PoisonedEncoder: Poisoning the Unlabeled Pre-training Data in Contrastive Learning

Hongbin Liu, Jinyuan Jia, and Neil Zhenqiang Gong,

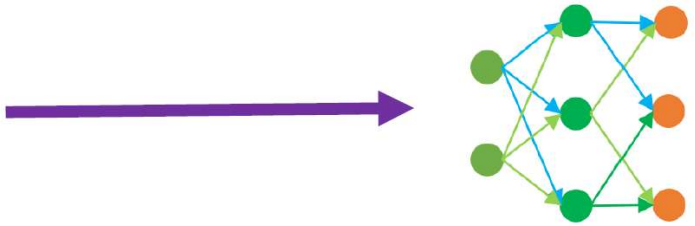
Duke University

Usenix Security 2022

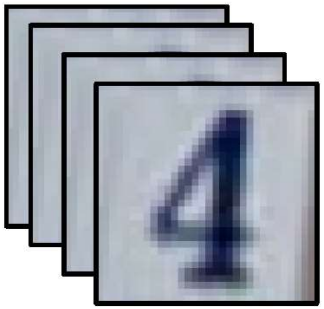
Conventional ML Paradigm

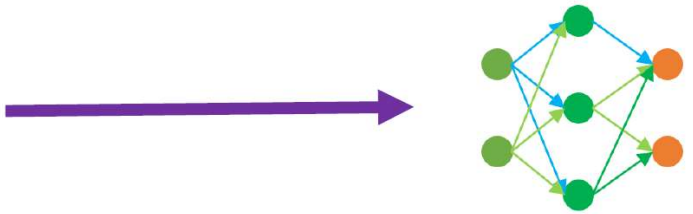
- **Supervised Learning**

Traffic sign recognition



