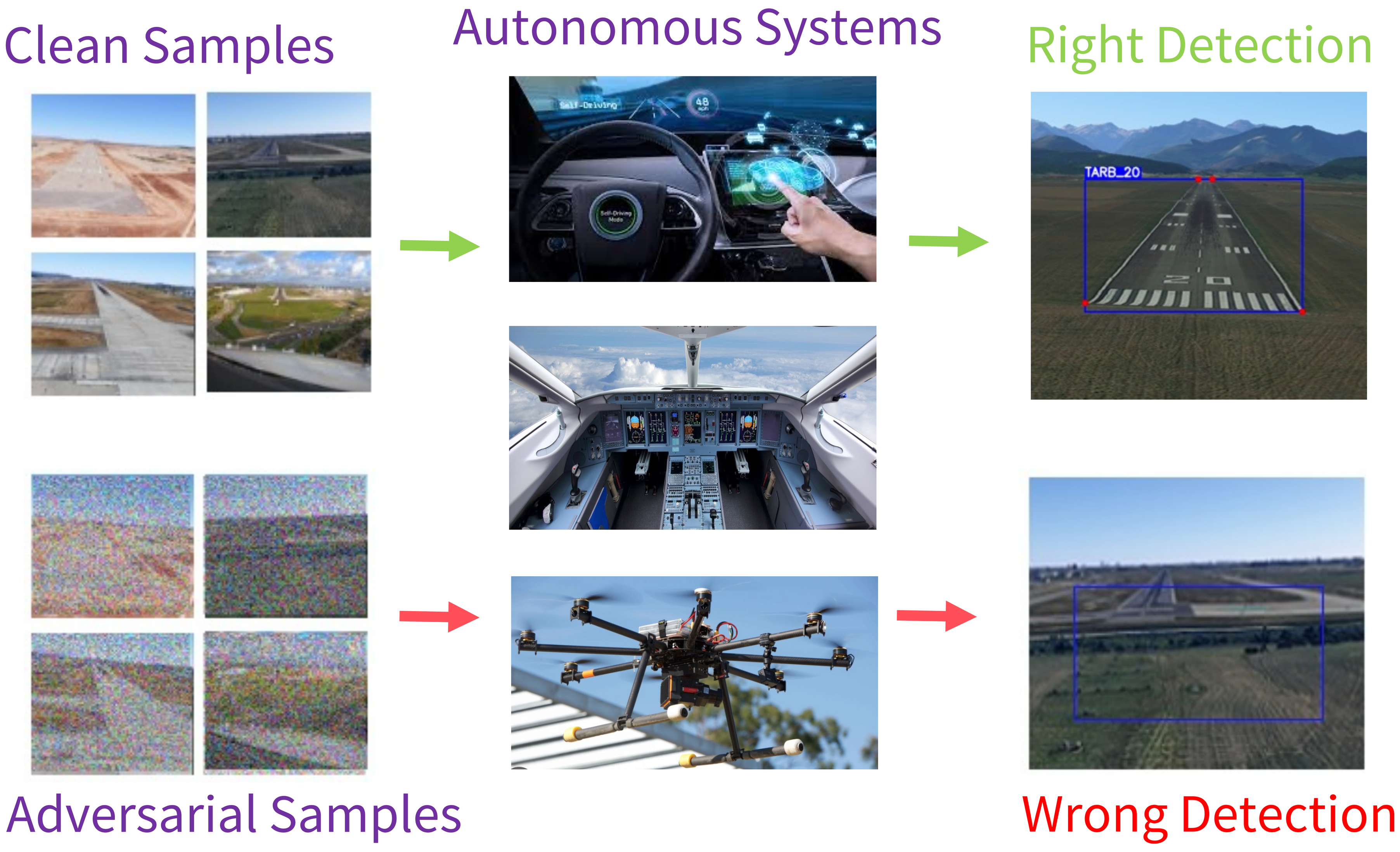
Rethinking Self-supervised Learning for Cross-domain Adversarial Image Recovery

Lancaster University

Research Fellow: Dr. Yi Li
Investigators: Prof. Plamen Angelov, Prof. Neeraj Suri

Adversarial Attacks to Autonomous Systems

Autonomous Systems (AS) are usually embodied as *Cyber-Physical Systems* (CPS) in which adversarial attacks can lead to catastrophic consequences, such as loss of life or serious injury, thus many autonomous systems are **safety-critical**.



