29TH USENIX SECURITY SYMPOSIUM

Interpretable Deep Learning under Fire

Xinyang Zhang[1] Ningfei Wang[2] Hua Shen[1] Shouling Ji[3,4] Xiapu Luo[5] Ting Wang[1]

[1]Pennsylvania State University, [2]UC Irvine,
[3]Zhejiang University, [4]Ant Financial, [5]Hong Kong Polytechnic University

DNN Interpretability

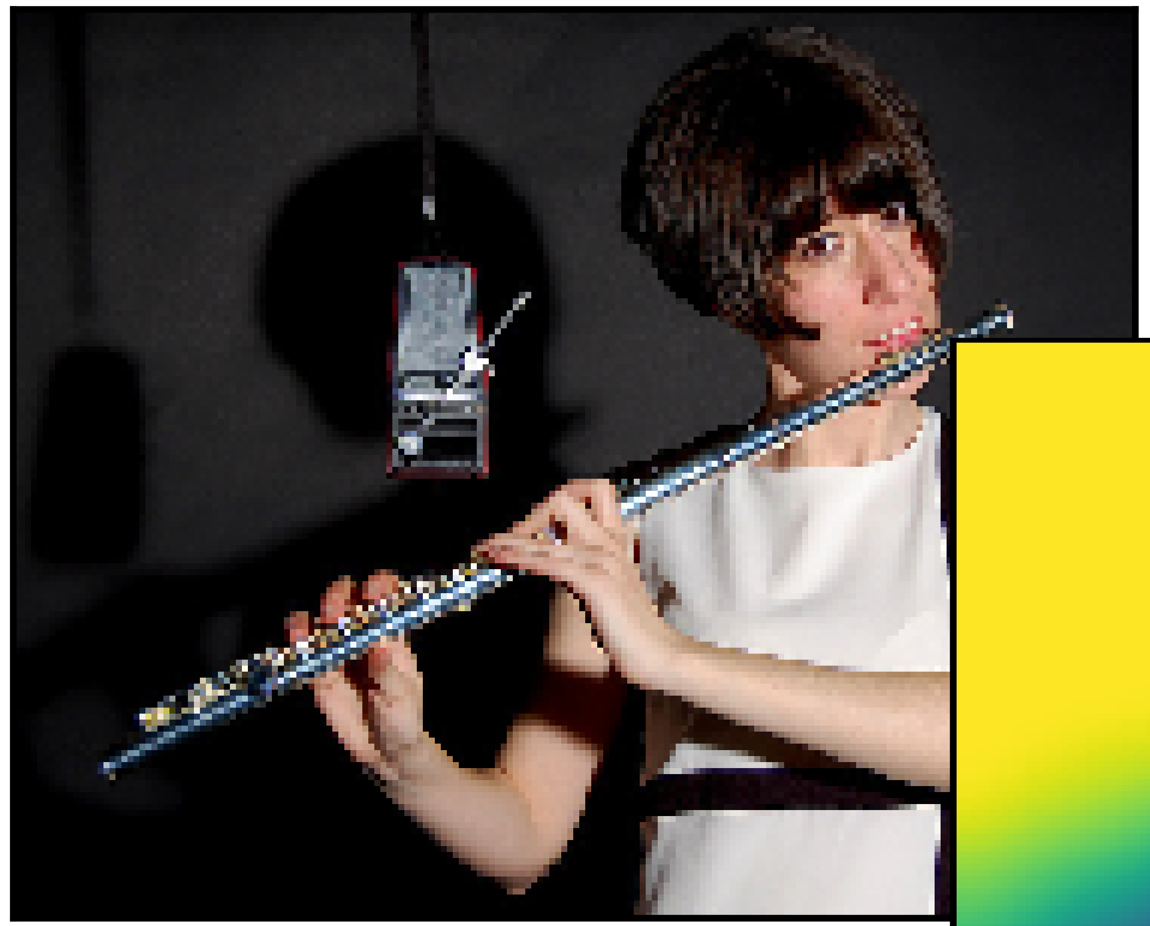
Lack of interpretability

- *How does a DNN arrive at a particular decision?*

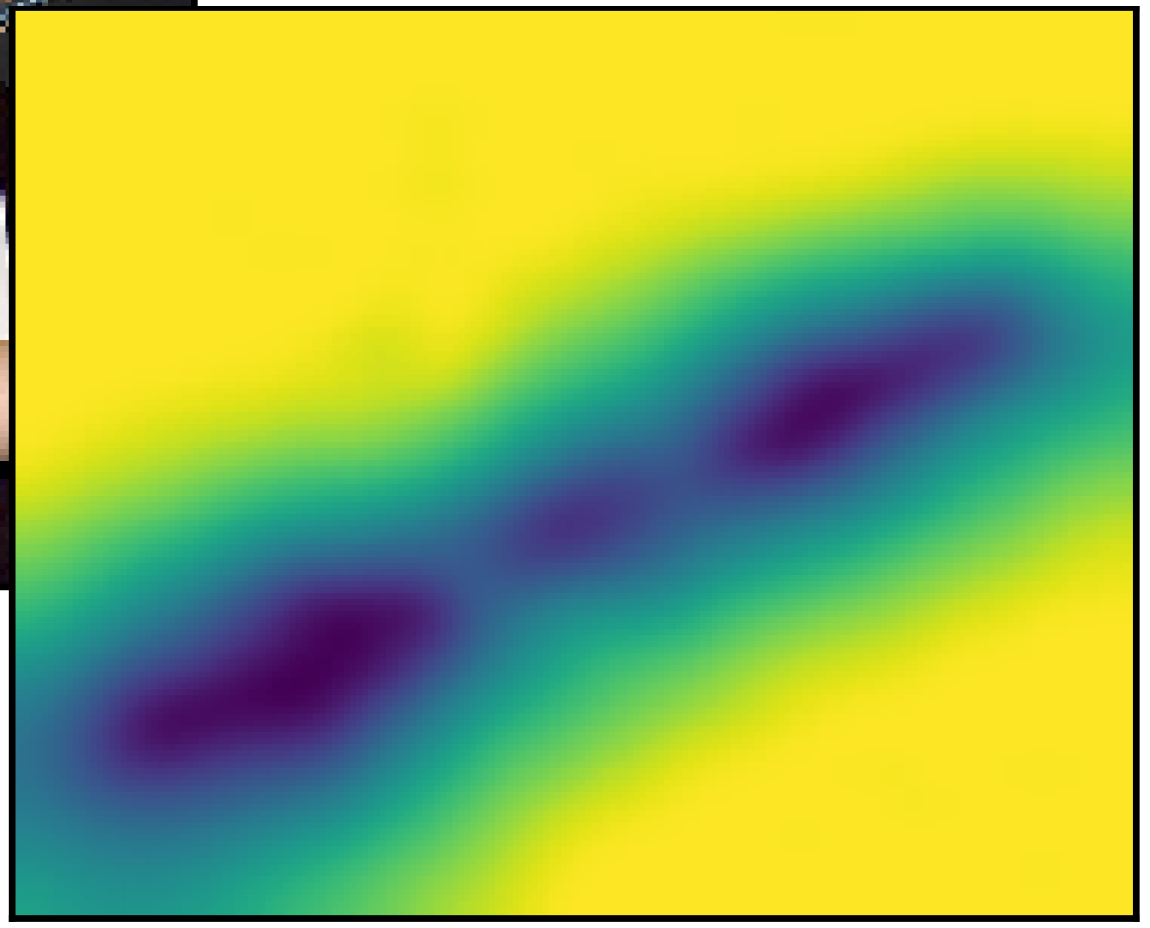
