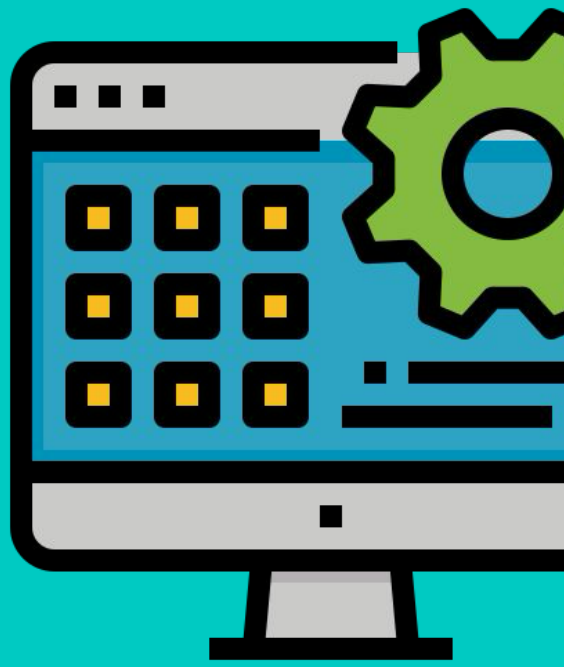
Robust Deep Learning for Computer Vision

A.A. 2024-2025

Lorenzo Cirillo, Ph.D Student at ALCOR Lab Sapienza.

Supervisor: Prof: Irene Amerini



SAPIENZA UNIVERSITÀ DI ROMA

ALCOR Lab

Overview

- What is a robust model?
- Robustness concept in deep learning
- Adversarial vs backdoor attacks
- Adversarial attacks and defenses
- Research topics at ALCOR Lab

What is a robust model?

What is a robust model?

Formal definition: In deep learning, robustness usually expresses the requirement that the network behaves smoothly, i.e., that **small input perturbations** or minor model modifications should not cause significant spikes in the output of models.

Less formal definition: A deep learning model is robust when it shows good performance in inconvenient conditions (e.g. when some images are manipulated to mislead the model).

Autonomous driving


Video surveillance

J. Liu, Y. Jin, *A comprehensive survey of robust deep learning in computer vision*, in Journal of Automation and Intelligence, 2023

Real-life example

Original image

Clean scenario

Deep learning model

prediction

Red light

Perturbed image

Perturbed scenario

Deep learning model

prediction

Green light

Robustness concept in Deep Learning

Local robustness definition

Given an original sample x_O, the perturbed version x', and a threshold δ, a model f is δ-locally robust at point (image) x_O iff:

$$\forall x_O.\ \|x' - x_O\| \leq \delta \ \Rightarrow\ \text{argmax}\, f(x') = \text{argmax}\, f(x_O)$$

- The higher δ, the more robust the model is
- Local because it takes into account single samples x_O in the dataset
- The threshold δ should be defined for each point in the dataset

G. Katz, et al., *Reluplex*, *An efficient SMT solver for verifying deep neural networks*, in: International Conference on Computer Aided Verification, Springer, 2017

Global robustness definition

Given an original sample x_O belonging to the input domain D, the set of labels L, the perturbed version x', a threshold δ, and a threshold ε, a model f is (δ,ε)-globally robust in input region D iff:

$$\forall x_O \in D.\ \|x' - x_O\| \leq \delta \Rightarrow \forall \ell \in L.\ |C(f, x', \ell) - C(f, x_O, \ell)| < \varepsilon$$

- C is the confidence level of f
- The higher δ, the more robust the model
- The lower the ε, the more robust the model
- Global because takes into account an entire input domain D

G. Katz, et al., *Reluplex, An efficient SMT solver for verifying deep neural networks*, in: International Conference on Computer Aided Verification, Springer, 2017

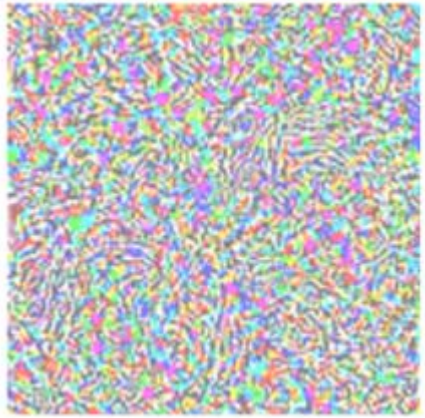
Adversarial vs Backdoor Attacks

Adversarial attack

An adversarial attack is an imperceptible manipulation of input data to mislead the model at **inference phase** (e.g. white-box, black-box)

