JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Can't Steal? Cont-Steal! Contrastive Stealing Attacks Against Image Encoders

Zeyang Sha[†] Xinlei He[†] Ning Yu[‡] Michael Backes[†] Yang Zhang[†]

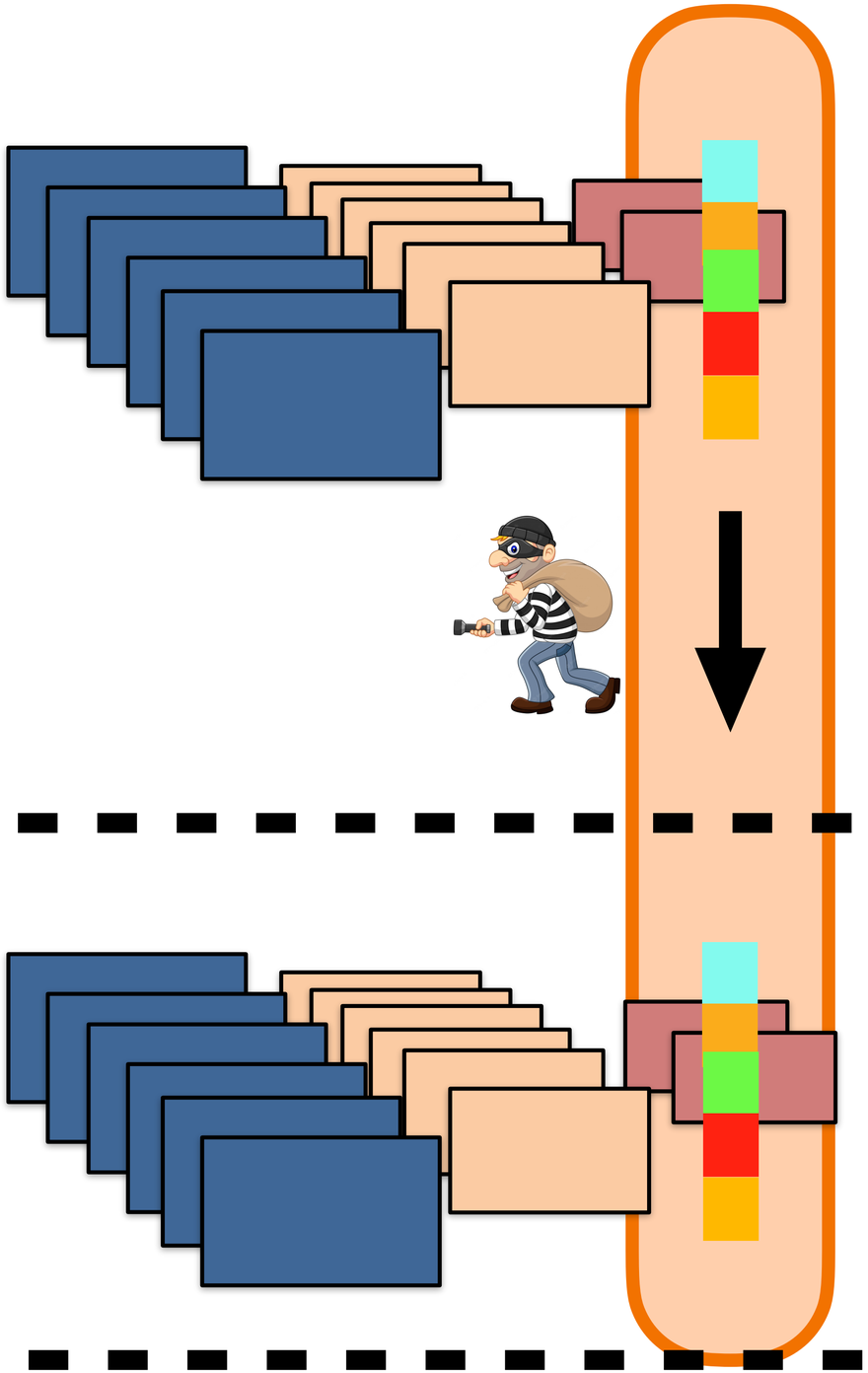
[†]*CISPA Helmholtz Center for Information Security* [‡]*Salesforce Research*

