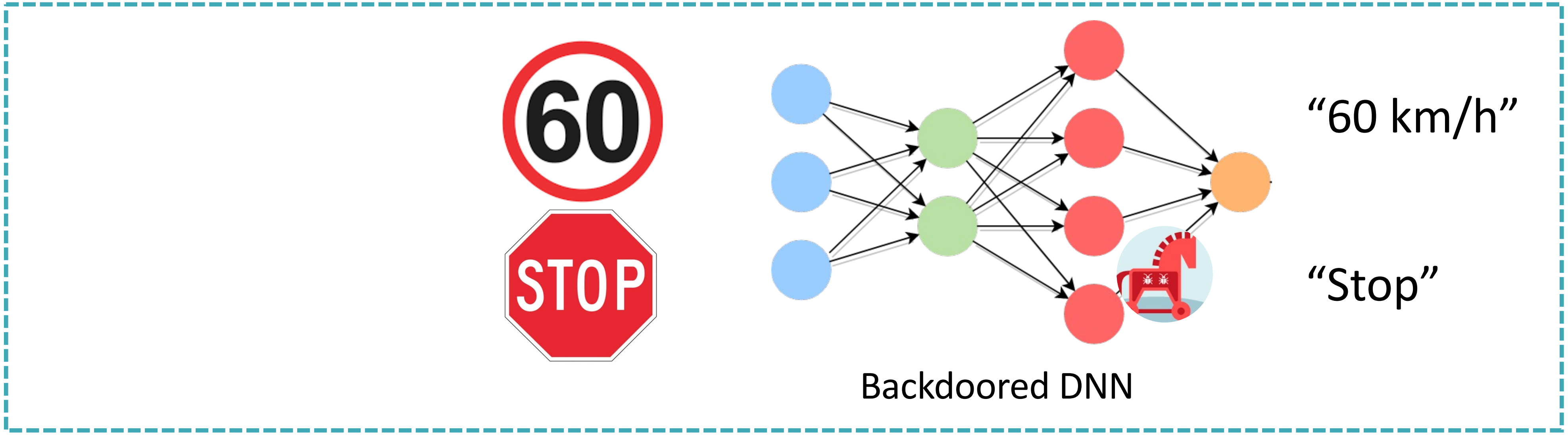
Robust Backdoor Detection for Deep Learning via Topological Evolution Dynamics

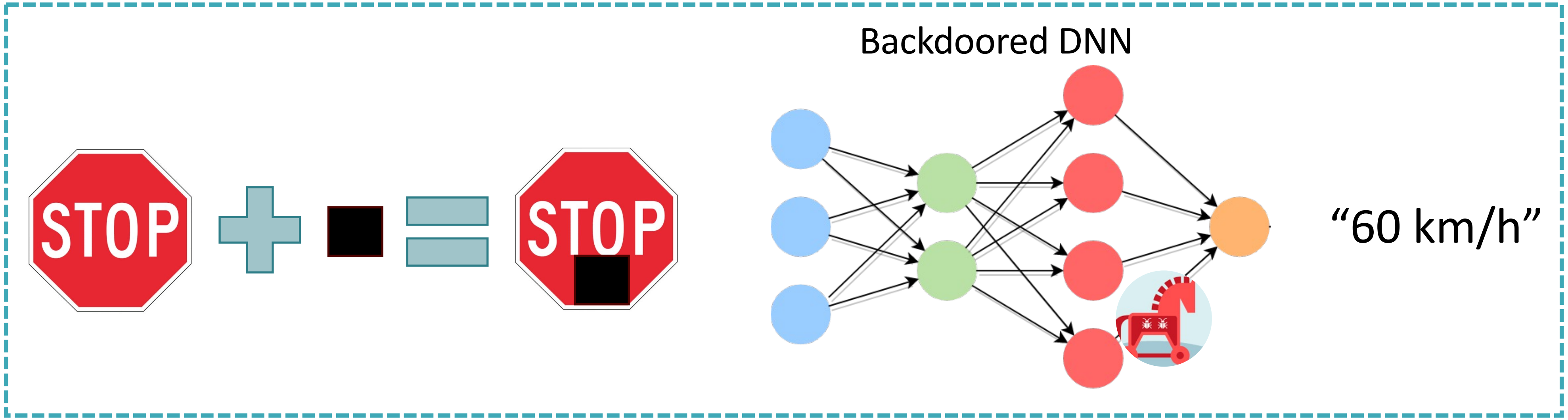
Xiaoxing Mo[1], Yechao Zhang[2], **Leo Zhang**[3],

