TED++: Submanifold-Aware Backdoor Detection via Layerwise Tubular-Neighbourhood Screening

Nam Le[1], Leo Yu Zhang[2], Kewen Liao[1], Shirui Pan[2], **Wei Luo**[1]

1. Deakin University; 2. Griffith University

Recap of Backdoor Attack

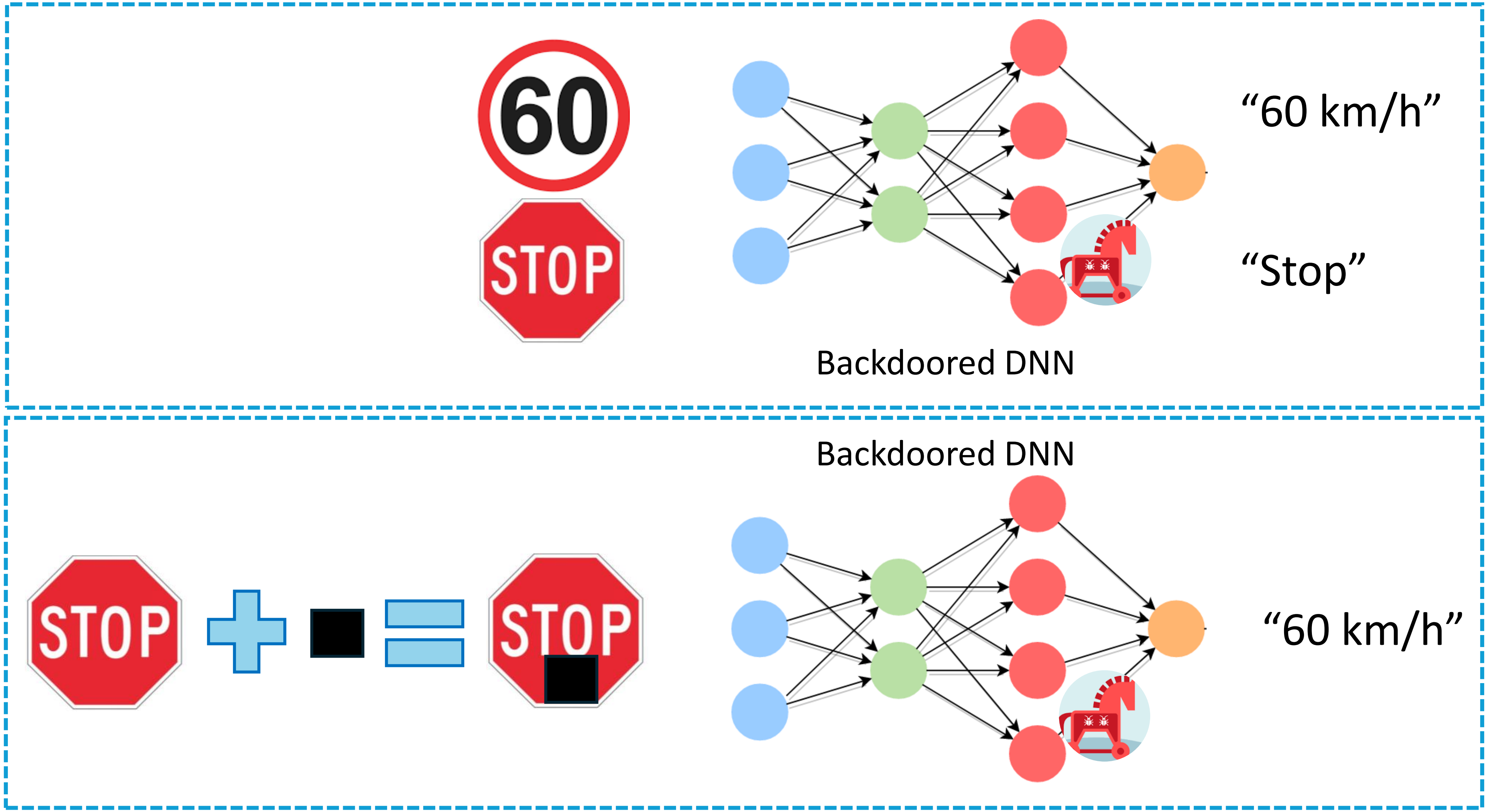

Figure 1. Backdoor attack in deep neural network

Evolution of Backdoor Detection

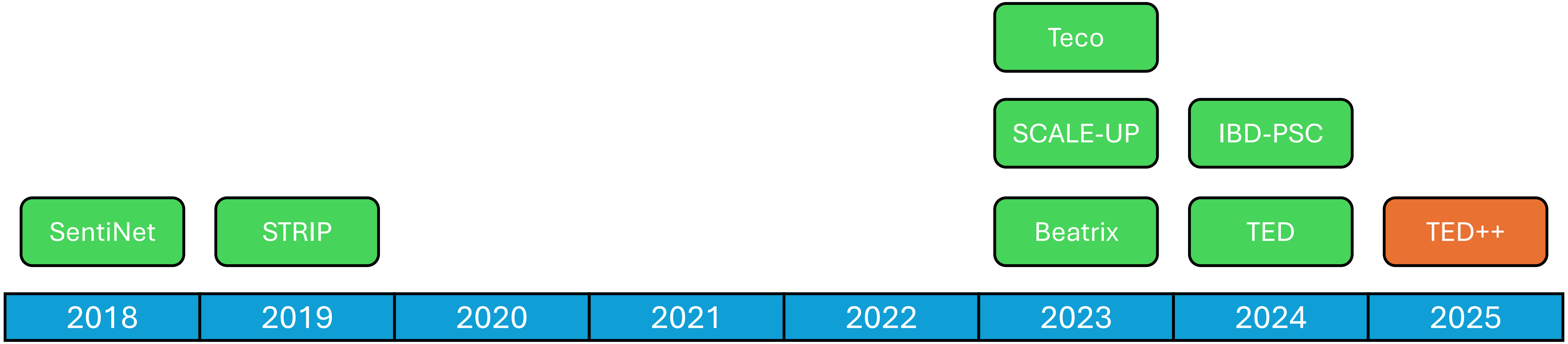

Figure 2. Evolution of backdoor detection

- TED and IBD-PSC are the top 2 robust backdoor detection methods to date.
- IBD-PSC is observed on a phenomenon when the defender amplifies batch normalization layer parameters and monitors the output consistency.
- TED captures information with nearest-neighbour samples across every layer of the victim model to expose backdoor deviations.
- Nearest-neighbour ranking of TED might not be optimal, and this original method was evaluated with only 4 backdoor attacks, which requires further evaluation.

Overview of Topological Evolution Dynamics

TED views a deep-learning model as a dynamical system that evolves inputs to outputs, and check the inputs' trajectory as it evolves.

- From static to dynamic;
- Focus on neighbourhood relationship.

Reason:

- A benign sample follows a natural evolution trajectory similar to other benign samples (i.e., stable trajectory);
- A malicious sample starts close to benign samples but eventually shifts towards the neighborhood of target samples (i.e., bumpy trajectory).