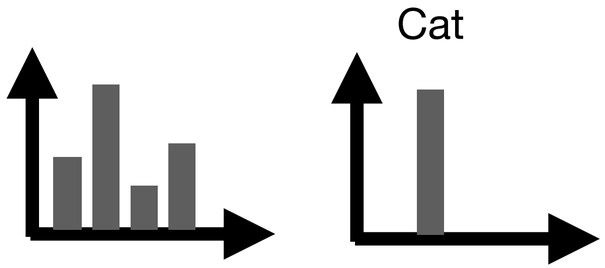
{zeyang.sha, xinlei.he, director, zhang}@cispa.de, ning.yu@salesforce.com

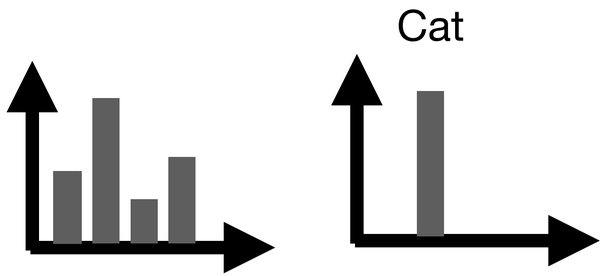
Poster: WED-PM-383

CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

salesforce




Cat


Cat

Threat Model

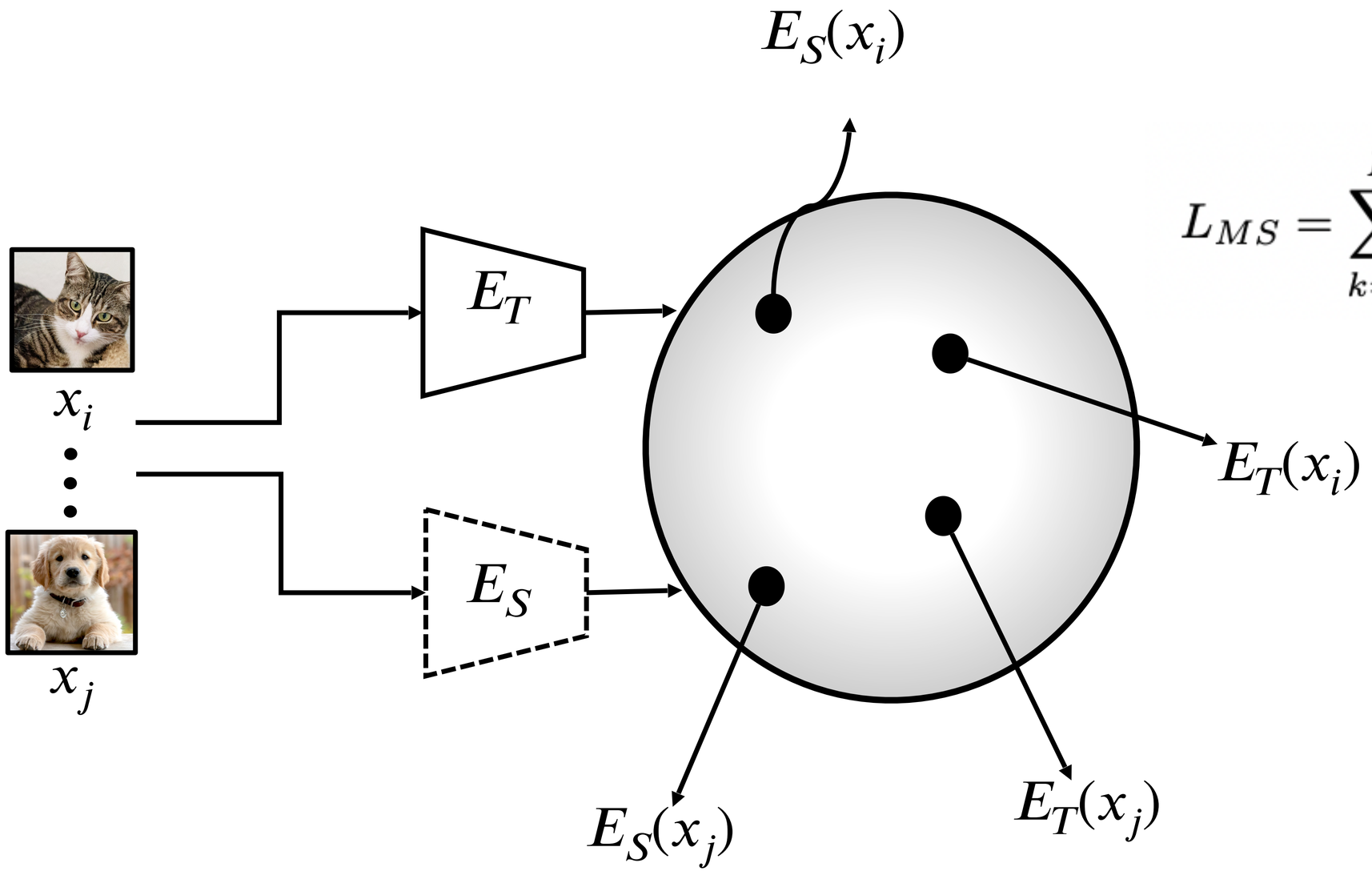
Adversary's Goal:

- **Theft:** The theft adversary aims to build a surrogate encoder that has similar performance on the downstream tasks as the target encode.
- **Utility:** The utility adversary is to construct a surrogate encoder that behaves normally on different downstream tasks.

Adversary's Background Knowledge:

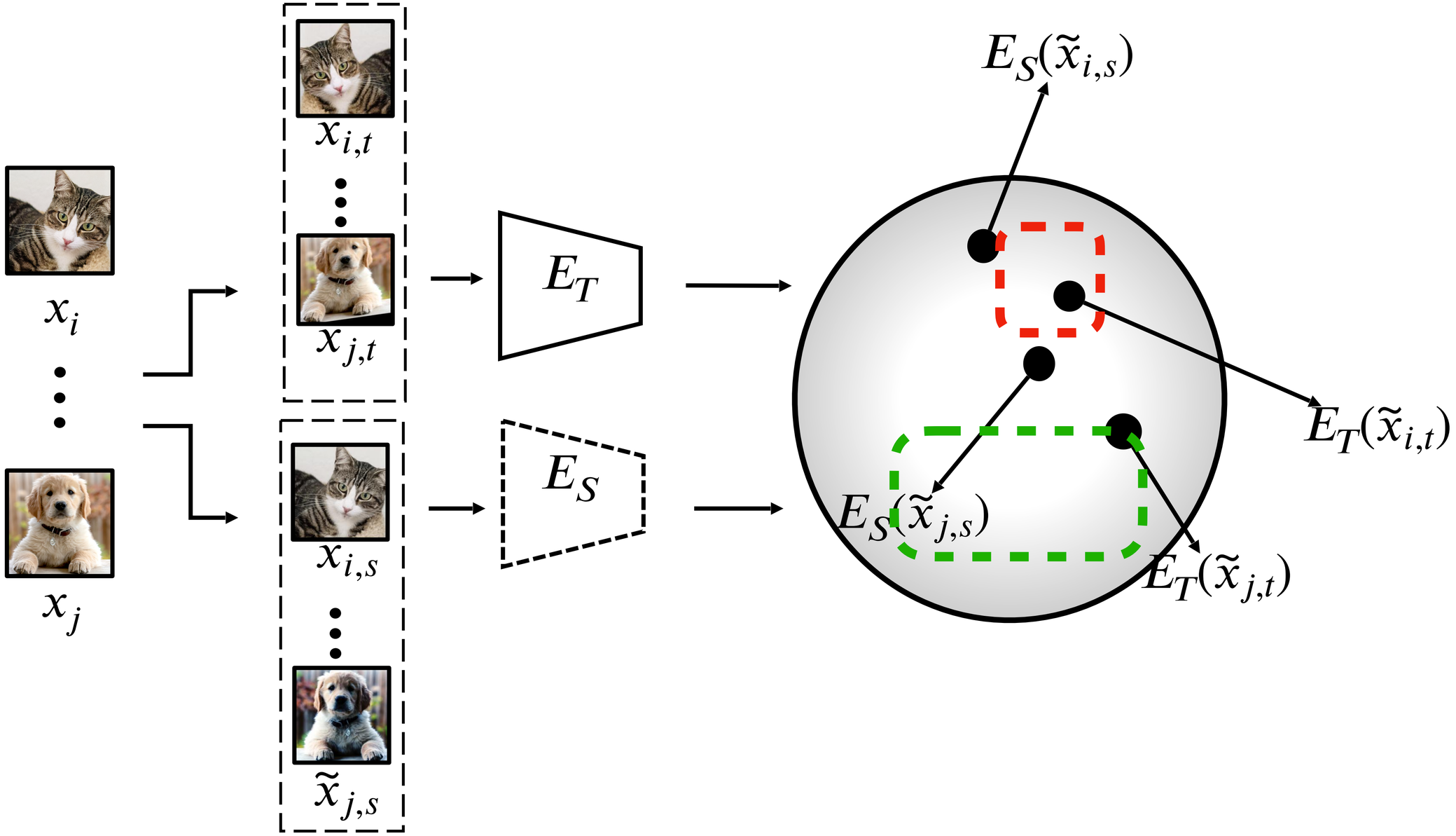
- **Knowledge About Target Model:** Only black-box access**.**
- **Knowledge About Train Data Distribution:** Two cases: (1) we assume the adversary has the same training dataset as the target encoder. (2) we assume that the adversary has totally no information about the target encoder's training dataset.

