UNICAD: A Unified Approach for Attack Detection, Noise Reduction and Novel Class Identification

Lancaster University

Researcher: Alvaro Lopez Pellicer

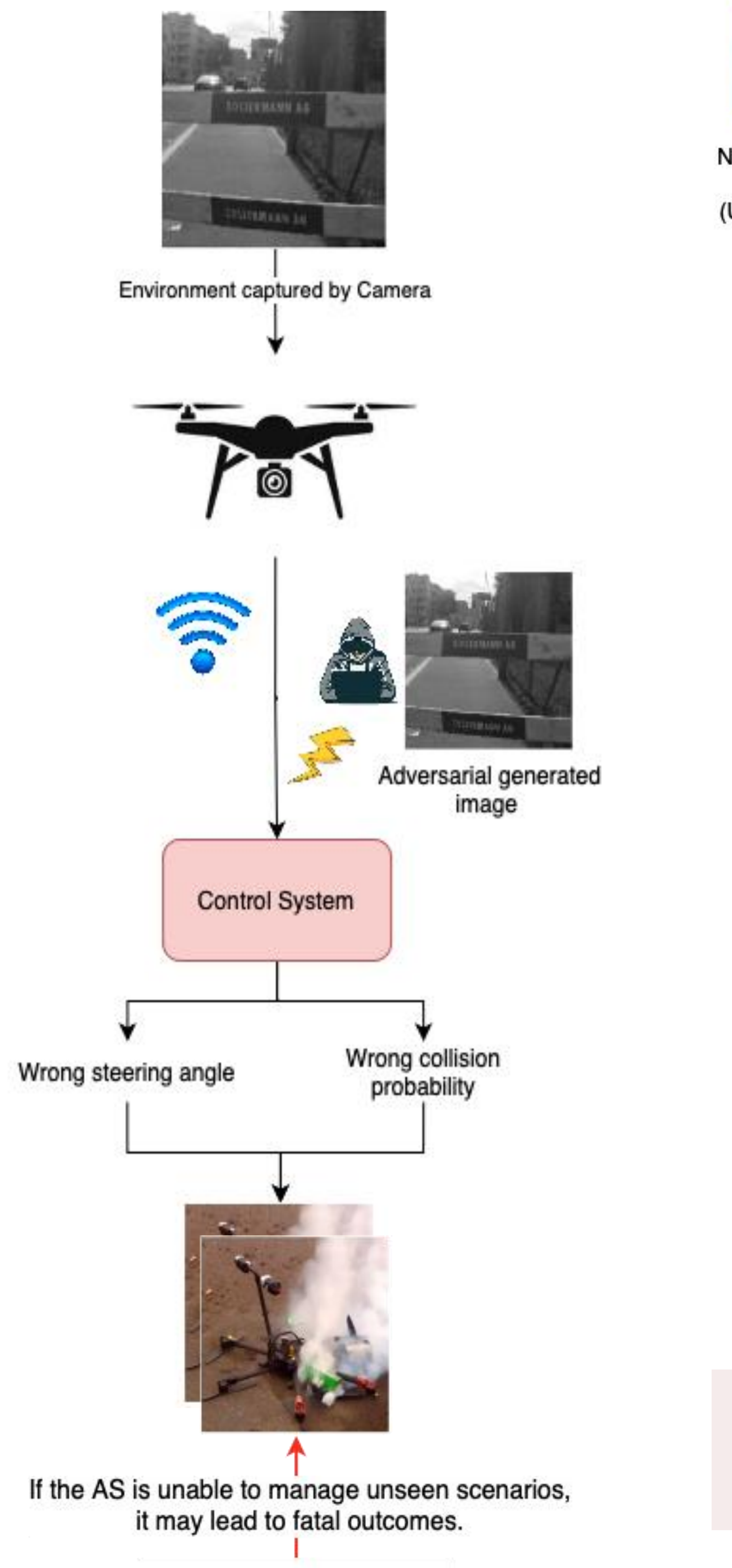
Supervisors: Prof. Plamen Angelov, Prof. Neeraj Suri