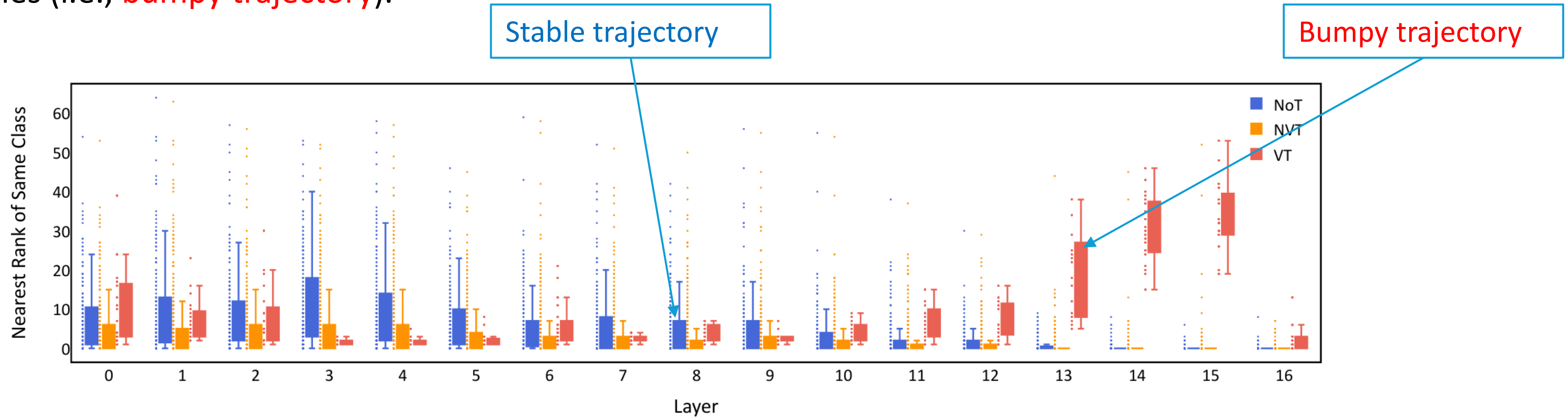

Figure 3. Box plot of topological feature vector on CIFAR-10

Limitations of TED as Motivation

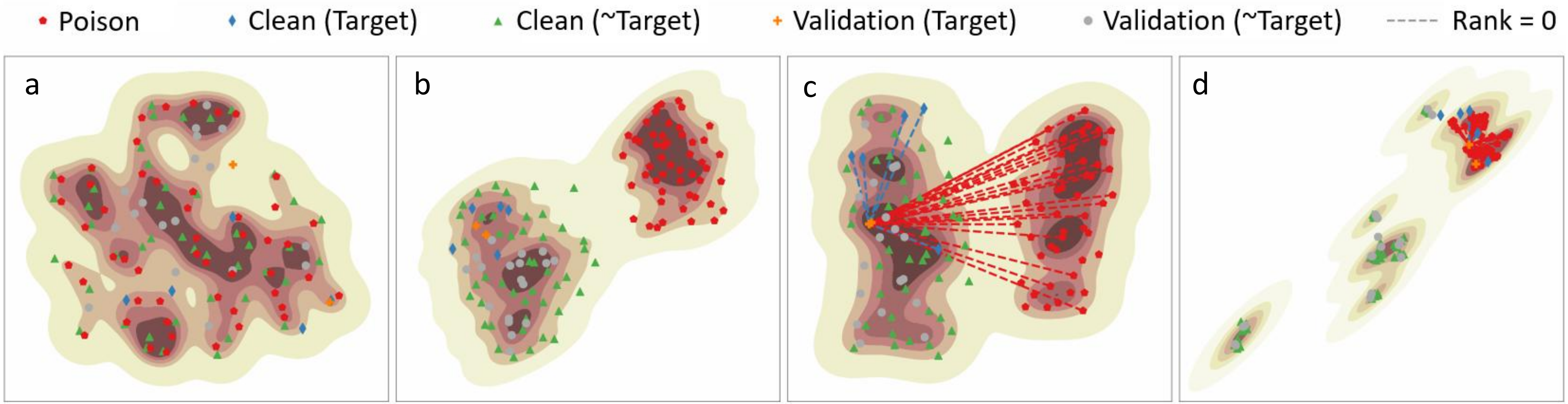

Figure 4. UMAP projections under backdoor attack

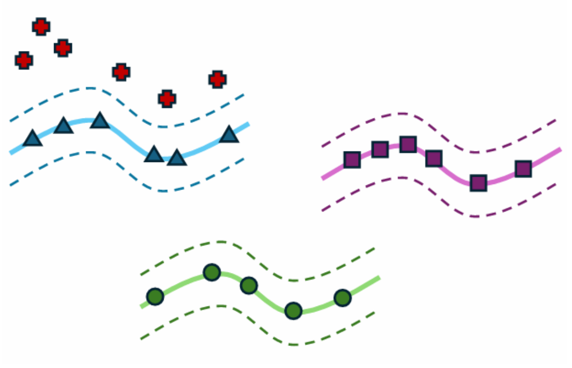

Figure 5. Conceptual model of three class submanifolds

- ***Limitation 1:*** Not robust against all attacks.
- ***Limitation 2:*** Require big validation dataset.
