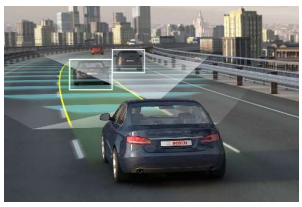
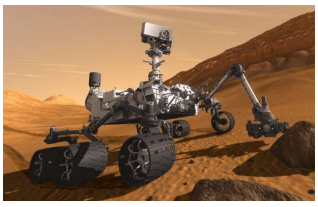
Support for Trustworthy AI in autonomous vehicles

Trustworthy AI

Due to their excellent performance, industry is going to use them in **safety-critical environments**:

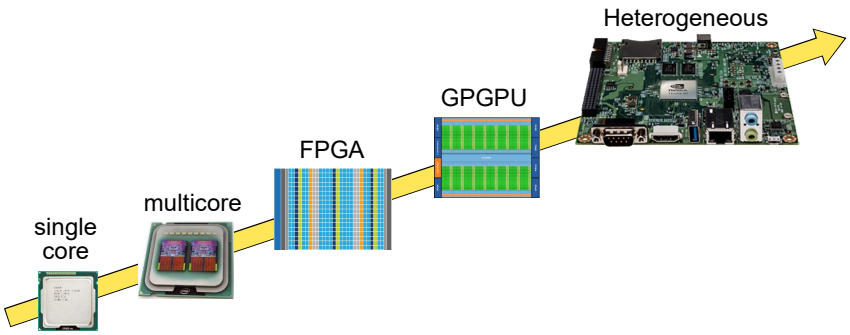
However, several important questions arise:

- Can we trust AI in all possible scenarios? **NO**
- Can AI algorithms respond in a predictable time? **NO**
- Can AI systems be certified? **NO**
- Are AI systems robust to cyber attacks? **NO**
- Are current CPS resilient to AI failures? **NO**

2

Computer evolution

To provide the required computational power, computer architectures evolved towards **heterogeneous platforms** that integrate multicore processors, FPGA, GPGPUs, TPUs, etc.



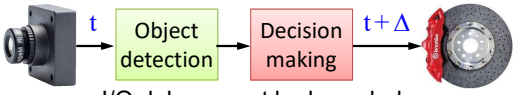
Developing safety-critical software on such platforms, while providing the required guarantees, is quite **challenging**.

3

Research addressed @ RETIS Lab

Architecture — HW acceleration and efficient use of heterogeneous platforms

Predictability



I/O delays must be bounded

Security — We must prevent, detect, and handle cyber attacks

Safety — We must plan for a backup solution if AI fails

Explainability — We must understand why a decision is taken

4

GPU vs. FPGA

	GPU	FPGA
Ease of development	✓	✗ ↗
Support for DNN frameworks	✓	✗ ↗
Performance (MIPS)	✓	✓
Energy consumption	✗	✓
Performance per Watt	✗	✓
Predictability	✗	✓
Weight & size	✗	✓

There is little to do with these drawbacks

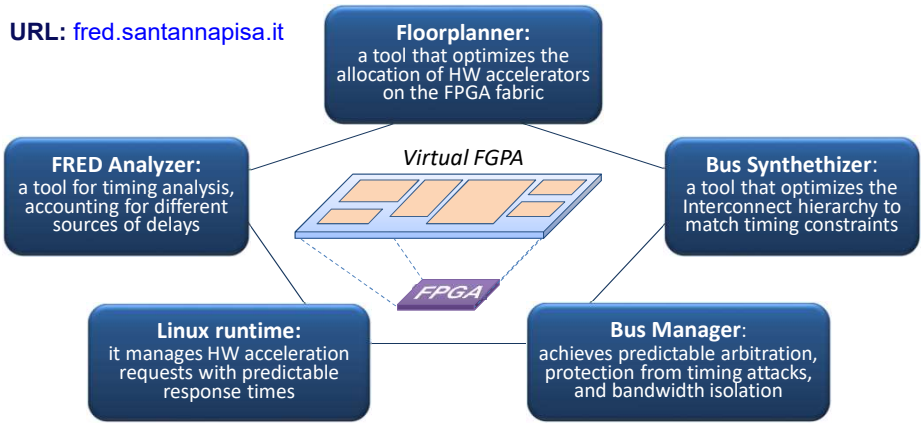
New solutions are coming to improve these drawbacks

