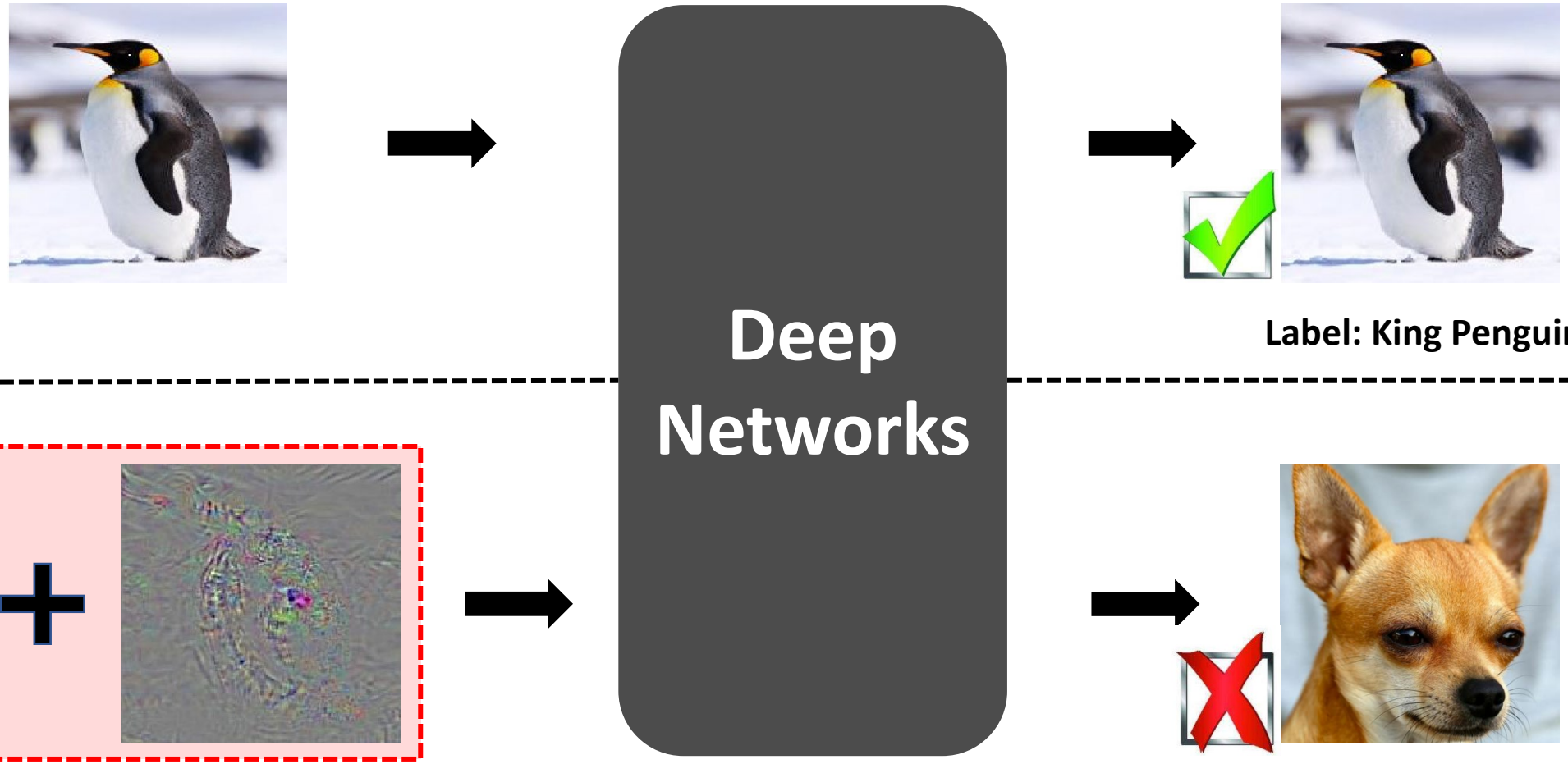
Alan Yuille (with Cihang Xie)
Johns Hopkins University

Deep networks are **Good**

Deep Networks

Label: King Penguin

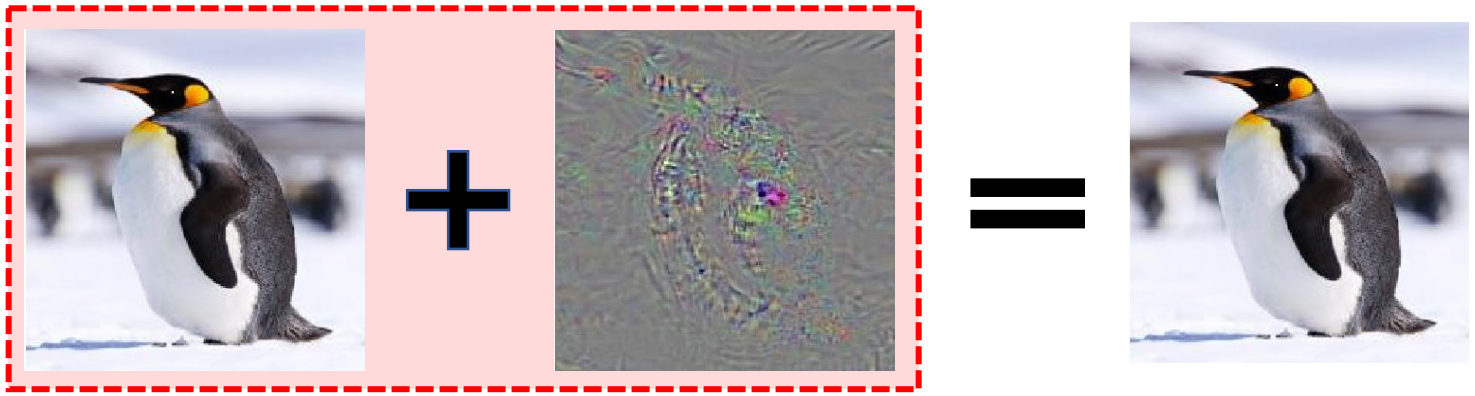
Deep networks are FRAGILE to small & carefully crafted perturbations



Deep networks are FRAGILE to small & carefully crafted perturbations

We call such images as

Adversarial Examples



Generating Adversarial Example is SIMPLE:

$$\textbf{maximize}\ \text{loss}(f(x+\mathbf{r}), y^{\text{true}}; \theta)$$

Maximize the loss function w.r.t. **Adversarial Perturbation r**

Generating Adversarial Example is SIMPLE:

maximize loss(f(x+**r**), y^{true}; θ)

Maximize the loss function w.r.t. **Adversarial Perturbation r**

minimize loss(f(x), y^{true}; **θ**);

