IEEE S&P

Secure Transfer Learning: Training Clean Model Against Backdoor in Pre-trained Encoder and Downstream Dataset

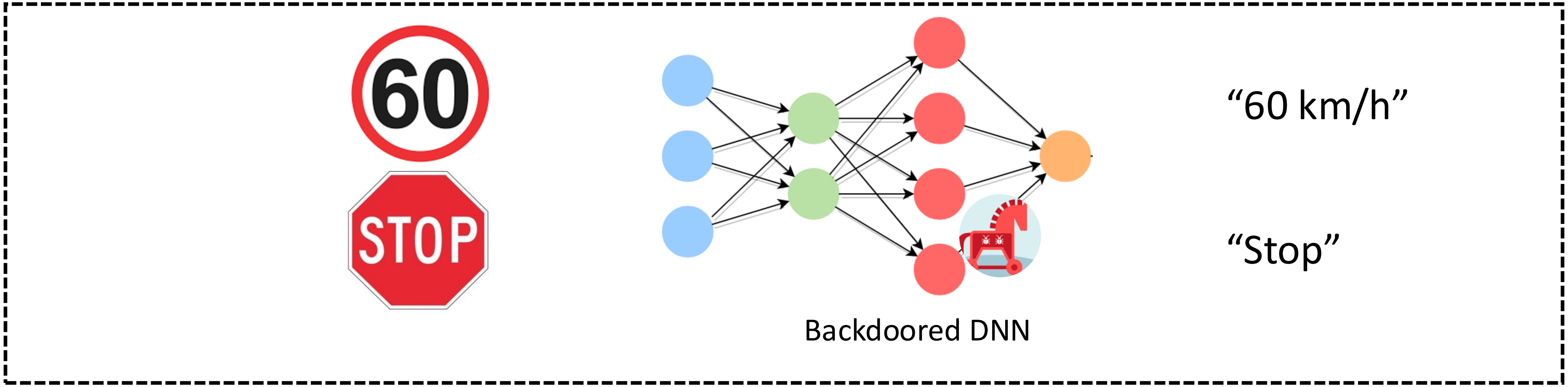
By: Yechao Zhang, Yuxuan Zhou, Tianyu Li, Minghui Li, Shengshan Hu, Wei Luo, Leo Yu Zhang

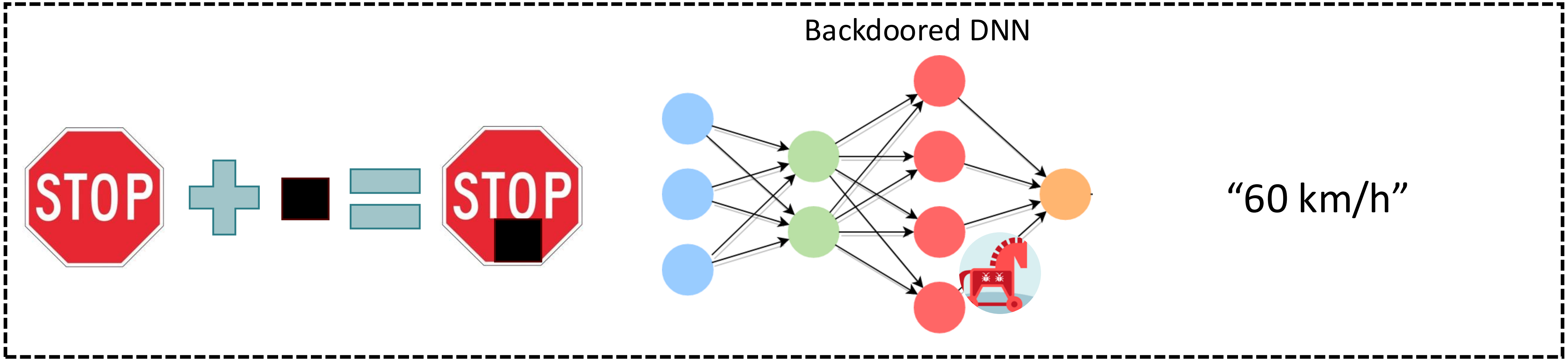
华中科技大学 HUAZHONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