Self-supervised Learning

What is self-supervised learning (SSL):

- Unlabeled data is processed to obtain useful representations that can help with downstream learning tasks.
- An intermediate form of unsupervised and supervised learning.

Why we need SSL-based adversarial attack recovery?

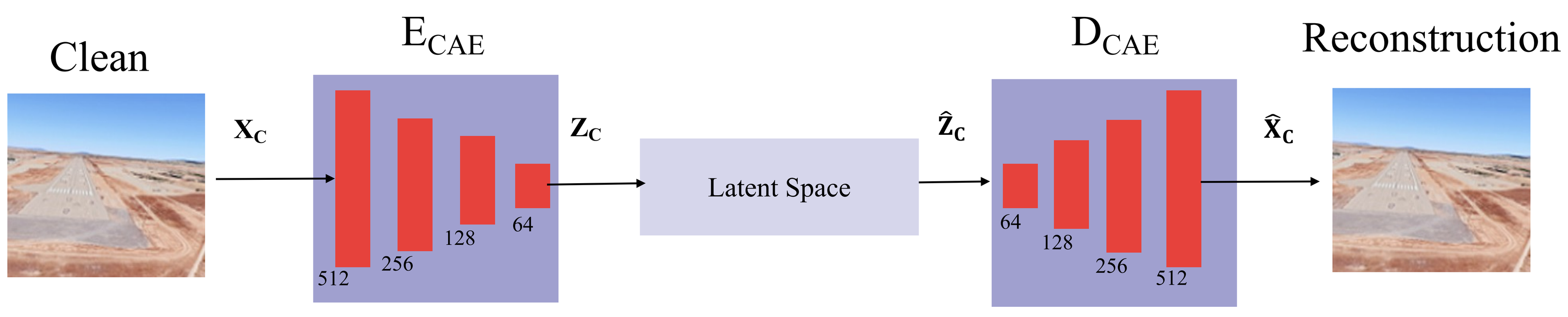
- Supervised training of the networks requires large sets of labelled paired data. However, these data is difficult or expensive to obtain.
- A trained model may suffer from performance degradation when deployed in previously unseen conditions e.g., a mismatch of attacks and datasets between the training and testing datasets.

What do we propose in this work?

- We propose the clean image autoencoder (CAE) to learn the latent representations of clean images.
- We propose the adversarial image autoencoder (AAE) to learn a shared latent space between the unpaired clean images and adversarial images to boost the generalization ability.
- The input of two autoencoders are clean images and adversarial images, respectively. However, they are unpaired, i.e., they are randomly selected different domains (datasets and attack algorithms).