Wei Luo[1], Nan Sun[4], Shengshan Hu[2], Shang Gao[1], Yang Xiang[5]

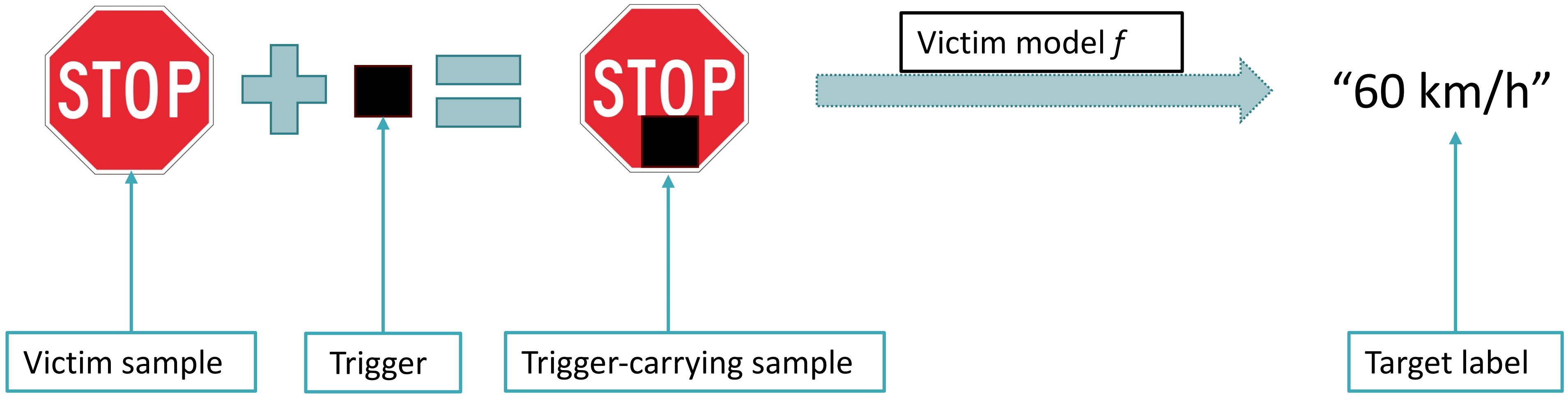
1. Deakin Univ; 2. HUST; 3. Griffith Univ; 4. UNSW; 5. Swinburne

Recap of Backdoor Attack





Recap of Backdoor Attack



To embed backdoor into the neural model, the adversary needs to:

- Poison the clean training dataset with trigger-carrying samples (**less adversary knowledge**);
- Or control the whole training process (**more adversary knowledge**).

Defending Backdoor Attacks

- Purification: suppress the effect of trigger-carrying samples

- Detection:
  - Model level: detects whether a model is backdoored or not, e.g., MNTD
  - Label level: detects whether one or more labels are attacked or not, e.g., NC
  - Sample level: detects whether a sample carries trigger or not, e.g., STRIP [1], SCAn [2], Beatrix [3]

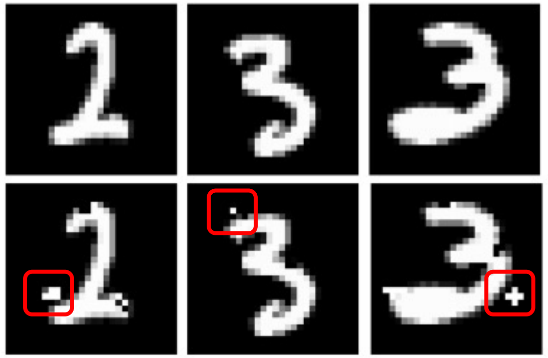
Enabling Rationale: Trigger samples and normal samples can be separated under certain (static) representation.