Intensive research on interpreting DNNs

- Backprop-guided
- Representation-guided
- Perturbation-guided
- Model-based

“flute”: 0.9973



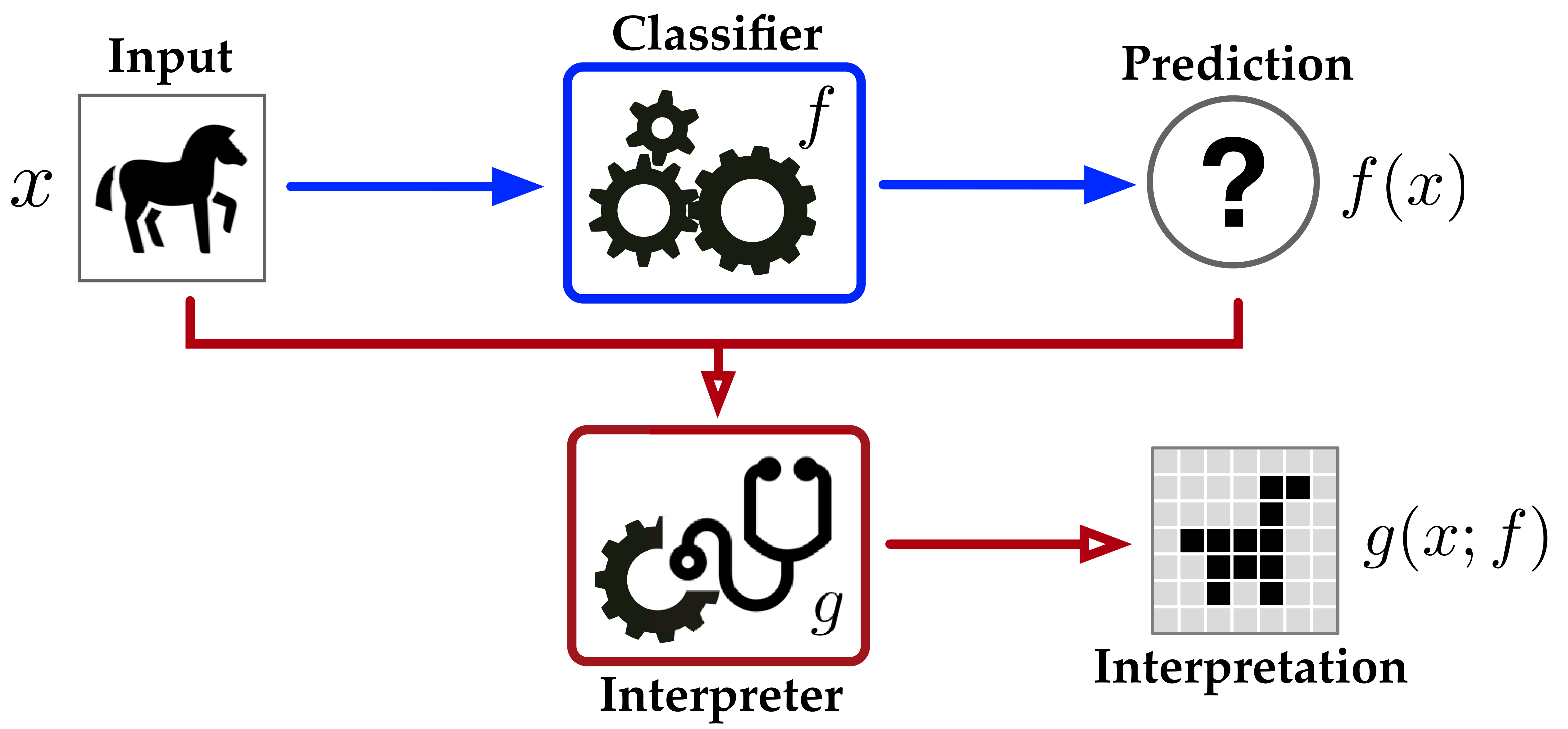
Attribution Map



Interpretable Deep Learning System

Interpretable deep learning system (IDLS)

- Consisting of DNN (classifier) and interpretation model (interpreter)
- Involving humans in the decision-making process
- Requiring the adversary to fool both classifier and interpreter