Minimize the loss function w.r.t. **Network Parameters θ Standard Training)**

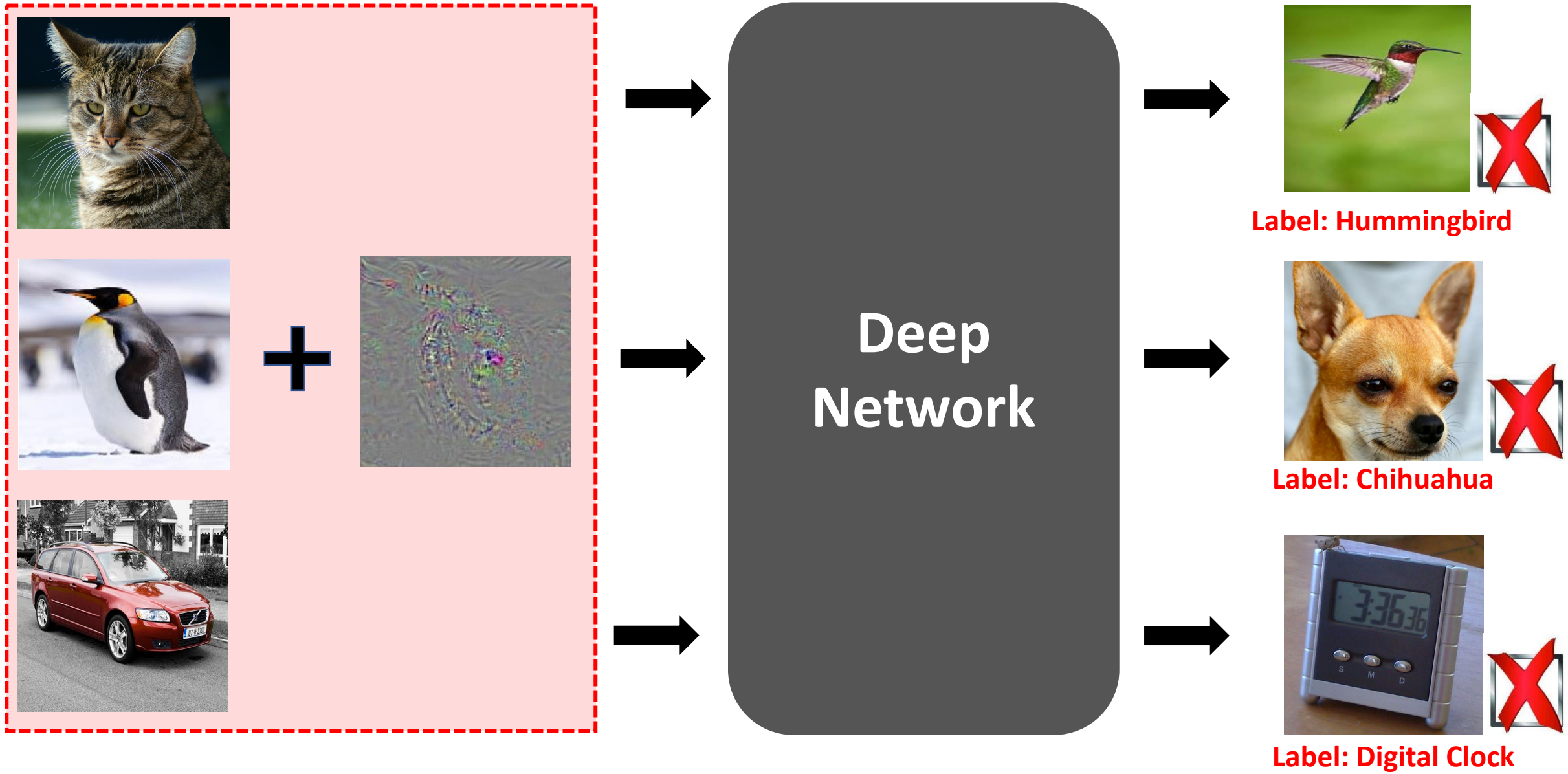
Part I: Simple Adversarial Attacks

- {Image, Model, Task}-Agnostic
- Beyond Pixel Perturbation

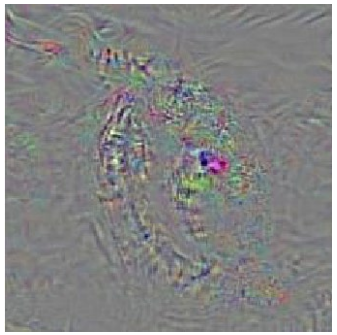
Part I: Simple Adversarial Attacks

- **{Image, Model, Task}-Agnostic**
- Beyond Pixel Perturbation

Adversarial Perturbations can be Image Agnostic



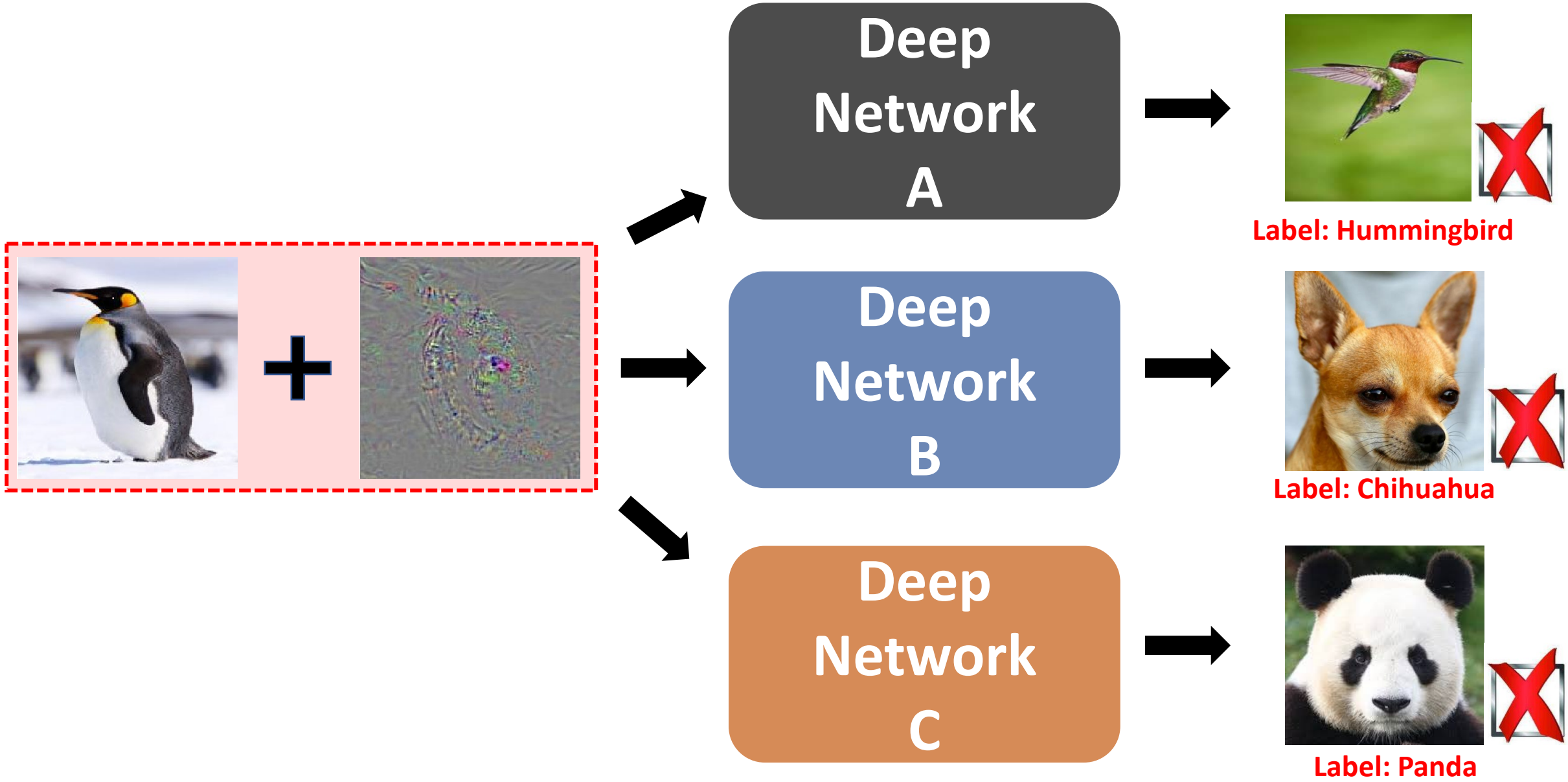
Adversarial Perturbations can be Image Agnostic



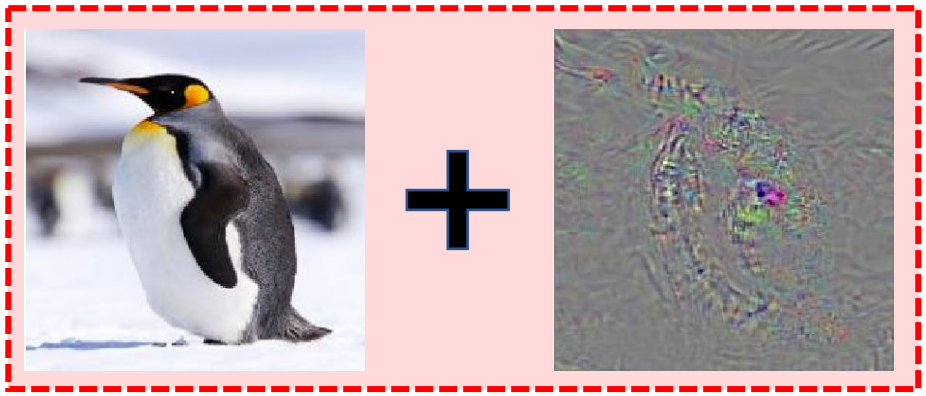
We call such perturbations as

Universal Adversarial Perturbations

Adversarial Examples can be Model Agnostic



Adversarial Examples can be Model Agnostic



We call such images as

Transferable Adversarial Examples

Adversarial Examples can be Task Agnostic

Adversarial examples **EXIST** on different tasks

Adversarial Examples can be Task Agnostic

Adversarial examples **EXIST** on different tasks

semantic segmentation

Adversarial Examples can be Task Agnostic

Adversarial examples **EXIST** on different tasks

semantic segmentation

pose estimation

Adversarial Examples can be Task Agnostic

Adversarial examples **EXIST** on different tasks

semantic segmentation

pose estimation

South Africa's historic Soweto township marks its 100th birthday on Tuesday in a mood of optimism.
57% **World**