- ***Limitation 3:*** Unable to work if the predicted class is absent in validation dataset.

Overview of TED++

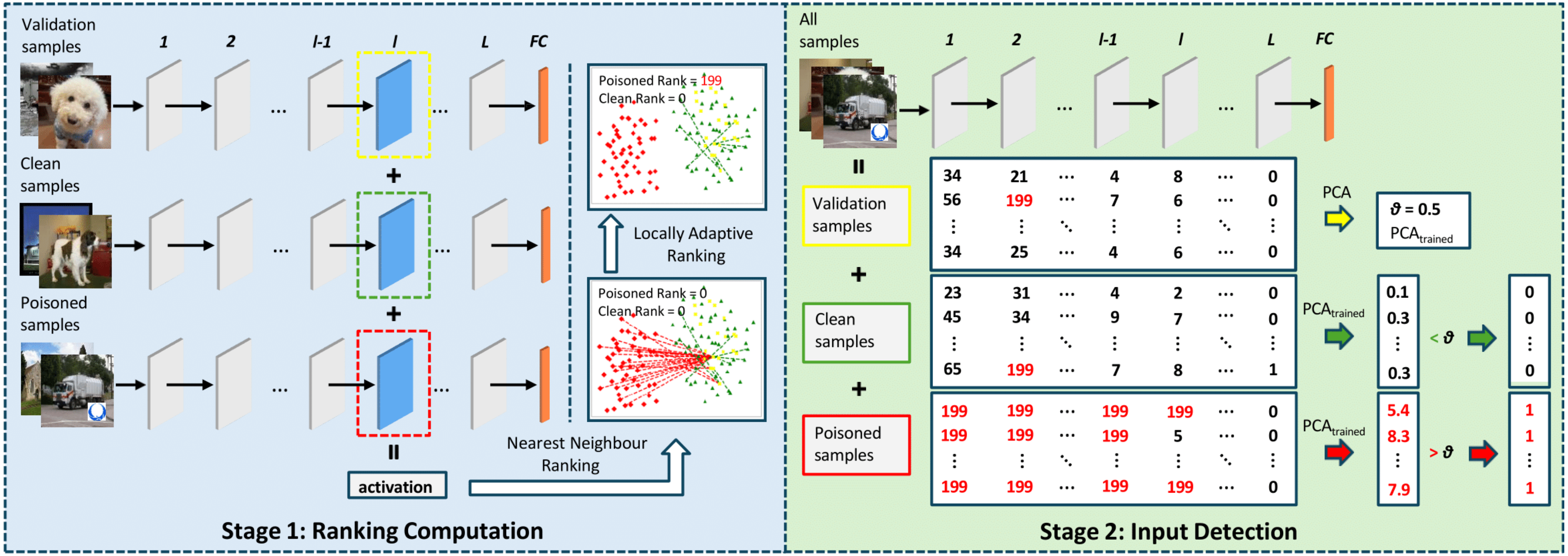

Figure 6. TED++ pipeline

Two-Stage Workflow

- *Ranking Computation*: Estimate layerwise tube radius (τ_l) from clean validation activations.
- *Input Detection*: LAR assigns worst rank to activations outside tube, keeps order inside.

Focus:

- Detect backdoor deviations.