Conventional Stealing Attacks

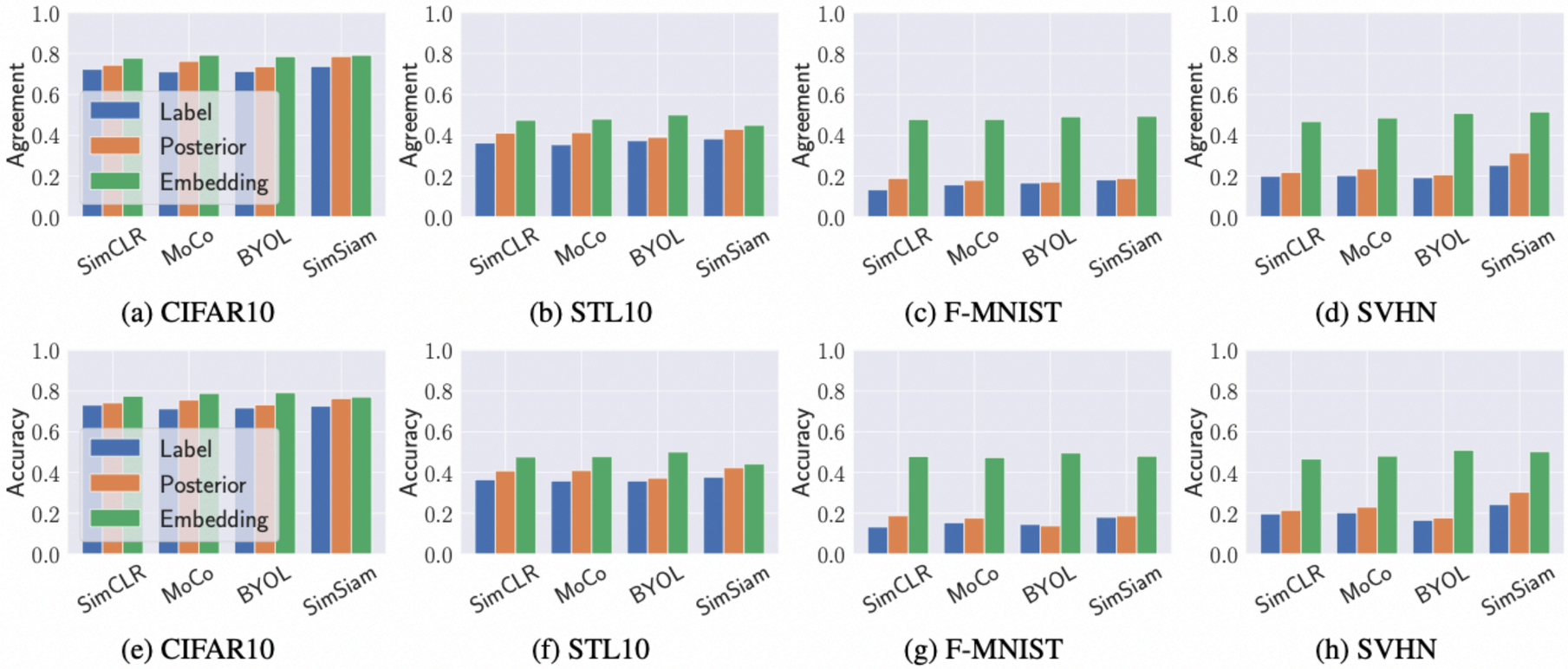


$$L_{MS} = \sum_{k=1}^{N} l(E_T(x_k), E_S(x_k))$$

Cont-Steal Attacks



Performance of Conventional Attacks

