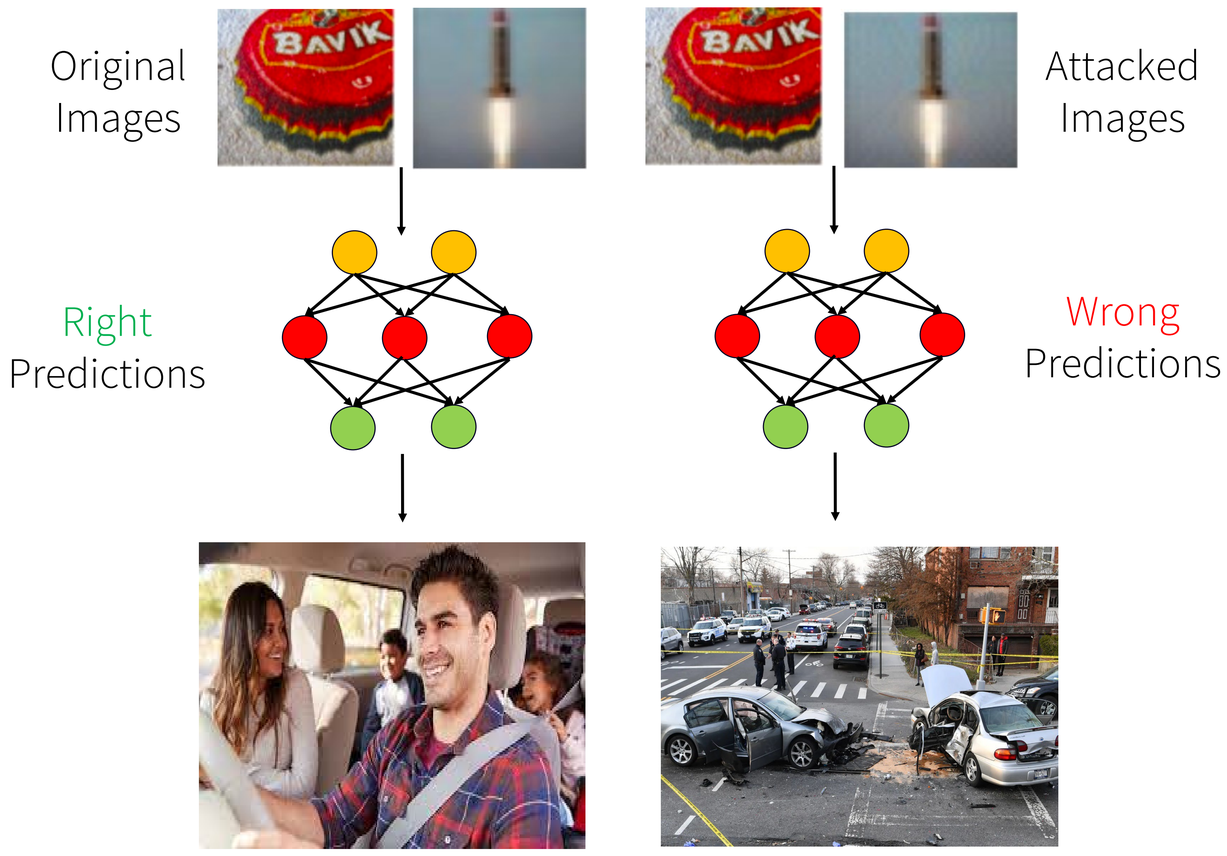
Domain Generalization and Feature Fusion for Cross-domain Imperceptible Adversarial Attack Detection