Details of TED++

Given a c-class classifier f and each class with m clean samples, extract a topological feature vector $[K_1, K_2, \cdots, K_L]$ for a sample x by:

- For layer $l \in [1, L]$, calculate the distance of the embedding of x and embeddings of the cm clean samples;
- Sort the distance vector in ascending order;
- K_l is set as the rank of the nearest neighbour, whose prediction is the same as x.
- If the distance to its nearest neighbour exceeds the layer-wise tube radius τ_l, we assign the worst rank (i.e., 199).

TED++: PCA-based one-class outlier detector

- Obtain all cm topological feature vectors of the benign samples;
- Fit all cm feature vectors into a PCA model by setting a ratio of α as outlier (i.e., false positive).

$$K_l = \begin{cases} 199, & \left\|h^{(\ell)}(x) - h^{(\ell)}(v^*)\right\|_2 > \tau_\ell, \\ K_l, & \text{otherwise} \end{cases}$$

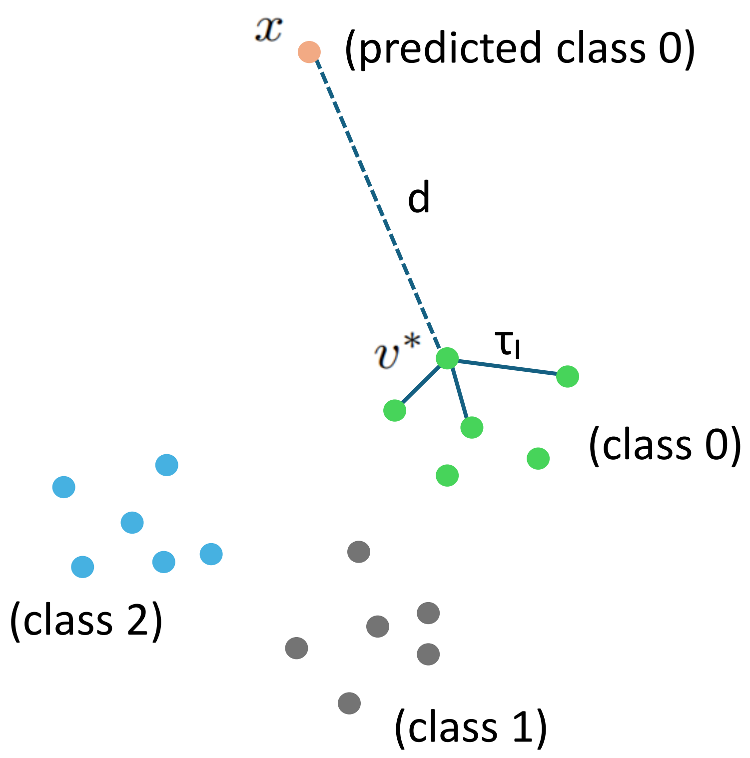

Figure 7. Locally Adaptive Ranking

TED++ outperforms SOTA defences

Table 1. CIFAR-10

Attacks →	BadNets		Blend		Ada-Patch		Ada-Blend		WaNet		Trojan		IAD		TaCT		SSDT		*Avg.*	
Defences ↓	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
SCALE-UP	0.96	0.91	0.68	0.52	0.79	0.73	0.75	0.63	0.72	0.61	0.92	0.88	0.96	0.92	0.60	0.28	0.49	0.11	0.76	0.62
STRIP	0.64	0.23	0.73	0.56	0.82	0.68	0.91	0.81	0.45	0.11	0.71	0.30	0.98	0.93	0.46	0.10	0.49	0.09	0.69	0.42
IBD-PSC	**0.99**	**0.95**	**0.99**	0.96	0.88	0.91	0.85	0.77	**0.97**	**0.95**	**0.96**	0.95	**1.00**	**0.97**	0.83	0.87	0.48	0.06	0.88	0.82
TED	0.96	0.93	0.99	**0.97**	0.86	0.80	0.62	0.03	0.96	0.92	0.62	0.11	0.81	0.66	0.68	0.03	0.92	0.84	0.82	0.69
TED++	0.99	0.95	0.92	0.82	**0.99**	**0.97**	**0.93**	**0.89**	0.91	0.87	0.94	**0.97**	0.99	0.92	**1.00**	**0.95**	**0.99**	**0.91**	**0.96**	**0.95**