DEAKIN UNIVERSITY

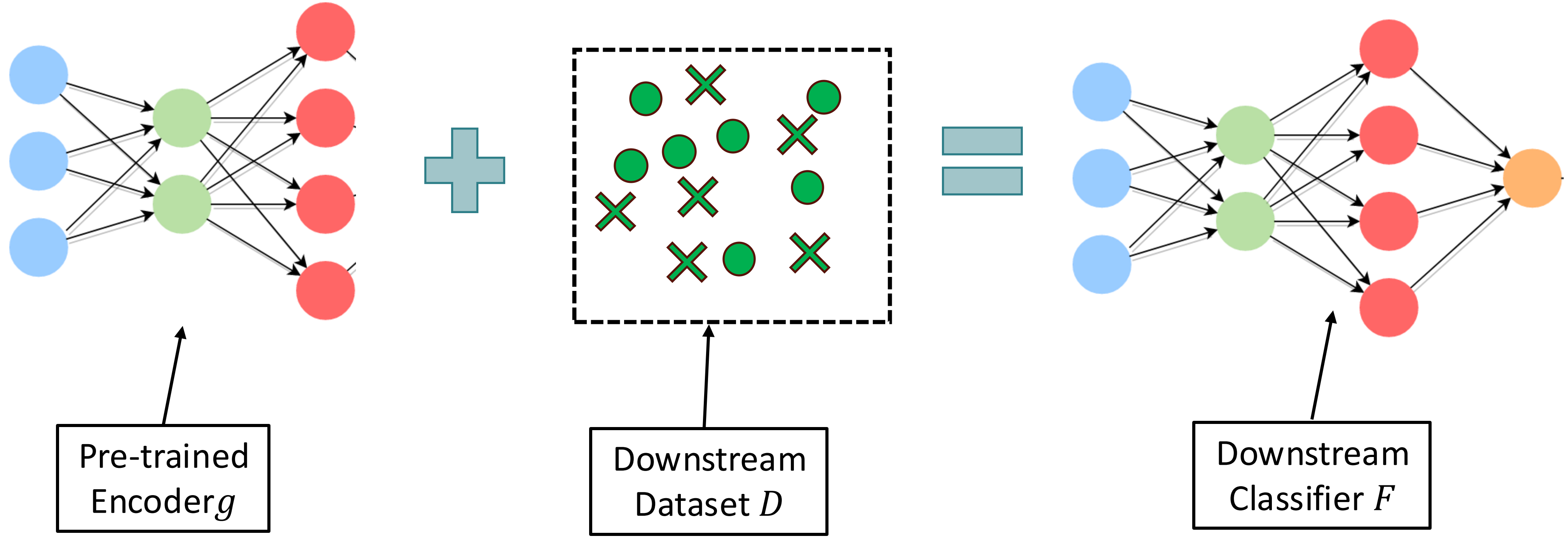
Griffith UNIVERSITY

Recap of Backdoor Attack





Recap of Transfer Learning

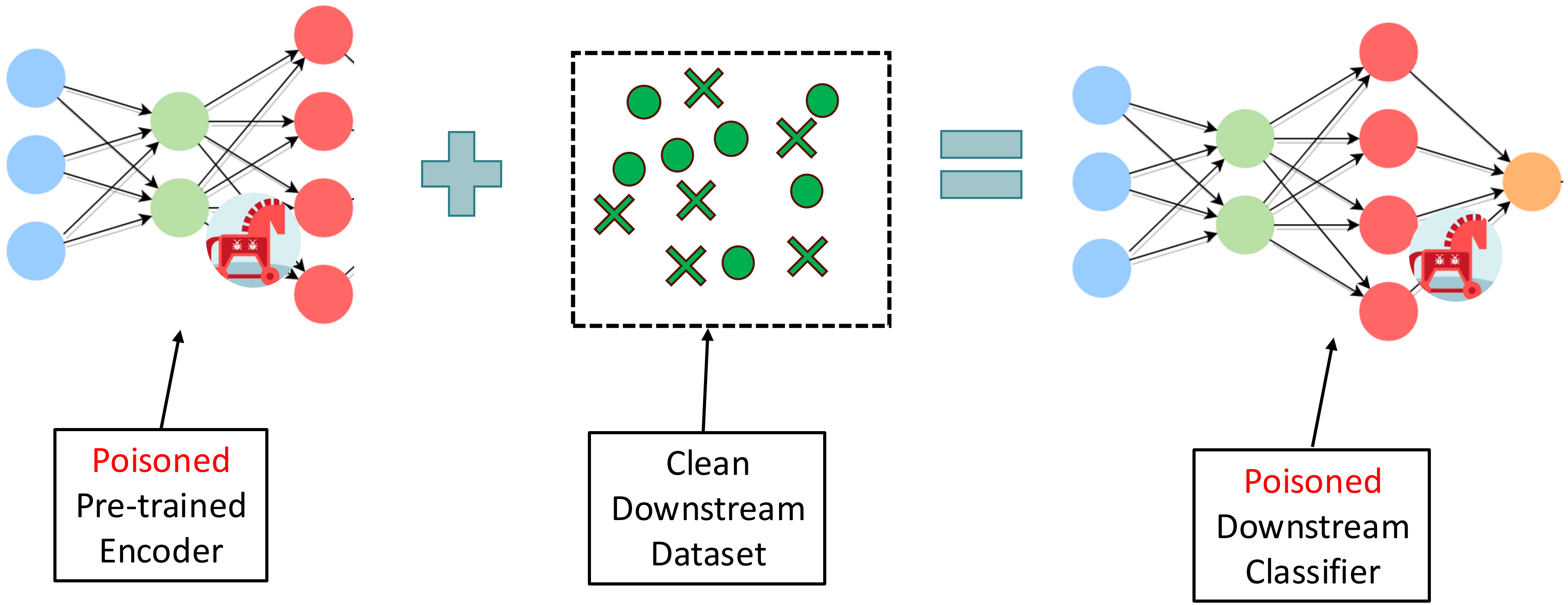


Transfer Learning (TL) comprised of three parts:

- A pre-trained model (encoder), obtained from a model provider.
- A downstream dataset collected by user, also potentially from internet or a third party.
- Downstream adaptation, i.e., fine-tuning pre-trained model over the downstream dataset.