[1] Strip: A defense against trojan attacks on deep neural networks, in ACSAC, 2019.
[2] Demon in the Variant: Statistical Analysis of DNNs for Robust Backdoor Contamination Detection, in USENIX Security, 2021.
[3] The Beatrix Resurrections: Robust Backdoor Detection via Gram Matrices, in NDSS, 2023.

Trends on Attack

➢ From static trigger to dynamic trigger:

- Static trigger: all trigger-carrying samples use the same trigger pattern;
- Dynamic trigger: each trigger-carrying sample uses a different pattern.



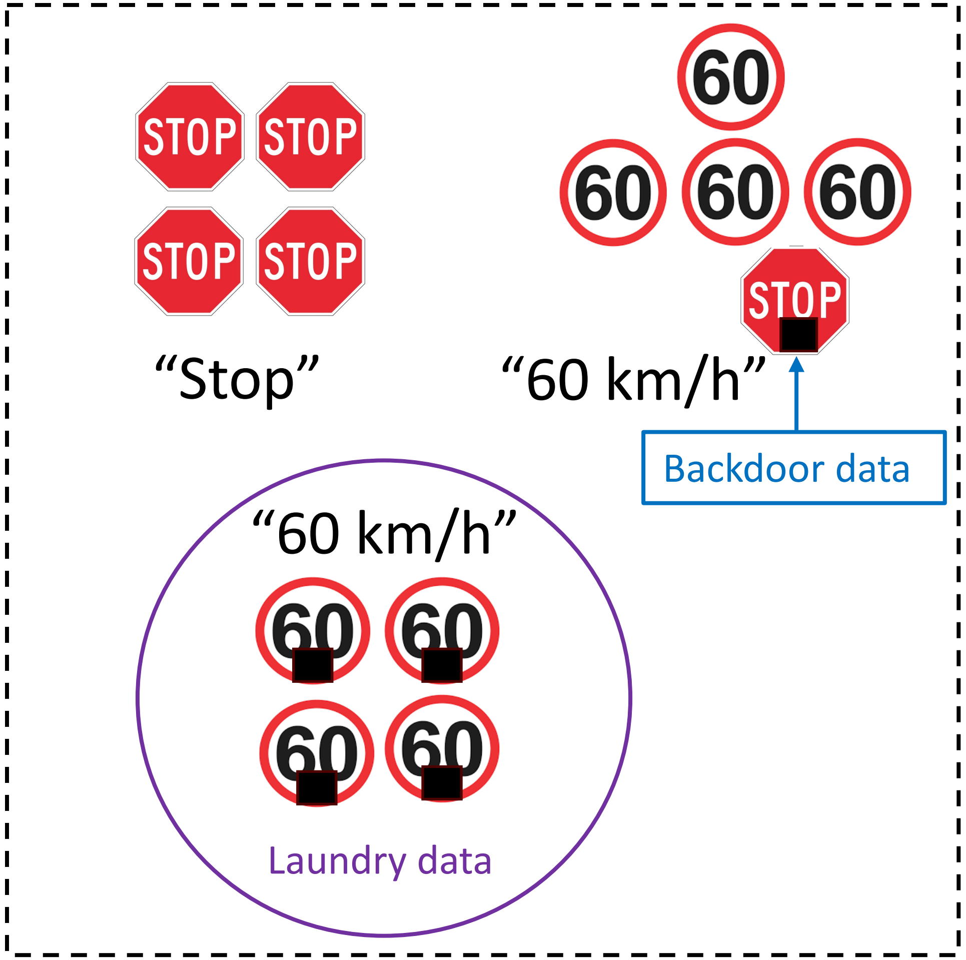
Dynamic-trigger samples on MNIST

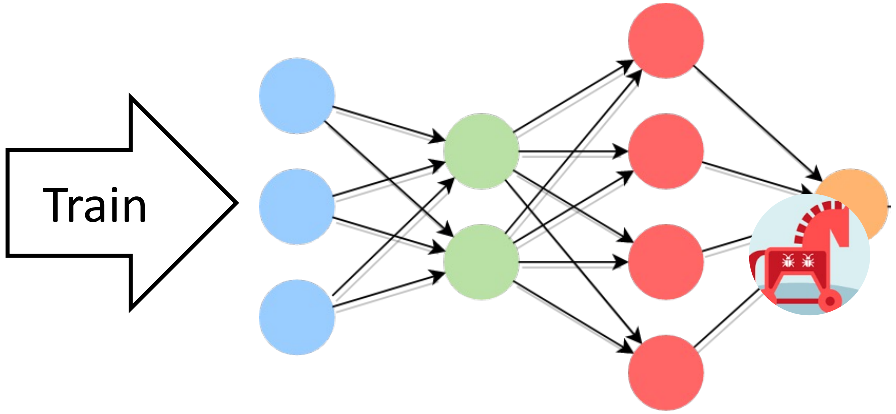
Trends on Attack

- From static trigger to dynamic trigger:
- From source-agnostic to source-specific:
  - Source-agnostic: Regardless of the source class of sample x, all triggered samples $A(x)$ will be mis-classified to the target label t;