Original image

Adversarial image

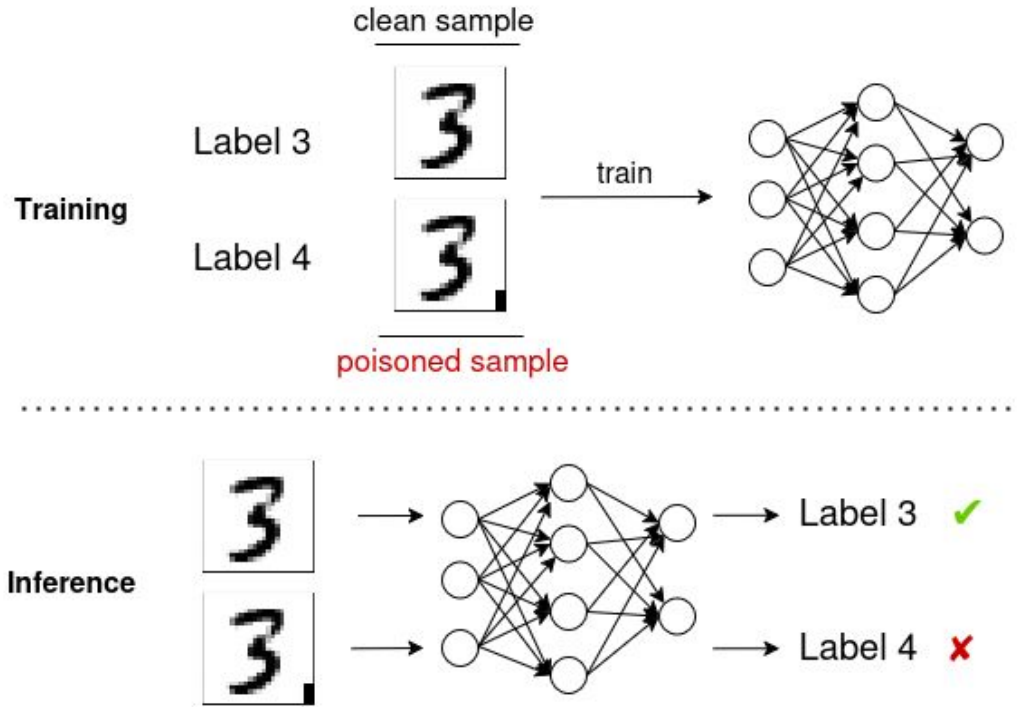


Adversarial - Original

C. Szegedy et al., *Intriguing properties of neural networks*. in: arXiv preprint arXiv:1312.6199, 2013

Backdoor attack

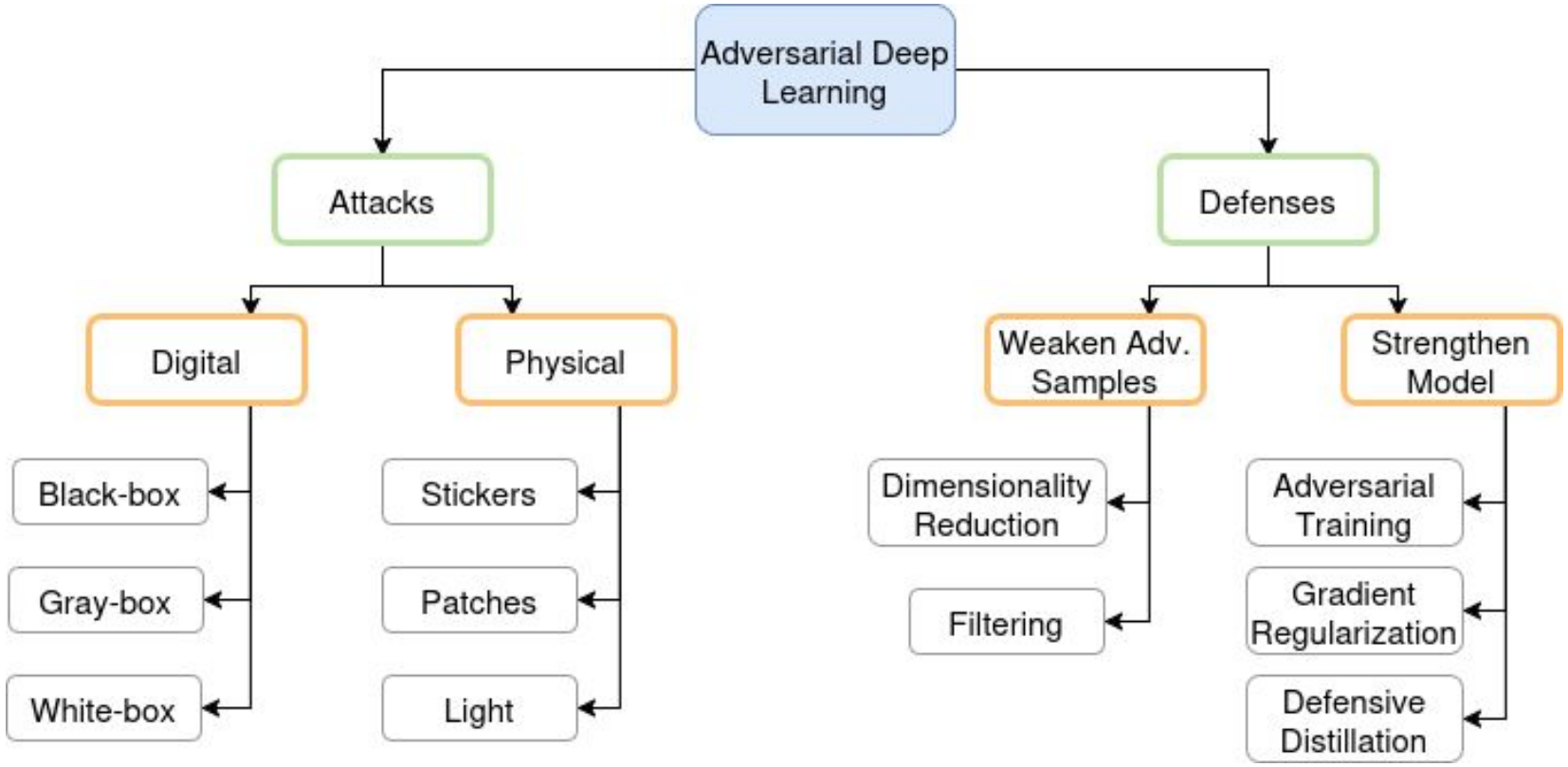
- A backdoor attack consists of embedding hidden patterns (called triggers) into the model during the **training phase** (e.g. clean-label, model-based)
- Triggers are like keys that open backdoor in the model to fool it