Backdoor Threat in Transfer Learning: Taxonomy of Threat Vectors

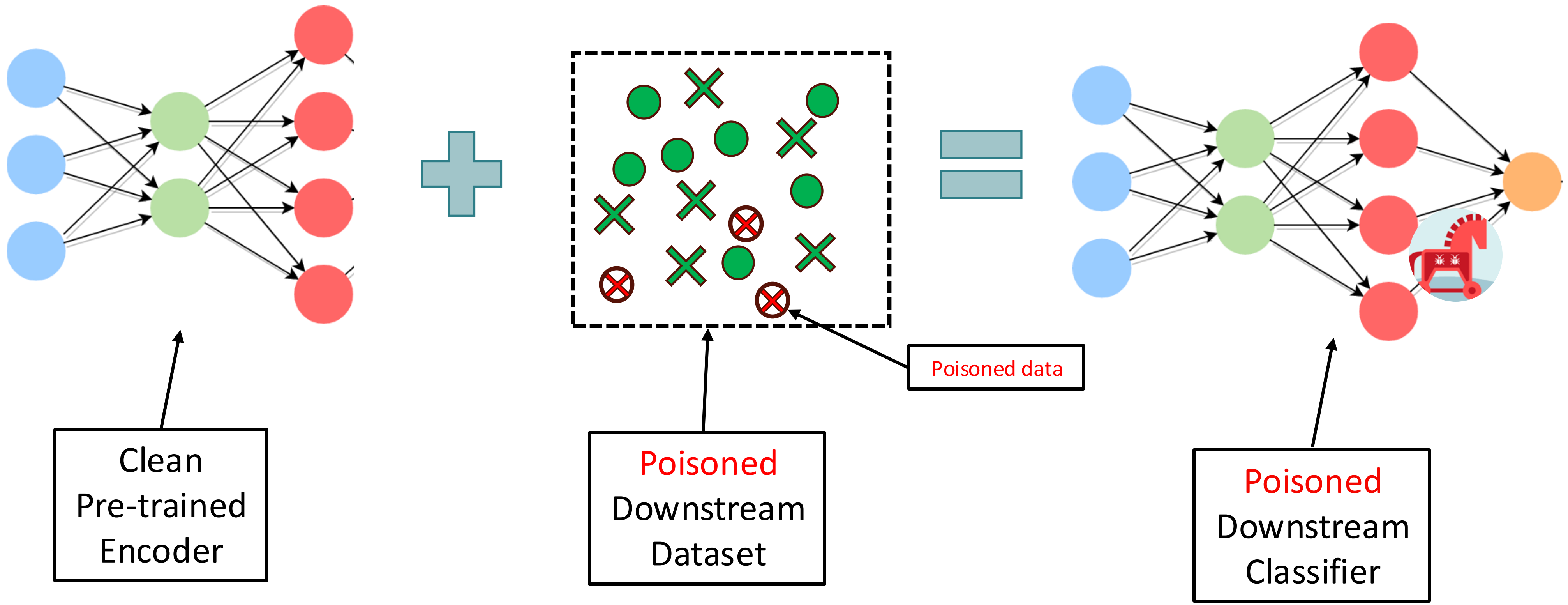
Threat-1: Encoder Poisoning



The attacker introduces a backdoor into the pre-trained encoder, either by directly tuning it to embed a trigger, or by poisoning pre-training data. The downstream classifier becomes poisoned.