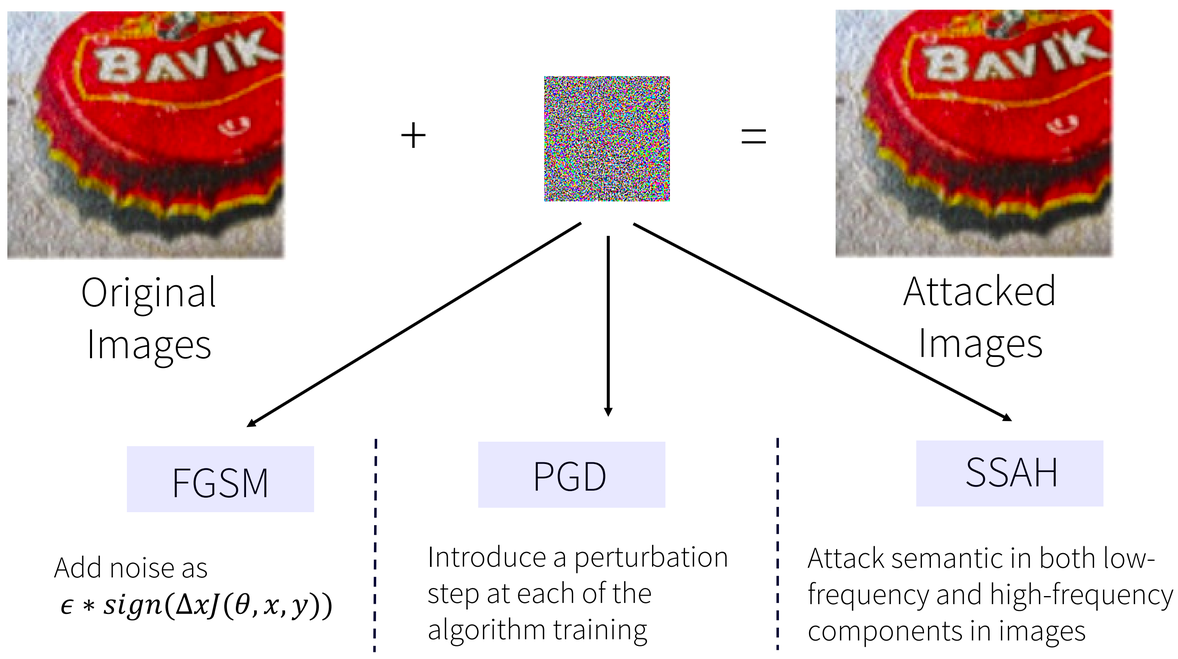
Lancaster University

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

Background: Imperceptible Adversarial Attack Detection

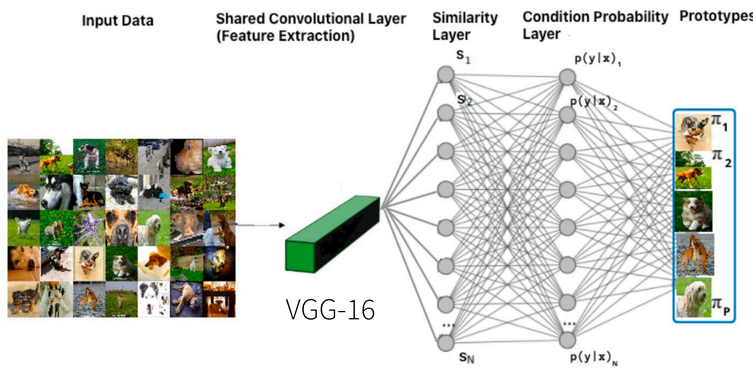


Attacks



Learning-Based Detection Methods: State-of-the-art

Sim-DNN



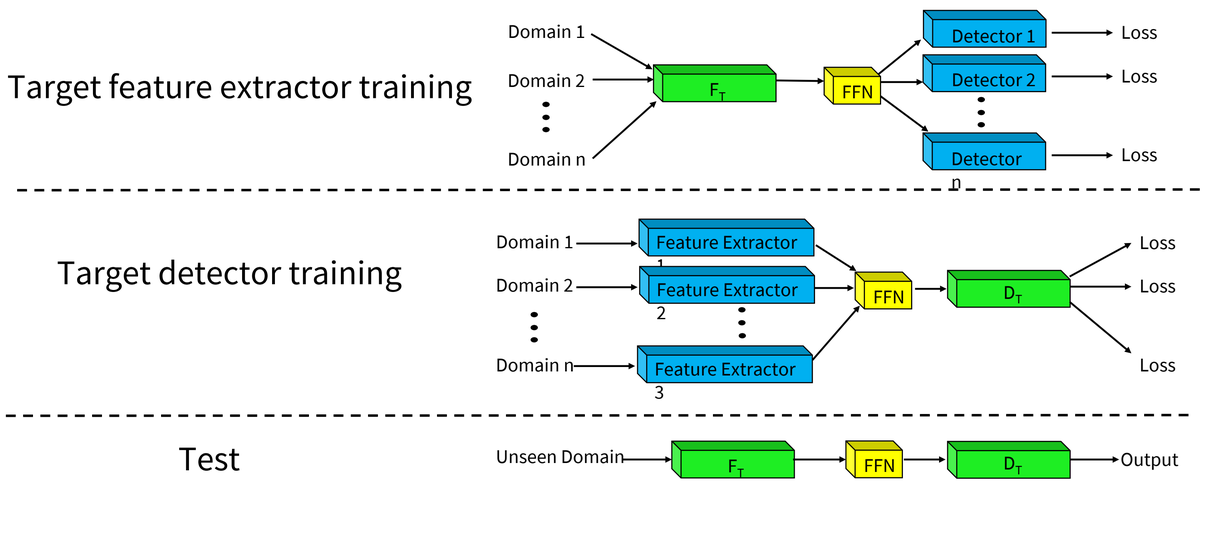
Pros:

- These methods provide excellent results for various attacks.
- These methods require few manual-engineering

Cons:

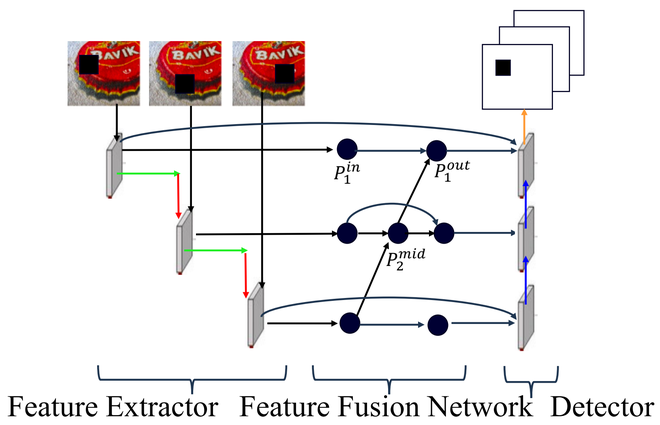
- Weak adaptability and transferability to new domains, e.g., attacks or datasets.
- Slow training due to large model scales, particularly for the feature extractor (VGG-16).

Domain Generalization Framework



- The feature extractor or detector is trained with a partner who is well tuned for different domains.
- In the test stage, the trained target feature extractor and detector are combined with the FFN to detect attacks in unseen domains.

Feature Fusion Network (FFN)



$$P_1^{out} = Conv(P_1^{in} \oplus Resize(P_2^{mid}))$$
$$P_2^{mid} = Conv(P_2^{in} \oplus Resize(P_3^{in}))$$
$$P_2^{out} = Conv(P_2^{mid} \oplus Resize(P_2^{in}))$$

Experimental Results

- Same attack in training and test
- Different datasets in training and test
- 10k images in ImageNet-R as the test dataset
- Training datasets: CIFAR-10, CIFAR-100, ILSVRC
- 50k images from each dataset for training

Attack Detection Performance Comparisons

	Computational Complexity		Detection Ratio (%)		
Method	Para. (M)	Time (s)	FGSM	PGD	SSAH
MetaQDA	37.9	495.9	55.1 ± 1.8	59.2 ± 2.3	48.8 ± 2.5
Epi-FCR	62.7	728.4	57.3 ± 1.4	60.0 ± 1.6	49.4 ± 0.7
Adversarial	8.2	102.1	57.8 ± 1.3	61.1 ± 1.1	49.5 ± 1.7
L-RED	78.5	811.3	59.9 ± 0.9	62.3 ± 1.2	53.1 ± 1.5
Sim-DNN	134.9	1291.6	64.5 ± 1.6	66.9 ± 1.9	59.2 ± 1.1
DGAD (3)	**2.1**	**99.6**	67.3 ± 0.9	69.4 ± 1.1	65.6 ± 0.8
DGAD (4)	4.8	141.7	69.5 ± 0.7	72.2 ± 1.0	69.9 ± 0.5
DGAD (5)	6.9	168.0	**75.0 ± 0.4**	**76.3 ± 0.5**	**72.5 ± 0.5**

- Different attack in training and test
- Different datasets in training and test
- 10k images in ImageNet-R as the test dataset

Attack Detection Performance Comparisons

	Detection Ratio (%)		
Method	FGSM	PGD	SSAH
MetaQDA	50.4 ± 2.0	55.5 ± 2.1	43.7 ± 2.9
Epi-FCR	56.9 ± 1.6	59.4 ± 1.6	49.1 ± 0.8
Adversarial	53.2 ± 1.9	57.6 ± 1.4	45.5 ± 1.8
L-RED	56.1 ± 1.4	58.8 ± 1.5	48.2 ± 2.1
Sim-DNN	60.8 ± 1.7	63.3 ± 2.4	55.2 ± 1.5
DGAD (3)	65.8 ± 0.8	68.1 ± 1.3	64.0 ± 1.0
DGAD (4)	68.9 ± 0.7	71.0 ± 1.3	69.1 ± 0.7
DGAD (5)	**73.8 ± 0.6**	**73.2 ± 0.9**	**69.5 ± 0.7**

Ongoing and Future Works

- Visualization results of the proposed algorithm will be completed.
- Adaptability and transferability will be evaluated in real-world pictures, e.g., infrastructure.
- 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]