South Africa's historic Soweto township marks its 100th birthday on Tuesday in a mooP of optimism.
95% **Sci/Tech**

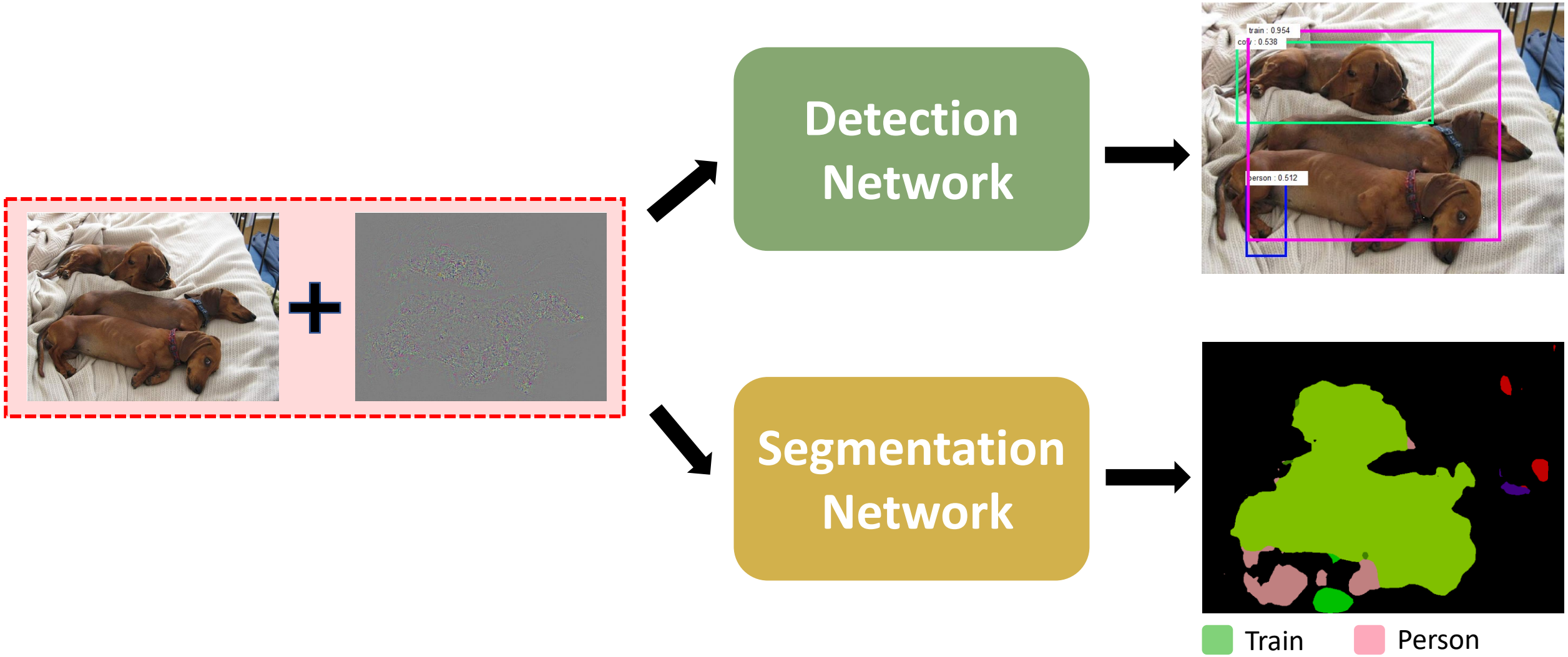
text classification

Adversarial Examples can be Task Agnostic

Adversarial examples **TRANSFER** between different tasks

Adversarial Examples can be Task Agnostic

Adversarial examples **TRANSFER** between different tasks



Quantitative Result of Transferability between Different Models [1]

Model	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-152
Inc-v3	FGSM	64.6%	23.5%	21.7%	21.7%
	I-FGSM	**99.9%**	14.8%	11.6%	8.9%
	DI^2-FGSM (**Ours**)	**99.9%**	35.5%	27.8%	21.4%
	MI-FGSM	**99.9%**	36.6%	34.5%	27.5%
	M-DI^2-FGSM (**Ours**)	**99.9%**	**63.9%**	**59.4%**	**47.9%**

Adversarial examples generated on Inc-v3 can attack Inc-v4, IncRes-v2 and Res-152 with high success rate.

[1] Xie, Cihang, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L. Yuille. "Improving transferability of adversarial examples with input diversity." In *CVPR*, 2019

Quantitative Result of Transferability between Different Models [1]

Model	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-152
Inc-v3	FGSM	64.6%	23.5%	21.7%	21.7%
	I-FGSM	**99.9%**	14.8%	11.6%	8.9%
	DI^2-FGSM (**Ours**)	**99.9%**	35.5%	27.8%	21.4%
	MI-FGSM	**99.9%**	36.6%	34.5%	27.5%
	M-DI^2-FGSM (**Ours**)	**99.9%**	**63.9%**	**59.4%**	**47.9%**

Adversarial examples generated on Inc-v3 can attack Inc-v4, IncRes-v2 and Res-152 with high success rate.

This transfer phenomenon indicates that

Different Networks Learn Similar Representations.

[1] Xie, Cihang, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L. Yuille. "Improving transferability of adversarial examples with input diversity." In *CVPR*, 2019

Simple Attacks are Agnostic to Network and Task

- This is insightful and has many motivated important research independent of attacks.

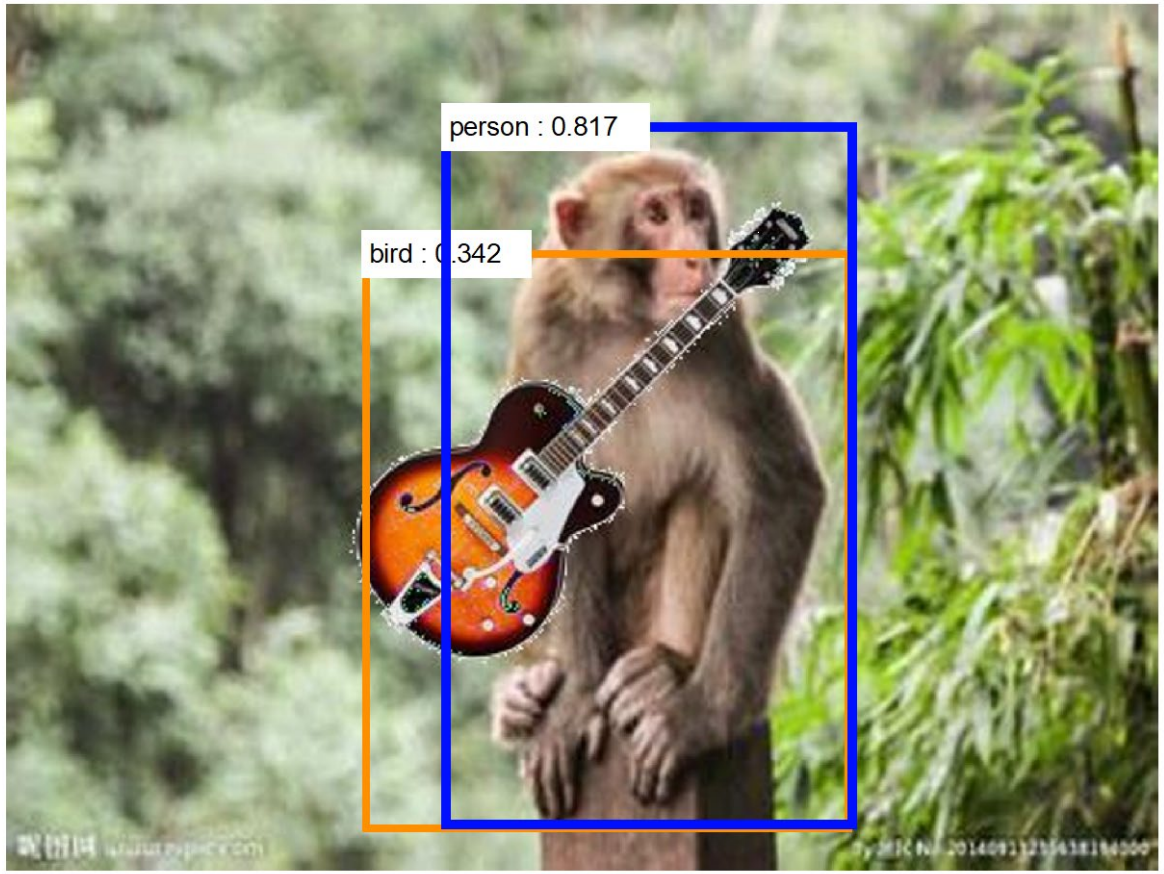
- Deep networks can be trained unsupervised on proxy tasks. The resulting hierarchical feature vectors form a backbone. Simple heads (single or multilevel perceptrons) can be trained on these features to perform different tasks.