Backdoor Threat in Transfer Learning: Taxonomy of Threat Vectors

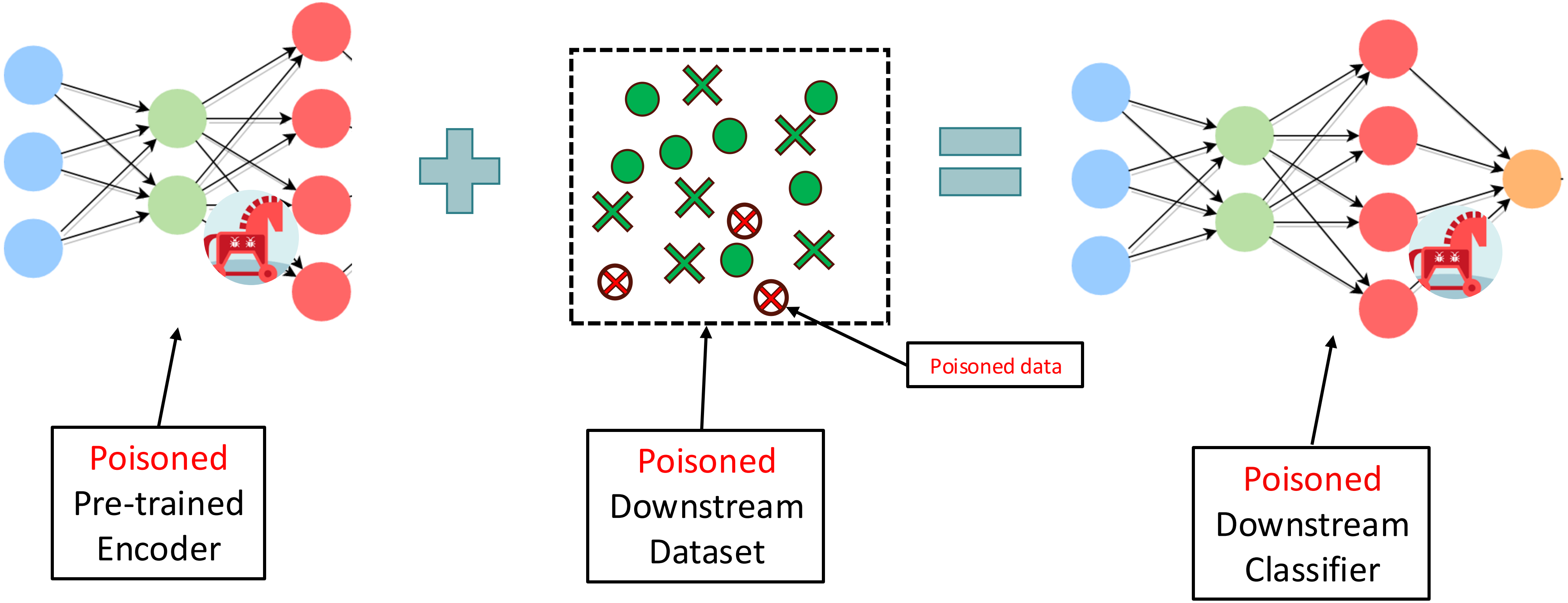
Threat-II: Dataset Poisoning



The attacker introduces a backdoor by poisoning the downstream dataset with injected trigger patterns. The downstream classifier becomes poisoned.

Backdoor Threat in Transfer Learning: Taxonomy of Threat Vectors

Threat-III: Adaptive Poisoning



The attacker introduces a backdoor by poisoning the pre-trained encoder and the downstream dataset with the **same backdoor trigger**. The downstream classifier becomes poisoned.

Defense Context in Transfer Learning

Defense Goal:

- **Utility**: ACC on the downstream task
- **Security**: low ASR
- **Generalizability**: different datasets, encoders, attack vectors, and hyperparameters

Defender's Capabilities and Constraints:

Limited Access to Data and Model:

- No access to pre-training data or hold-out clean data.
- Full control over encoder g and downstream dataset D: access, analysis, and modification allowed.

Ignorance of Threat Model:

- Defender is unaware of the specific backdoor threat.
- Both g and D must be treated as untrustworthy.

Computational Constraints:

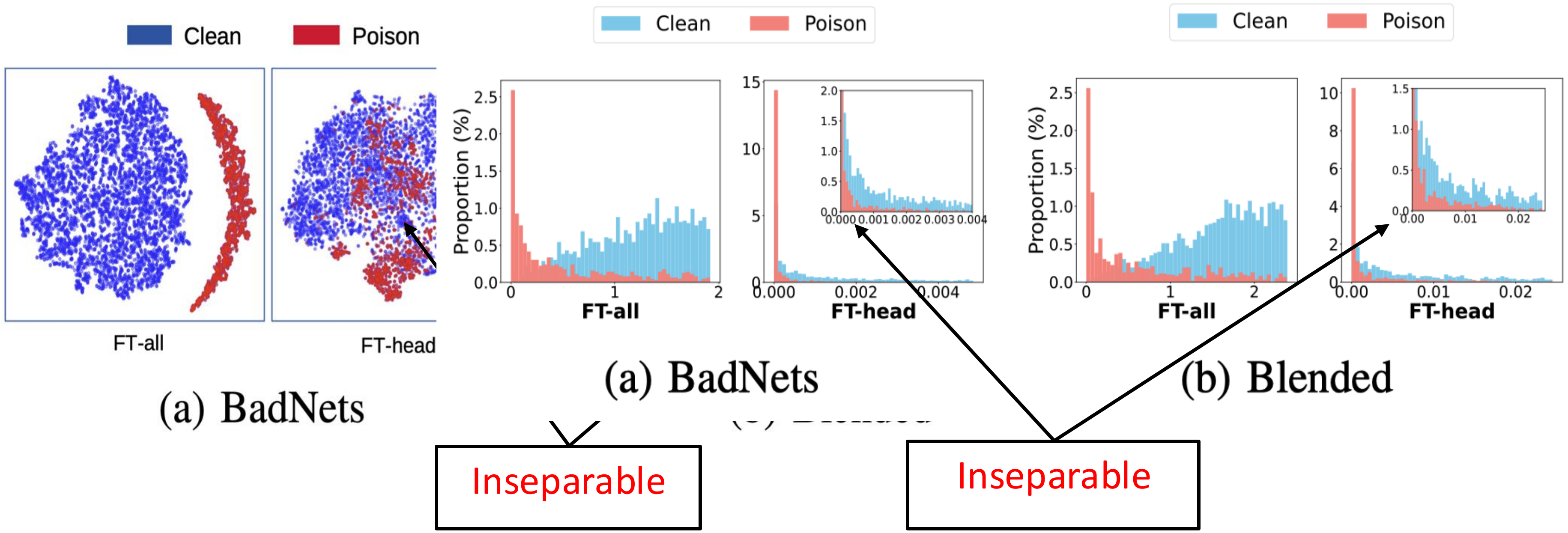
- Defense should be memory-efficient.
- Defense process can span a relatively long period.