T. Gu, B. Dolan-Gavitt, S. Garg, *Badnets: Identifying vulnerabilities in the machine learning model supply chain,* in: arXiv preprint arXiv:1708.06733, 2017

Adversarial Attacks and Defenses

Adversarial methods taxonomy



Adversarial attacks and defenses techniques taxonomy

White-box attacks

Giving x as the original sample, l as the associated label, x' the adversarial sample, $\mathcal{L}(\cdot)$ as the loss function, and ε as the attack intensity, we define the following standard gradient-based attacks:

- Fast Sign Gradient Method (FGSM)

$$x' = x + \varepsilon \, sign(\nabla_x \mathcal{L}(x,l))$$

- Iterative FGSM (I-FGSM)

$$x'_0 = x$$

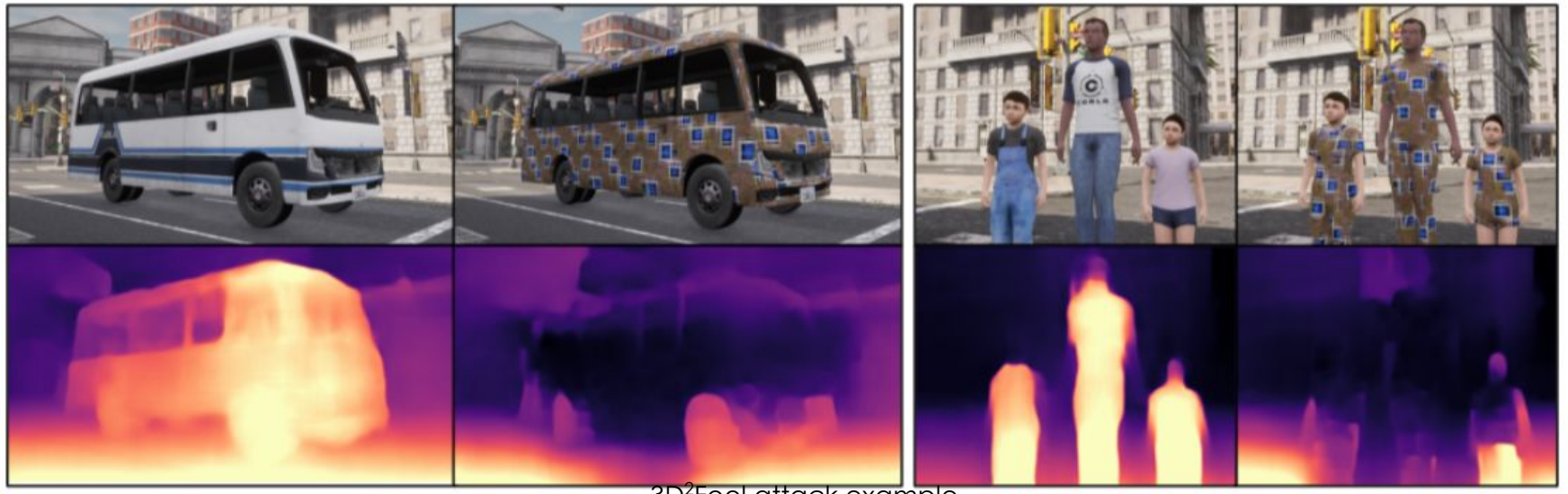
$$x'_{n+1} = Clip\{x'_n + \varepsilon \, sign(\nabla_x \mathcal{L}(x'_n,l))\}$$

- Projected Gradient Descent (PGD)