Digit recognition

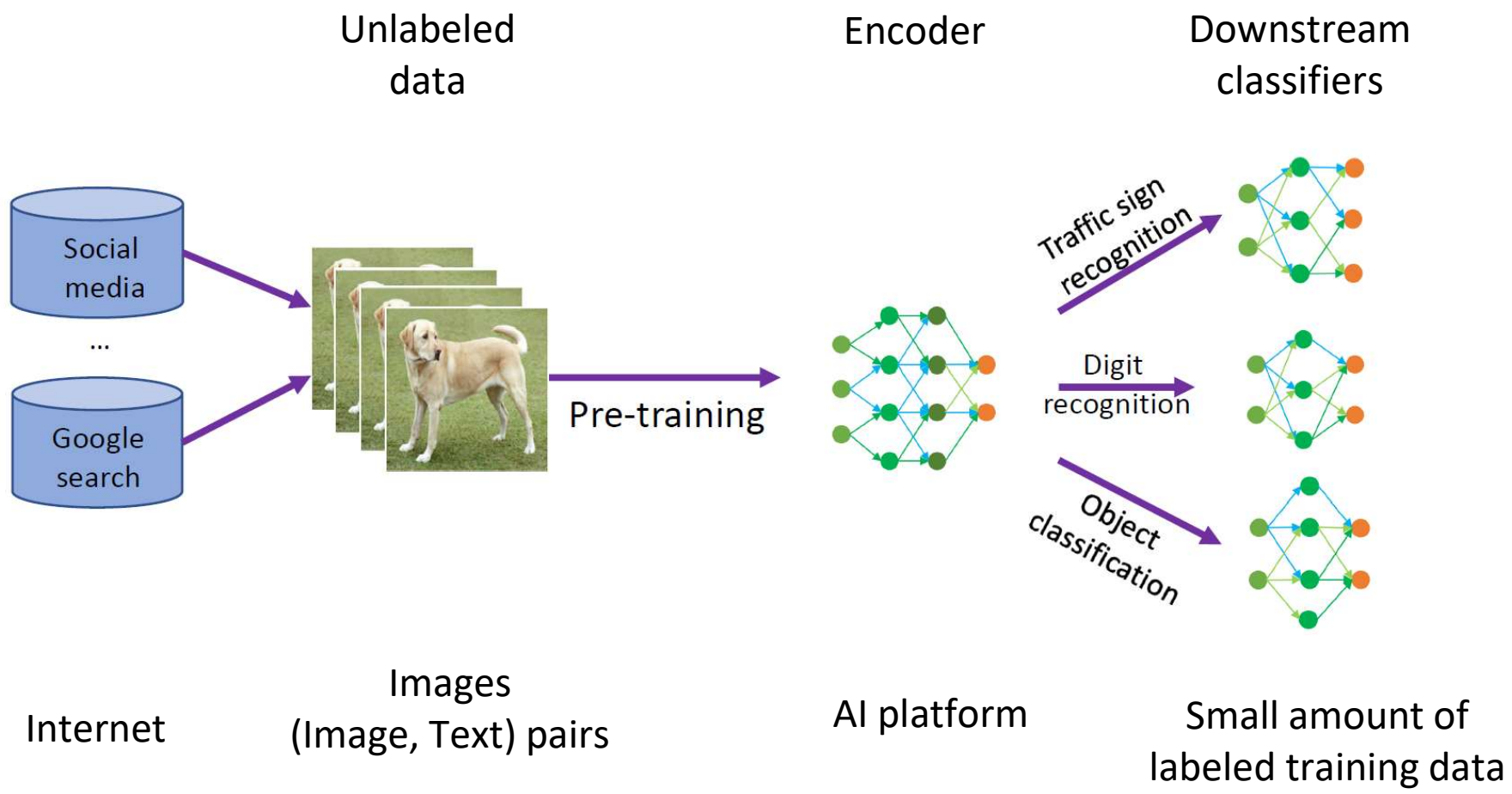




Key Challenge: require lots of labeled training data for each task

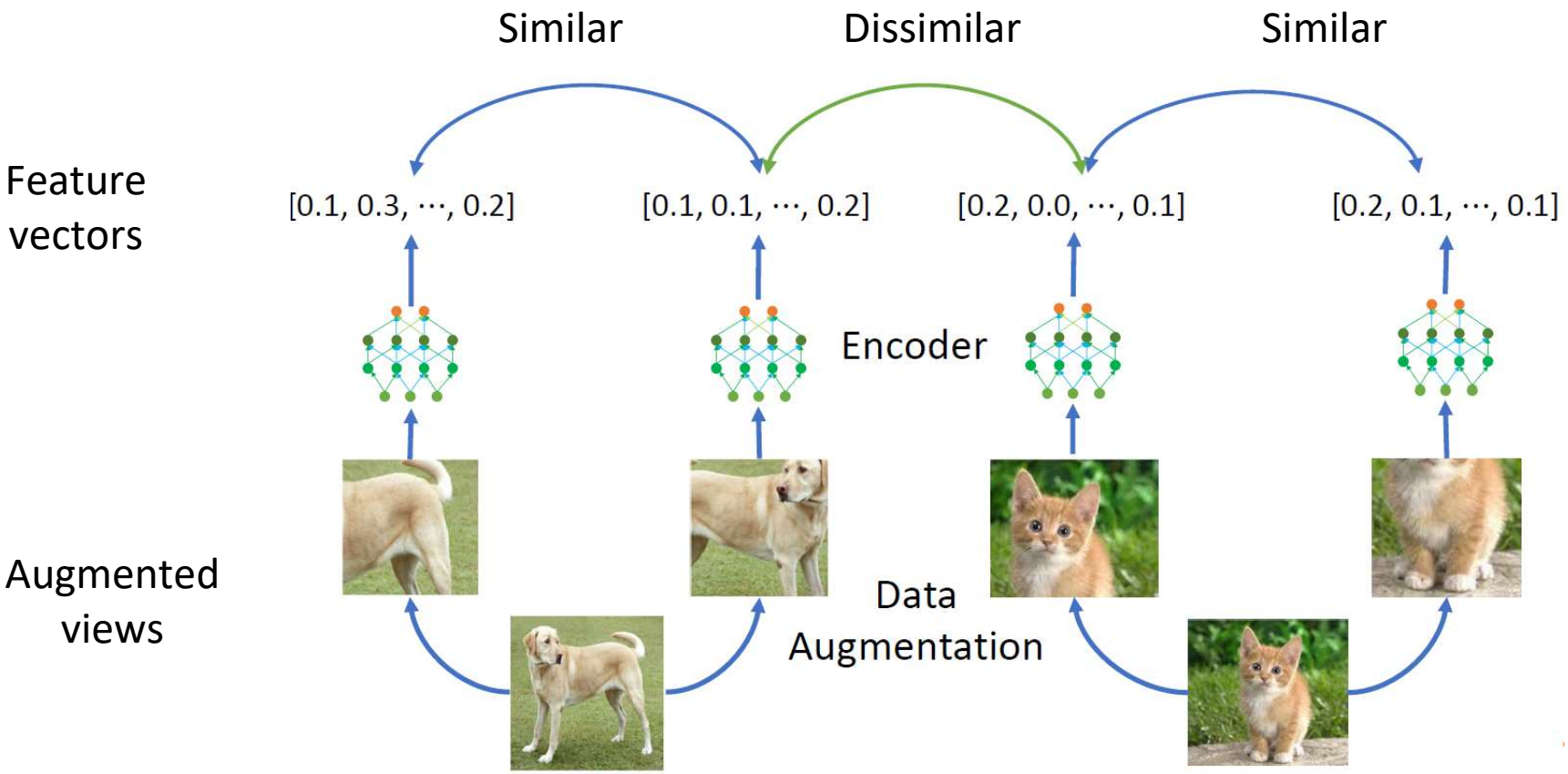
General-Purpose AI

- **Pre-trained Encoder**



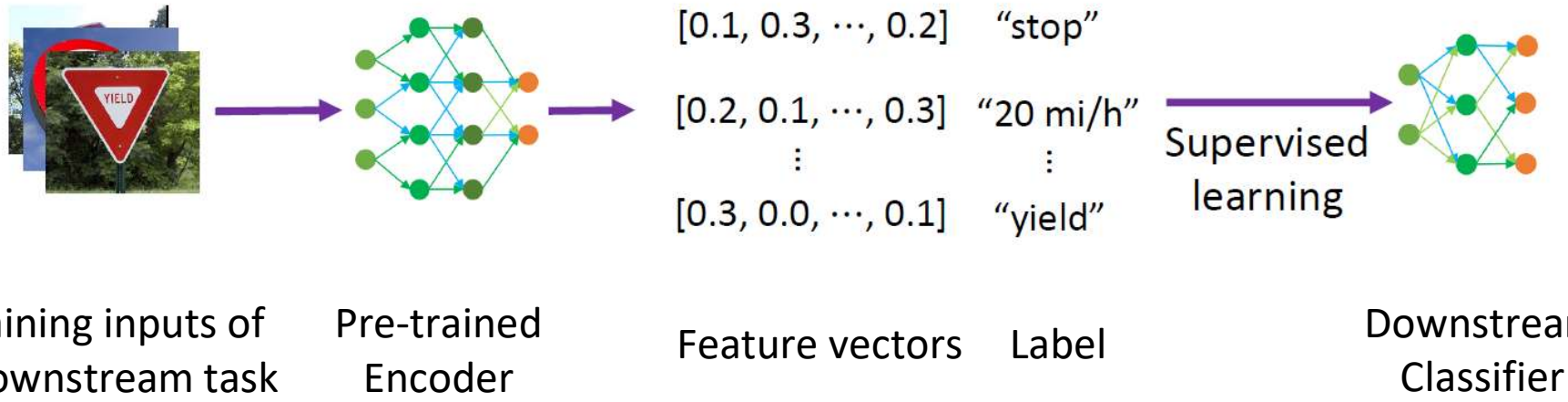
Contrastive Learning

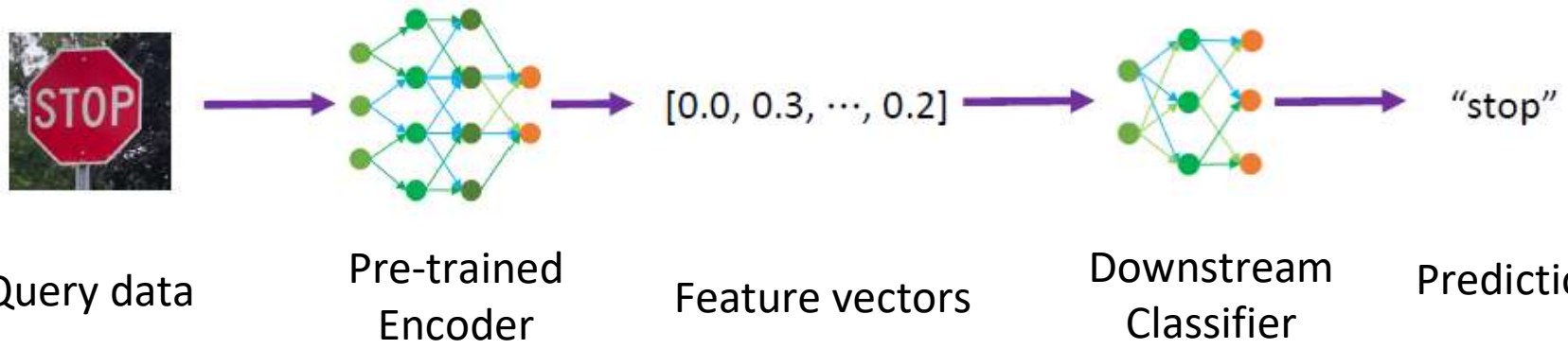
- **Pre-training an Encoder– SimCLR [ICML'20]**