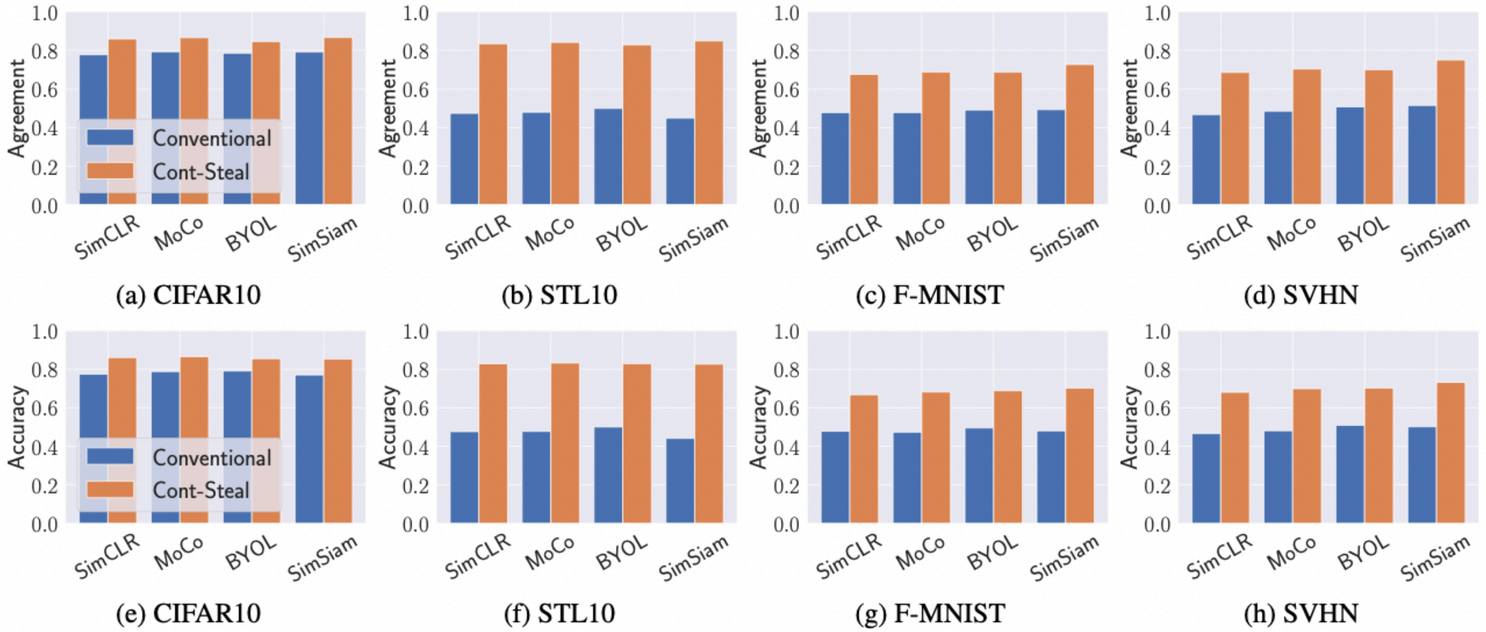


(a) CIFAR10 (b) STL10 (c) F-MNIST (d) SVHN

(e) CIFAR10 (f) STL10 (g) F-MNIST (h) SVHN

Image encoders are more vulnerable to model stealing attacks then traditional classifier