5

FPGA support

The FRED Framework

URL: fred.santannapisa.it



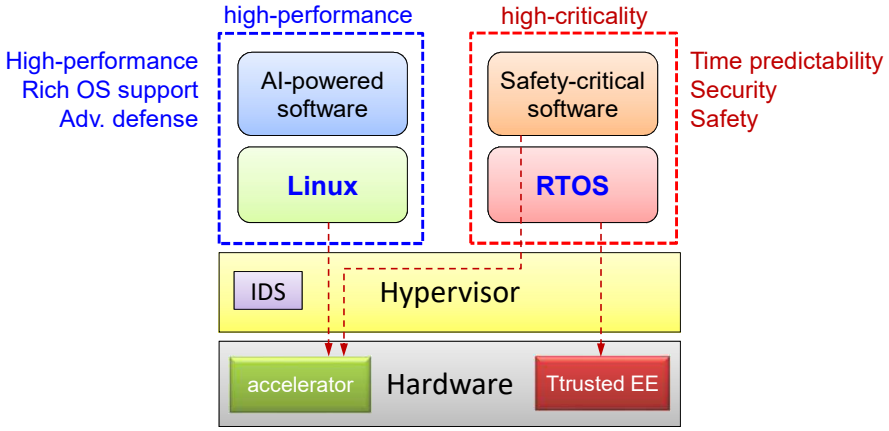
Paper

A. Biondi et al., "A Framework for Supporting Real-Time Applications on Dynamic Reconfigurable FPGAs", Proc. of the IEEE Real-Time Systems Symposium, 2016.

6

Architecture support

AI processing and RT control have different requirements:

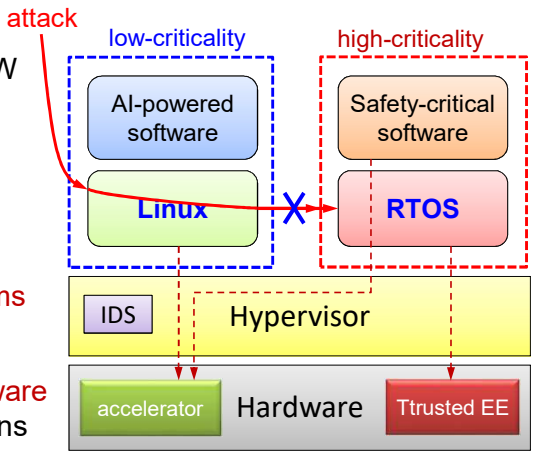


7

Architecture support

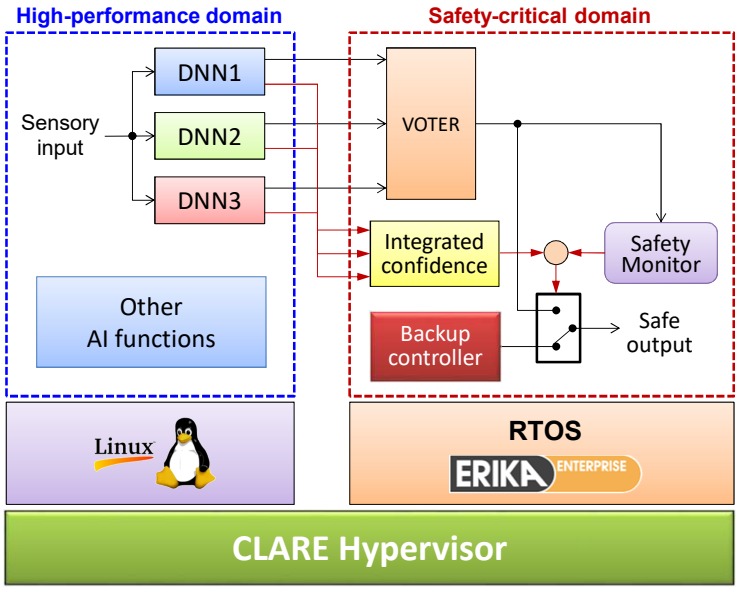
Security must be achieved at different architecture levels:

- Separate low-criticality SW from high-criticality one
- Hypervisor technology to isolate low-criticality SW from high-criticality ones
- Intrusion detection systems to detect cyber attacks
- Make use of trusted firmware to protect critical operations



8

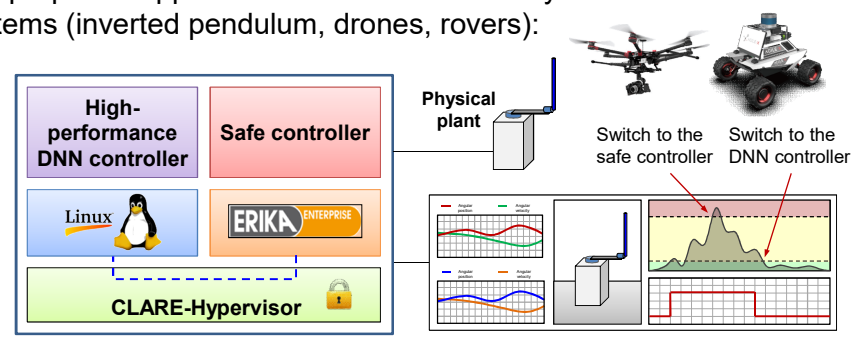
Architecture support

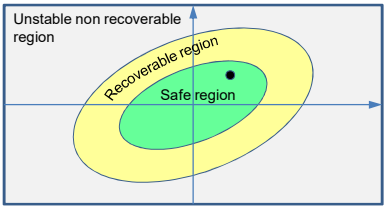


9

Proof of concept

The proposed approach has been successfully tested on different systems (inverted pendulum, drones, rovers):



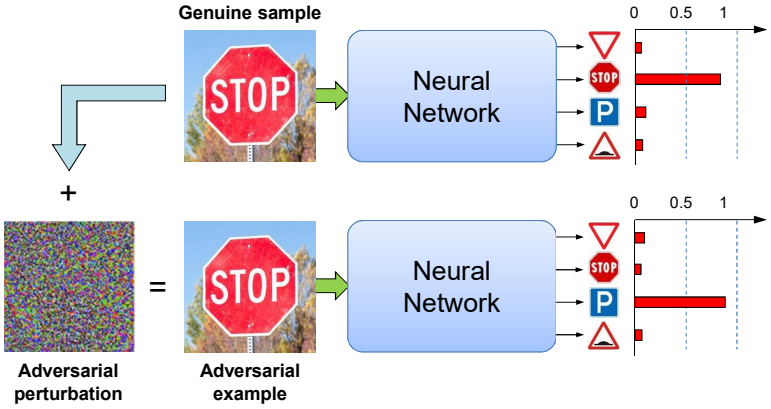


The safety monitor uses a Lyapunov approach to detect when the system state exits a safe region (green zone) and switch to the safe controller.

10

Adversarial attacks

Neural networks are prone to **adversarial examples**, i.e., malicious inputs containing imperceptible perturbations that can make a neural network to make wrong predictions.



11

Real-world AEs

Classic AEs modify the pixels of an image, hence the attacker must have access to the AI system (DNN input, memory, or camera).

Real-world AEs are directly applied to objects in the physical world, without accessing the AI system.