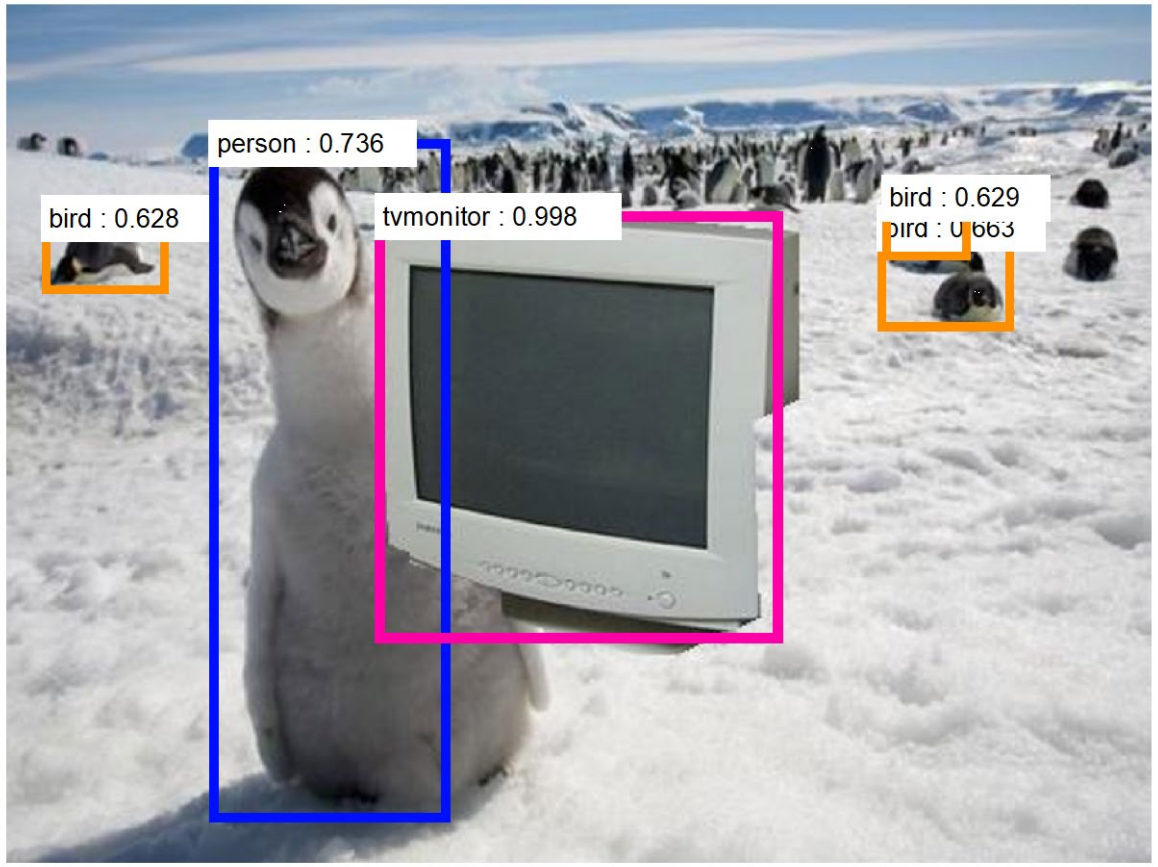
- Unsupervised learning can exploit the enormous amounts of unannotated data. Foundation Models trained on huge datasets to perform many tasks.

Part I: Adversarial Attacks

- {Image, Model, Task}-Agnostic
- **Beyond Pixel Perturbation**

Beyond Pixel Perturbations --- Patch Attacks

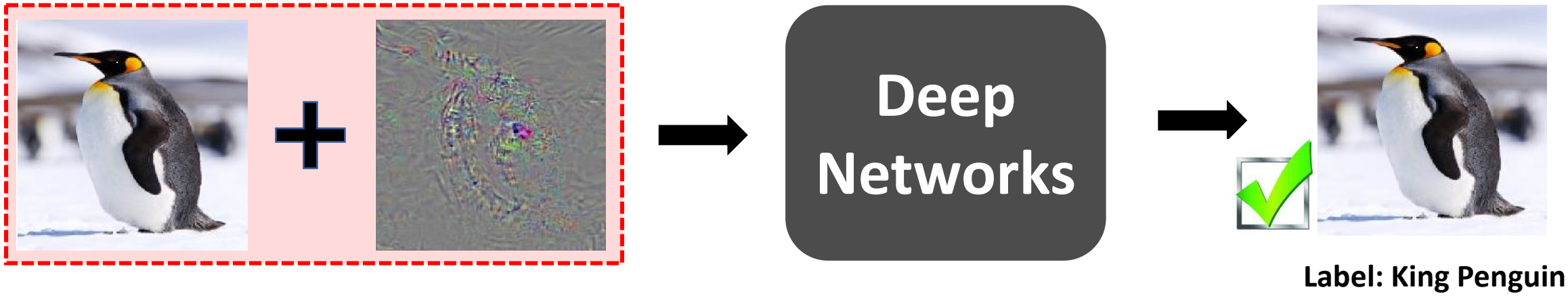




[4] Wang, Jianyu, Zhishuai Zhang, Cihang Xie, et al. "Visual concepts and compositional voting." In *Annals of Mathematical Sciences and Applications.* 2018 .

Part II: Adversarial Defense

- Min-Max learning .
- Denoising the network features



Min-Max Learning

- The natural defense is to find adversarial examples and retrain the deep network using them.
- This can be elegantly formalized by a min-max formulation (Madry et al).