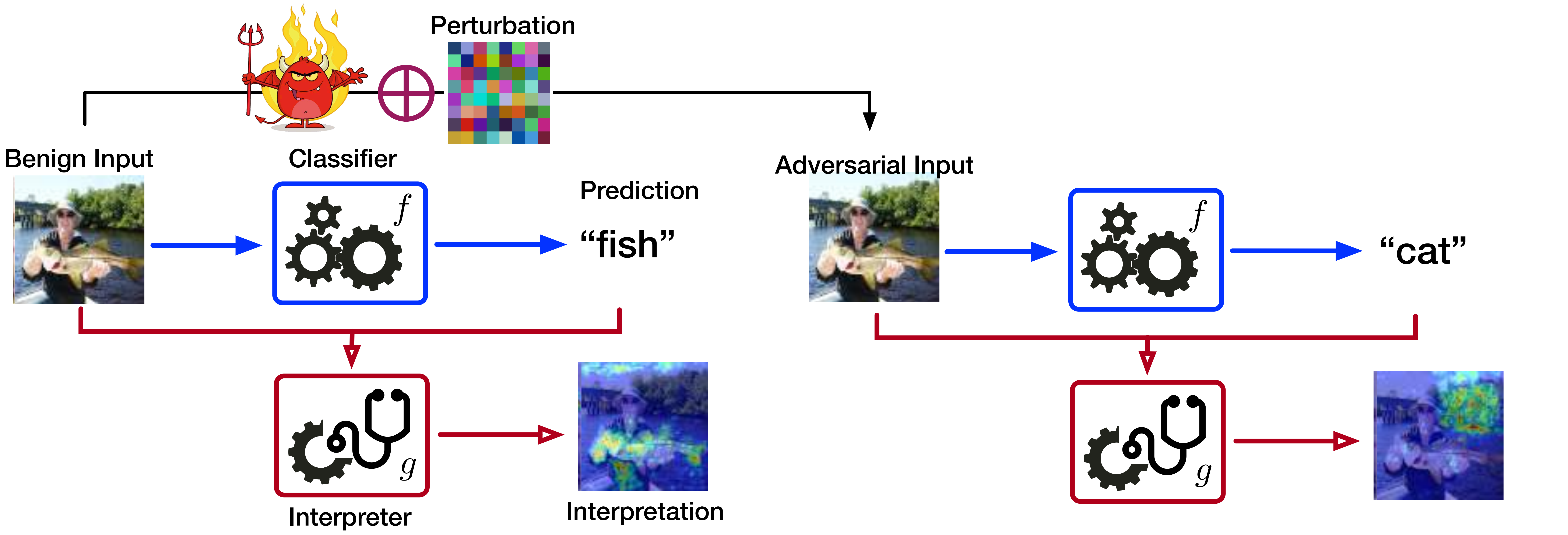
Interpretability = Security?

Goal

- Understanding the security vulnerabilities of IDLSes

Approach

- Developing attacks that simultaneously fool classifier and interpreter



29TH USENIX SECURITY SYMPOSIUM

ADV² Attack

Overall formulation

1. Triggering target prediction c_t and target interpretation m_t
2. Minimizing perturbation magnitude $\Delta(x, x_\circ)$

$$\min_x \Delta(x, x_\circ) \quad \text{s.t.} \begin{cases} f(x) = c_t \\ g(x; f) = m_t \end{cases}$$

Regularized optimization

$$\begin{aligned} \min_x \quad & \ell_{\text{prd}}(f(x), c_t) + \lambda \ell_{\text{int}}(g(x; f), m_t) \\ \text{s.t.} \quad & \Delta(x, x_\circ) \leq \varepsilon \end{aligned}$$