Contrastive Learning

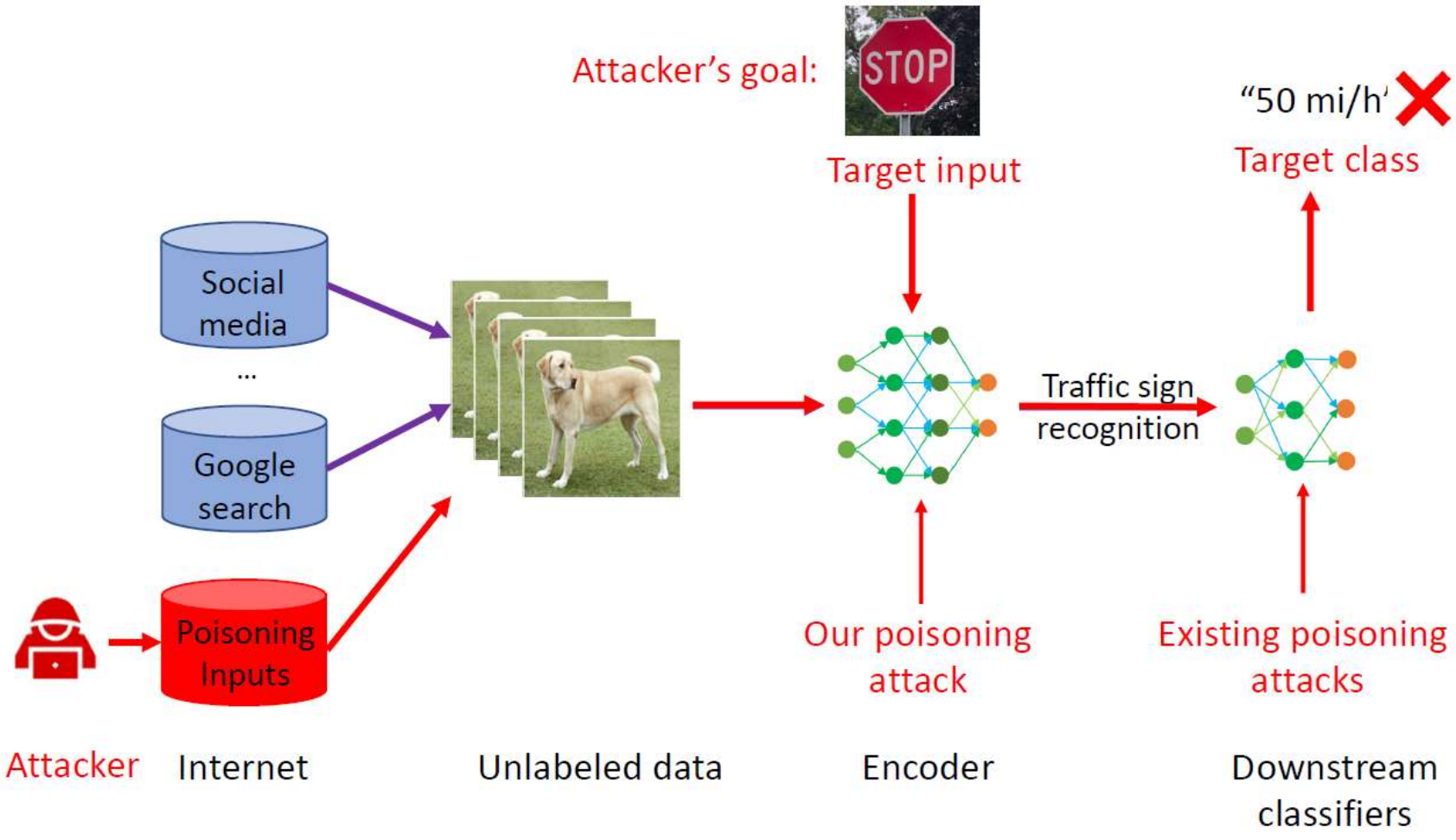
- **Downstream Classifier**





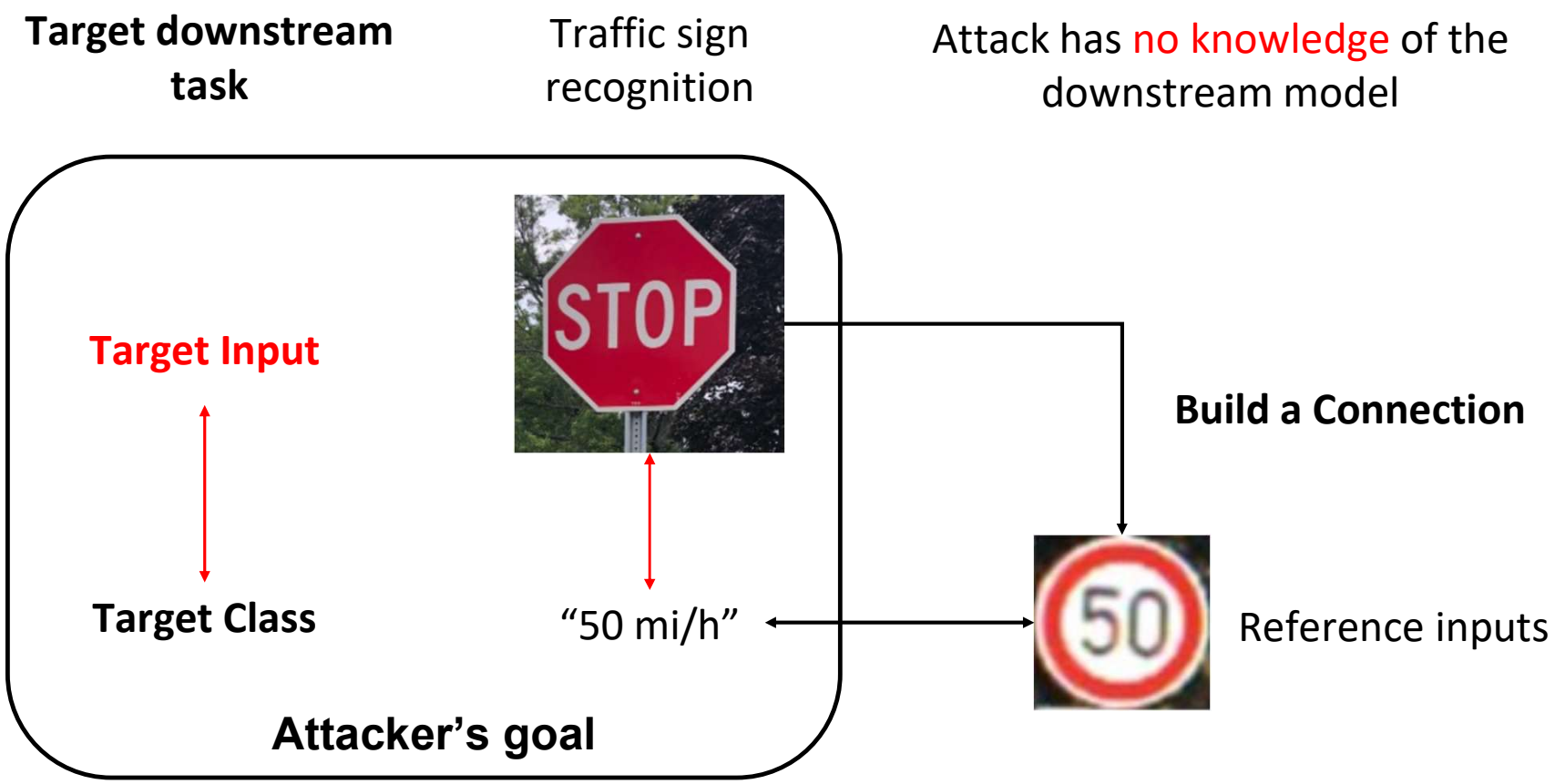
Poisoning Attacks

- **Vulnerable to Poisoning Attacks**



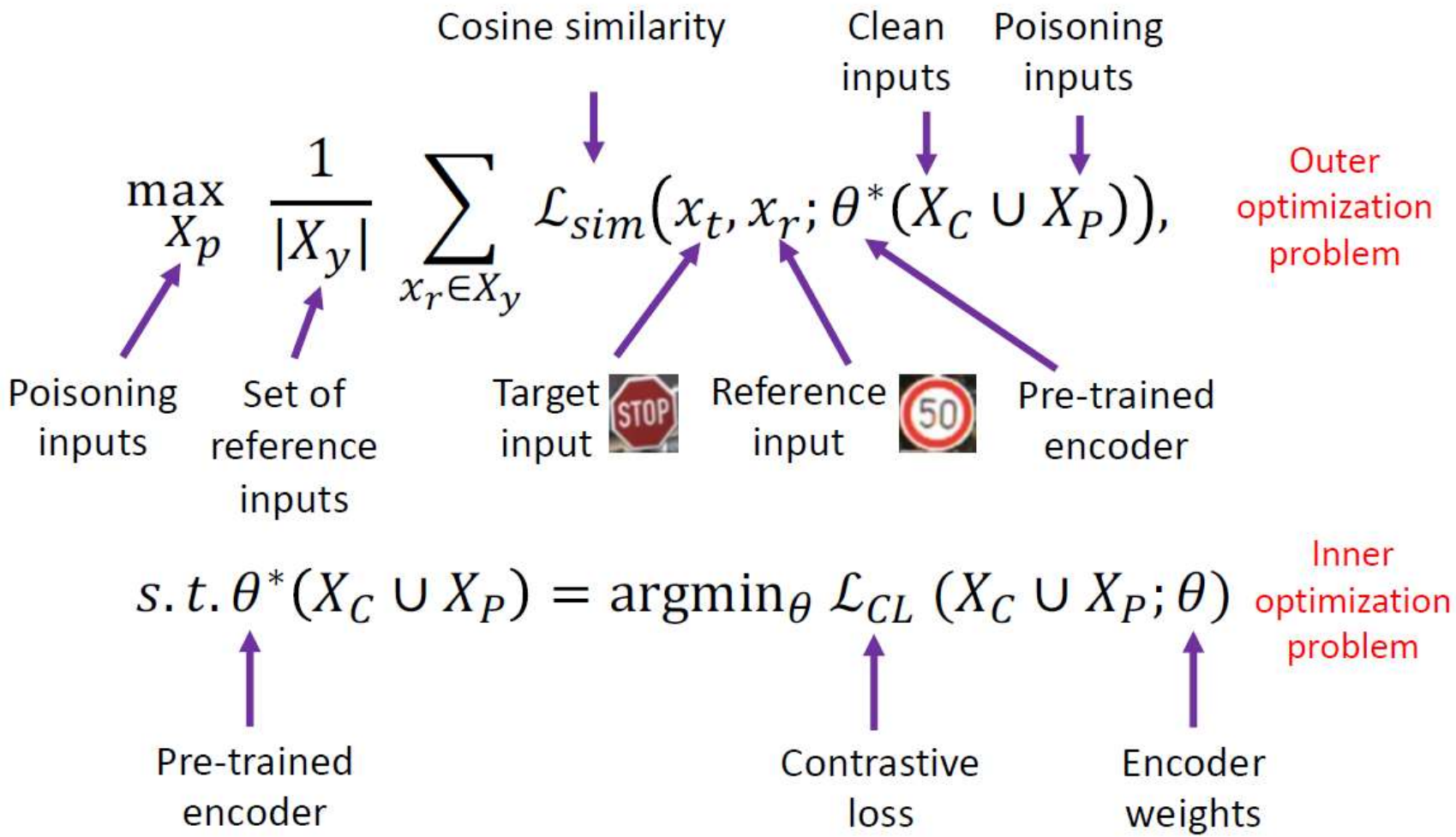
Threat Model