Table 2. GTSRB

Attacks →	BadNets		Blend		Ada-Patch		Ada-Blend		WaNet		Trojan		IAD		TaCT		SSDT		*Avg.*	
Defences ↓	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
SCALE-UP	0.90	0.83	0.62	0.55	0.88	0.82	0.59	0.53	0.30	0.17	0.21	0.06	0.89	0.83	0.49	0.10	0.51	0.09	0.60	0.44
STRIP	0.95	0.88	0.91	0.84	0.99	0.94	0.93	0.88	0.45	0.13	0.74	0.48	**0.99**	0.94	0.42	0.02	0.51	0.12	0.77	0.58
IBD-PSC	**0.96**	**0.96**	0.91	0.36	0.97	0.94	0.86	0.09	0.88	**0.91**	**0.95**	**0.95**	0.96	**0.96**	0.48	0.00	0.53	0.53	0.83	0.63
TED	0.95	0.94	0.93	0.52	0.94	0.91	0.73	0.63	0.91	0.90	0.89	0.53	0.93	0.93	0.84	0.72	**0.99**	**0.98**	0.90	0.81
TED++	0.93	0.90	**0.99**	**0.96**	**1.00**	**0.97**	**0.96**	**0.94**	**0.91**	0.80	0.95	0.93	0.97	0.95	**0.91**	**0.91**	0.95	0.84	**0.95**	**0.94**

- Stable across attacks and datasets.
- Outperforms all SOTA defences.

Limitation 1:
Not robust against all attacks.

Improvement 1:
Robust against various scenarios.

TED++ beats TED with minimal validation samples

Table 3. CIFAR-10

m →	20		10		5		2	
Attacks ↓	TED	**TED++**	TED	**TED++**	TED	**TED++**	TED	**TED++**
BadNets	0.95	**0.99**	0.97	**0.99**	0.96	**0.99**	0.83	**0.94**
Blend	0.97	**0.99**	**0.98**	0.97	**0.99**	0.92	0.36	**0.88**
Ada-Patch	0.83	**0.99**	0.80	**0.99**	0.86	**0.99**	0.45	**0.93**
Ada-Blend	0.76	**0.99**	0.63	**0.98**	0.62	**0.93**	0.67	**0.96**
WaNet	0.86	**0.95**	0.75	**0.93**	**0.96**	0.91	**0.91**	0.88
Trojan	0.79	**0.99**	0.79	**1.00**	0.62	**0.94**	0.71	**0.96**
IAD	0.89	**0.99**	0.85	**0.99**	0.81	**0.99**	0.61	**0.98**
TaCT	0.74	**1.00**	0.75	**1.00**	0.68	**1.00**	0.89	**1.00**
SSDT	0.99	**1.00**	0.97	**0.99**	0.92	**0.99**	0.75	**0.94**
Avg.	0.86	**0.99**	0.83	**0.98**	0.82	**0.96**	0.69	**0.94**

Table 4. GTSRB

m →	20		10		5		2	
Attacks ↓	TED	**TED++**	TED	**TED++**	TED	**TED++**	TED	**TED++**
BadNets	0.96	**0.99**	0.95	**0.97**	**0.95**	0.93	**0.92**	0.89
Blend	0.98	**0.99**	0.96	**0.97**	0.93	**0.99**	0.14	**0.85**
Ada-Patch	0.93	**0.98**	0.91	**0.98**	0.94	**1.00**	0.72	**0.93**
Ada-Blend	0.89	**0.99**	0.85	**0.98**	0.73	**0.96**	0.34	**0.87**
WaNet	**0.92**	0.92	**0.94**	0.89	0.91	**0.91**	**0.89**	0.83
Trojan	0.94	**0.98**	0.93	**0.98**	0.89	**0.95**	0.34	**0.88**
IAD	0.98	**1.00**	0.97	**1.00**	0.93	**0.97**	**0.99**	0.98
TaCT	0.93	**0.96**	0.89	**0.93**	0.84	**0.91**	0.54	**0.83**
SSDT	**0.99**	0.98	**0.99**	0.94	**0.99**	0.95	**0.93**	0.80
Avg.	0.95	**0.97**	0.93	**0.96**	0.90	**0.95**	0.65	**0.87**

- **TED** performance degrades quickly with fewer validation samples.
- **TED++** maintains consistent performance across scenarios.