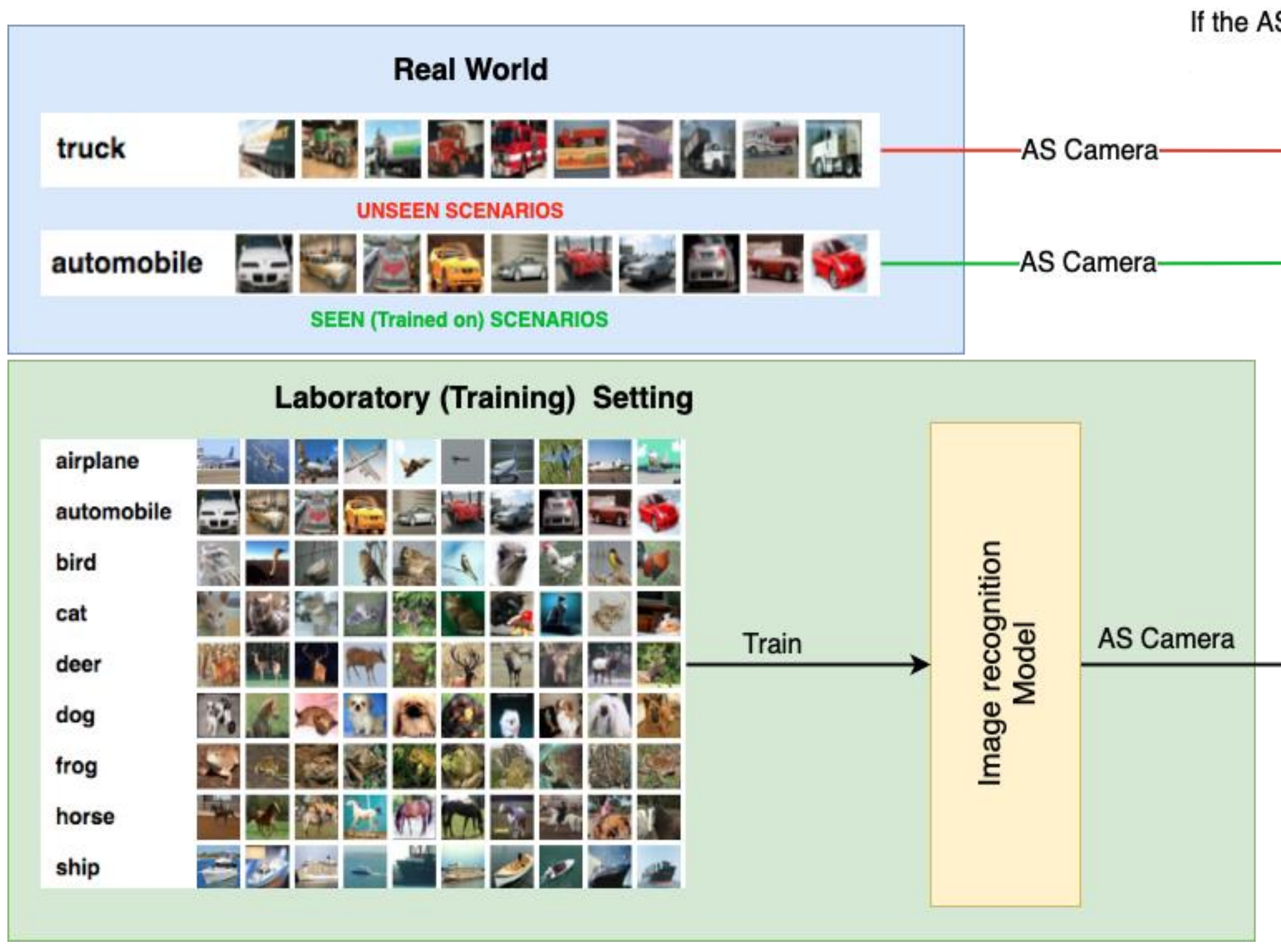
Challenges for Autonomous Systems

Autonomous systems face numerous challenges in their operation due to the uncertain and dynamic multi-layer attack surfaces

Critical Challenges

- **Accurate operation**: Autonomous Systems (AS) consist of complex ensembles of interconnected components including sensors, actuators, communication modules and control algorithms, that collaboratively perform tasks with minimal to no human intervention. Often, these systems rely on image sensing for perception and decision-making in the physical environment. Each discrete component needs to individually perform at the requisite level of accuracy in order to result in a collectively stable AS operation.
- **Safety**: Autonomous Systems are safety-critical and increasingly utilize Deep Neural Networks for multiple tasks (DNNs). Adversarial attacks are one of the most critical challenges for DNNs and AS. These attacks can take various forms, such as data poisoning, model inversion, or evasion, and can have serious consequences for the safety, reliability, and privacy.
- **Unknown scenarios**: DNNs are often trained in a set of known scenarios. For instance, they may be trained to identify different objects in aerial images, however, when a new object appears in which the systems hasn't been trained on, often It would be misclassified. AS need to be able to detect and handle unknown scenarios, failure to do so could lead to catastrophic consequences.