- **Attacker's goal and background knowledge**



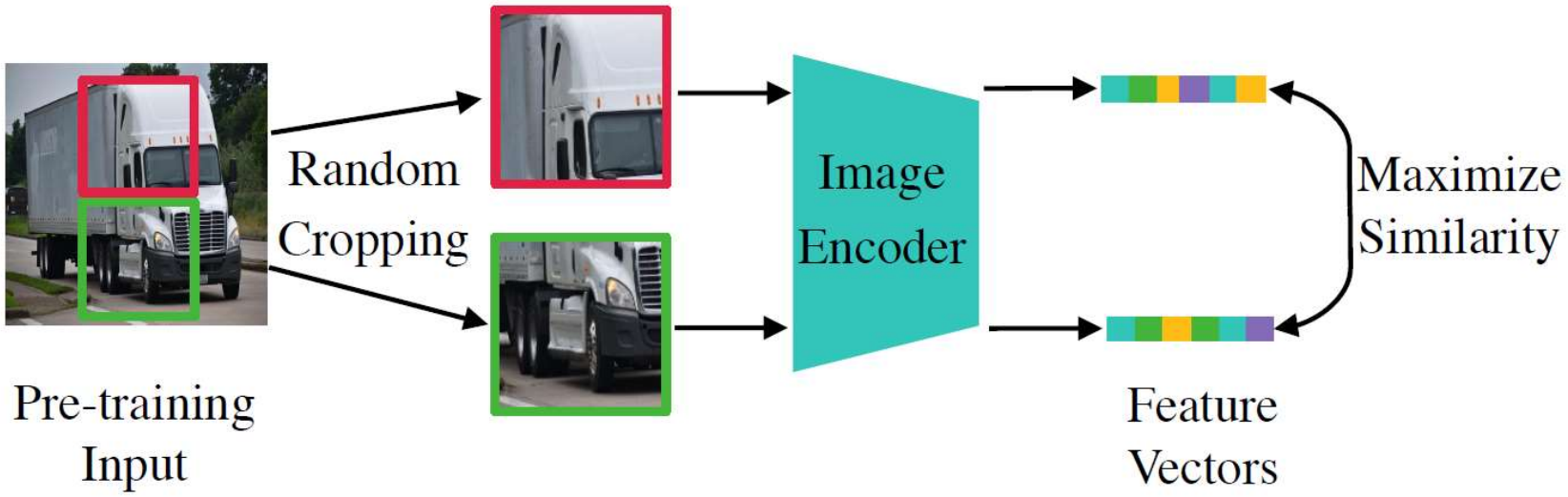
Poisoning attack

- **A bi-level optimization problem**



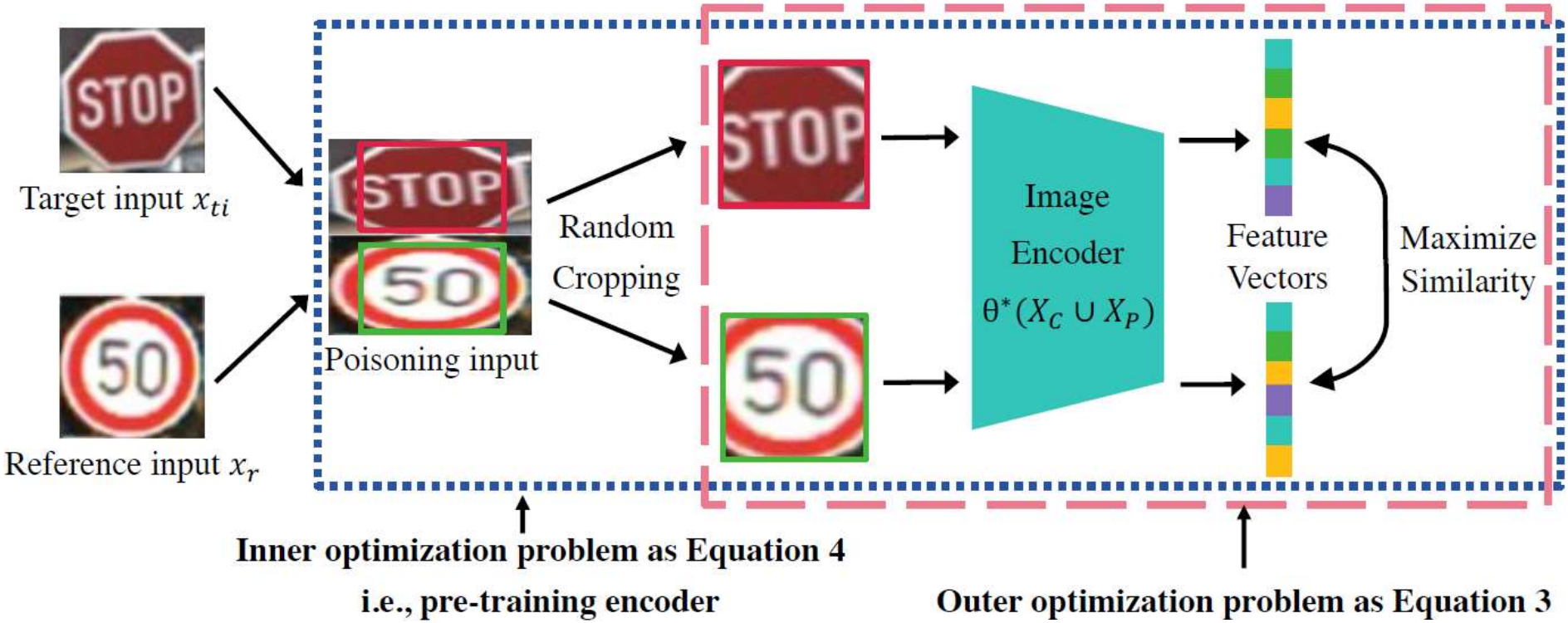
PoisonedEncoder: Heuristic solution

- **Revisit the Contrastive Learning**



PoisonedEncoder: Heuristic solution

- **Image Combination**



$$\theta^*(X_c \cup X_p) = \underset{\theta}{\operatorname{argmin}} \mathcal{L}_{\mathrm{CL}}(X_c \cup X_p; \theta) \qquad \mathcal{L}_{\mathrm{sim}}(x_{ti}, x_r; \theta^*(X_c \cup X_p))$$