Attack Instantiation

Backprop-guided interpretation

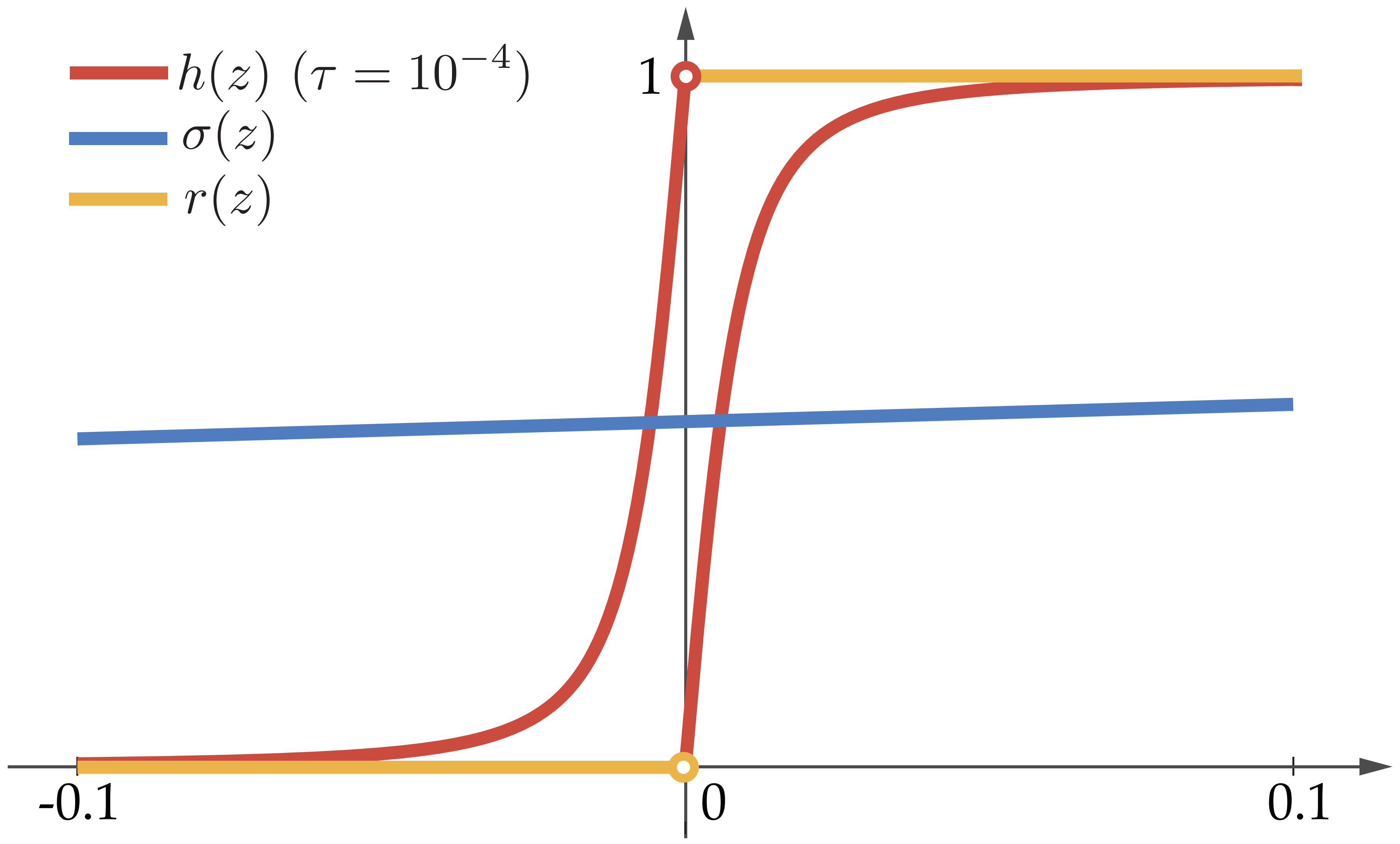
- Gradient saliency (GRAD) interpreter

$$m = \left| \frac{\partial f_c(x)}{\partial x} \right|$$

- Gradient enhancement for ReLU

$$h(z) \triangleq \begin{cases} (z + \sqrt{z^2 + \tau})' = 1 + z/\sqrt{z^2 + \tau} & (z < 0) \\ (\sqrt{z^2 + \tau})' = z/\sqrt{z^2 + \tau} & (z \geq 0) \end{cases}$$

- Label smoothing to avoid gradient saturation



Attack Instantiation (cont.)

Perturbation-guided interpretation

- MASK interpreter

$$\min_{m} f_c(\phi(x;m)) + \lambda \|1-m\|_1 \quad \text{s.t. } 0 \le m \le 1$$

- A bi-level optimization formulation

$$\begin{aligned} \min_{x} \quad & \ell_{\text{adv}}(x, m_*(x)) \\ \text{s.t.} \quad & m_*(x) = \arg\min_{m} \ell_{\text{map}}(m;x) \end{aligned}$$

- Updating m_* estimate and x alternatively
- Stabilizing optimization with imbalanced update and periodical reset

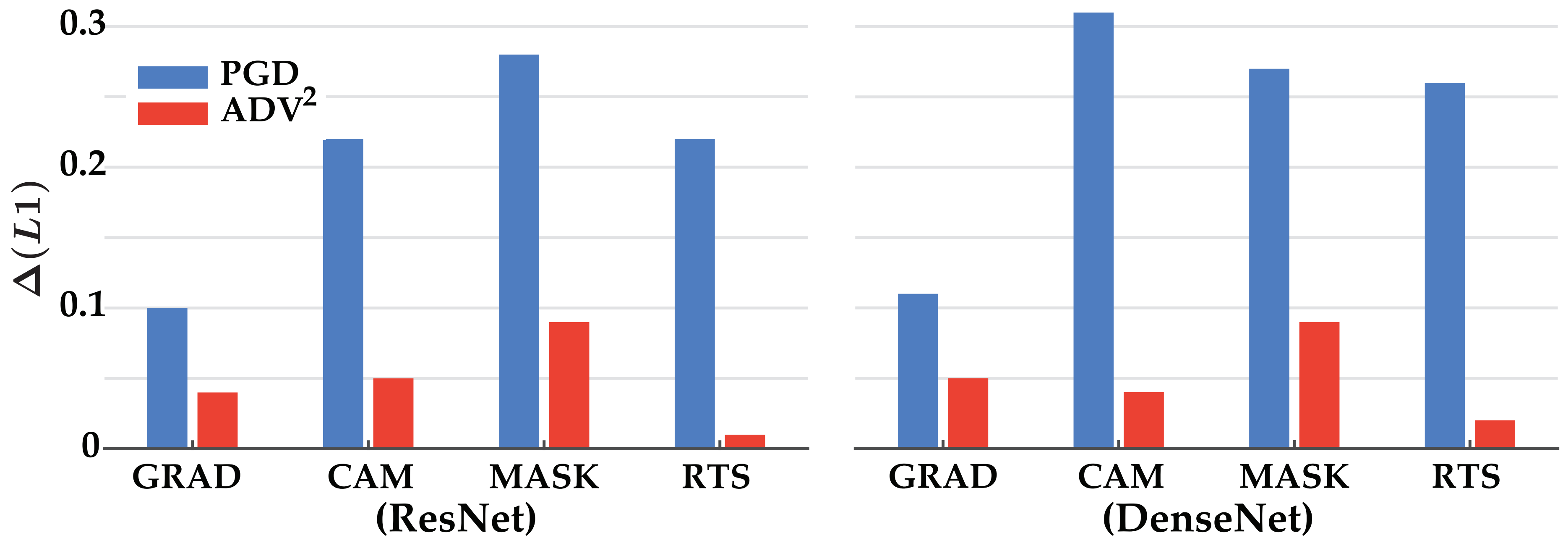
Evaluation

- Attack effectiveness (misclassification)

