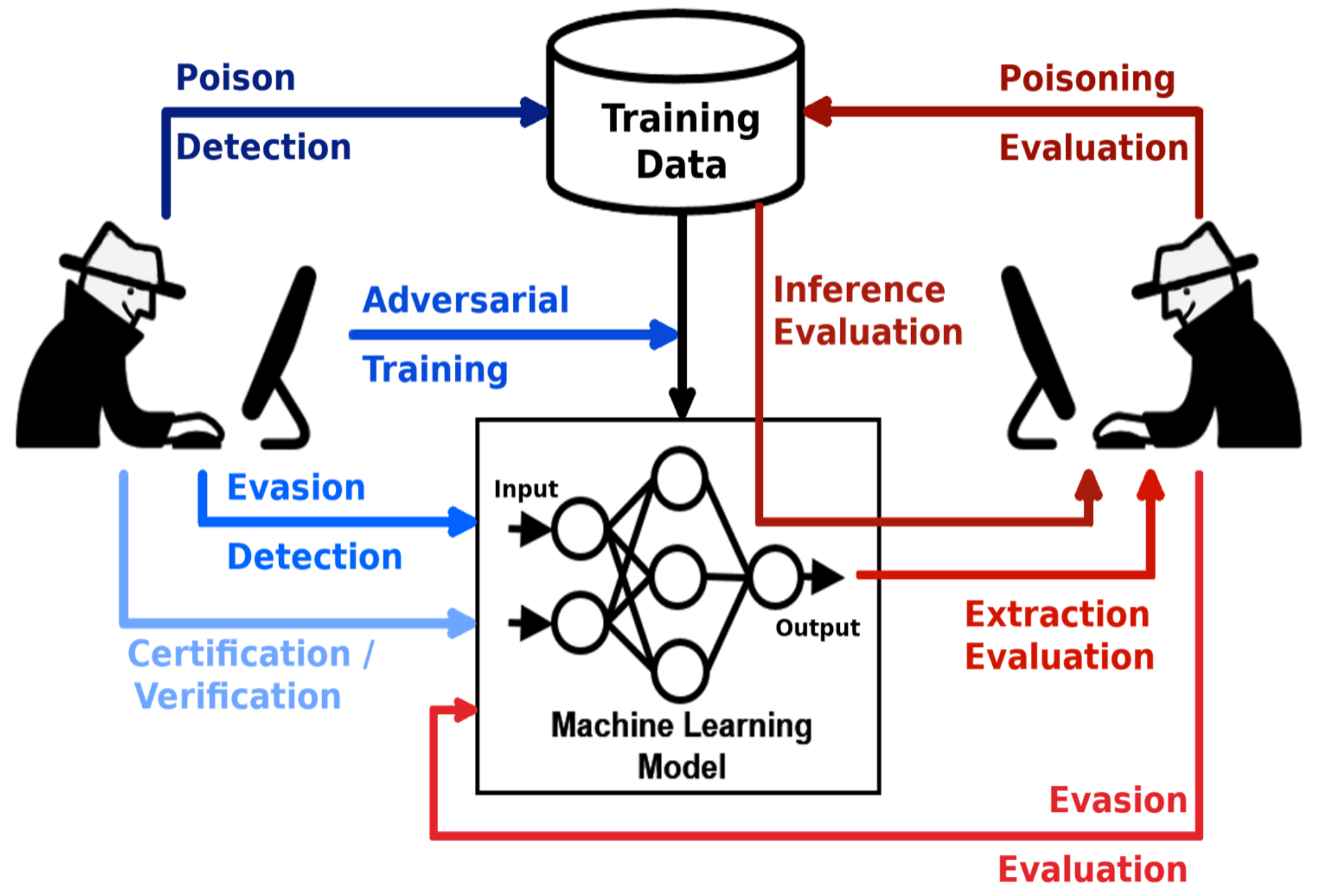
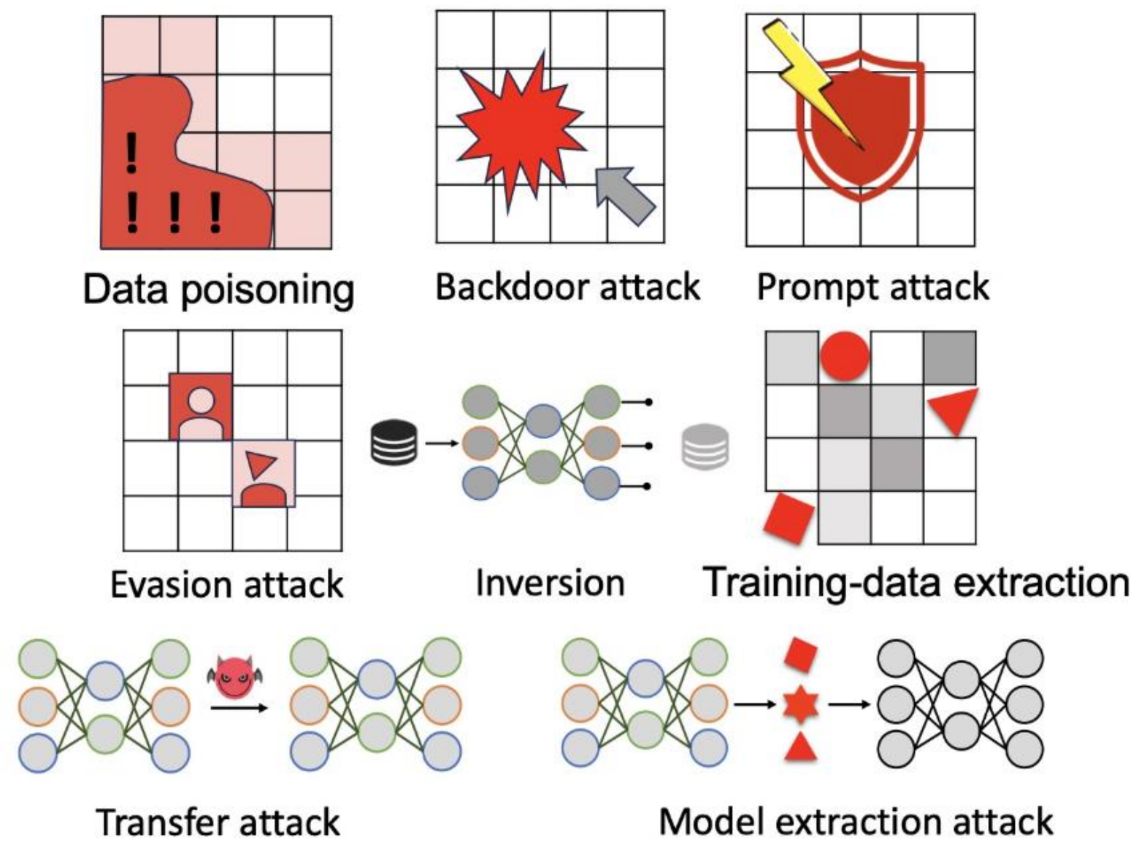
Self-Supervised Representation Learning for Adversarial Attack Detection