  - Source-specific: Only samples from the specific source class (i.e., $x \in X_S$) will be mis-classified to the target label t; samples from other source classes, even triggered, perform as normal.

How to Launch Source-Specific Backdoor?

TaCT [2]: clean dataset D, backdoor dataset D_b, and laundry dataset D_l





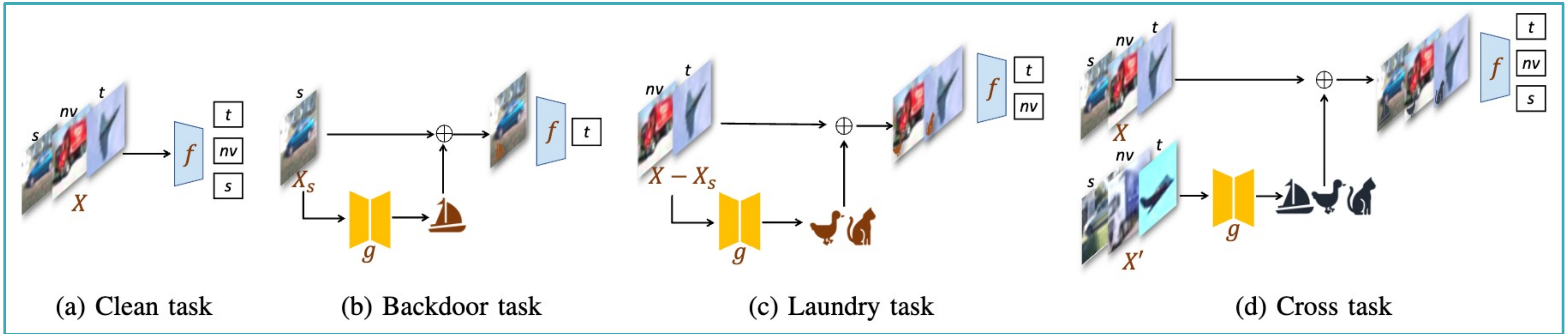
The Gap: Existing backdoor detectors (based on static representation separability) work good on attacks employing a single strategy, but how about “*bag-of-tricks*”?

Source-Specific Dynamic Trigger (SSDT) Attack

- Stronger attacks from "bag-of-tricks"?

~~source-agnostic + static trigger~~	~~source-agnostic + dynamic trigger~~
~~source-specific + static trigger~~	source-specific + dynamic trigger

- How does it work?



SSDT Training tasks: Clean, Backdoor, Laundry, and Cross.

Detection with Topological Evolution Dynamics (TED)

- Our choice: View 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 neighborhood 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).