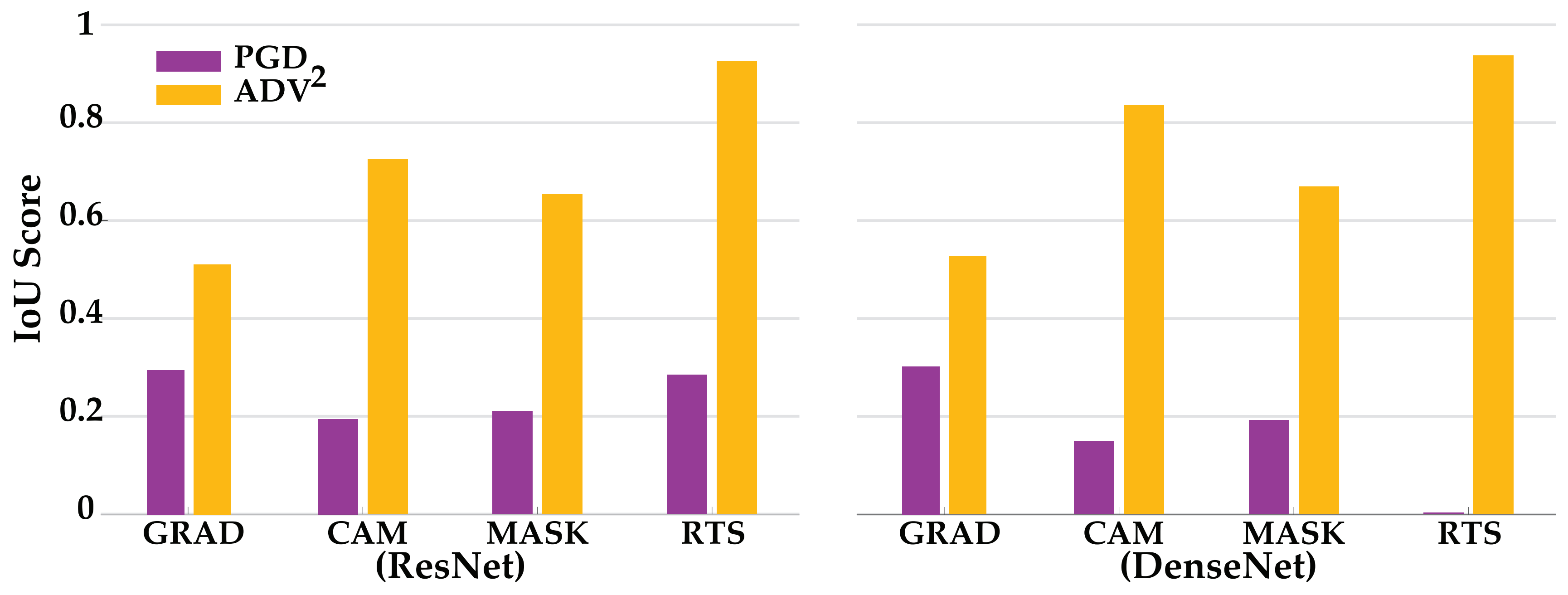
Setting:
- Dataset — ImageNet
- Classifier — ResNet-50, DenseNet-169
- Interpreter — GRAD, CAM, MASK, RTS
- Attack model — PGD, ADV^2
- Target interpretation — benign attribute map

Classifier	ResNet				DenseNet			
Interpreter	GRAD	CAM	MASK	RTS	GRAD	CAM	MASK	RTS
PGD	100% (1.0)				100% (1.0)			
ADV2	100% (0.99)	100% (1.0)	98% (0.99)	100% (1.0)	100% (0.98)	100% (1.0)	96% (0.98)	100% (1.0)

- Attack effectiveness (misinterpretation)



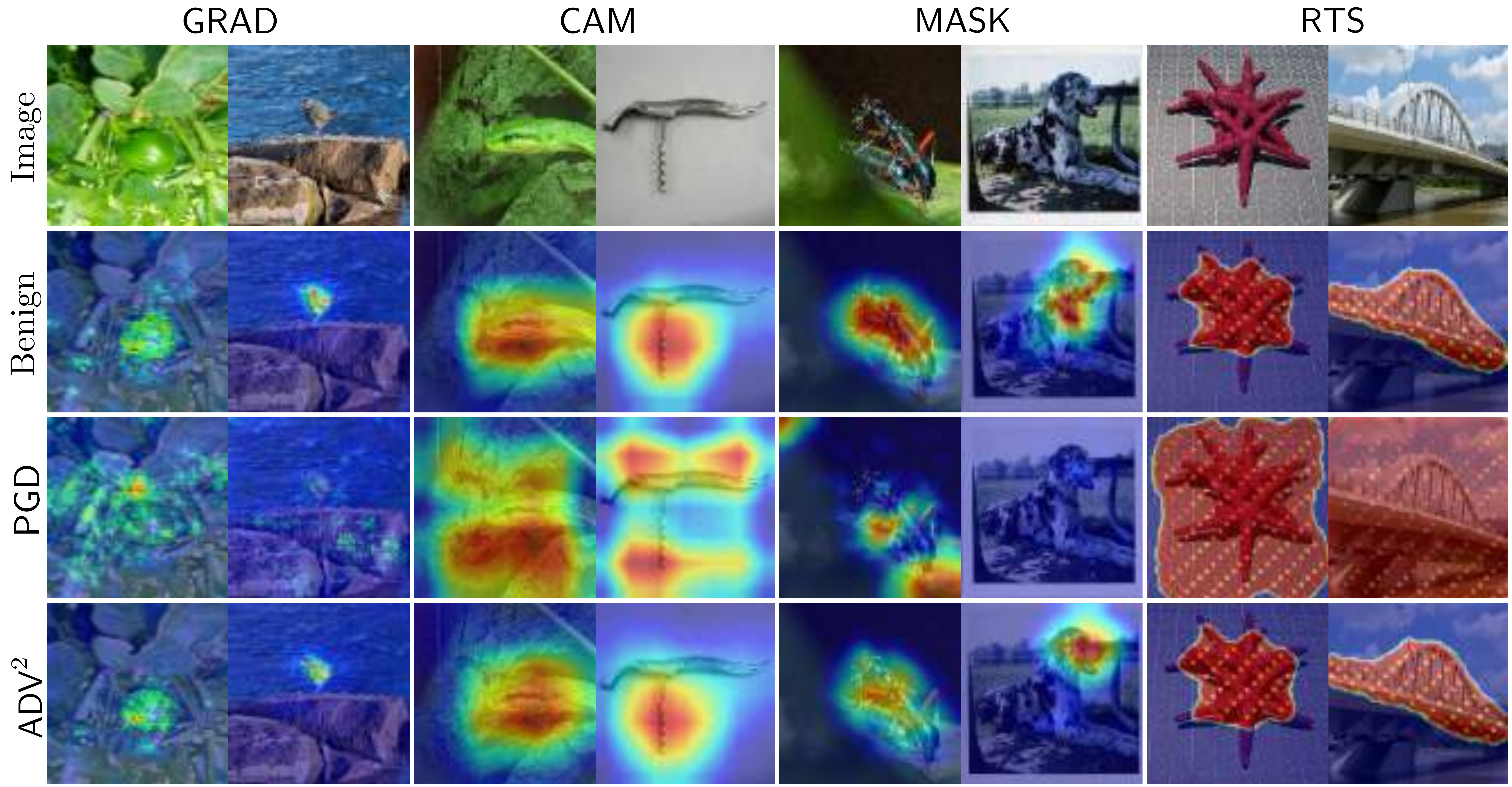
L1 distance between benign and adversarial attribution maps.