Examples of Combined Images

- **Image Combination**

(a) Combination 1 (b) Combination 2

(c) Combination 3 (d) Combination 4

Four combination methods

Real-world examples

Experiments

- **Experimental Setup**

Pre-training encoders

- Pre-training algorithm: SimCLR, MoCo
- Pre-training dataset: CIFAR10 ImageNet

Building downstream classifiers

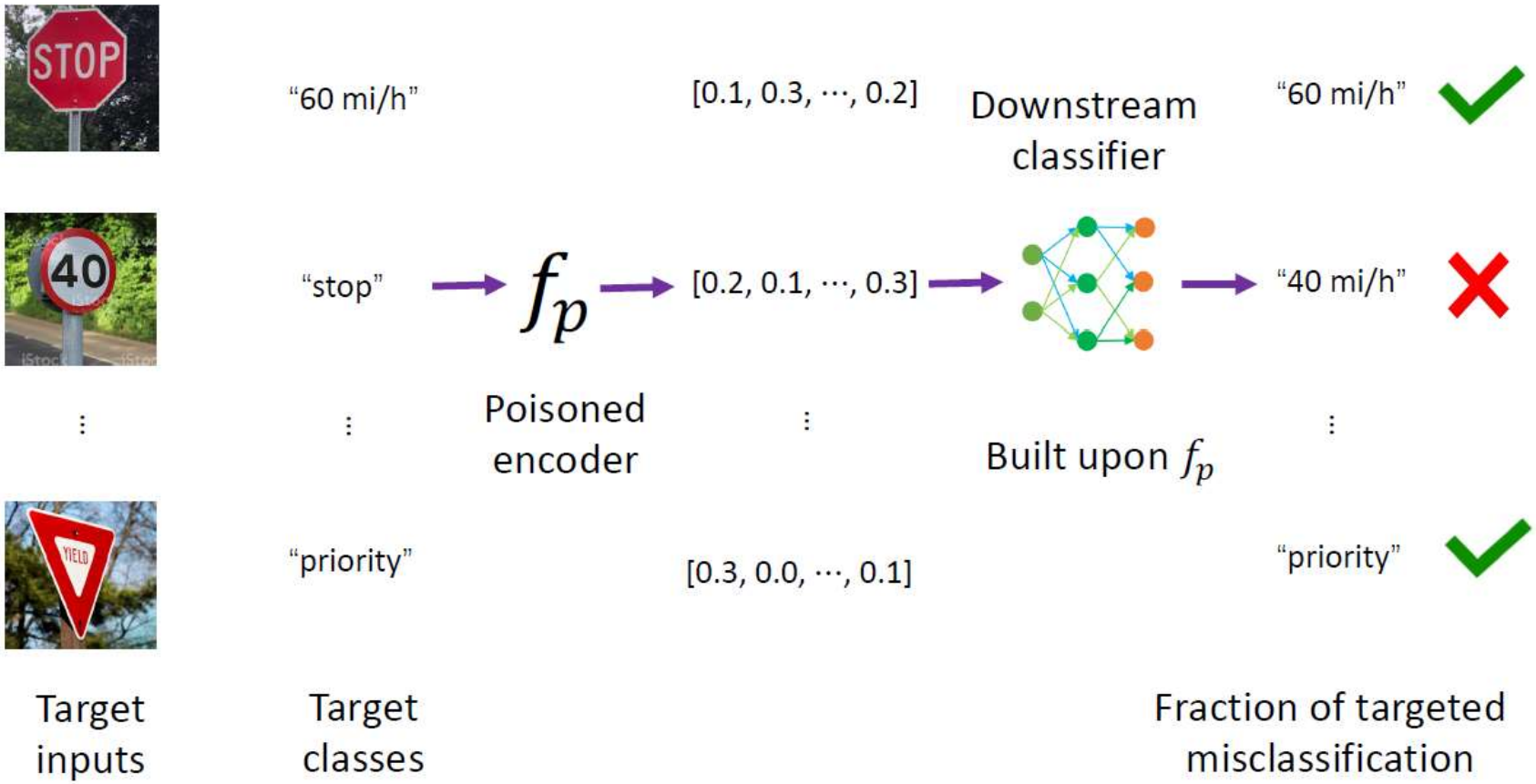
- Downstream tasks
  STL10, Facemask, EuroSAT
- Downstream classifier
  A fully connected neural network

Parameter settings

- # reference inputs = 50
- Poisoning rate = 1%
- # random experimental trails = 10

Experiments

- **Attack Success Rate**



Experiments

- **Comparison**

Pre-training Dataset	Target Downstream Dataset	Witches' Brew	ICP	Ours
CIFAR10	STL10	0.1	0.5	0.8
	Facemask	0.1	0.6	0.9
	EuroSAT	0.0	0.2	0.5
Tiny-ImageNet	STL10	0.0	0.4	0.7
	Facemask	0.1	0.8	1.0
	EuroSAT	0.0	0.2	0.4

No Attack	Witches' Brew	ICP	Ours
0.183	0.257	0.463	0.689

Experiments

- **PoisonedEncoder preserves utility**

Pre-training Dataset	Target Downstream Dataset	Downstream Dataset	CA	PA
CIFAR10	STL10	STL10	0.718	0.715
		Facemask	0.947	0.952
		EuroSAT	0.815	0.821
	Facemask	STL10	0.718	0.716
		Facemask	0.947	0.937
		EuroSAT	0.815	0.820
	EuroSAT	STL10	0.718	0.724
		Facemask	0.947	0.953
		EuroSAT	0.815	0.797
Tiny-ImageNet	STL10	STL10	0.635	0.637
		Facemask	0.965	0.968
		EuroSAT	0.816	0.853
	Facemask	STL10	0.635	0.633
		Facemask	0.965	0.977
		EuroSAT	0.816	0.855
	EuroSAT	STL10	0.635	0.633
		Facemask	0.965	0.970
		EuroSAT	0.816	0.844

CA: clean accuracy
PA: poisoned accuracy

Experiments

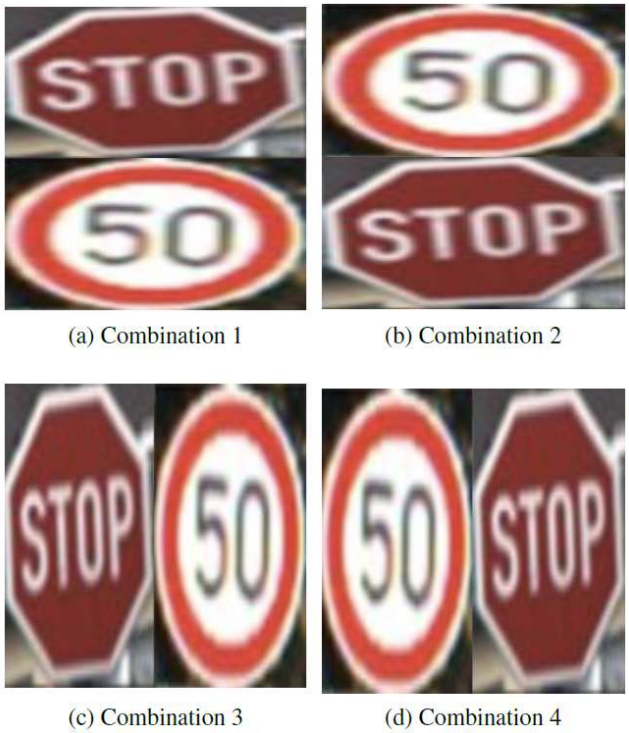
- **Impact of different combination methods**