- Replace standard learning:
- *min loss(f(x), ytrue; ϑ); with respect to ϑ*

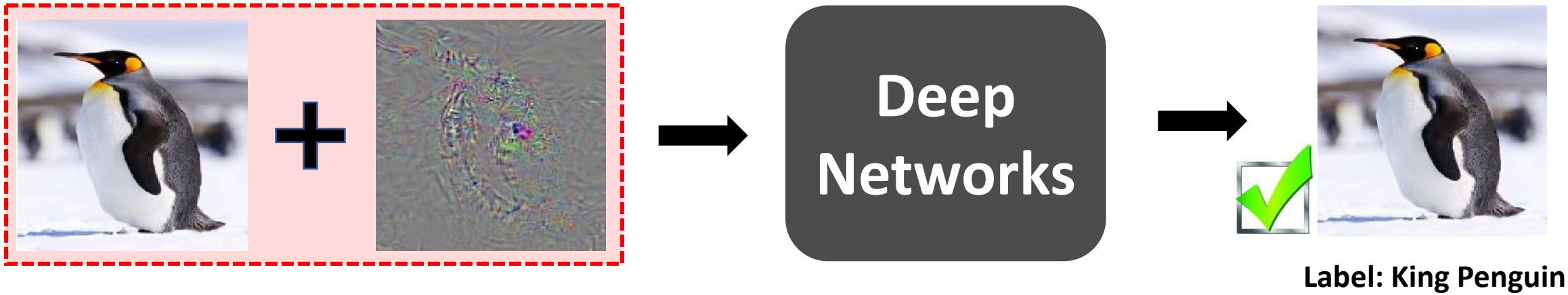
- By min-max learning:
- *min _ϑ max _r loss(f(x+r), ytrue; ϑ); min wrt ϑ and max wrt r.*

- Metaphor: datapoints are tiny islands (x) given territorial waters (radius r).
- Min max is requires nontrivial algorithms (game theory)

Denoising the Features

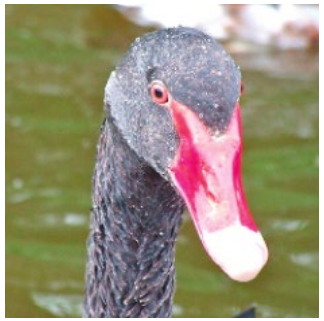
- **Robust Network Features**

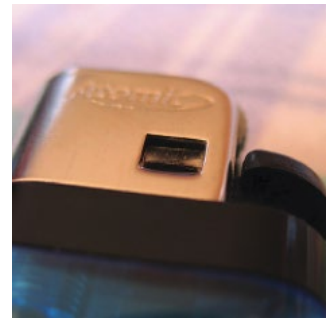
want to **learn** feature representations robust to adversarial images



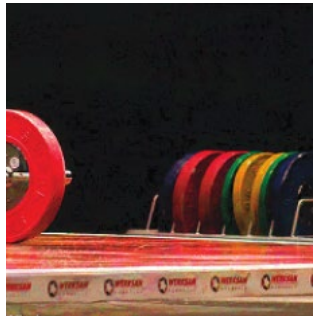
Observation: Adversarial perturbations are SMALL on the pixel space

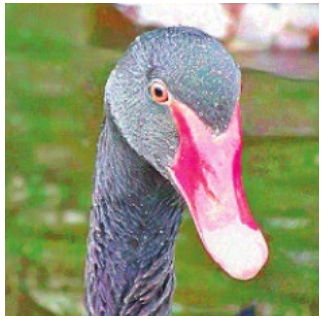
Clean





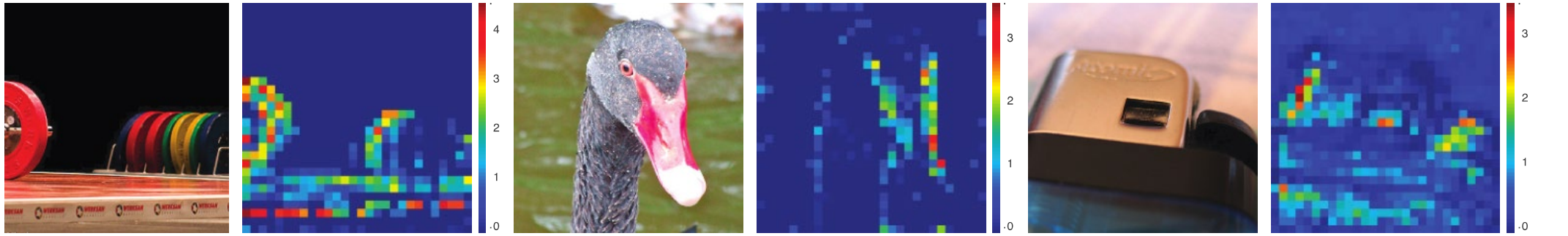
Adversarial



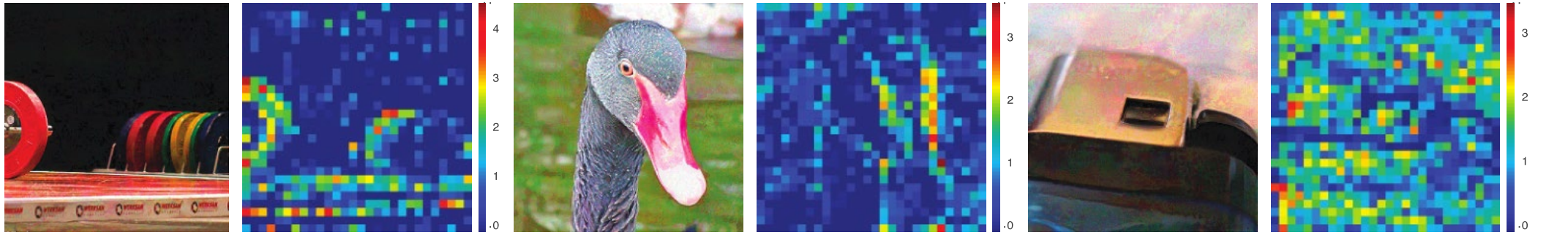


Observation: Adversarial perturbations are BIG in the feature space

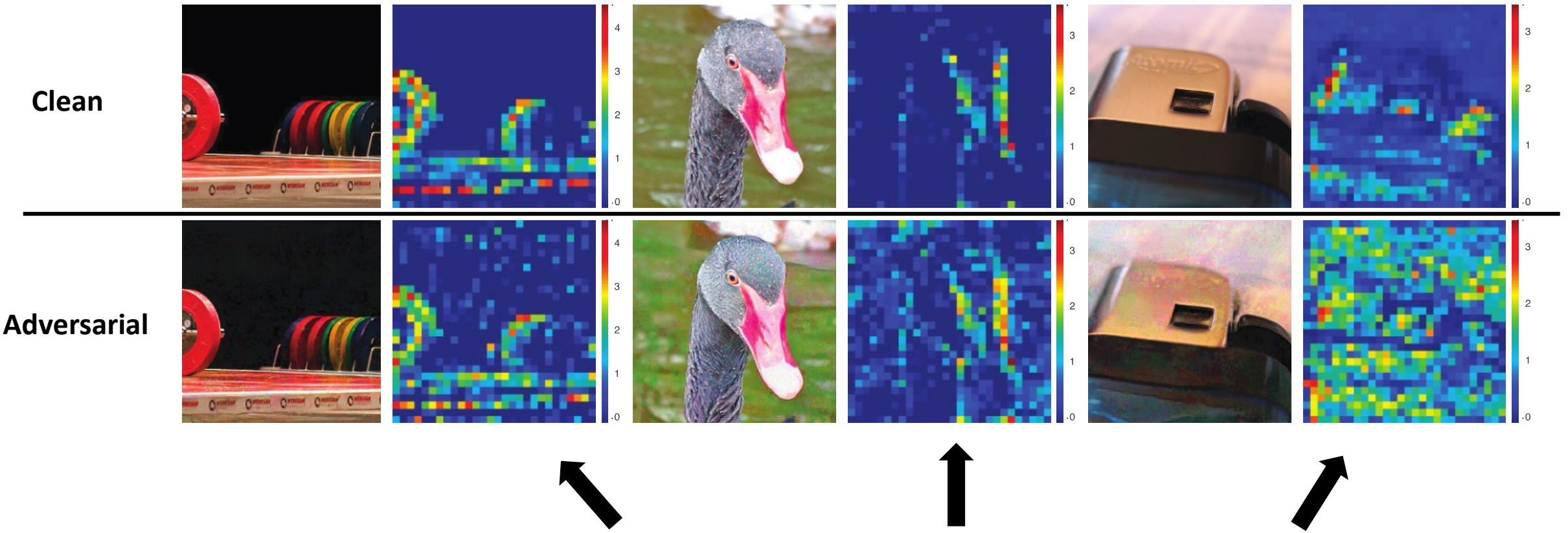
Clean



Adversarial



Observation: Adversarial perturbations are BIG in the feature space



One Solution: Denoising at feature level (C. Xie et al).