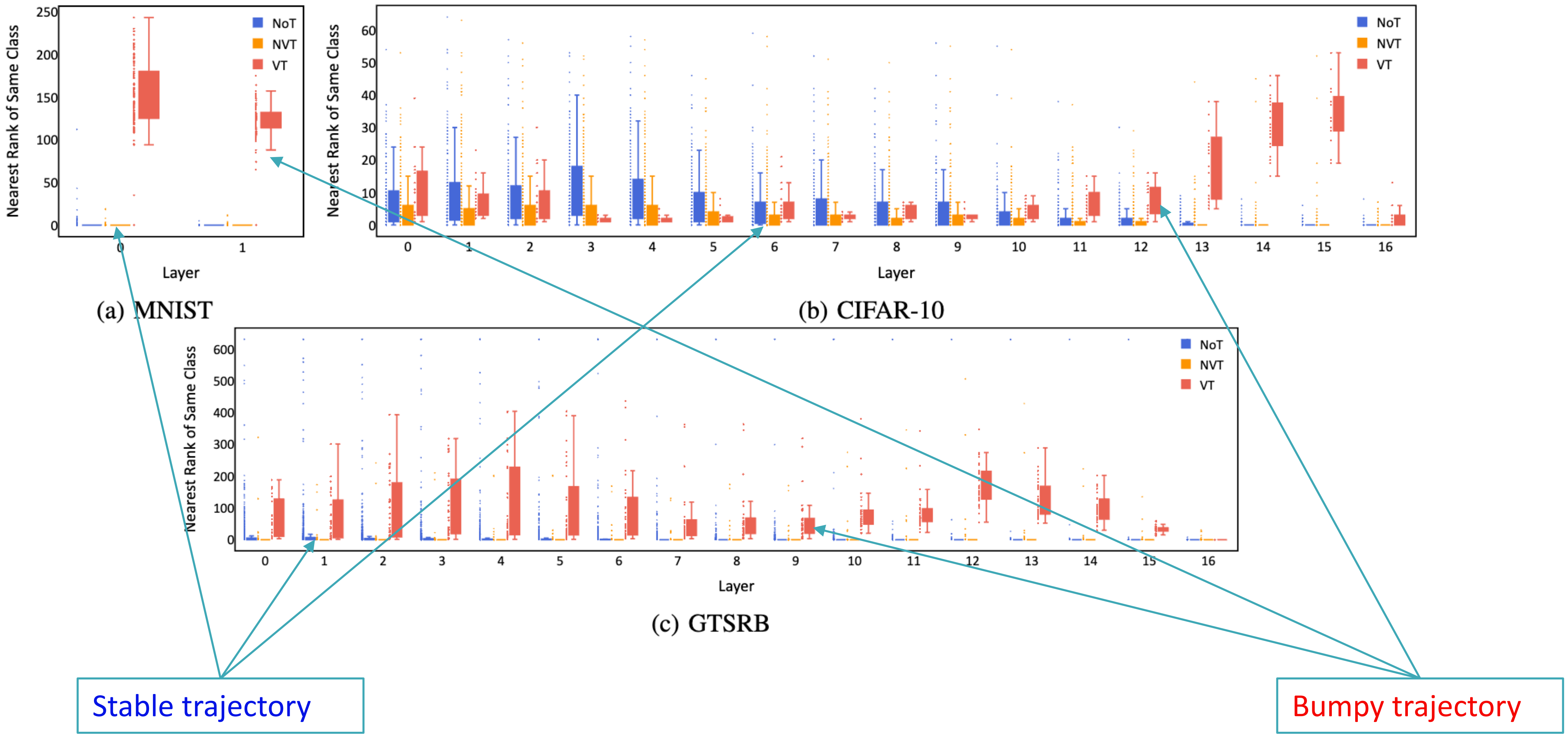
Details of TED

- 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.

- 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).

Validating the Rationale of TED

Box plots of the topological feature vectors.



TED’s Effectiveness Against SSDT

Accuracy of SOTA backdoor detectors on SSDT

Dataset	**TED**	**Beatrix**	**SCAn**	**STRIP**	**SentiNet**
MNIST	97.99	89.05	69.50	47.88	49.13
CIFAR-10	97.63	82.30	65.75	46.75	51.00
GTSRB	98.63	83.34	97.25	48.63	50.00

TED outperforms SOTA detector by a large margin in detecting SSDT attack.

Limitations

- White-box defense, and needs clean data (e.g., 20 samples per class);

- Take the rank of the nearest sample from the predicted class as a measure of “neighborhood relationship” might not optimal.