Regarding all these constraints, where are we yet?

Current Defense Type I: Poison Detection in SL vs TL

Poison Detection: Identifying and removing abnormal samples from a poisoned dataset (**Threat-II**).

- Rely on **latent separability** or believe poison samples are **low-loss data**.



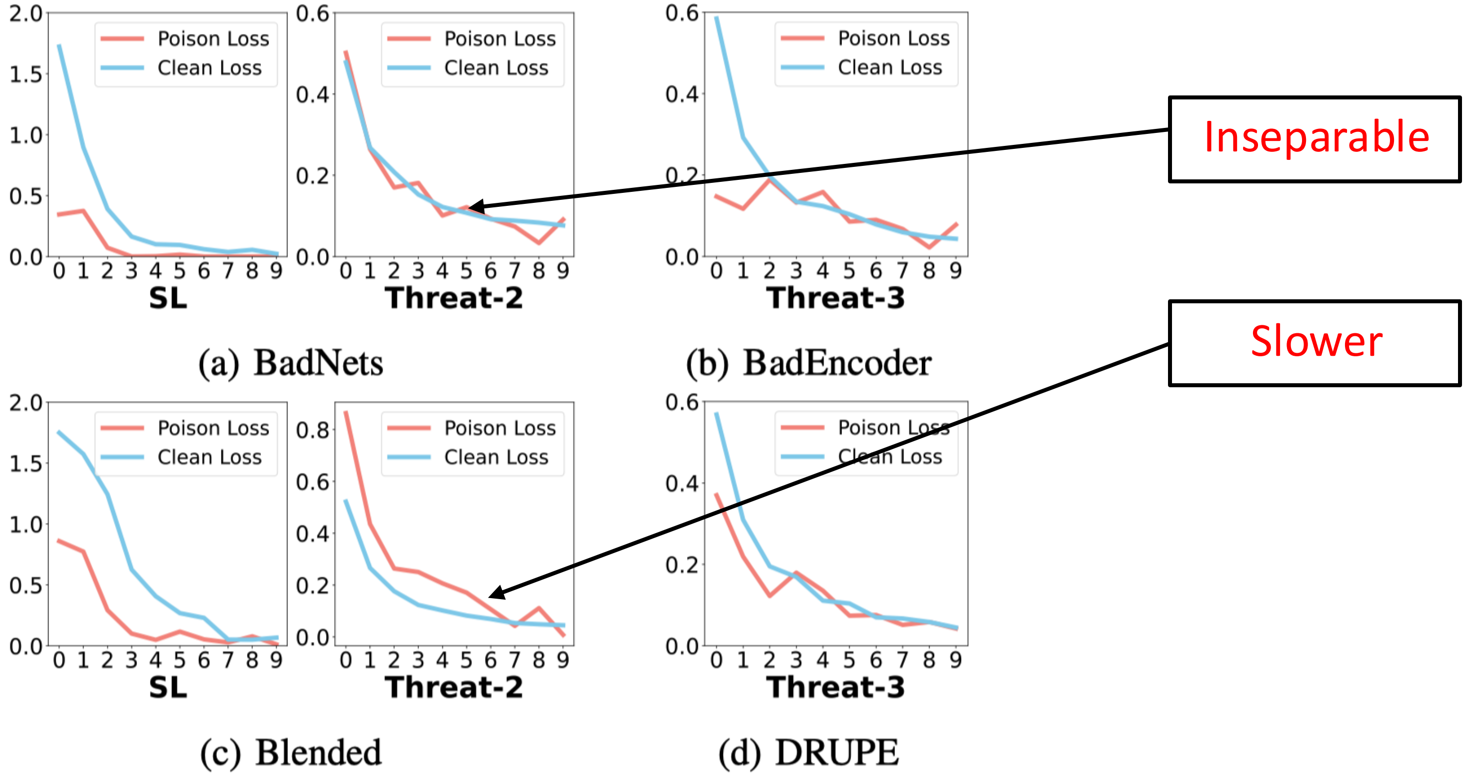
Under transfer learning (even assumes a clean validation dataset):

- **latent separability** assumption does not hold, the poison samples and benign samples are not easily separable.
- **low-loss data** are not excessively poison samples.

Current Defense Type II: Poison Suppression in SL vs TL

Poison Suppression: Train a clean model from poisoned dataset by suppressing backdoor feature (**Threat-II and III**).

- Current poison suppression believes backdoor feature learn faster than benign feature.



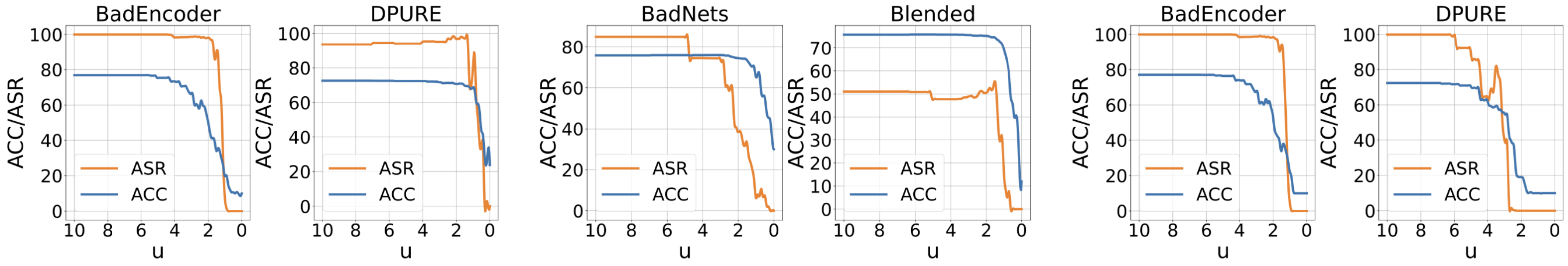
Under transfer learning,

- backdoor feature does not necessarily learn faster than benign feature.