$$x'_0 = x$$

$$x'_{n+1} = Proj\{x'_n + \varepsilon \, sign(\nabla_x \mathcal{L}(x'_n,l))\}$$

MDE physical attacks: 3D^{2}Fool



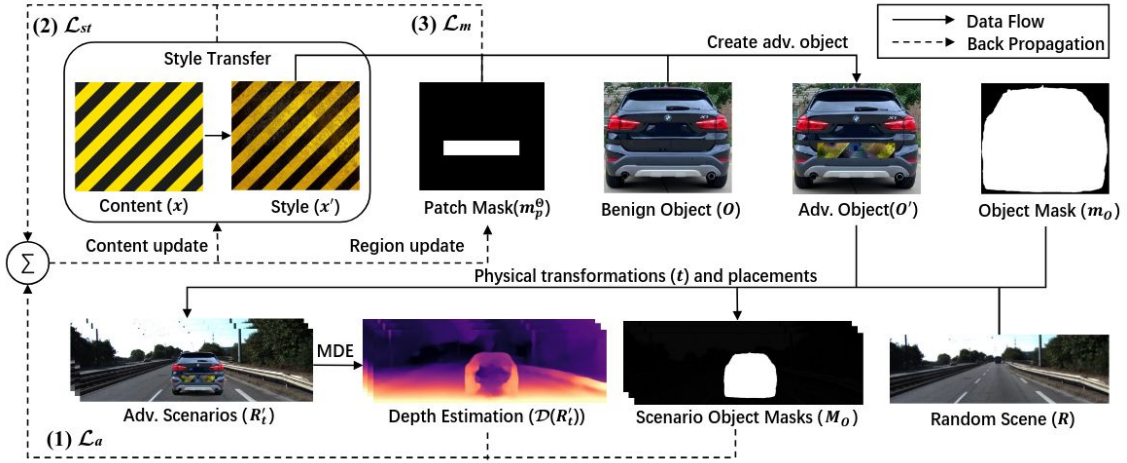
3D^{2}Fool attack example

- Model agnostic (it works regardless the model architecture)
- Valid with weather changes
- MDE error of over 10 meters

J. Zheng et al., *Physical 3D Adversarial Attacks against Monocular Depth Estimation in Autonomous Driving,* in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

MDE physical attacks: SPOO

SPOO attack example



SPOO pipeline

- Physical-object-oriented adversarial patches
- MDE error of over 6 meters
- 0.93 of attack success rate

Z. Cheng et al., *Phyical attack on monocular depth estimation with optimal adversarial patches*, in: Computer Vision – ECCV, 2022

Adversarial defenses

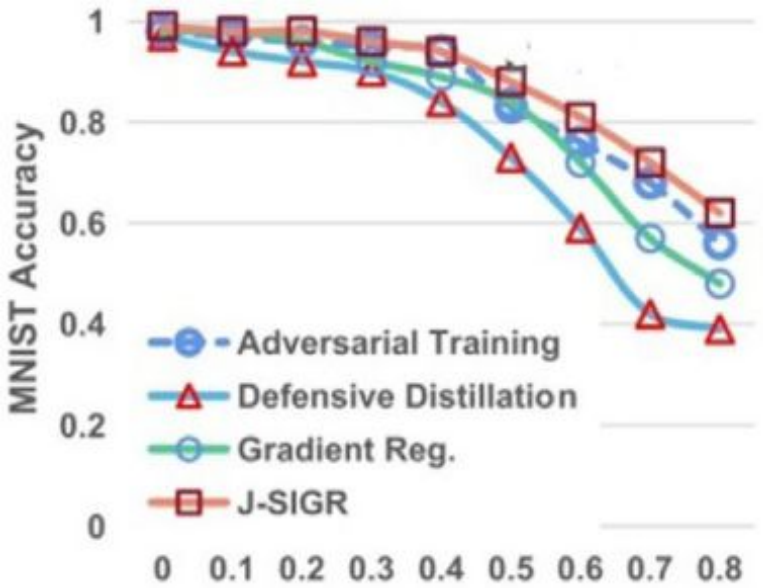
Adversarial Training
Train on adversarial samples to increase the learned data distribution

Gradient Regularization
Add a regularization to the loss function to penalize the gradient wrt to specific inputs

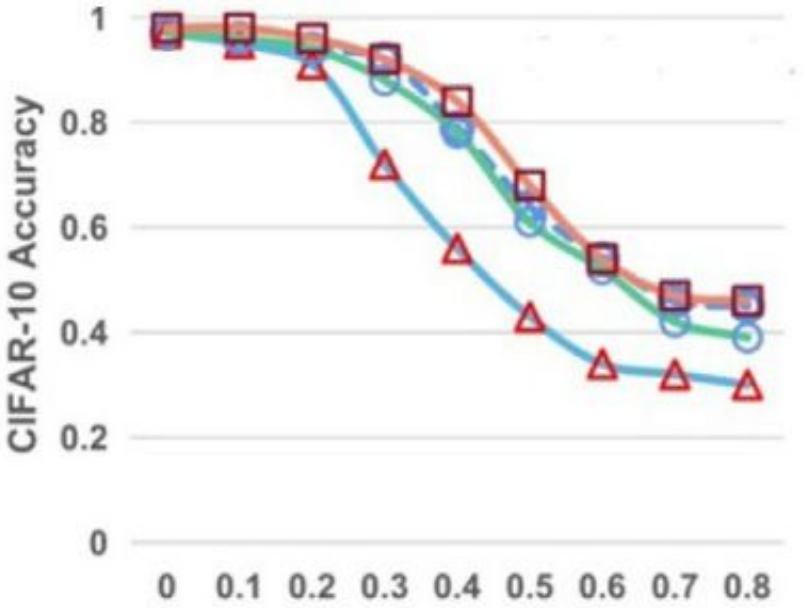
Defensive distillation
train a student model with soft labels generated from a teacher model to reduce sensitivity to small input perturbations.