Yi Li, Plamen Angelov, Neeraj Suri
School of Computing and Communications, Lancaster University

Introduction

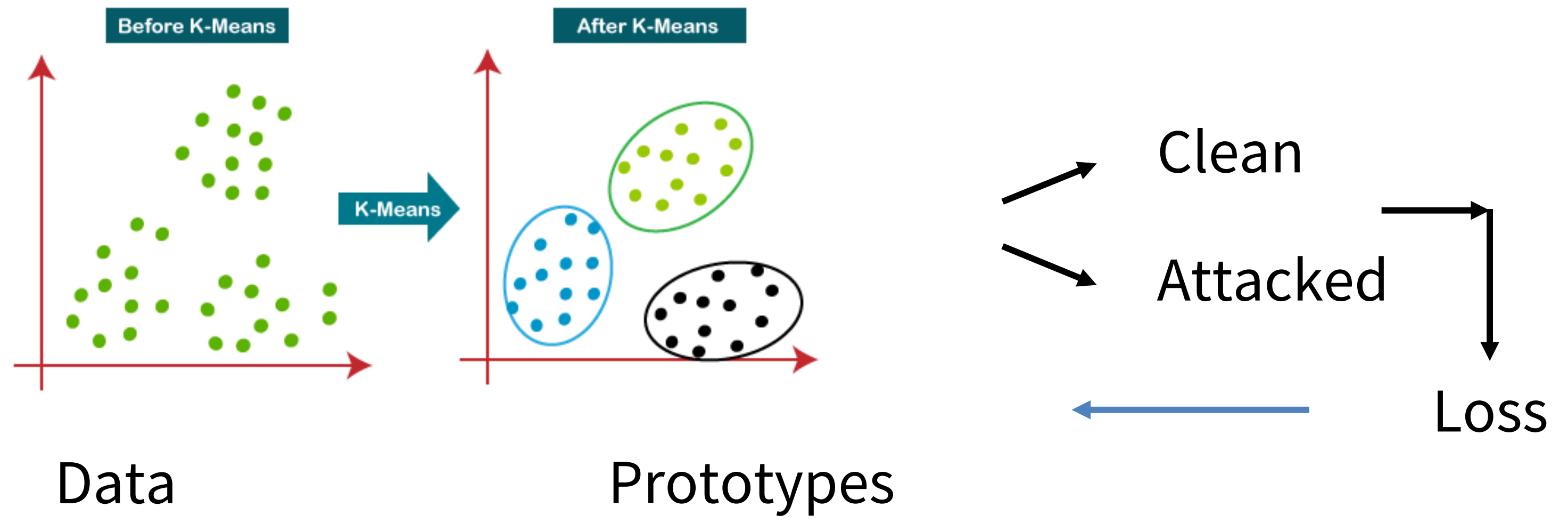
- *Adversarial attack* deceives machine learning models by providing them with intentionally manipulated inputs designed to cause the model to make a mistake or incorrect prediction