(a) Pre-trained on CIFAR10

Combination Method	ASR
1	0.4
2	0.5
3	0.5
4	0.6
1+2	0.5
1+2+3	0.6
1+2+3+4	0.8

(b) Pre-trained on Tiny-ImageNet

Combination Method	ASR
1	0.3
2	0.3
3	0.2
4	0.4
1+2	0.5
1+2+3	0.6
1+2+3+4	0.7

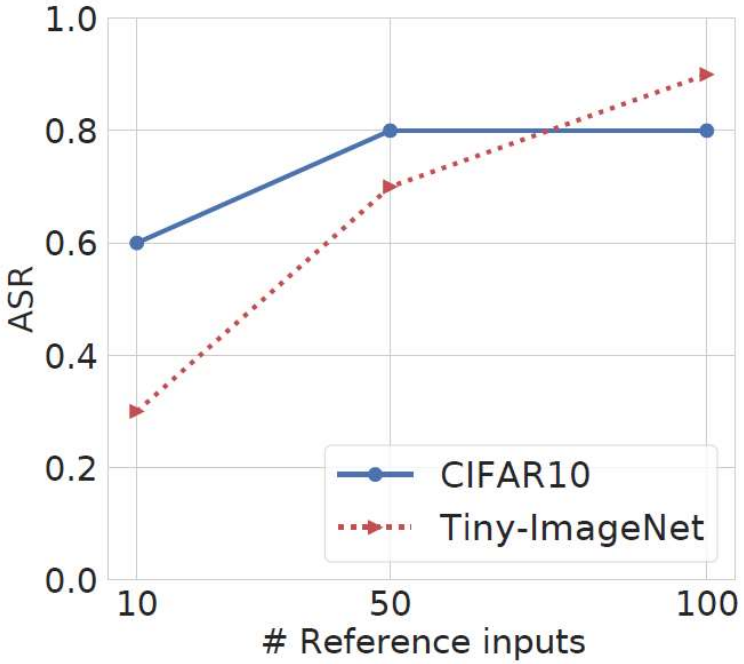


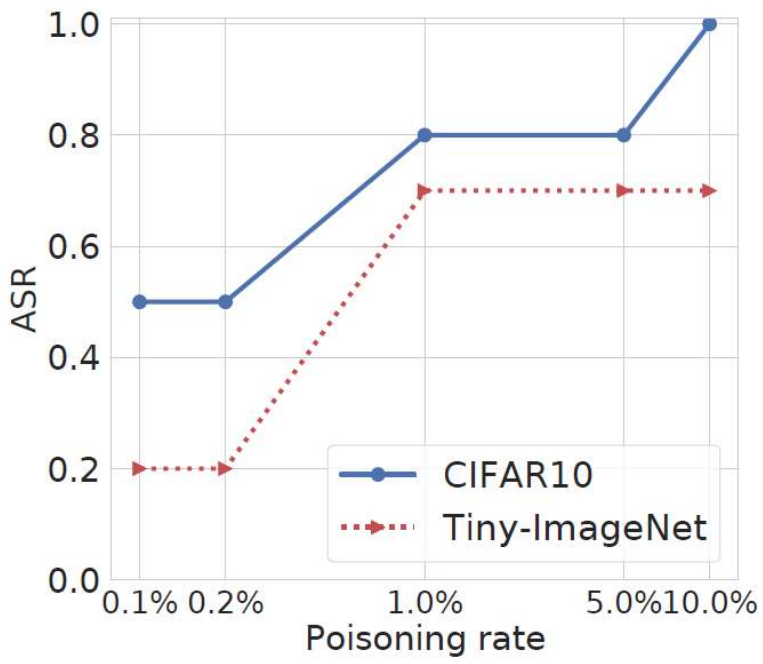
(a) Combination 1 (b) Combination 2

(c) Combination 3 (d) Combination 4

Experiments

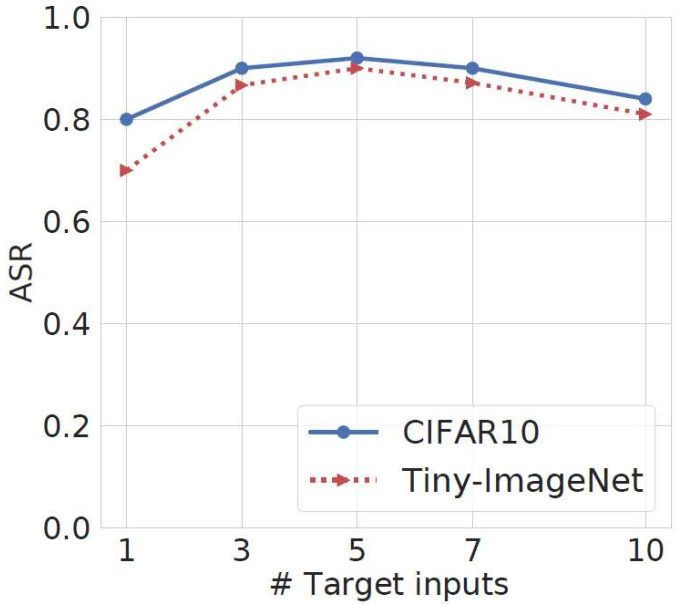
- **Impact of the number of reference inputs and poisoning rate**