PARKING (92%)

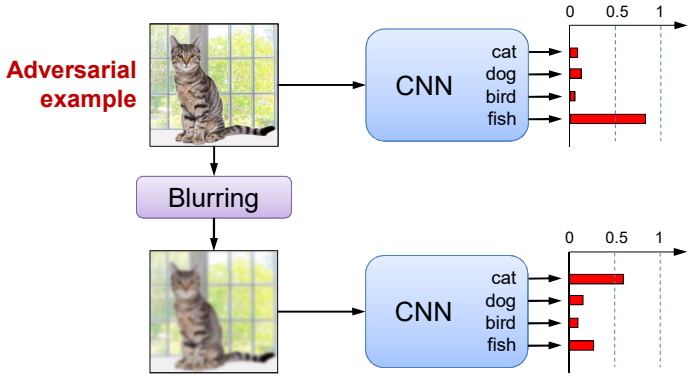
BRAD PITT (93%)

RIFLE (91%)

12

Input transformations

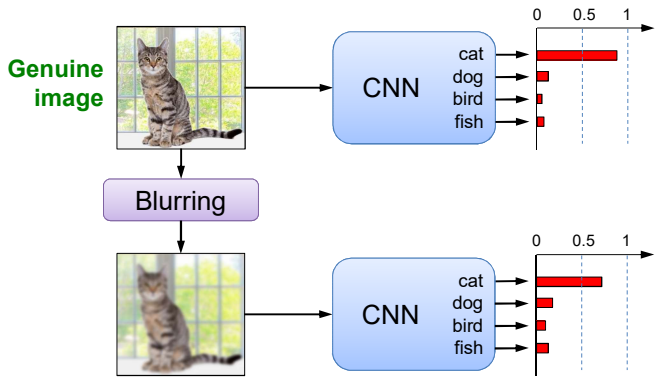
Another approach exploits the fact that standard AEs loose their effect when they are subject to certain **input transformations** (e.g., blurring, translation, rotations):



13

Input transformations

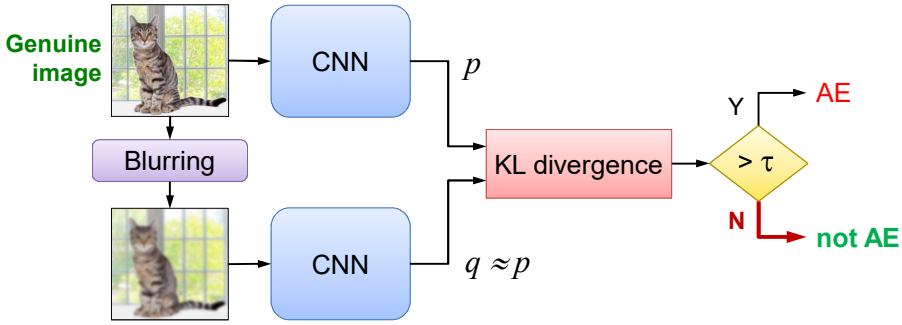
For genuine images, the same transformations do not cause a strong degradation in the prediction:



14

Input transformations

Nesti et al. [TNNLS21] proposed a detection method that compares the two distributions using a KL-divergence: a sample is considered to be AE if the two predictions are "distant" from each other:



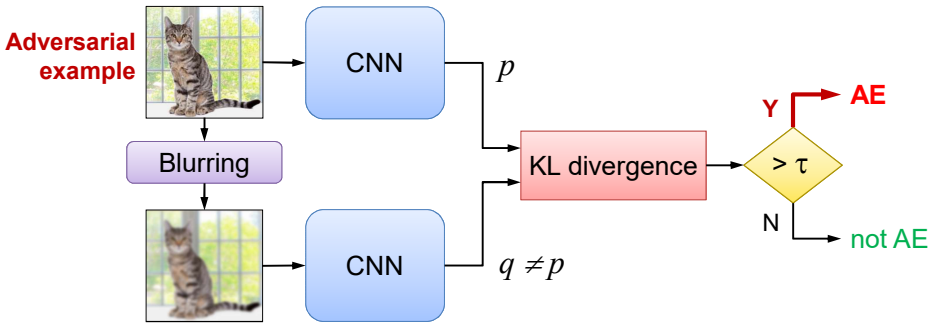
Paper

F. Nesti, A. Biondi, and G. Buttazzo, "Detecting Adversarial Examples by Input Transformations, Defense Perturbations, and Voting", IEEE Transactions on Neural Networks and Learning Systems, 2021.