J-SIGR

leverages Jacobian normalization to improve robustness and introduces regularization of perturbation-based saliency maps



Comparison of ResNet 20 accuracy on the MNIST dataset trained with the four different defense techniques and attacked with FGSM



Comparison of ResNet 20 accuracy on the CIFAR-10 dataset trained with the four different defense techniques and attacked with FGSM

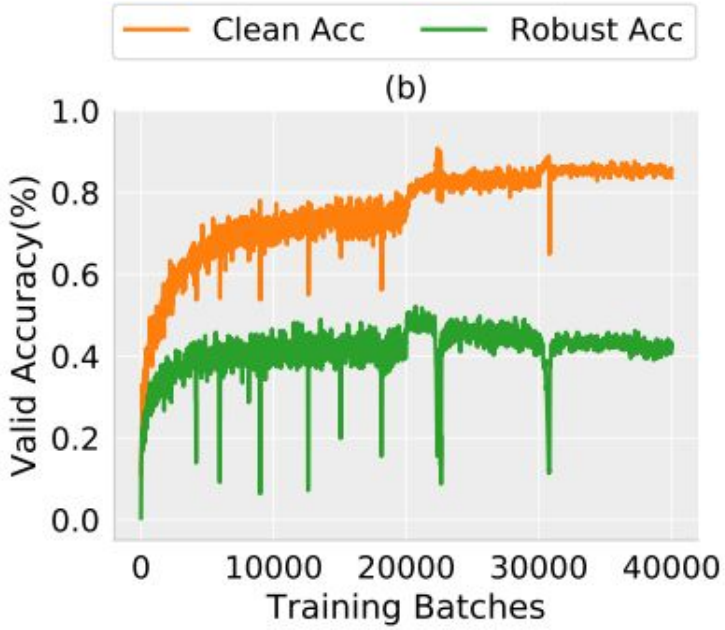
Liu. D et all, *Jacobian norm with Selective Input Gradient Regularization for interpretable adversarial defense*, in: Pattern Recognition 145, 2024

Evaluation metrics

- Accuracy on clean and adversarial samples
- Attack success rate

$$\text{ASR} = \frac{\#\text{ flipped predictions}}{\#\text{ total samples}}$$

- Distortion: use the l_p-norm to compute the distance between the clean and adversarial images
- Average confidence of adversarial class
- Average confidence of true class