Current Defense Type III : Poison Removal in SL vs TL

Poison Removal: reconstructing a clean model by direct modifying, regardless of how the backdoor was injected (**Threat-I, II and III**).

- Current poison removal requires a hold-out clean dataset or assumes certain property to determine backdoor-related neurons.



(a) **Threat-1** (b) **Threat-2** (c) **Threat-3**

ASR and ACC descend almost together.

Under transfer learning (without access to clean data),

- Blindly making assumptions on what kind of neurons are more likely to be responsible for backdoor, is also unreliable.

Why Existing Defenses Fail in Transfer Learning

Reactive vs Proactive:

Reactive solution: Identifying what constitutes poisoned features or characteristics (followed by eliminating these poison elements).

- Known threats
- What if the threats are unknown: e.g., novel types of attacks, different training paradigms.

Proactive mindset: identifying and amplifying clean elements to defend against unknown backdoor threats.

Our Proactive Design: Trusted Core Bootstrapping

Identifying clean elements (data and neuron/channel):

- **Sifting a Clean Sub-Set:**
  - Majority Rule: A high-credible sample should belong to the majority group of samples in a DNN layer.
  - Consistency Rule: A high-credible sample should have consistent nearest neighbors from its class across different DNN layers.

- **Filtering the Encoder Channel**

Our Proactive Design: Trusted Core Bootstrapping

Identifying clean elements (data and neuron/channel):

- **Sifting a Clean Sub-Set:**
  - Majority Rule: A high-credible sample should belong to the majority group of samples in a DNN layer.
  - Consistency Rule: A high-credible sample should have consistent nearest neighbors from its class across different DNN layers.

- **Filtering the Encoder Channel:**
  - Selective Unlearning: $\max_{\theta_{\text{norm}}} \mathbb{E}_{(x,y)\in\mathcal{D}} \left[\ell\left(f(\phi_{\text{down}}) \circ g(x;\theta_{\text{pre}}), y\right)\right]$
  - Filter Recovering: $\min_{\mathbf{m}^{\kappa}} \mathbb{E}_{(x,y)\in\mathcal{D}_{\text{sub}}} \left[\ell\left(f(\phi_{\text{down}}) \circ g(x;\mathbf{m}^{\kappa} \odot \hat{\theta}_{\text{pre}}), y\right)\right]$
  - Channel Filtering: keep the channels with larger mask values.

Our Proactive Design: Trusted Core Bootstrapping (T-Core)

Bootstrapping Learning (amplifying clean elements):

- Optimization of Untrusted Channels: $\min_{\phi,\psi} \mathbb{E}_{(x,y)\in\mathcal{D}_{\text{clean}}} \left[\ell\left(f(\phi) \circ g(x;\psi \cup \chi), y\right)\right]$

- Clean Data Pool Expansion with Loss Guidance: Incorporate samples with the lowest loss from the entire set into the clean pool.

- Clean Pool Expansion with Meta Guidance:

$$\texttt{Loss}_1 \leftarrow \{\ell(f(\phi) \circ g(x;\phi \cup \chi), y) \mid (x,y) \in \mathcal{D} \setminus \mathcal{D}_{\text{clean}}\};$$
$$\texttt{Loss}_2 \leftarrow \{\ell(f(\phi') \circ g(x;\phi' \cup \chi), y) \mid (x,y) \in \mathcal{D} \setminus \mathcal{D}_{\text{clean}}\};$$

Incorporate samples with the smallest loss reduction $\texttt{Loss}_1 - \texttt{Loss}_2$ into the clean pool.