15

Input transformations

Nesti et al. [TNNLS21] proposed a detection method that compares the two distributions using a KL-divergence: a sample is considered to be AE if the two predictions are "distant" from each other:



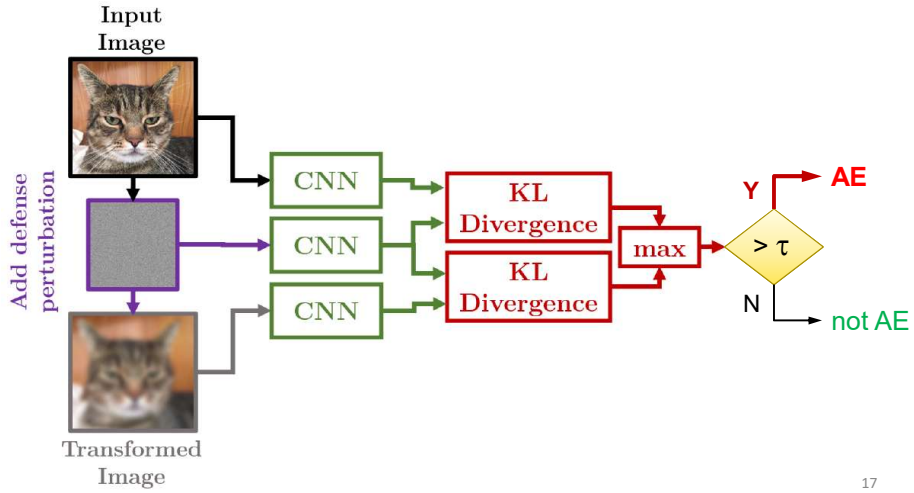
Paper

F. Nesti, A. Biondi, and G. Buttazzo, "Detecting Adversarial Examples by Input Transformations, Defense Perturbations, and Voting", IEEE Transactions on Neural Networks and Learning Systems, 2021.

16

Detecting robust AEs

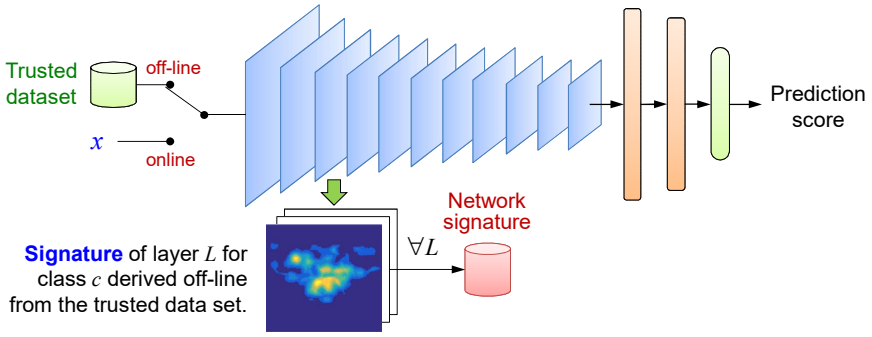
Exploiting more complex architectures, the method is also able to detect AEs that are **robust** to different transformations and DNN models:



17

Coverage analysis

We proposed a coverage-based method to analyze the activations produced by a trusted dataset ∀ class and ∀ layer and aggregate them in a sort of **signature**, representing what the DNN has learned:



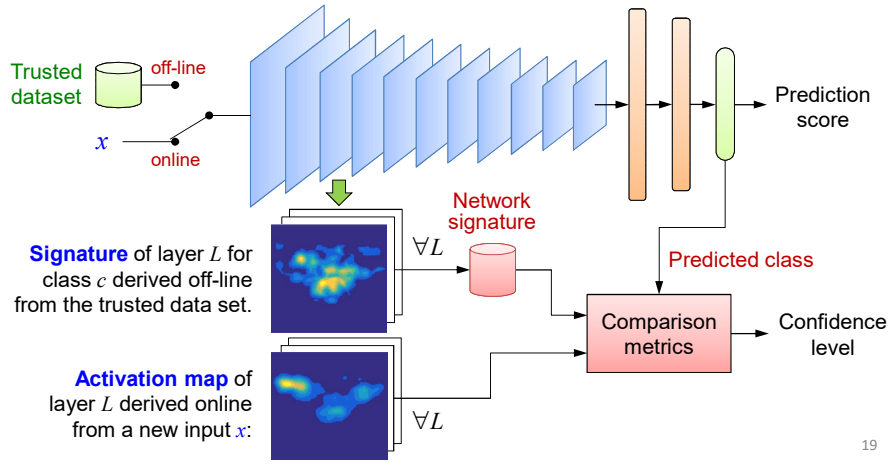
Paper

G. Rossolini, A. Biondi, and G. Buttazzo, "Increasing the Confidence of Deep Neural Networks by Coverage Analysis", arXiv:2101.12100 [cs.LG], January 2021.

18

Coverage analysis

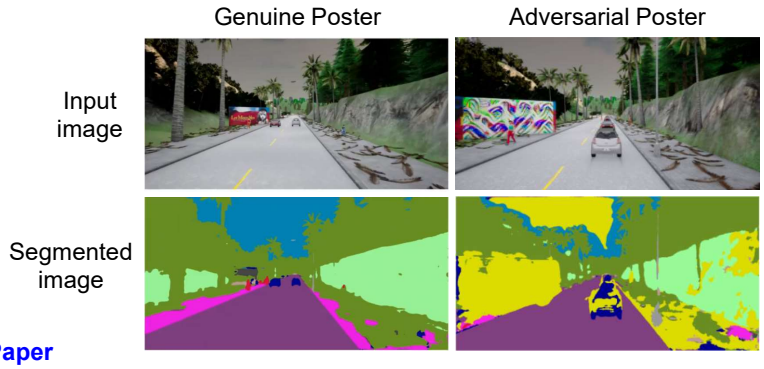
Then, for a new input x, the current activation state is compared with the stored **signature** corresponding to the predicted class. The higher the matching, the higher the confidence.



19

Real-world AE for SS

We performed an extensive experimental study to evaluate the robustness of semantic segmentation networks against real-world adversarial attacks, based on patches and physical posters:



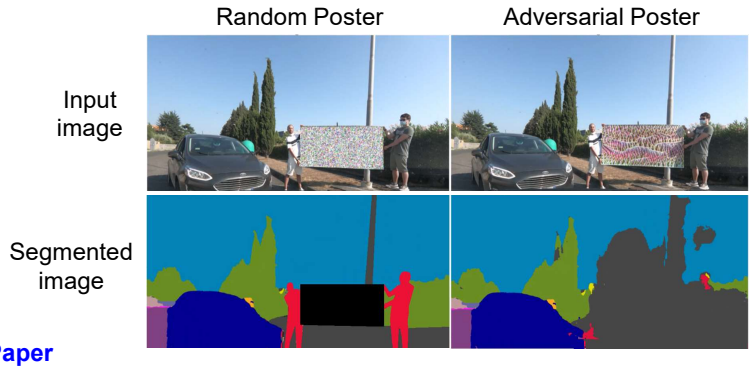
Paper

F. Nesti, G. Rossolini, S. Nair, A. Biondi, and G. Buttazzo, "Evaluating the Robustness of Semantic Segmentation for Autonomous Driving against Real-World Adversarial Patch Attacks", Proc. of WACV 2022.