Experiments

- **Impact of the # of target inputs and target downstream tasks**



(a) Pre-trained on CIFAR10

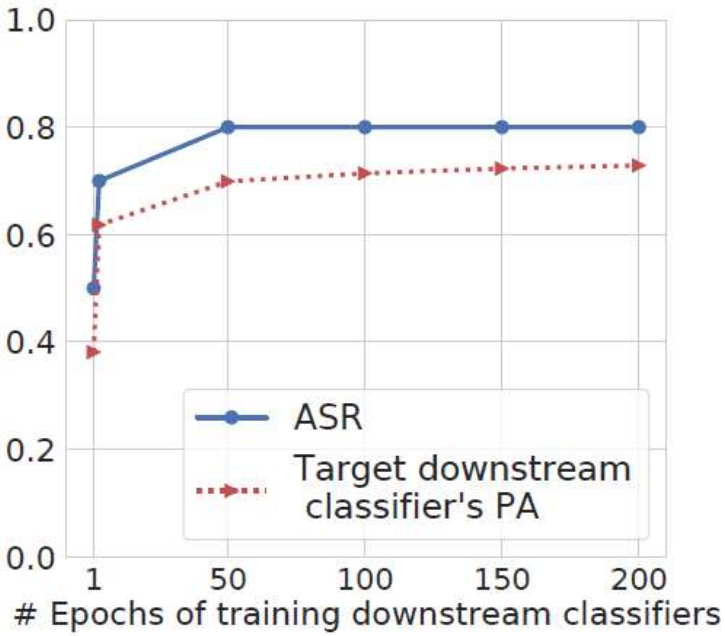
Target Downstream Tasks	ASR
STL10	0.8
Facemask	0.9
EuroSAT	0.5

(b) Pre-trained on Tiny-ImageNet

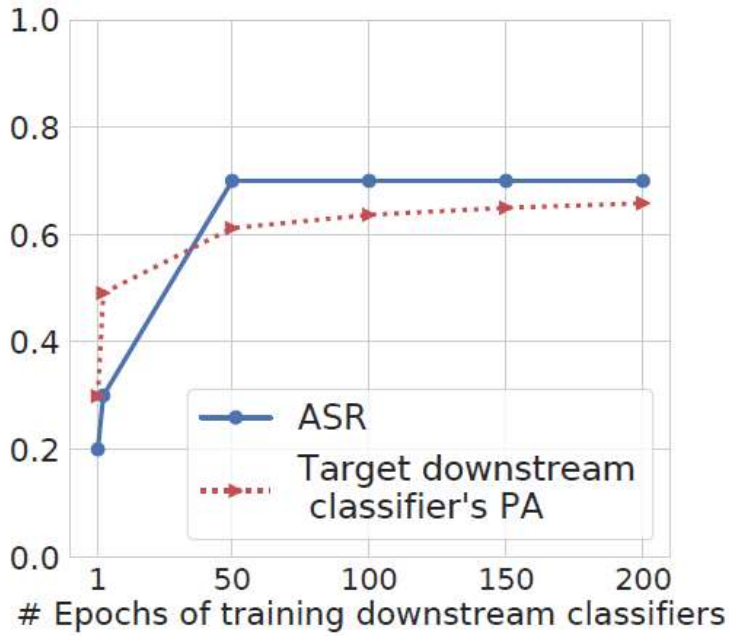
Target Downstream Tasks	ASR
STL10	0.6
Facemask	1.0
EuroSAT	0.4

Defense

- **Early Stopping**



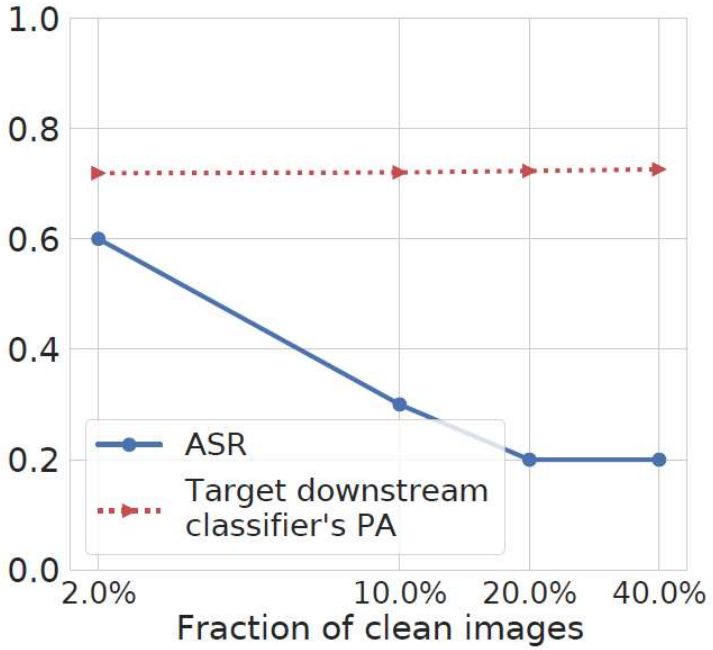
(c) Pre-trained on CIFAR10



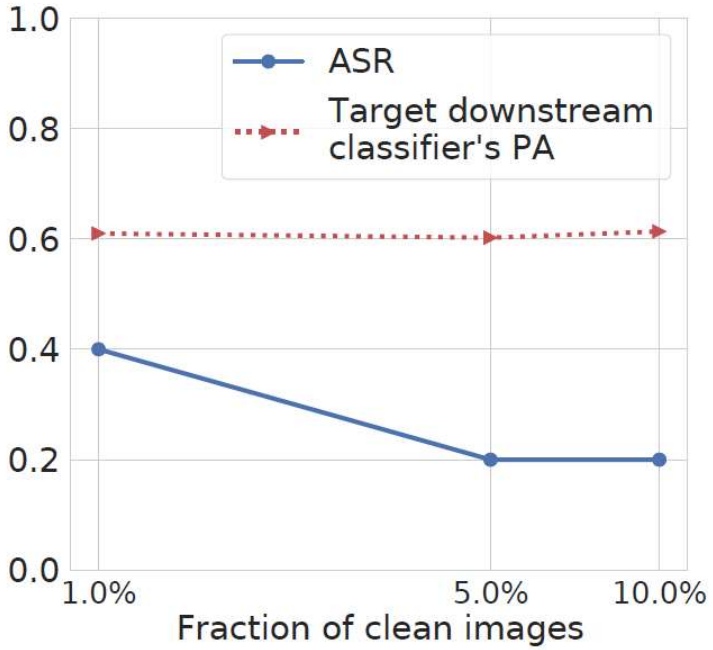
(d) Pre-trained on Tiny-ImageNet

Defense

- **Early Stopping**



(a) Pre-trained on CIFAR10



(b) Pre-trained on Tiny-ImageNet