Intersection-of-union (IOU) of benign and adversarial attribution maps.

Evaluation (cont.)

- Sample inputs, predictions, and interpretations

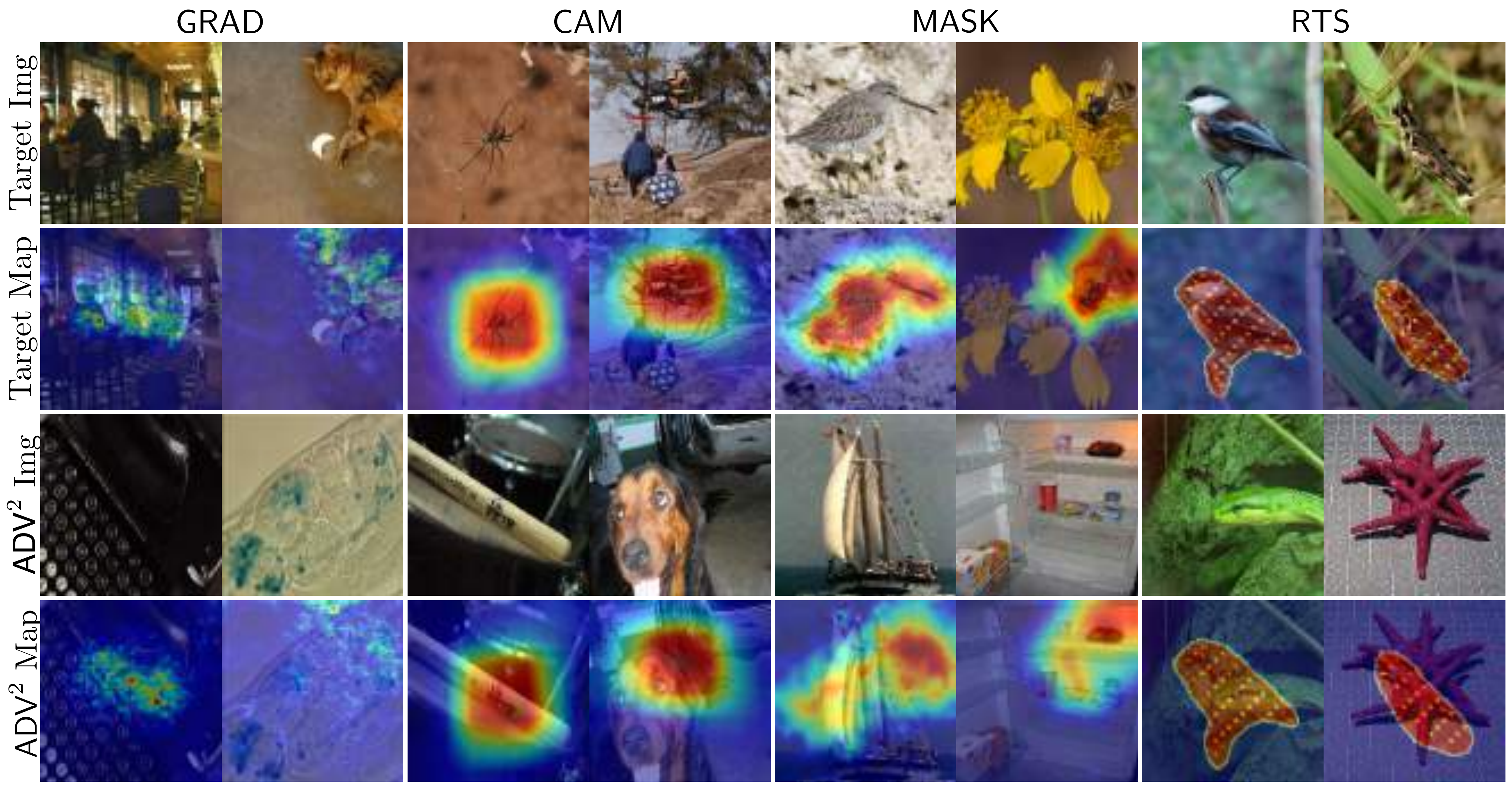


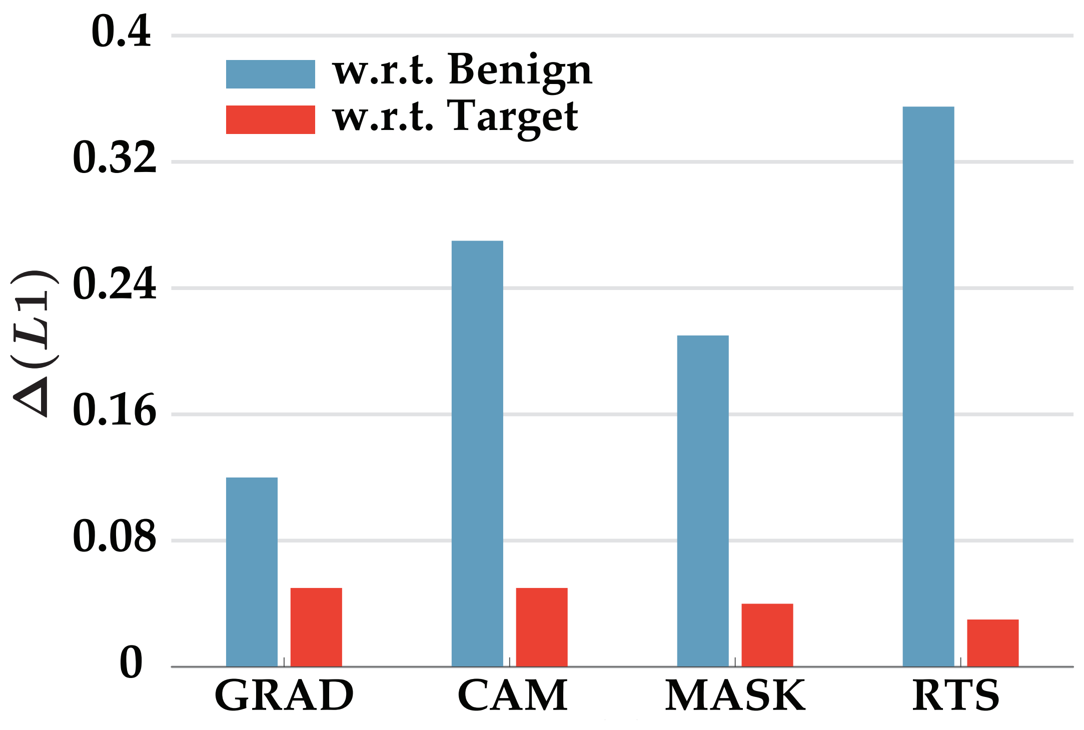