Why do we need a unified solution?

Autonomous Systems need to:

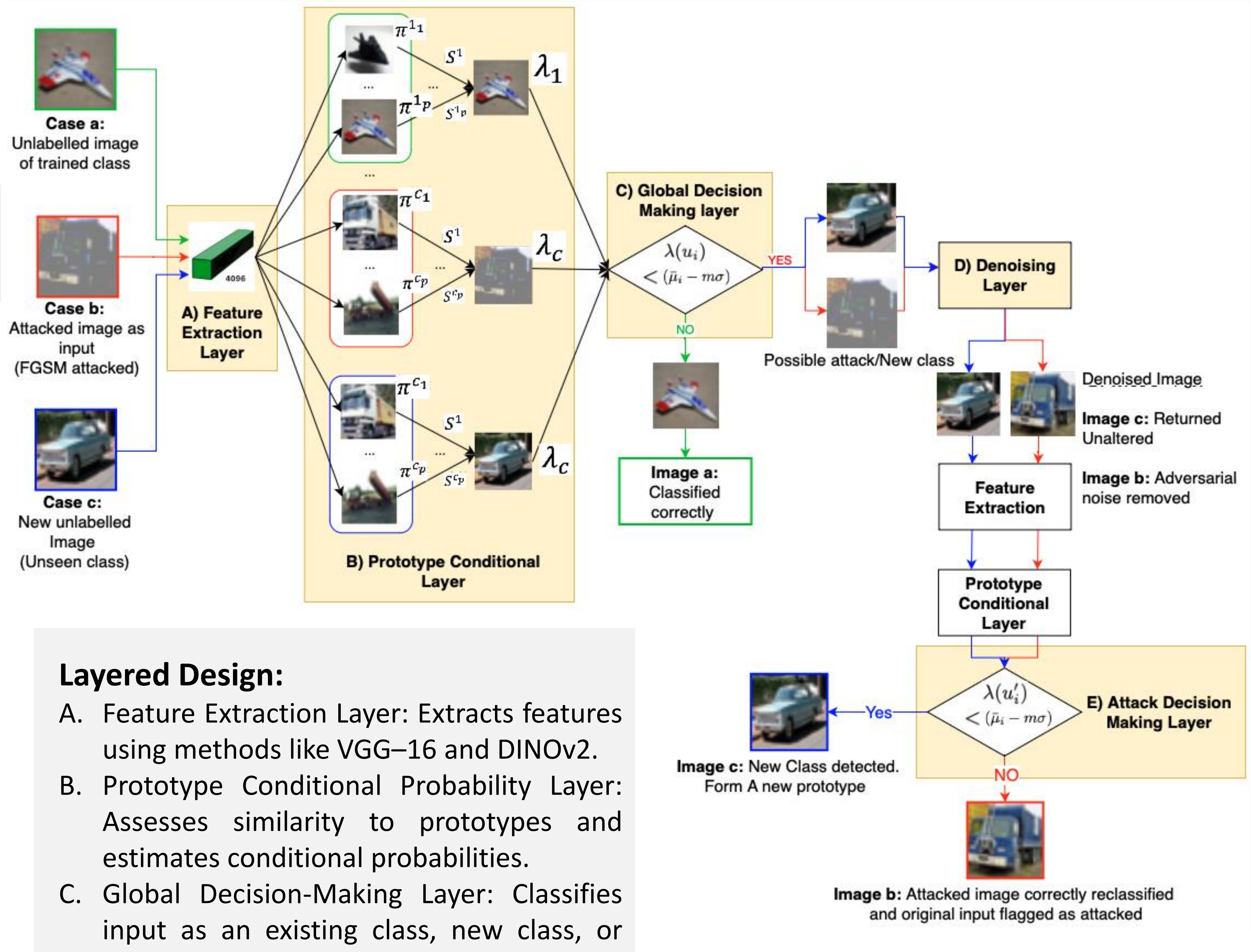
- ✓ Operate correctly under trained scenarios
- ✓ Be able to recognize (detect) Adversarial Attacks in real time
- ✓ React to adversarial attacks often by mitigating their impact
- ✓ Detect unseen scenarios out of the scope of original training
- ✓ React to unseen scenarios