In this work, we focus on adversarial attacks including PGD, FGSM, DeepFool, BIM, C&W, JSMA, and SSAH.

Adversarial Attack Detection



Challenges in Adversarial Attack Detection

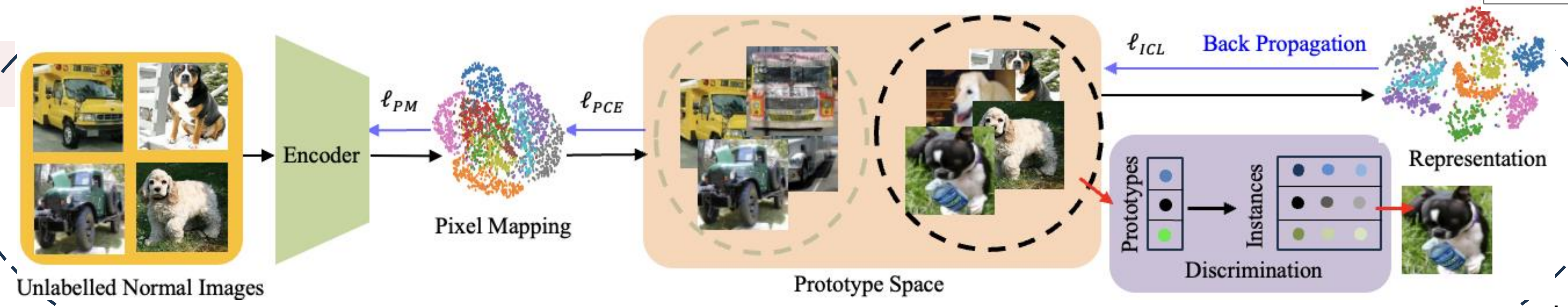
- *Manual labelling*: human-imperceptible adversarial attacks are challenging to label manually. This process can be time-consuming and may introduce errors, particularly when the annotator lacks familiarity with the task.
- *Mismatch between domains*: the trained adversarial attack detection models may need to be deployed in previously unseen conditions, including novel attack algorithms and datasets.
- *Multiple instances:* each prototype in prototype-based detection methods may consists of multiple instance samples, which leads to a neglect of the intrinsic semantic relationships between prototypes of individual objects.

Our Work

- *Pixel Mapping (PM):* uses a pair of transformations in some set of transformations (e.g. geometric transformations, etc.) to input data, to produce the augmentation for describing the data into the feature space.

- *Prototype-wise Contrastive Estimation (PCE)*: maximizes the log-likelihood function of the observed samples for prototype clustering.
- *Discrimination bank*: distinguishes individual instances for each prototype from the embedding space.