Proposed Framework

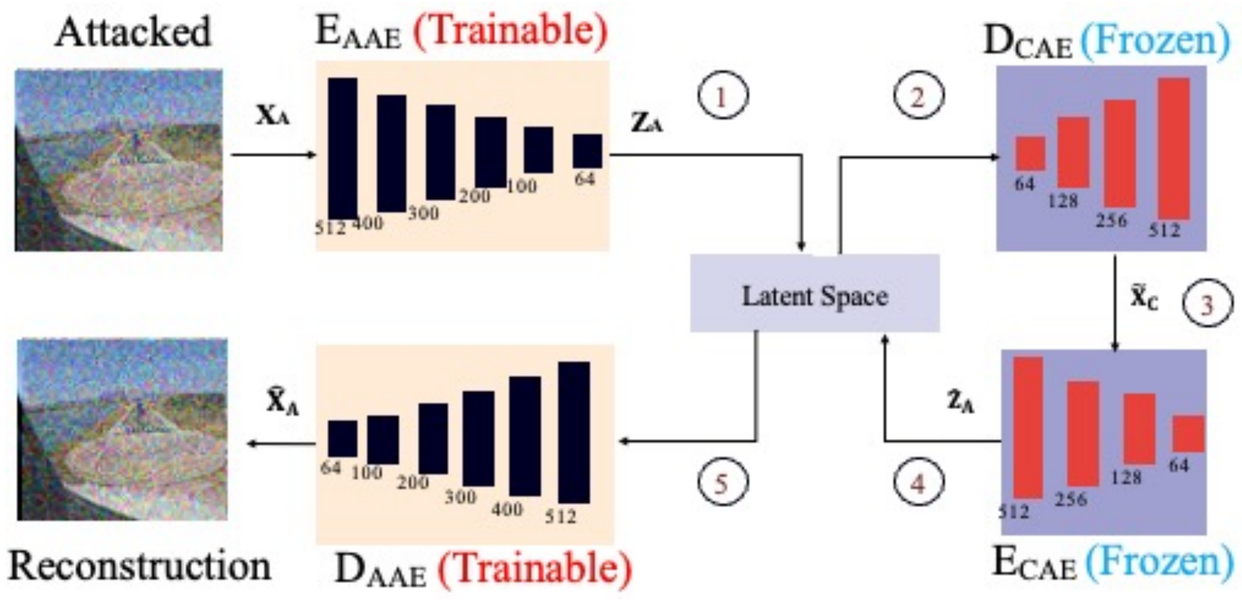
Clean image autoencoder training



- The clean images X_c from the public landing runway dataset are fed into the CAE to learn the features Z_c in the latent space.
- In CAE, both E_{CAE} and D_{CAE} consist of four 1-D convolutional layers. In E_{CAE}, the size of the hidden dimension decreases sequentially from 512 -> 256 -> 128 -> 64. Accordingly, the dimension of the latent space is set to 64, with the stride of 1 and the kernel size of 7 used for the convolutions. Different from E_{CAE}, the decoder D_{CAE} scale up the latent dimensions sequentially.

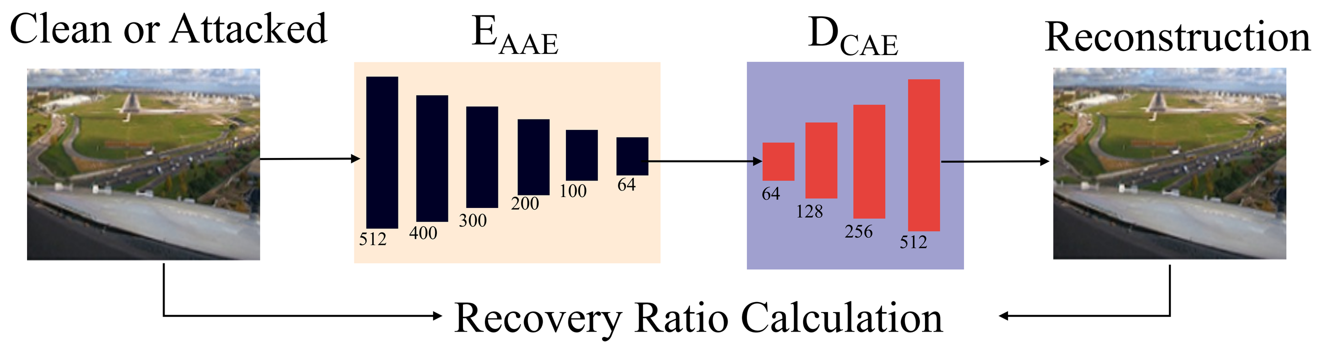
Proposed Framework

Adversarial image autoencoder training



- The weights of the CAE are frozen in this stage.
- The AAE learns a shared latent space between clean images and adversarial images.