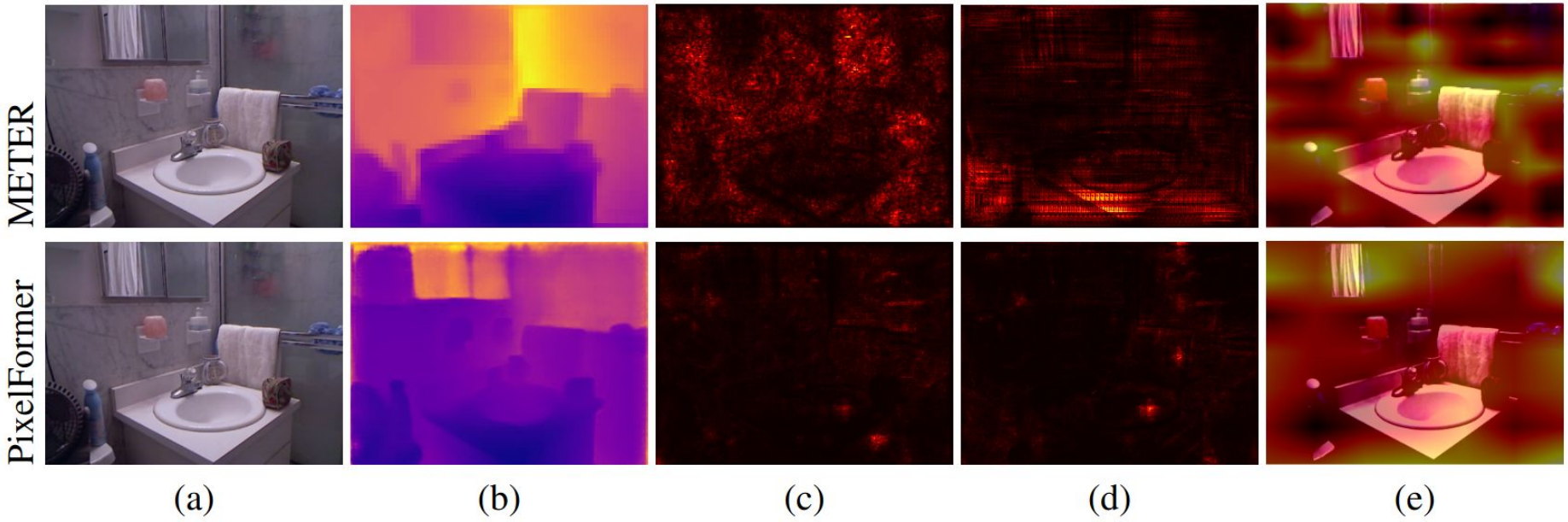
Clean vs adversarial accuracy

B. Li et all, *Towards Understanding Fast Adversarial Training*, in: arXiv preprint arXiv:2006.03089, 2020

Research topics at ALCOR Lab

Our research - MDE Explainability

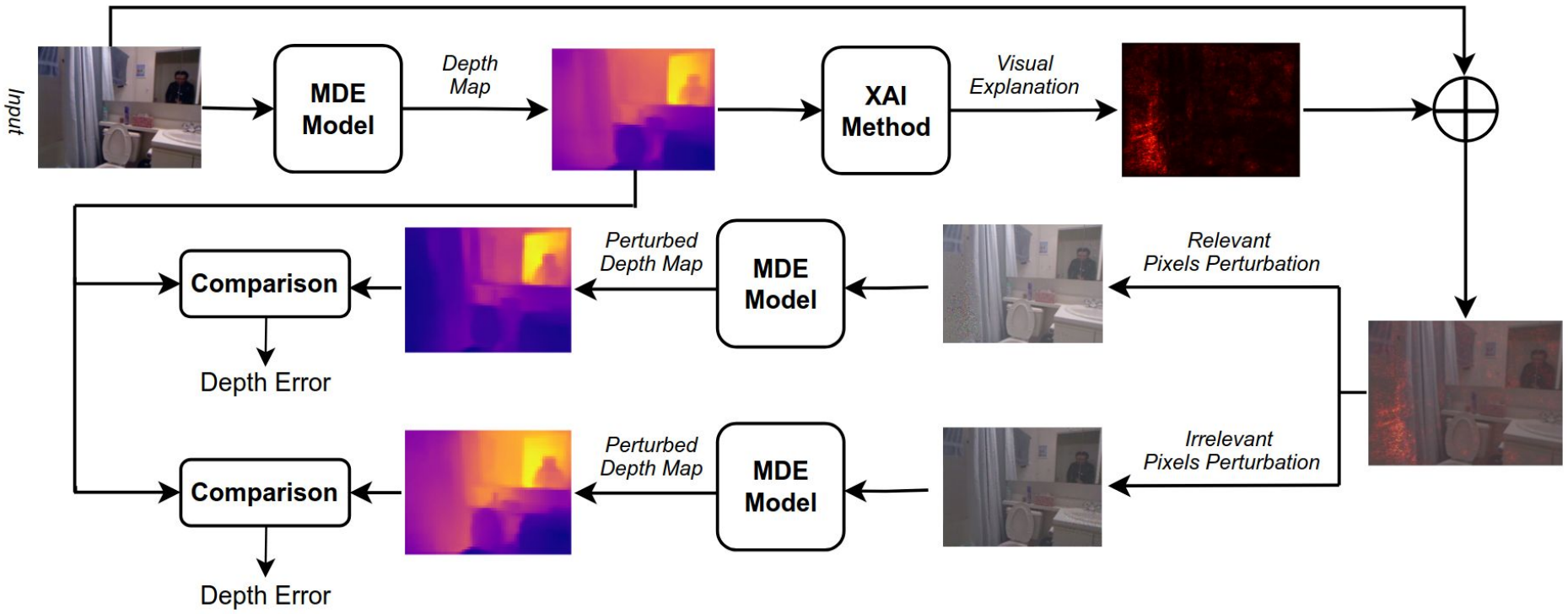
Study the explainability of METER and PixelFormer for Monocular Depth Estimation (MDE) through black mask, gaussian mask, and FGSM



Graphic comparison of the generated visual explanations for METER (first row) and PixelFormer (second row). (a) Input image. (b) Predicted depth map. (c) Saliency Map explanation. (d) Integrated Gradients explanation. (e) Attention Rollout explanation.

L. Cirillo, C. Schiavella, L.Papa, P. Russo, I. Amerini, *Shedding Light on Depth: Explainability Assessment in Monocular Depth Estimation*, in: Proceeding of the International Joint Conference on Neural Networks, 2025

How do we evaluate the explanations?



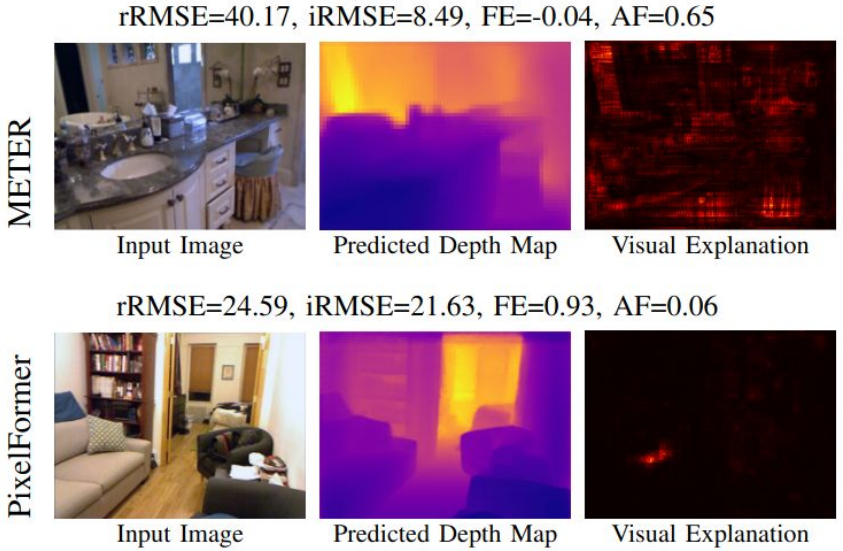
Pipeline of the exploited evaluation framework

L. Cirillo, C. Schiavella, L.Papa, P. Russo, I. Amerini, *Shedding Light on Depth: Explainability Assessment in Monocular Depth Estimation*, in: Proceeding of the International Joint Conference on Neural Networks, 2025