Traditional Image Denoising Operations:

Local filters (predefine a local region $\Omega(i)$ for each pixel i):

- Bilateral filter $y_i = \frac{1}{C(x_i)} \sum_{\forall j \in \Omega(i)} f(x_i, x_j) x_j$
- Median filter $y_i = median\{\forall j \in \Omega(i) : x_j\}$
- Mean filter $y_i = \frac{1}{C(x_i)} \sum_{\forall j \in \Omega(i)} x_j$

Non-local filters (the local region $\Omega(i)$ is the whole image I):

- Non-local means $y_i = \frac{1}{C(x_i)} \sum_{\forall j \in I} f(x_i, x_j) x_j$

Training Strategy: Adversarial training

- Core Idea: train with adversarial examples

- Implementation: distributed on 128 GPUs, 32 images per GPU
  (since finding adversarial examples is computationally expensive)

Two Ways for Evaluating Robustness

Defending Against White-box Attacks

- Attackers know everything about models
- Directly maximize loss(f(x+**r**), y^{true}; θ)

Two Ways for Evaluating Robustness

Defending Against White-box Attacks

- Attackers know everything about models
- Directly maximize loss(f(x+**r**), y^{true}; θ)

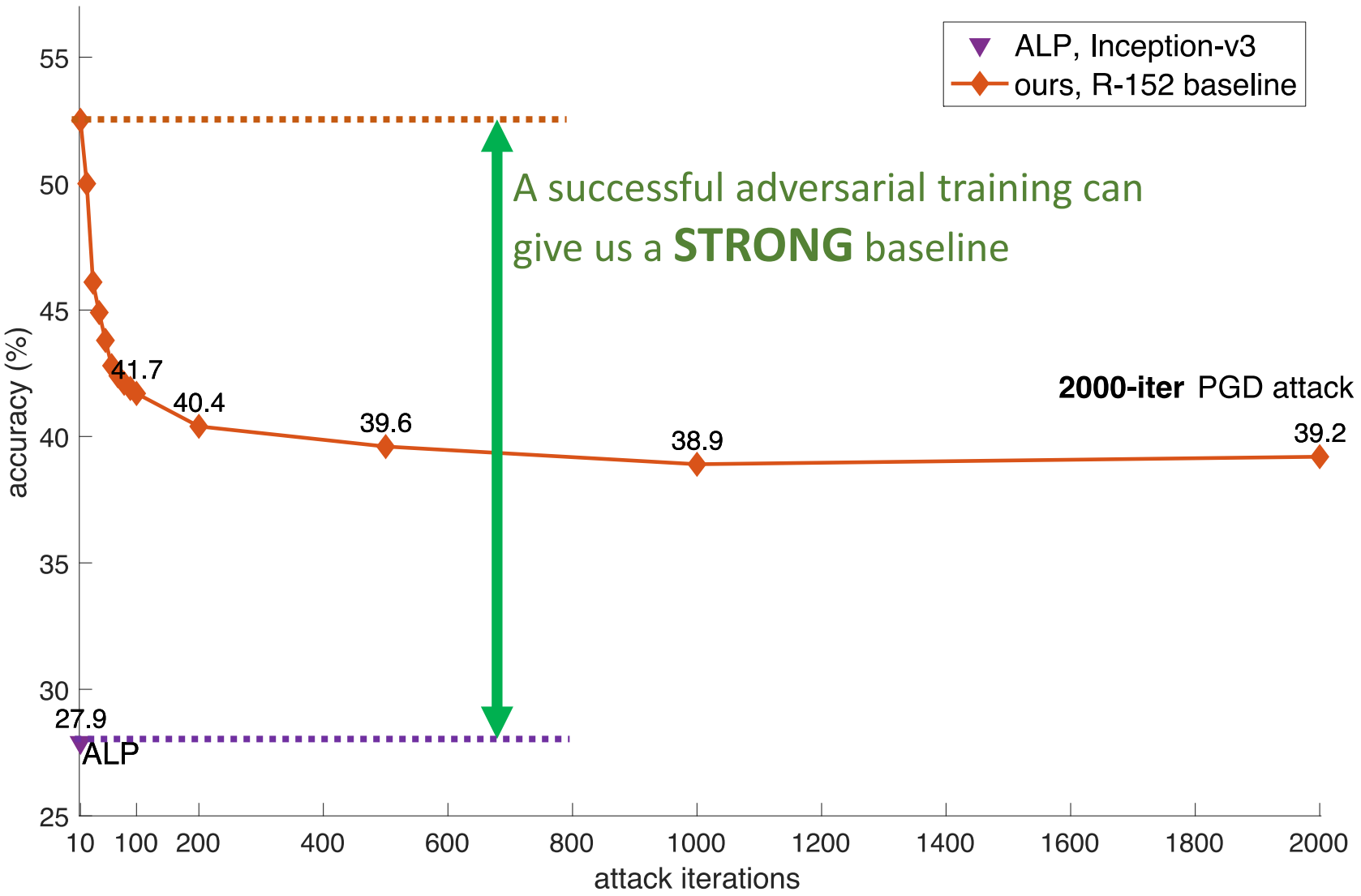
Defending Against Blind Attacks

- Attackers know nothing about models
- Attackers generate adversarial examples using substitute networks
  (**rely on transferability**)

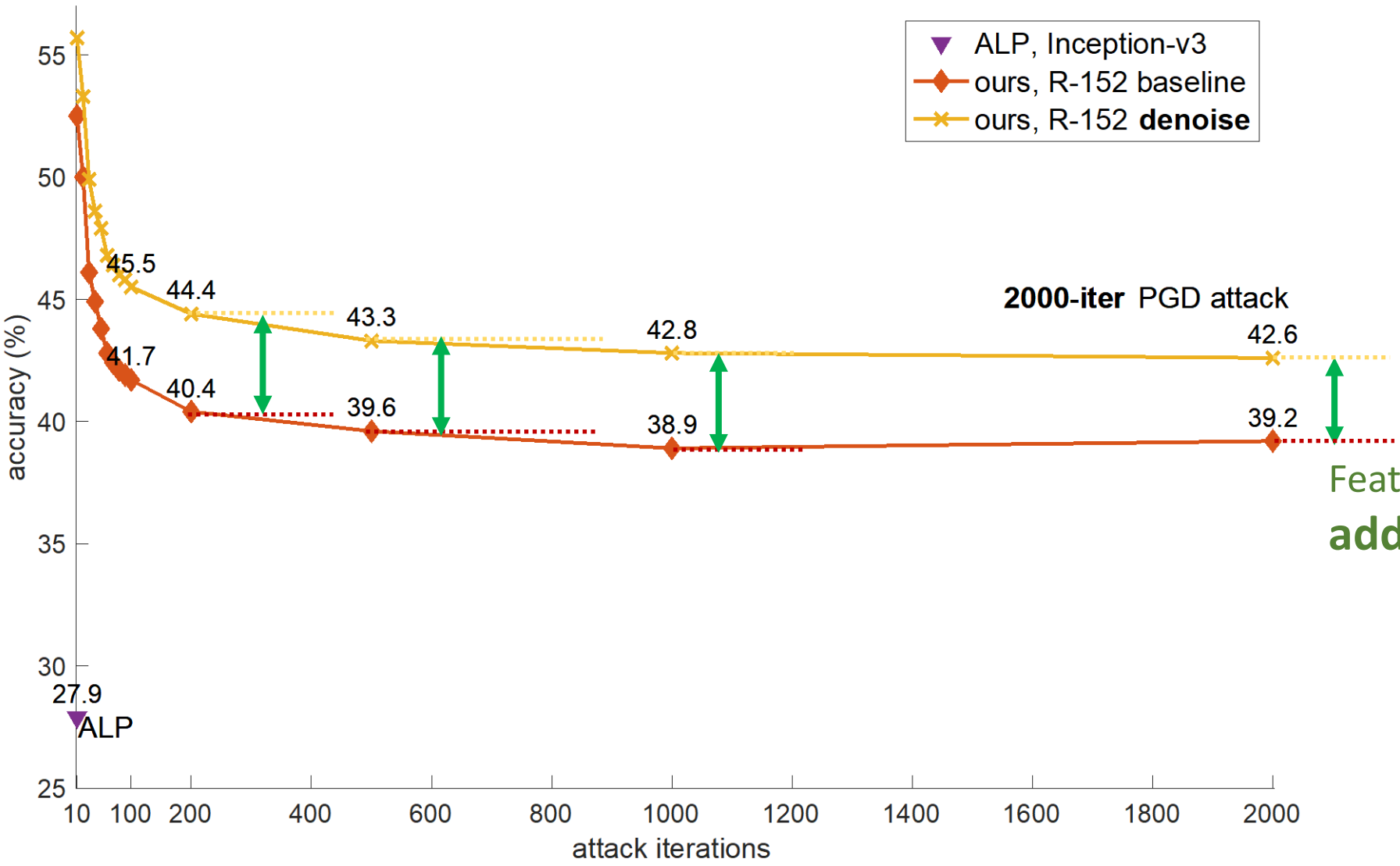
Defending Against White-box Attacks

- Evaluating against adversarial attackers with attack iteration up to 2000
  (**more attack iterations indicate stronger attacks**)