Our proposed solution (UNICAD) compared to others:

Method	Unseen Class detection	Attack detection	Attack Recovery
Previous work on unseen class detection (xClass)	✓	✗	✗
Previous work on attack detection (simDNN)	✗	✓	✗
Denoising Autoencoders (DAE)	✗	✗	✓
Ours	✓	✓	✓

UNICAD Framework Overview

A novel architecture integrating state-of-the-art techniques for efficient adversarial attack detection, noise reduction, and novel class recognition



Layered Design:

A. Feature Extraction Layer: Extracts features using methods like VGG–16 and DINOv2.
B. Prototype Conditional Probability Layer: Assesses similarity to prototypes and estimates conditional probabilities.
C. Global Decision-Making Layer: Classifies input as an existing class, new class, or adversarial attack.
D. Denoising Layer: A denoising autoencoder that removes adversarial noise while preserving the integrity of clean inputs.
E. Attack Decision Making Layer: Re-evaluates denoised images to determine if they represent a new class or an attack.

Key Features

- Effective in detecting adversarial attacks and data concept drifts (Unseen scenarios).
- Reduces Adversarial noise using advanced Denoising Autoencoders.
- Identifies new classes using similarity-based neural networks.
- Maintains performance in seen and not attacked scenarios (Normal scenarios)

Results and Discussion

Scenario	Accuracy (%)					
	UNICAD (VGG-16 FE)	UNICAD (DINOv2 FE)	xClass (VGG-16 FE)	Traditional DAE (Defence)	VGG-16 (No defence)	DINOv2 (No defence)
Clean	80.86	92.93	80.86	81.4	92.0	**97.63**
PGD (ε = 0.01)	74.81	**77.77**	49.3	60.00	0.05	56.8
PGD (ε = 0.3)	72.63	**82.29**	14.2	61.10	32.12	0.7
FGSM (ε = 0.01)	70.01	**77.37**	49.0	63.00	31.06	58.9
FGM (ε = 0.03)	64.9	**76.02**	47.1	61.00	22.56	17.3
FGM (ε = 0.3)	73.09	**81.10**	15.6	57.10	0.11	12.4
C&W (L2 norm)	73.2	**79.33**	0.6	36.70	0.00	0.8
Unseen Class detection	62.30	**83.38**	62.30	0.00	0.00	0.00

Experimental Setup

- Framework Validation: CIFAR-10 datasets to evaluate UNICAD's robustness.
- Unseen class setting: UNICAD and comparative methods trained on CIFAR-10 classes 0-8, leaving class 9 (trucks) unseen.
- Performance Assessment Criteria: Analysing the classification accuracy of comparative methods in clean settings, against FGSM, PGD, C&W attacks and unseen scenarios. Accuracy measured on the CIFAR-10 testing dataset and CIFAR-9 for unseen class detection.
- Unseen Class Detection metric: Detection(%) = (TP + TN) / (TP + FP + TN + FN) x 100

Key Results

- UNICAD with VGG-16 FE: Clean image classification slightly higher than DDSA, over 70% accuracy in adversarial attacks (FGSM, PGD, C&W), comparable unseen class detection to xClass.
- UNICAD with DINOv2 FE: Enhanced feature extraction leading to superior performance, significant lower accuracy drop in adversarial attacks, robust in unseen class detection.
- Performance Comparison: Demonstrates robustness in adversarial attack scenarios, effective in unseen class detection, balanced approach to classification accuracy and security against attacks.

Future work

- **Exploring Latent Space Similarities**: Investigate the similarity of different classes in the latent space and its relevance, especially for closely related classes like trucks and automobiles.
- **Optimizing Denoising Layer**: Work on optimizing the denoising layer to better adapt to a variety of changing and evolving adversarial attacks.
- **Exploring Real-world scenarios:** Test UNICAD on closer to real-world scenarios such as simulations or more realistic datasets such as LARD dataset.

This work is supported by the Engineering and Physical Sciences Research Council [Grant number: EP/V026763/1]