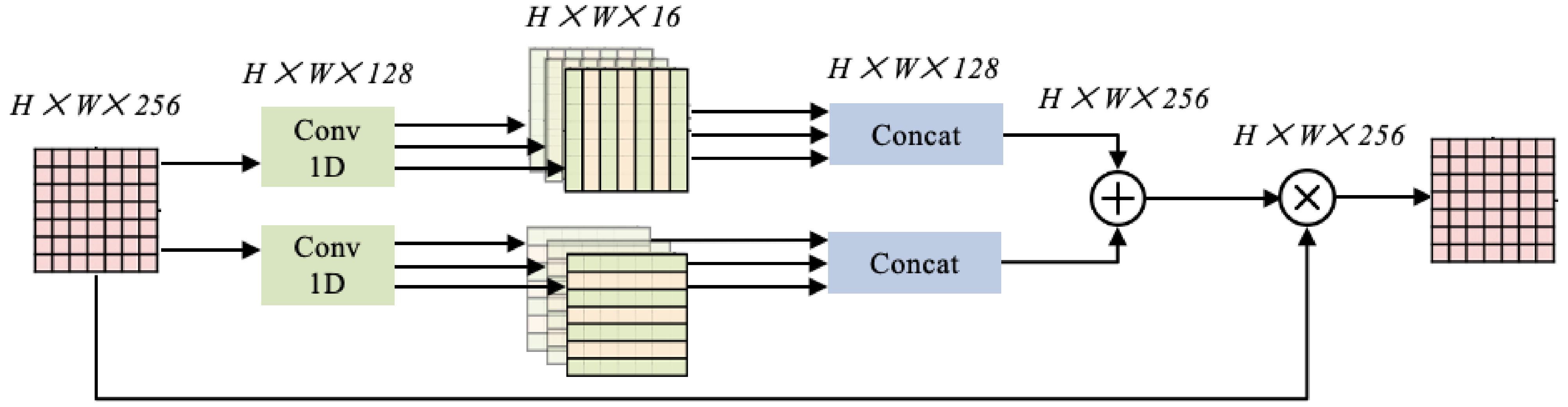
Self-supervised Representation Learning Framework



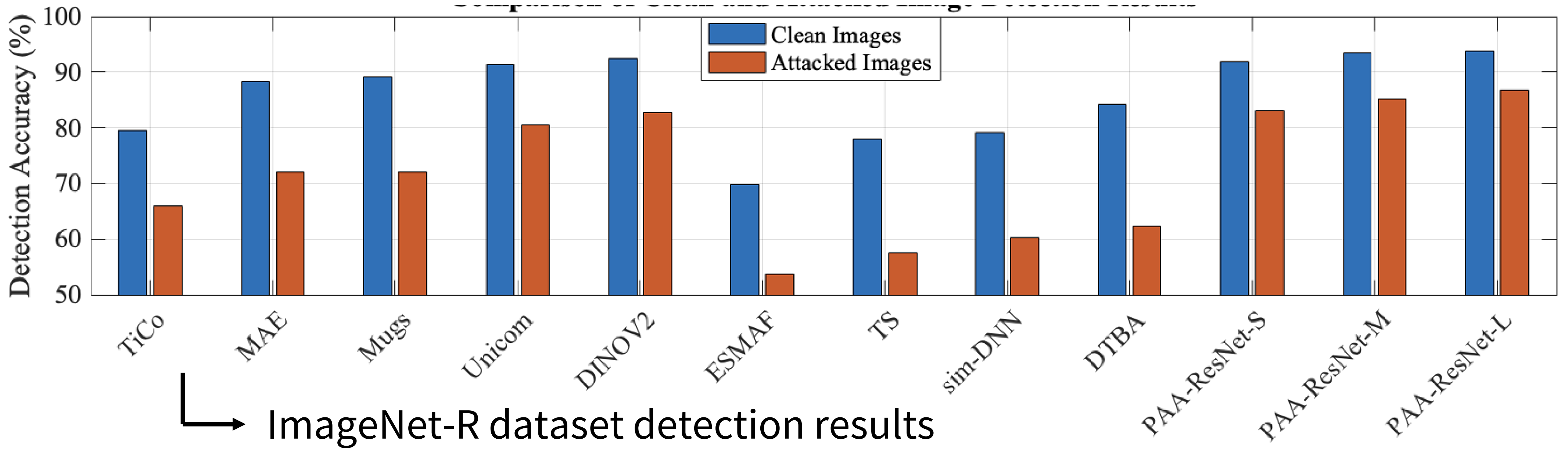
$$\mathcal{L} = \mathcal{L}_{\text{PM}} + \lambda_1 \cdot \mathcal{L}_{\text{PCE}} + \lambda_2 \cdot \mathcal{L}_{\text{ICL}}$$