T-Core's Effectiveness against Dataset Poisoning

Dataset	Dataset Poisoning	BadNets		Blended		SIG		WaNet		TaCT		Adap-Blend		Adap-Patch	
		ACC↑	ASR↓	ACC↑	ASR↓	ACC↑	ASR↓	ACC↑	ASR↓	ACC↑	ASR↓	ACC↑	ASR↓	ACC↑	ASR↓
STL-10	No Defense	75.64	90.24	75.65	50.35	76.51	59.97	76.21	4.76	75.19	64.13	75.75	9.04	76.43	1.92
	Ours	64.08	2.15	65.59	1.60	62.85	6.00	64.55	1.60	66.26	1.00	65.93	3.24	62.55	1.08
CIFAR-10	No Defense	85.04	92.21	84.84	89.12	84.72	89.10	84.40	9.11	84.28	82.60	83.39	34.34	84.16	5.66
	Ours	87.38	3.48	87.35	5.90	87.31	2.54	87.58	0.23	89.04	0.10	87.31	2.54	87.38	3.48
GTSRB	No Defense	81.79	95.02	81.30	90.39	81.90	74.37	80.74	8.81	81.95	89.20	80.85	69.73	78.54	28.20
	Ours	92.03	1.31	91.37	3.04	94.13	0.38	91.10	1.31	91.82	1.87	90.87	0.62	92.25	1.09
SVHN	No Defense	59.80	99.42	60.11	98.30	59.83	97.58	59.65	15.77	59.91	91.90	59.84	89.90	59.87	70.86
	Ours	91.19	4.14	90.88	6.82	91.09	3.22	90.11	1.45	91.25	2.92	90.22	1.31	90.95	1.23
ImageNet-10	No Defense	85.06	92.85	85.00	40.42	86.29	55.33	85.71	3.33	85.88	95.00	86.35	24.06	85.71	6.48
	Ours	80.46	3.86	81.65	2.42	82.00	2.85	83.71	0.94	84.53	3.33	80.24	1.94	81.71	2.48

T-Core consistently yield a low ASR and high ACC.

T-Core's Effectiveness against Encoder Poisoning or Adaptive Poisoning

Encoder Poisoning	Pre-training Dataset	Downstream Dataset	Methods	Threat-1 ACC↑	Threat-1 ASR↓	Threat-3 ACC↑	Threat-3 ASR↓
BadEncoder	CIFAR-10	STL-10	No Defense	76.58	98.51	76.79	100.00
			Ours	55.23	4.29	66.24	1.40
		GTSRB	No Defense	80.77	99.63	78.45	99.97
			Ours	90.86	3.90	91.92	0.01
		SVHN	No Defense	65.35	97.56	67.93	99.44
			Ours	85.93	3.76	92.52	0.65
	STL-10	CIFAR-10	No Defense	70.57	98.93	69.66	99.96
			Ours	60.65	5.22	62.90	6.80
		GTSRB	No Defense	70.83	98.99	66.67	99.83
			Ours	87.08	4.93	90.43	0.76
		SVHN	No Defense	64.89	98.98	63.55	99.57
			Ours	86.76	6.09	87.34	0.54
DRUPE	CIFAR-10	STL-10	No Defense	71.85	97.72	72.39	99.94
			Ours	54.54	6.28	66.38	5.19
		GTSRB	No Defense	76.39	98.10	75.22	99.20
			Ours	93.28	4.50	90.65	3.73
		SVHN	No Defense	72.99	92.71	71.34	99.87
			Ours	87.27	6.47	89.57	3.60
	STL-10	CIFAR-10	No Defense	71.14	80.49	71.21	99.66
			Ours	63.93	1.61	63.07	5.70
		GTSRB	No Defense	65.11	85.03	64.90	99.18
			Ours	84.51	3.97	85.82	0.86
		SVHN	No Defense	58.43	96.28	58.35	99.66
			Ours	87.37	5.58	83.91	0.37
CTRL	STL-10	STL-10	No Defense	52.15	9.88	53.08	9.81
			Ours	48.01	0.18	48.56	1.41
	CIFAR-10	CIFAR-10	No Defense	75.31	44.90	75.63	53.56
			Ours	56.66	3.07	59.35	3.72
	GTSRB	GTSRB	No Defense	66.78	6.54	64.29	26.11
			Ours	82.42	0.87	88.11	1.91
SSLBackdoor	ImageNet	ImageNet-10	No Defense	82.85	36.48	83.29	87.94
			Ours	72.35	0.42	81.35	1.76
CorruptEncoder	ImageNet	ImageNet-10	No Defense	82.35	58.46	82.47	92.12
			Ours	72.82	1.03	81.47	4.79

T-Core consistently yield a low ASR and high ACC.