Defending Against White-box Attacks – Part I

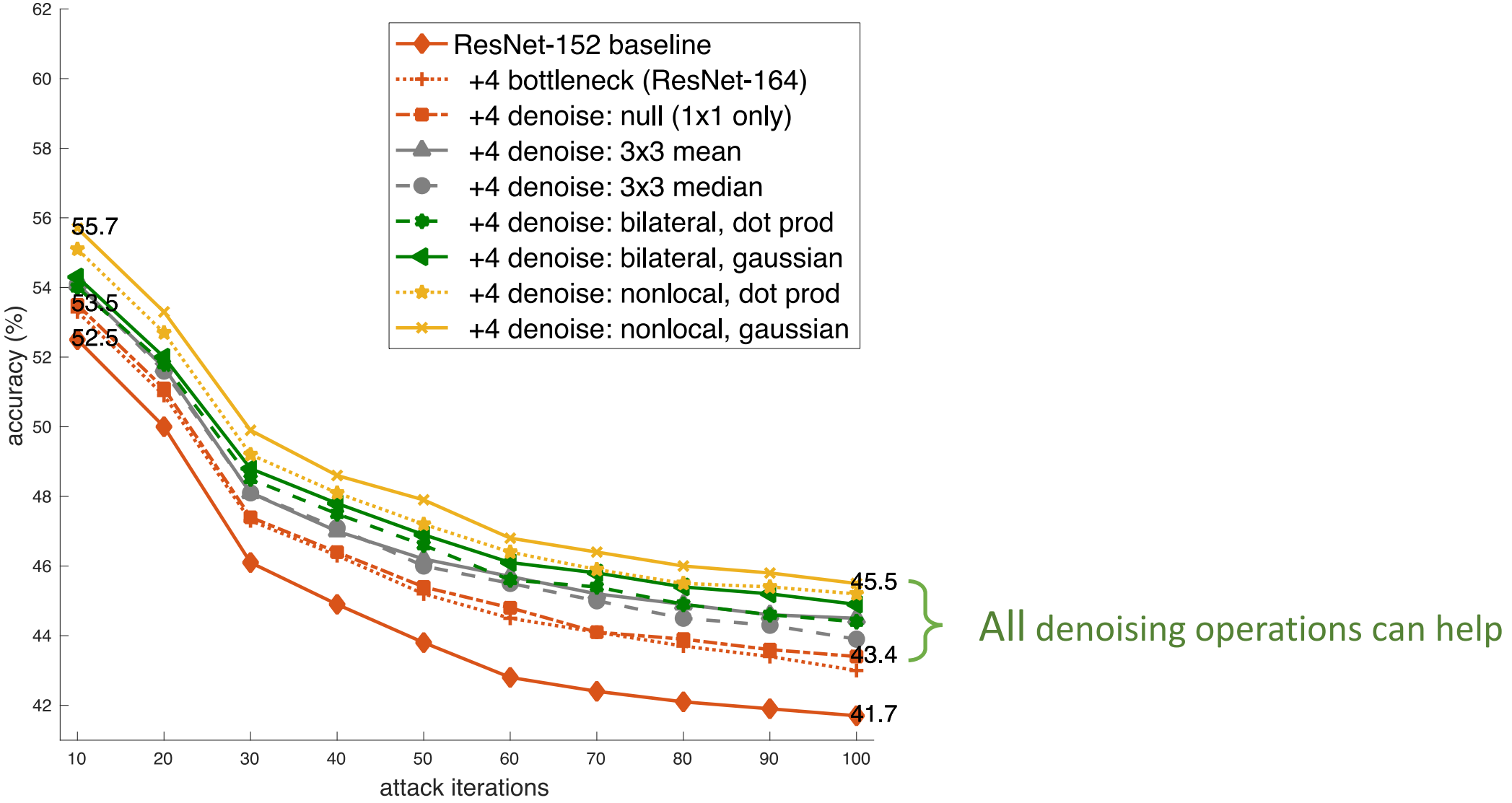


Defending Against White-box Attacks – Part I



Feature Denoising can give us **additional** benefits

Defending Against White-box Attacks – Part II



All denoising operations can help

Defending Against Blind Attacks

- Offline evaluation against 5 BEST attackers from NeurIPS Adversarial Competition 2017
- Online competition against 48 UNKNOWN attackers in CAAD 2018

CAAD 2018 “all or nothing” criterion: an image is considered correctly classified only if the model correctly classifies all adversarial versions of this image created by all attackers

Defending Against Blind Attacks --- CAAD 2017 Offline Evaluation

model	accuracy (%)
CAAD 2017 winner	0.04
CAAD 2017 winner, under 3 attackers	13.4
ours, R-152 baseline	43.1
+4 denoise: null (1×1 only)	44.1
+4 denoise: non-local, dot product	46.2
+4 denoise: non-local, Gaussian	**46.4**
+all denoise: non-local, Gaussian	**49.5**

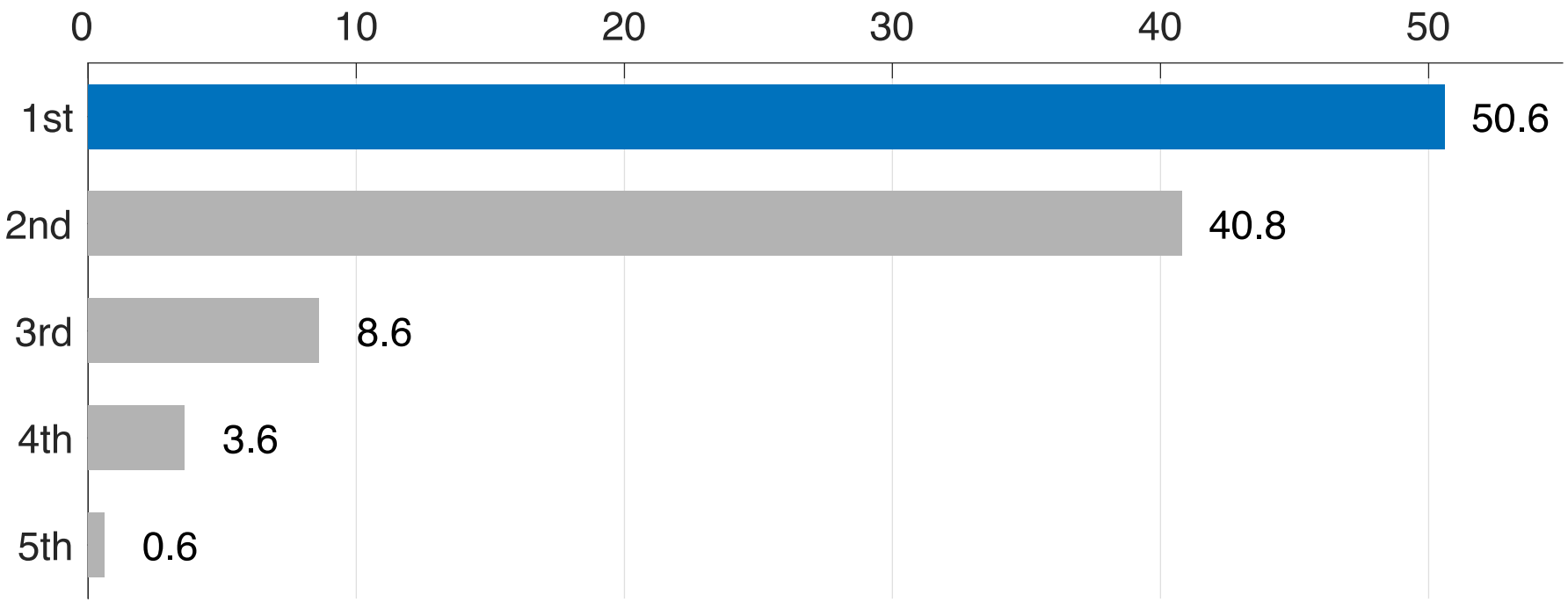
Defending Against Blind Attacks --- CAAD 2017 Offline Evaluation

model	accuracy (%)
CAAD 2017 winner	0.04
CAAD 2017 winner, under 3 attackers	13.4
ours, R-152 baseline	43.1
+4 denoise: null (1×1 only)	44.1
+4 denoise: non-local, dot product	46.2
+4 denoise: non-local, Gaussian	**46.4**
+all denoise: non-local, Gaussian	**49.5**