Root of Attack Vulnerability

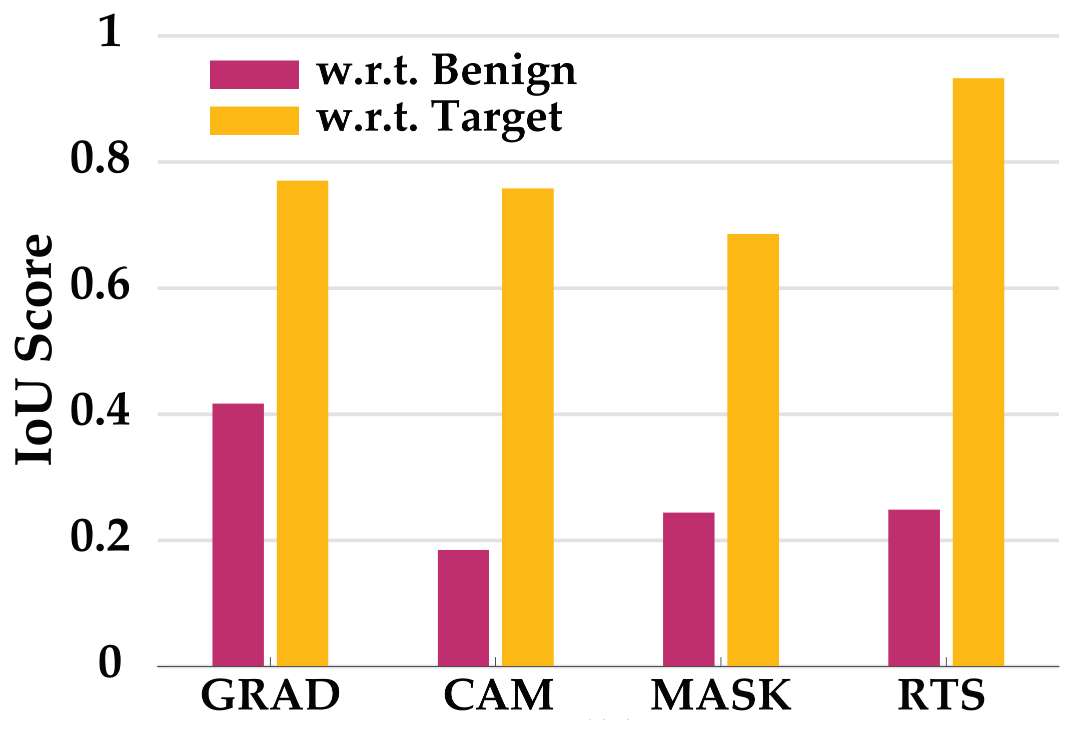
Conjecture: prediction-interpretation gap

- Interpreter’s explanations only partially describe classifier’s predictions, making it practical to exploit both models simultaneously.