T-Core’s Effectiveness against Encoder and Dataset Poisoning

Encoder Poisoning	Pre-training Dataset	Downstream Dataset	Dataset Poisoning	BadNets			Blended			SIG			WaNet			TaCT			Adap-Blend			Adap-Patch		
				ACC	ASR-E	ASR-D	ACC	ASR-E	ASR-D	ACC	ASR-E	ASR-D	ACC	ASR-E	ASR-D	ACC	ASR-E	ASR-D	ACC	ASR-E	ASR-D	ACC	ASR-E	ASR-D
BadEncoder	CIFAR-10	STL-10	No Defense	76.30	99.51	91.50	76.28	99.96	60.10	76.51	99.99	59.36	76.43	99.56	4.51	75.71	99.90	62.75	76.19	96.54	10.14	76.93	99.99	1.57
			Ours	67.75	4.67	1.00	67.04	6.85	6.68	53.10	3.88	2.53	67.54	5.11	1.82	67.46	5.72	4.25	68.75	6.65	1.40	68.28	6.03	6.22
		GTSRB	No Defense	72.60	99.24	93.75	73.22	99.77	86.36	73.16	99.15	74.81	78.17	99.94	6.09	73.86	99.20	91.73	72.98	95.95	65.60	72.22	99.69	28.43
			Ours	90.54	0.01	1.38	88.27	0.31	5.05	91.69	0.00	0.98	91.88	0.04	0.66	92.60	0.80	0.00	87.79	0.00	3.30	93.90	0.27	0.29
		SVHN	No Defense	68.47	98.80	99.27	67.98	98.95	98.11	68.19	98.70	96.63	67.99	98.78	11.86	68.19	98.80	94.12	68.07	98.81	90.81	68.26	97.90	71.75
			Ours	92.19	4.29	3.79	92.19	4.29	0.10	92.80	4.80	0.65	90.20	7.94	2.76	91.51	2.49	0.75	90.30	4.23	0.14	92.72	4.86	0.07
	STL-10	CIFAR-10	No Defense	69.56	97.88	78.00	70.33	98.39	71.98	69.72	99.83	77.42	69.94	99.82	9.12	69.66	99.66	70.00	69.84	99.77	16.28	70.03	99.76	5.78
			Ours	63.27	5.76	4.76	62.73	6.28	4.97	68.42	8.29	3.64	62.63	6.61	4.47	65.47	6.36	0.00	64.38	7.71	2.03	63.05	6.08	0.13
		GTSRB	No Defense	70.67	97.52	83.43	69.59	98.77	82.33	70.86	99.19	74.56	69.63	99.80	4.33	68.33	98.05	81.07	68.56	99.10	54.45	69.58	98.95	12.30
			Ours	85.65	0.11	5.45	86.03	0.70	0.87	85.18	1.73	0.24	85.27	0.22	4.39	86.03	0.05	1.06	85.58	1.10	5.13	87.05	1.80	1.52
		SVHN	No Defense	67.44	85.95	98.85	66.29	85.93	98.93	67.45	88.96	93.92	64.88	84.07	11.91	67.78	87.69	94.53	67.60	81.29	89.94	66.77	80.30	26.85
			Ours	83.90	4.30	10.10	86.63	3.72	5.32	85.96	9.18	2.55	88.96	5.10	1.01	86.34	3.15	0.31	86.40	4.87	2.09	86.92	6.15	4.50
DRUPE	CIFAR-10	STL-10	No Defense	71.94	99.43	75.22	71.09	98.00	53.97	72.49	93.63	35.50	72.08	90.18	10.14	71.78	97.54	49.75	71.34	99.39	11.42	71.63	98.35	1.89
			Ours	63.16	14.90	10.92	68.30	10.89	5.89	64.34	7.49	0.49	64.59	6.38	4.29	63.63	11.24	13.00	64.74	7.92	2.67	65.00	7.39	2.96
		GTSRB	No Defense	74.35	73.36	94.19	74.57	72.99	87.63	74.95	74.70	69.57	74.48	73.02	6.58	74.67	72.91	87.07	73.95	73.01	61.30	73.76	72.97	14.79
			Ours	87.98	7.05	3.16	90.17	7.23	6.66	88.16	3.18	0.74	89.14	3.61	0.47	89.93	5.82	6.82	89.14	5.05	7.63	89.87	3.10	1.85
		SVHN	No Defense	71.35	75.53	99.45	71.37	75.74	97.60	71.21	75.81	94.45	71.04	76.95	11.60	71.31	72.91	96.35	71.26	77.03	85.17	71.09	76.30	51.23
			Ours	89.54	9.64	6.78	88.73	6.92	4.90	89.02	9.88	4.32	87.19	6.66	3.66	92.34	3.60	2.77	89.20	5.10	1.01	89.70	5.04	2.97
	STL-10	CIFAR-10	No Defense	70.26	78.54	74.24	70.71	77.58	74.19	70.83	79.10	69.62	70.87	78.66	9.27	70.62	78.55	69.00	70.81	78.63	14.13	71.15	78.63	4.93
			Ours	64.74	6.87	7.43	63.46	7.53	7.69	67.31	4.94	1.91	66.18	4.02	1.73	66.28	5.49	0.10	62.63	4.96	3.40	63.56	3.31	6.31
		GTSRB	No Defense	63.40	78.25	90.50	63.71	84.92	88.70	64.29	85.40	74.55	63.99	78.12	6.09	63.47	86.80	78.54	61.18	80.32	67.40	62.00	79.83	18.46
			Ours	86.10	0.21	3.94	87.08	1.42	5.85	86.44	2.82	0.03	84.47	1.00	3.18	82.18	0.25	5.45	81.90	1.61	2.95	81.32	0.62	7.58
		SVHN	No Defense	59.12	94.66	96.56	59.77	97.48	97.43	58.03	92.94	91.53	59.77	95.08	15.17	59.47	97.46	92.33	60.02	98.69	84.58	59.74	96.81	16.52
			Ours	82.13	5.95	6.25	83.22	4.03	4.56	83.75	9.64	2.77	82.76	2.45	3.59	83.85	2.93	0.98	81.13	9.01	5.10	83.17	3.05	1.65

T-Core consistently yield a low ASR and high ACC.

Summary

- We identify a complex and challenging yet general backdoor threat model within the transfer learning scenario that previous research has overlooked.

- We conduct an exhaustive analysis of the existing backdoor and reveal their limitations under the transfer learning scenario.

- We propose a proactive mindset as an alternative and introduce a Trusted Core Bootstrapping framework as an instantiation, providing concrete designs that are more robust and generalizable.

Thanks!