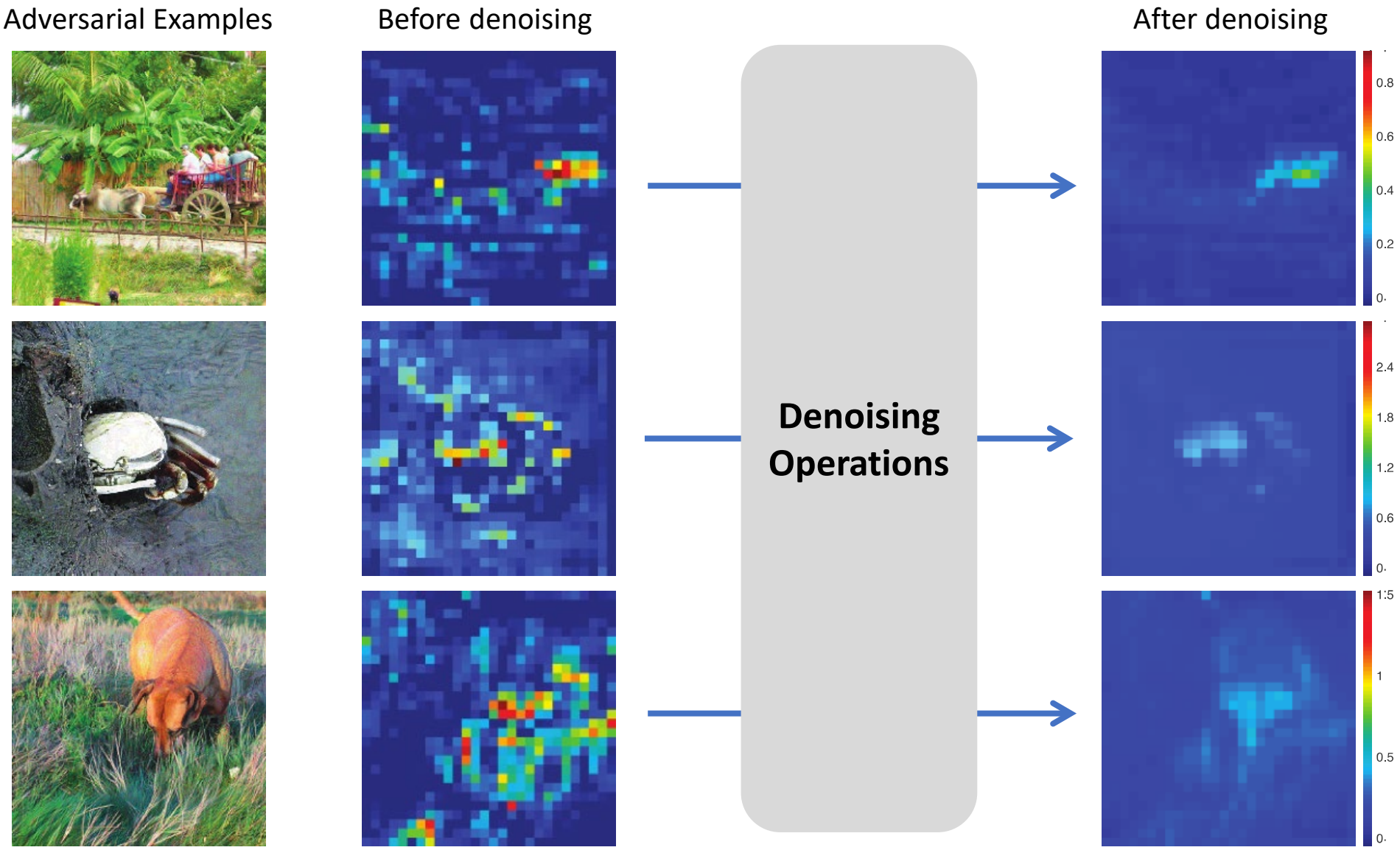
Defending Against Blind Attacks --- CAAD 2017 Offline Evaluation

model	accuracy (%)
CAAD 2017 winner	0.04
CAAD 2017 winner, under 3 attackers	13.4
ours, R-152 baseline	43.1
+4 denoise: null (1×1 only)	44.1
+4 denoise: non-local, dot product	46.2
+4 denoise: non-local, Gaussian	**46.4**
+all denoise: non-local, Gaussian	**49.5**

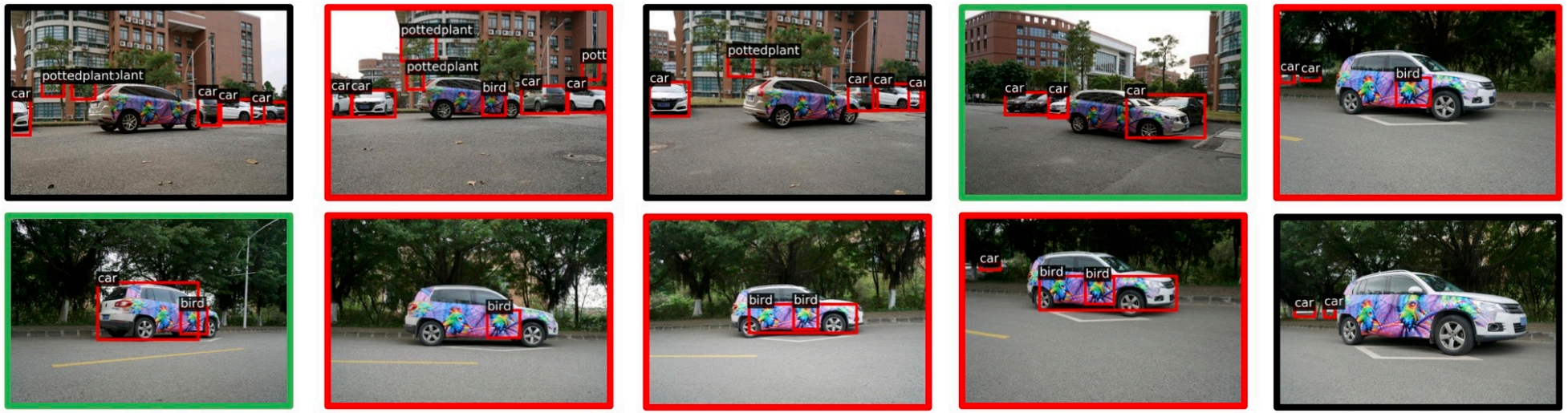
Defending Against Blind Attacks --- CAAD 2018 Online Competition

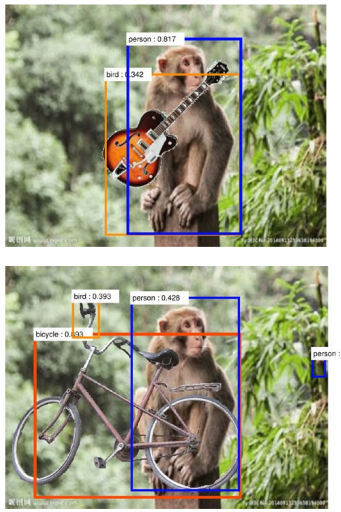


Visualization



Defending against adversarial attacks is still a long way to go…





Questions?