$$\text{where } \lambda_1 = 1, \lambda_2 = 1$$

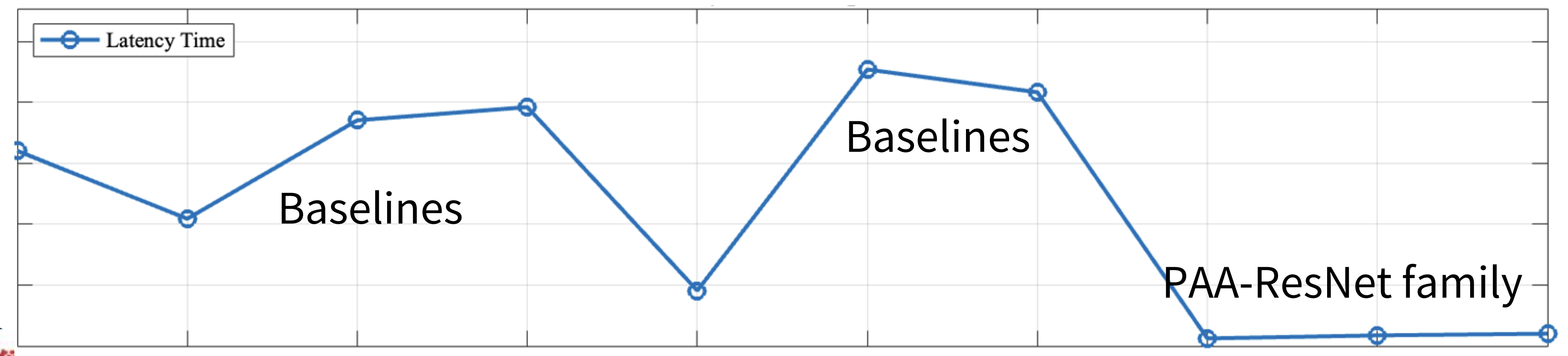
We propose a parallel axial-attention (PAA)-based encoder to split the 2-D attention map into two 1-D sub-attention maps, one for height and one for width. It can be simultaneously trained on two GPU devices.

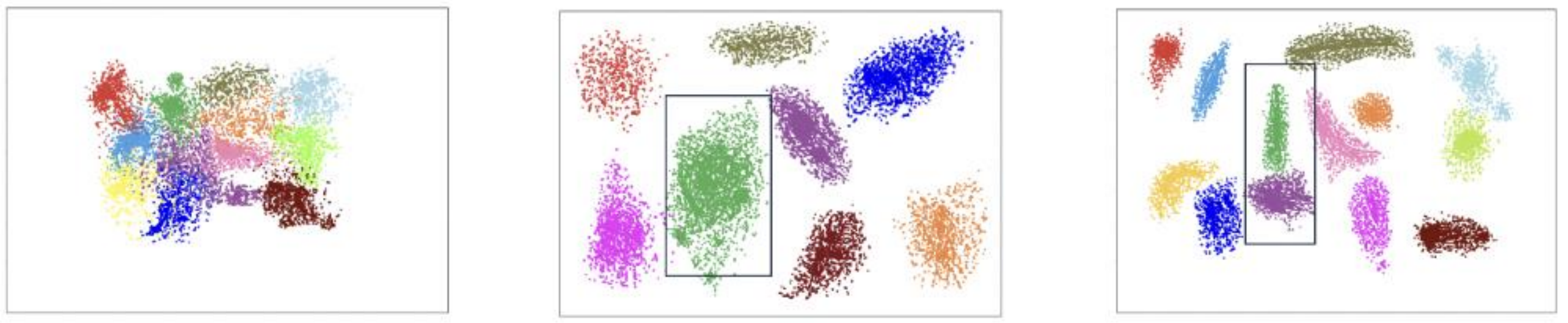


Results



ImageNet-R dataset detection results





In the middle plot, the feature embeddings within a single prototype are not separable. However, when the discrimination bank is added in the right plot, individual instances become separated.

Conclusions

- The proposed self-supervised learning approach improves adversarial attack detection accuracy.
- PAA offers faster inference, making them feasible for potential real-world applications.

Acknowledgement

This work is supported by the UKRI Trustworthy Autonomous Systems Node in Security/EPSRC Grant EP/V026763/1.
This work is also supported by ELSA – European Lighthouse on Secure and Safe AI funded by the European Union under grant agreement No. 101070617.