Novelty: Attribution Fidelity

Given a model f, the input image x, the relevant perturbed image x_{rel}, the irrelevant perturbed image x_{irr}, and a metric to compute the distances between the model predictions $d(\cdot)$, the Attribution Fidelity (*AF*) is:

$$AF = \frac{|\, d(f(x_{rel}), f(x)) \,| - |\, d(f(x_{irr}), f(x)) \,|}{|\, d(f(x_{rel}), f(x)) \,| + |\, d(f(x_{irr}), f(x)) \,|}$$



AF comparison on visual explanations for METER and PixelFormer

L. Cirillo, C. Schiavella, L.Papa, P. Russo, I. Amerini, *Shedding Light on Depth: Explainability Assessment in Monocular Depth Estimation*, in: Proceeding of the International Joint Conference on Neural Networks, 2025

Our research - Deepfake Detection

Performed attacks:

- FGSM
- I-FGSM
- PGD (Projected Gradient Descent)

Explainability-Driven Adversarial Robustness Assessment for Generalized Deepfake Detectors

Context

Generative models produce high-quality deepfake images hard to detect

Generalized and robust deepfake detectors

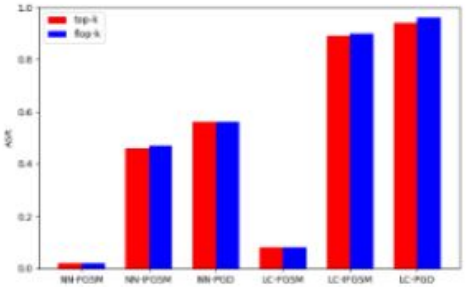
Methodology

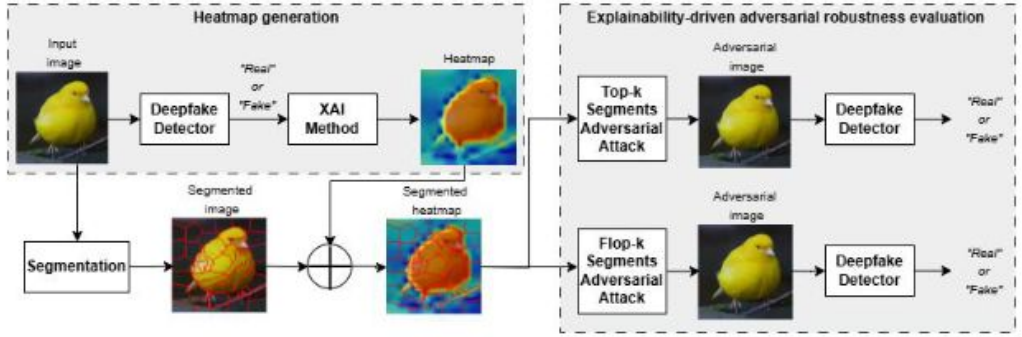
Explainability heatmaps to perform targeted adversarial attacks

Accuracy drop and attack success rate measurement

Findings

53% and 65% of attack success rate on the adopted models



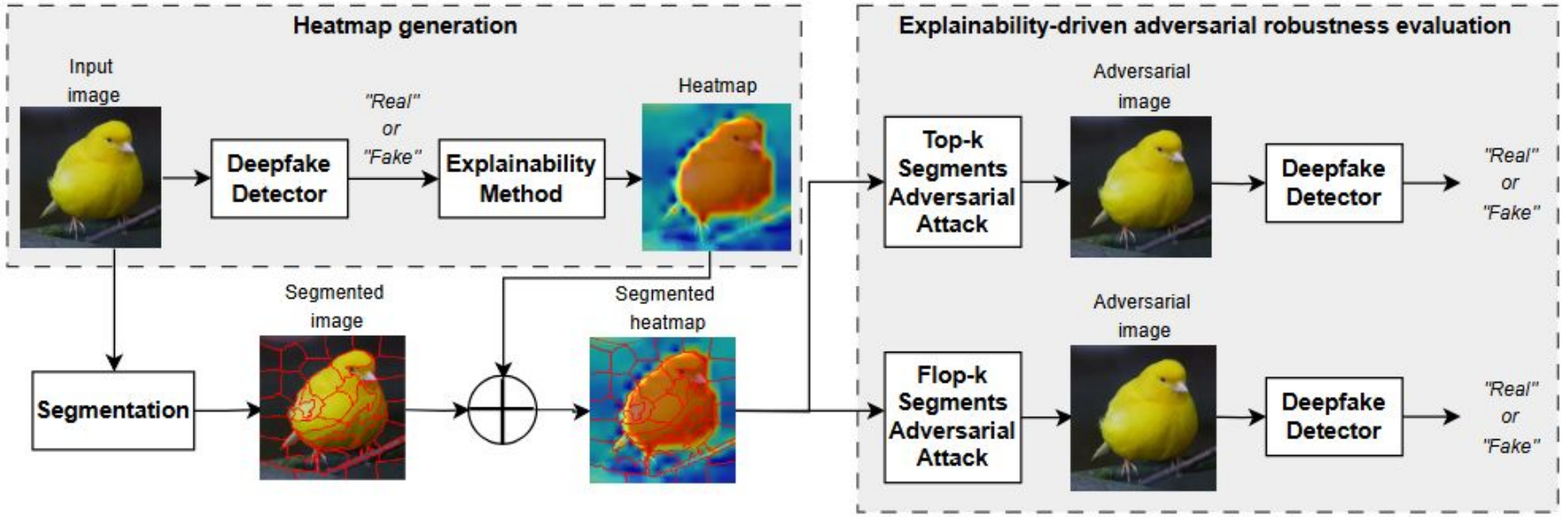


Conclusion: explainability analysis as a tool to reveal deepfake detector vulnerabilities to adversarial attacks

L. Cirillo,A. Gervasio, I. Amerini, *Explainability-Driven Adversarial Robustness Assessment for Generalized Deepfake Detectors*, in press.