Performance of Cont-Steal Attacks



(a) CIFAR10 (b) STL10 (c) F-MNIST (d) SVHN

(e) CIFAR10 (f) STL10 (g) F-MNIST (h) SVHN

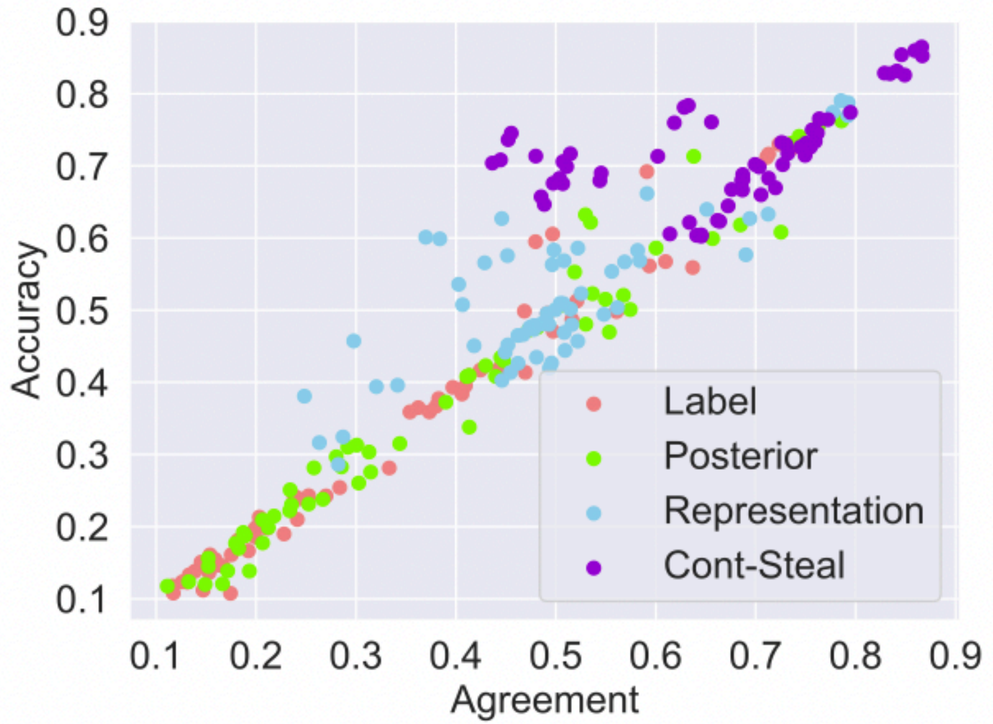
(a) Target encoder (b) Cont-Steal (c) Conventional

Cont-Steal can achieve much better performance then conventional steal attacks

Cost Analysis