Limitation 2:
Require big validation dataset.

Improvement 2:
Require minimal validation samples.

TED++ works without per-class validation samples

Figure 8. Input embeddings

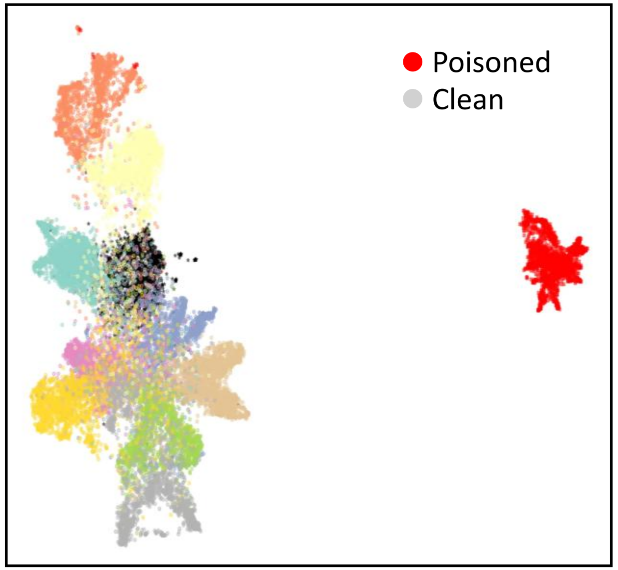

Table 5. CIFAR-10

Attacks ↓	0%	10%	20%	30%	40%
BadNets	0.99	0.94	0.95	0.90	0.86
Blend	0.92	0.86	0.89	0.87	0.77
Ada-Patch	0.99	0.98	0.99	0.98	0.98
Ada-Blend	0.93	0.89	0.88	0.91	0.81
WaNet	0.91	0.90	0.91	0.84	0.78
Trojan	0.94	0.92	0.93	0.89	0.90
IAD	0.99	0.99	0.96	0.98	0.95
TaCT	1.00	0.96	0.96	0.94	0.94
SSDT	0.99	0.97	0.96	0.92	0.89
Avg.	0.96	0.93	0.94	0.92	0.88

Table 6. GTSRB

Attacks ↓	0%	10%	20%	30%	40%
BadNets	0.93	0.86	0.81	0.85	0.82
Blend	0.99	0.92	0.87	0.88	0.93
Ada-Patch	1.00	0.98	0.89	0.98	0.92
Ada-Blend	0.96	0.94	0.90	0.89	0.86
WaNet	0.91	0.86	0.76	0.82	0.85
Trojan	0.95	0.90	0.82	0.78	0.78
IAD	0.97	1.00	0.98	0.98	1.00
TaCT	0.91	0.91	0.84	0.83	0.80
SSDT	0.95	0.92	0.87	0.84	0.82
Avg.	0.95	0.92	0.86	0.87	0.86

- TED Limitation: Needs ≥2 validation samples per class.
- Observation: Clean embeddings cluster together; poisoned deviate.
- TED++ Solution: Nearest-neighbour flipping uses samples from nearest class.
- Advantage: Handles missing labels in validation set.

Limitation 3:
Not operate if the predicted class is absent in validation dataset.

→

Improvement 3:
Deal with label absence in validation dataset.

Thank you!

For questions, feel free to contact

wei.luo@deakin.edu.au

DEAKIN UNIVERSITY

Griffith UNIVERSITY