With funding from the:

Federal Ministry of Research, Technology and Space

ConnRAD – Connectivity and Resilience for Automated Driving

Final Presentation **23.10.2025**

ConnRAD

Explainable and Trustworthy AI

Dennis Eisermann, Artur Hermann, Frank Kargl

Relevance for Autonomous Driving

Perception

- **Sensory:** Observation of Environment for Proper Decisions
- **Artificial Intelligence (AI):** Abstraction from Concrete Values to Symbols and Decisions

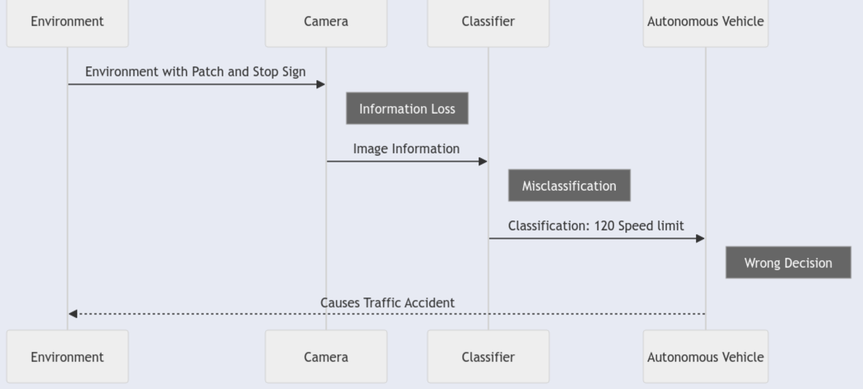
Secure and Resilient

- **Vulnerable to Evasion Attacks**: Intended Manipulation of Input to Cause Misclassification
- **Adversarial Patches** : Special Attack form to Manipulate Camera Feed with Patches
- **Universal Naturalistic Patches**: Patches with Natural Motifs instead of Noise

Explainable and Interpretable

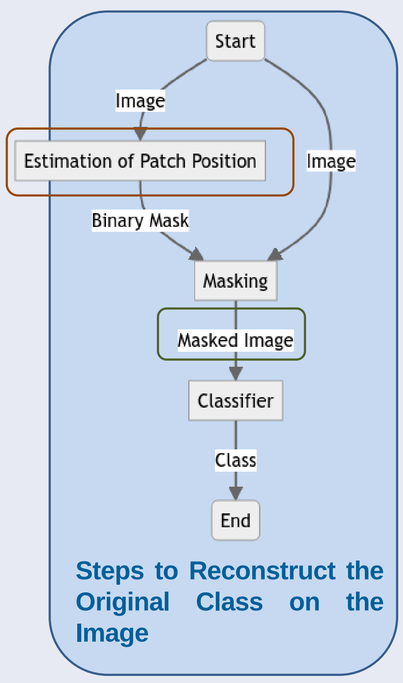
- **Decision Extraction:** Necessary to Estimate Trustworthiness
- **Saliency Maps:** Indicator for Relevant Areas of the Decision Process
- **Saliency Sentinel:** Our New Method to Mitigate Adversarial Patches

Adversarial Attacks against Autonomous Vehicles

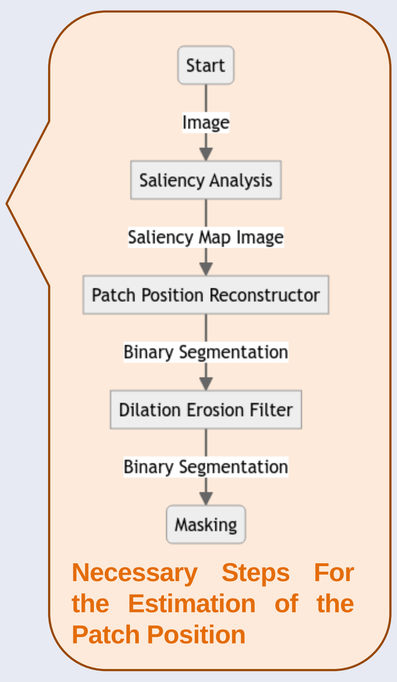


Sequence Diagram Illustrating the Step-by-Step Realization and Execution of an Evasion Attack Against the Perception and Decision-Making Systems of an Autonomous Vehicle

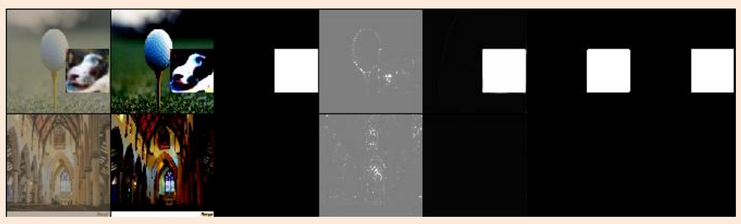
Saliency Sentinel: Defense System Against Universal Naturalistic Adversarial Patches



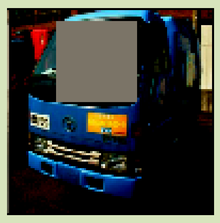
Steps to Reconstruct the Original Class on the Image



Necessary Steps For the Estimation of the Patch Position



Example Visualization of each step to predict the position



The Saliency Sentinel covers the area of the Universal Naturalistic Patch to mitigate its tampering. This modified image will be shown to the classifier. The classifier predicts it as Garbage Truck correctly.

Experimental Results

Model	Attack Sucess Rate	Clean Accuracy
VGG16	83.85%	98.52%
Saliency Sentinel	19.64%	98.49%

Comparison of VGG16 and Saliency Sentinel in Case of Attacks with Untargeted Universal Naturalistic Patches and without any Attack on the Imagenette Dataset

Trustworthy AI Based on NIST Artificial Intelligence Risk Management Framework

- **Valid and Reliable**: Systems should perform consistently across conditions, contexts, and datasets in line with their intended purpose.
- **Safe**: AI must avoid behavior that causes harm to humans, systems, or environments.
- **Secure and Resilient**: Systems must withstand malicious interference like adversarial attacks
- **Accountable and Transparent**: Actors in the AI lifecycle should be traceable and governance structures documented.
- **Explainable and Interpretable**: AI should provide comprehensible rationales for outputs.
- **Privacy-Enhanced**: AI systems should minimize personal data usage
- **Fair – with Harmful Bias Managed**: Systems must detect and mitigate bias at dataset, algorithmic, and outcome levels.