Observation: random class interpretation





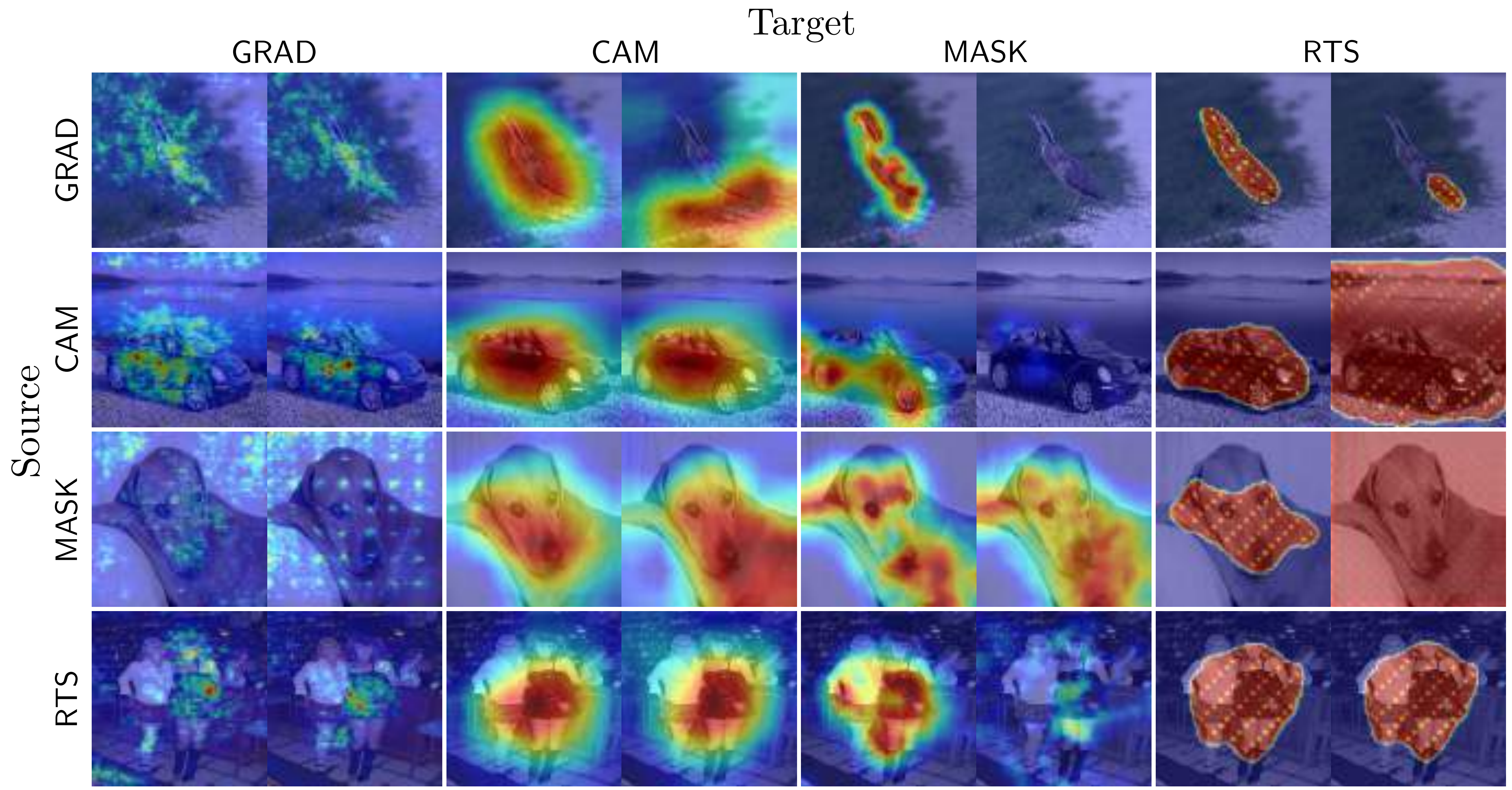


Root of Prediction-Interpretation Gap

Conjecture: limitations of existing interpretation models

- Different interpreters focus on distinct aspects of DNN behaviors (e.g., gradient, intermediate representations, etc.)

Observation: low attack transferability



29TH USENIX SECURITY SYMPOSIUM

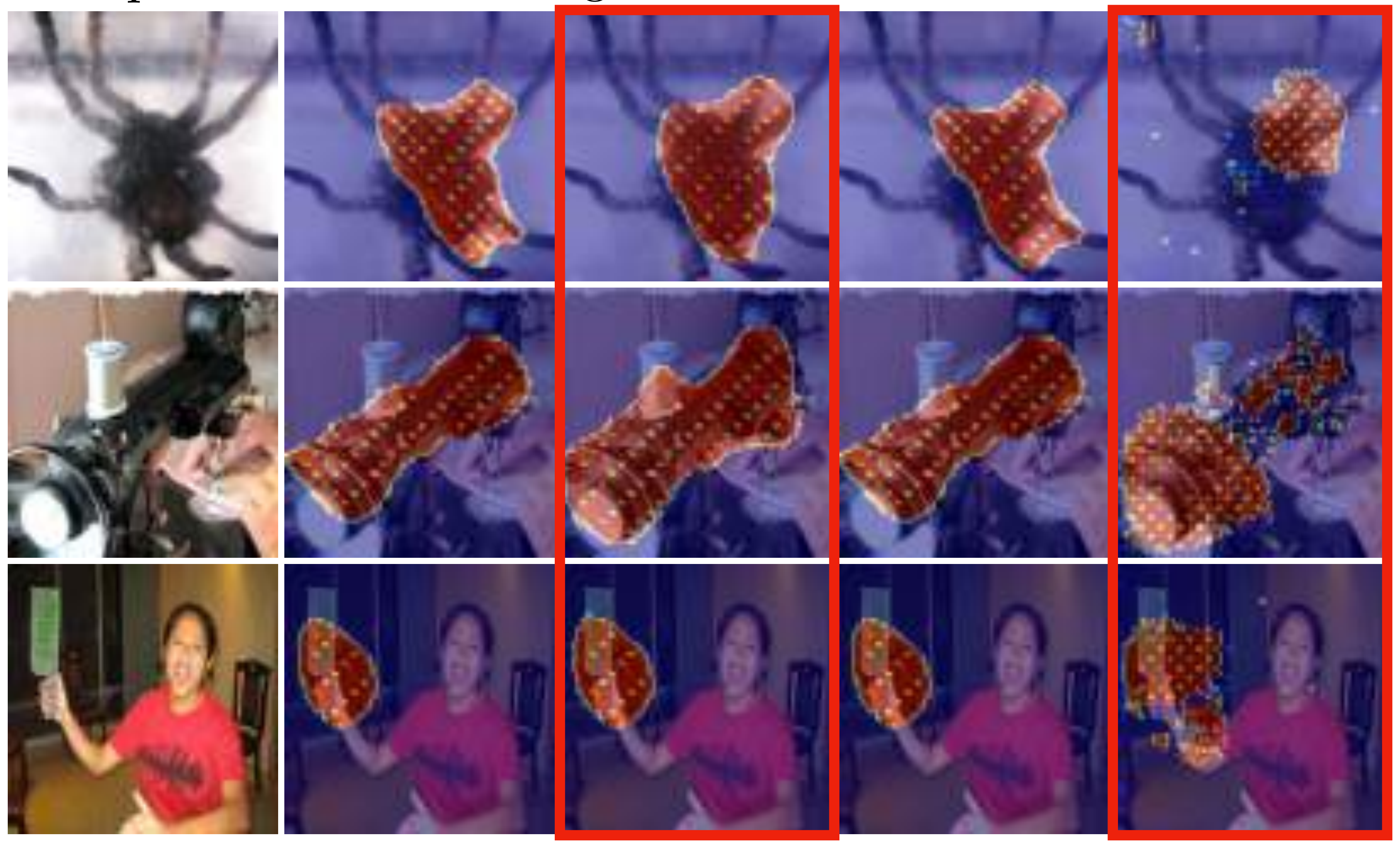
Potential Countermeasures

Ensemble interpretation

- Multiple, complimentary interpreters to fully cover DNN behaviors

Adversarial interpretation

- Minimizing prediction-interpretation gap using adversarial examples



	RTS	RTS^A
Benign	╲	0.03
ADV^2	0.01	0.10

$\mathcal{L}_1$ measures

Key Findings

Finding 1

- The interpretability of existing interpretable deep learning systems merely provides limited security assurance.

Finding 2

- The prediction-interpretation gap is one possible cause that the adversary is able to exploit both classifier and interpreter simultaneously.

Finding 3

- Adversarial training aiming to minimize the prediction-interpretation gap potentially improves the robustness of interpreters.

Thank You!

Please direct your questions to
zxydi1992@hotmail.com