20

Real-world AE for SS

We performed an extensive experimental study to evaluate the robustness of semantic segmentation networks against real-world adversarial attacks, based on patches and physical posters:



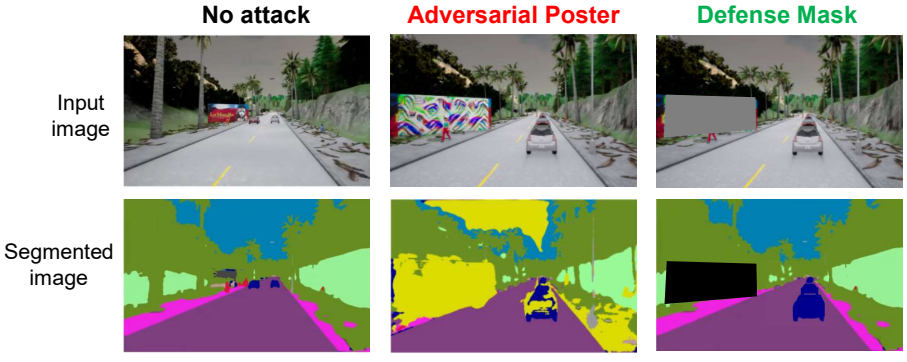
Paper

F. Nesti, G. Rossolini, S. Nair, A. Biondi, and G. Buttazzo, "Evaluating the Robustness of Semantic Segmentation for Autonomous Driving against Real-World Adversarial Patch Attacks", Proc. of WACV 2022.

21

Real-world Adv. attacks

Rossolini et al. developed a new defense method for real-world adversarial attacks identifying and **masking the adversarial region**:



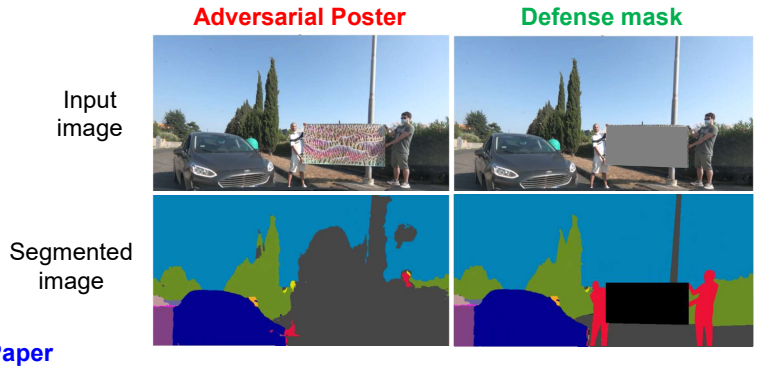
Paper

G. Rossolini, F. Nesti, F. Brau, A. Biondi, and G. Buttazzo. "Defending from physically-realizable adversarial attacks through internal over-activation analysis", arXiv:2203.07341 2022.

22

Real-world Adv. attacks

Nesti et al. performed an extensive experimental study to evaluate the robustness of semantic segmentation networks against real-world adversarial attacks, based on patches and physical posters:



Paper

F. Nesti, G. Rossolini, S. Nair, A. Biondi, and G. Buttazzo, "Evaluating the Robustness of Semantic Segmentation for Autonomous Driving against Real-World Adversarial Patch Attacks", WACV 2022.

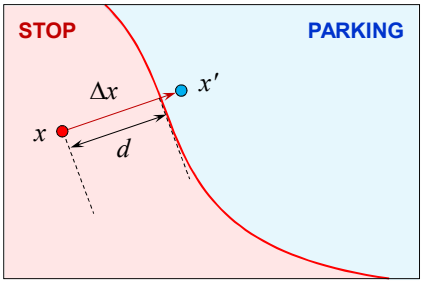
23

Provable robustness

The problem of defense methods is that they do not provide any guarantee.

To address this, we are also investigating theoretical methods to formally prove the robustness of a DNN against adversarial attacks.

If we can compute the distance d of an input x to the closest classification boundary, then we can guarantee that input x is safe for any attack with magnitude less than d.



Paper

F. Brau, G. Rossolini, A. Biondi, and G. Buttazzo, "On the Minimal Adversarial Perturbation for Deep Neural Networks with Provable Estimation Error ", arXiv:2201.01235, 2022.

24