Test stage



- The trained E_{AAE} and D_{CAE} are combined as the final model

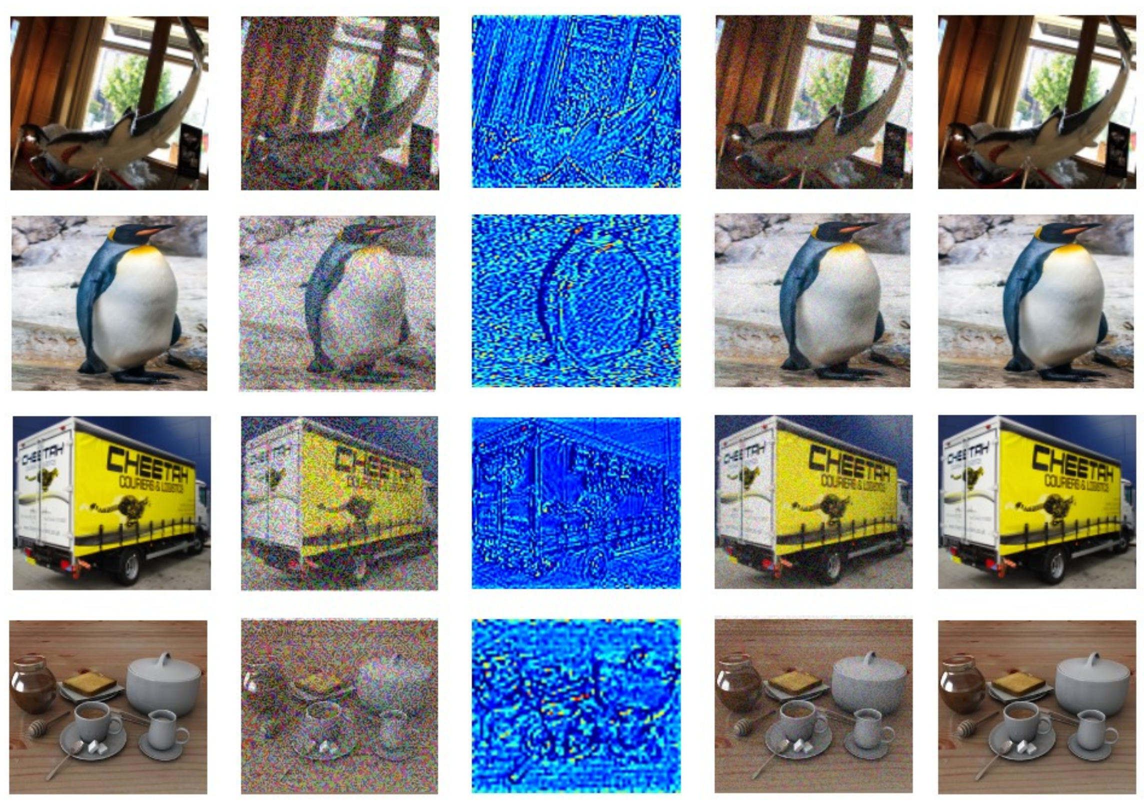
Experimental Settings

- CAE Training: 10k images from the COCO dataset
- AAE Training: 40k images from the CIFAR-10 dataset
- Test: 10k images from the ImageNet-R dataset
- Backbones: CNN
- Attack algorithms: FGSM, PGD, SSAH, DeepFool, BIM, CW, JSMA