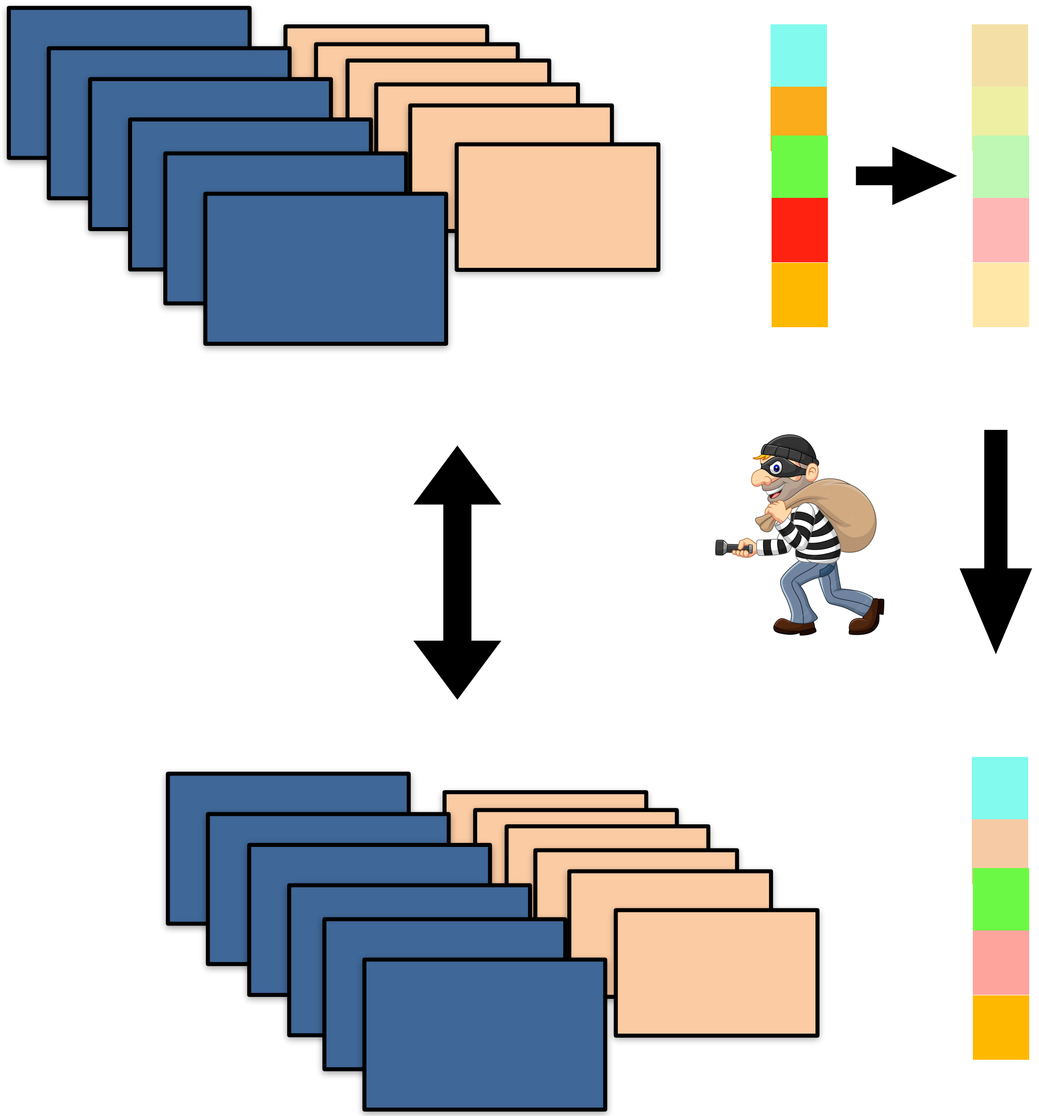
	Monetary Cost		Time Cost	
Model	**Normal ($)**	**Cont-Steal ($)**	**Normal (h)**	**Cont-Steal (h)**
SimCLR	58.68	11.83 (1.83 + 10)	20.01	0.62
MoCo	54.83	12.13 (2.13 + 10)	18.69	0.73
BYOL	61.46	12.08 (2.08 + 10)	20.96	0.71
SimSiam	57.14	12.00 (2.00 + 10)	19.46	0.68