Novelty: Explainability-Driven Attacks

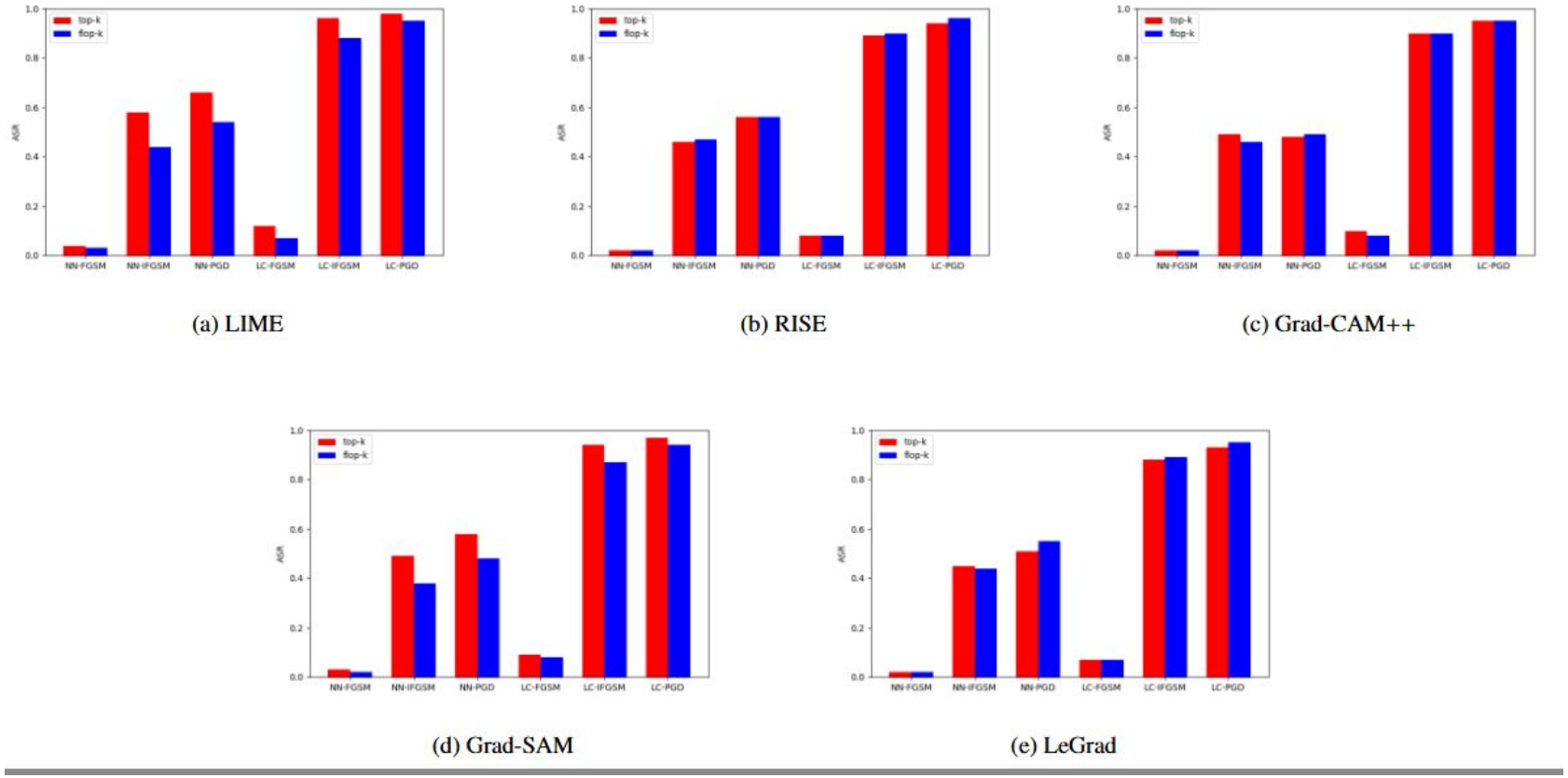
Study the adversarial robustness of generalized deepfake detectors by attacking the most and least relevant regions of the input image according to the visual explanation.



Proposed framework

L. Cirillo,A. Gervasio, I. Amerini, *Explainability-Driven Adversarial Robustness Assessment for Generalized Deepfake Detectors*, in press.

Results: XAI shows model vulnerabilities



(a) LIME (b) RISE (c) Grad-CAM++

(d) Grad-SAM (e) LeGrad

Attack Success Rate (ASR) on the analyzed deepfake detectors obtained by attacking the top-1 and flop-1 segments with FGSM, I-FGSM, and PGD.

L. Cirillo,A. Gervasio, I. Amerini, *Explainability-Driven Adversarial Robustness Assessment for Generalized Deepfake Detectors*, in press.

Robust Deep Learning for Computer Vision

Lorenzo Cirillo