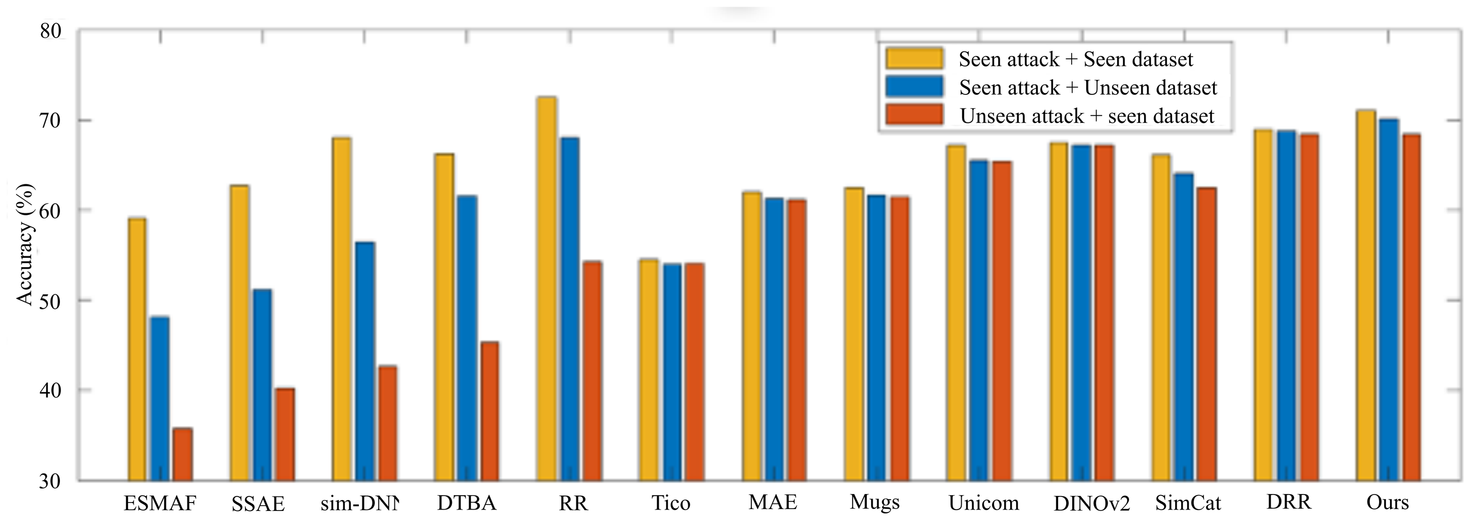
Experimental Results

	Recovery Ratio (%)								
	Clean	FGSM	PGD	SSAH	DeepFool	BIM	CW	JSMA	Avr
ESMAF	70.8	52.7	67.5	62.9	39.7	35.9	37.2	41.0	51.0
SSAE	74.0	58.5	67.0	69.2	41.5	39.4	41.6	41.3	54.1
Sim-DNN	76.2	60.7	72.3	71.0	44.8	46.7	49.9	50.2	59.0
DTBA	79.2	59.2	75.5	74.9	51.4	53.8	56.0	59.9	63.8
RR	86.5	62.7	79.0	76.2	67.1	58.7	60.9	71.3	70.3
TiCo	74.5	53.6	68.6	65.2	45.1	44.5	43.1	57.9	56.6
MAE	82.2	59.6	75.5	74.4	54.2	50.3	51.4	62.8	63.8
Mugs	83.4	57.2	75.9	76.7	56.0	51.1	50.8	64.3	64.4
Unicom	86.4	59.8	76.2	79.3	61.0	55.5	58.4	68.2	68.1
DINOv2	87.5	61.6	79.4	78.3	64.5	57.1	57.9	71.6	69.7
SimCat	85.1	58.0	75.2	77.0	56.4	56.5	55.3	69.6	66.6
DRR	87.2	64.8	79.6	78.2	66.9	60.7	60.1	70.3	71.0
Ours	**87.9**	**65.9**	**80.0**	**79.7**	**69.1**	**61.5**	**61.8**	**72.4**	**72.3**

Visualizations



- Results on the Image-R dataset.
- Supervised: ESMAF, SSAE, sim-DNN, DTBA, RR
- Self-supervised: Tico, MAE, Mugs, Unicom, DINOv2, SimCat, DRR



Ongoing and Future Works

- The proposed framework is potentially applied in other downstream tasks, e.g., road condition detection.
- Ablation study of the proposed algorithm will be provided.

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