Relationship between Accuracy and Agreement

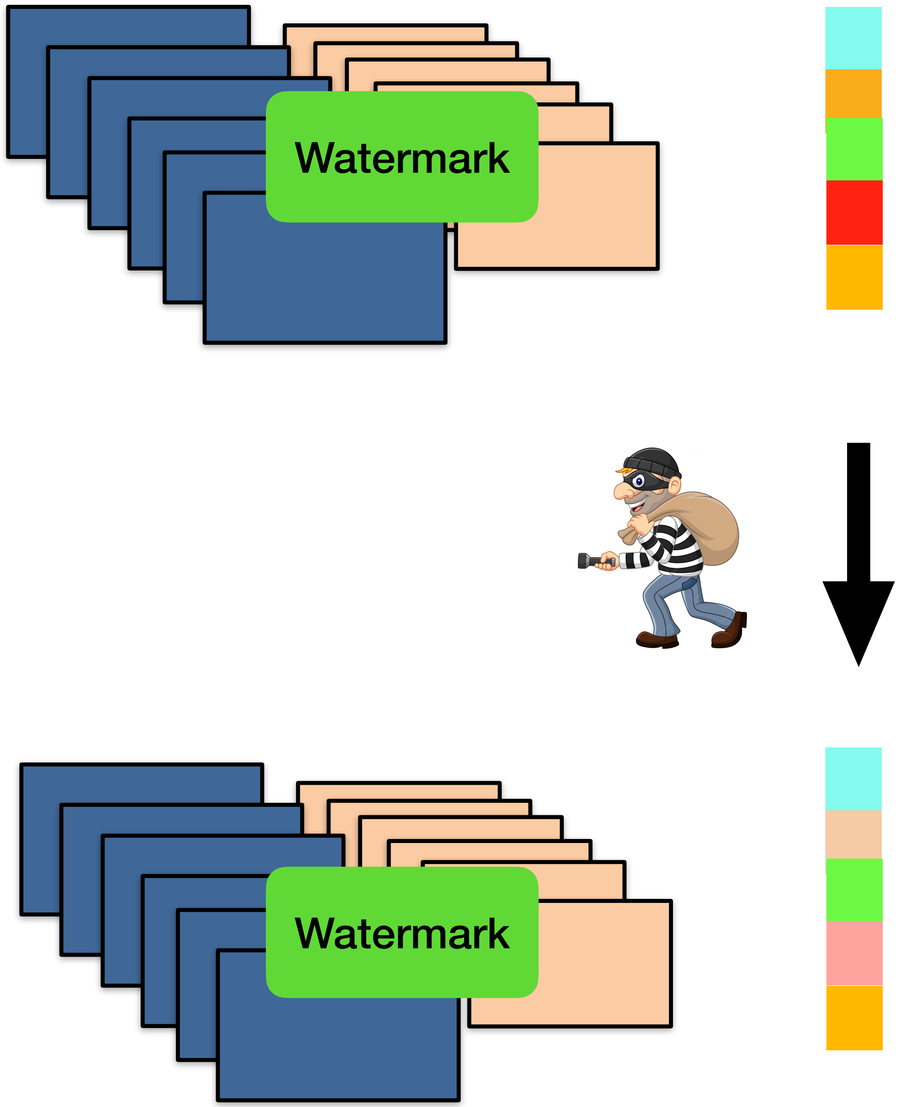


Defense

Perturbation-based Defense

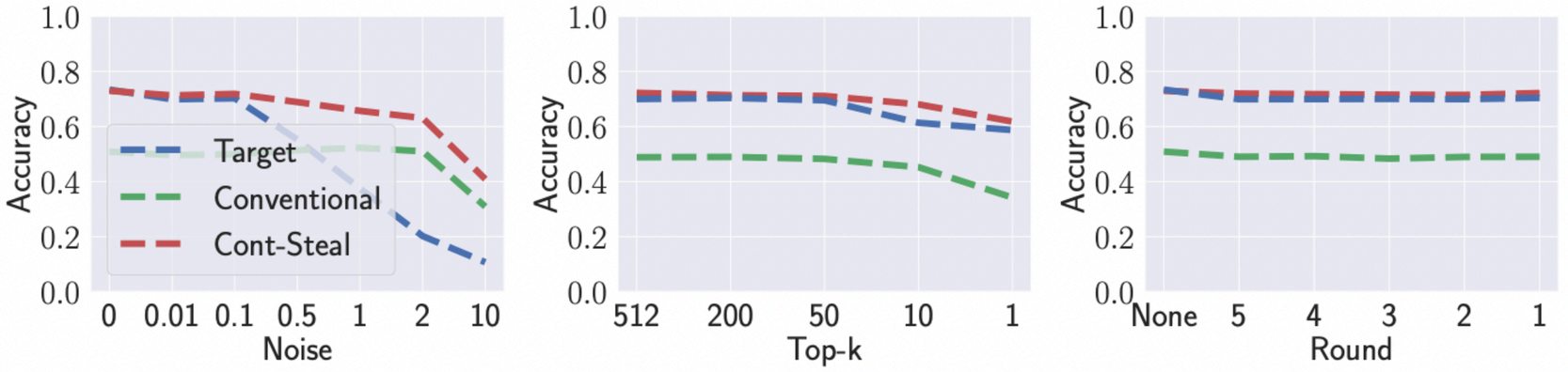


Watermark-based Defense



Defense

Perturbation-based Defense



(a) Adding noise (b) Top-k (c) Rounding

Watermark-based Defense

Dataset	Target model (acc/wr)	Cont-Steal (acc/wr)	Baseline (acc/wr)
CIFAR10	0.864 / 0.998	0.769 / 0.130	0.871 / 0.095
STL10	0.721 / 0.999	0.702 / 0.034	0.733 / 0.111
SVHN	0.501 / 0.999	0.535 / 0.303	0.492 / 0.103
F-MNIST	0.857 / 0.999	0.813 / 0.061	0.850 / 0.099

Conclusion

In summary, we make the following contributions:

(1)We pioneer the investigation of the vulnerability of unsupervised image encoders against model stealing attacks. We discover that encoders are more vulnerable than classifiers;

(2) We propose Cont-Steal, the first contrastive learning-based stealing attack against encoders that outperforms the conventional attacks to a large extent;

(3) Extensive evaluation shows that the advantageous performance of Cont-Steal is consistently amplified in various settings, especially when the adversary suffers from zero information of the target dataset, limited amount of data, or restricted query budgets.