CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY
salesforce

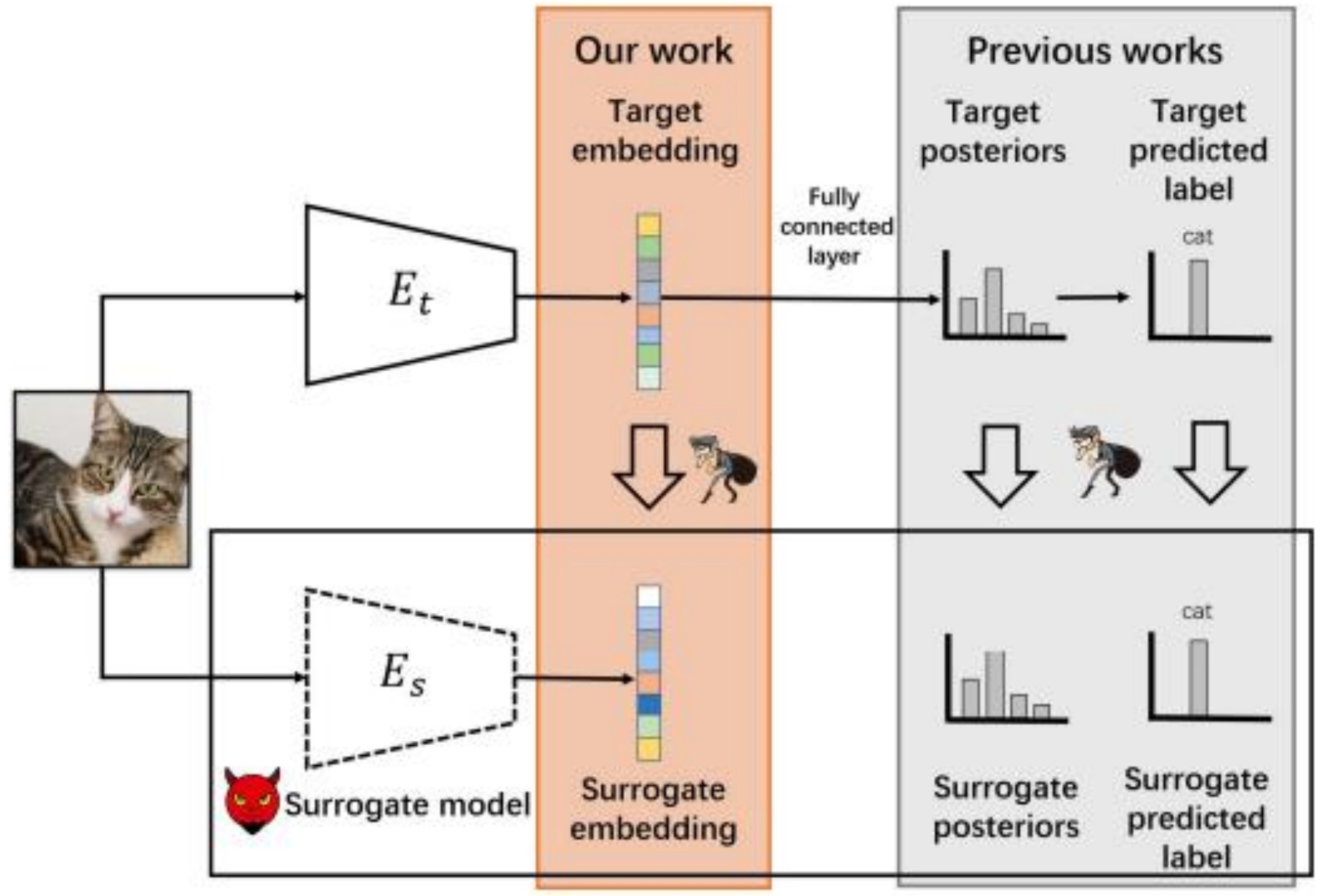
Can’t Steal? Cont-Steal! Contrastive Stealing Attacks Against Image Encoders

Zeyang Sha[1] Xinlei He[1] Ning Yu[2] Michael Backes[1] Yang Zhang[1]

[1] CISPA Helmholtz Center for Information Security [2] Salesforce Research

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Motivation



Model stealing attacks aims at stealing the **functionality of the target model**. So far, model stealing attacks concentrate on the supervised classifiers, i.e., the model responses are **prediction posteriors or labels** for a specific downstream task. The vulnerability of self-supervised image encoders **is unexplored**.

Threat Model

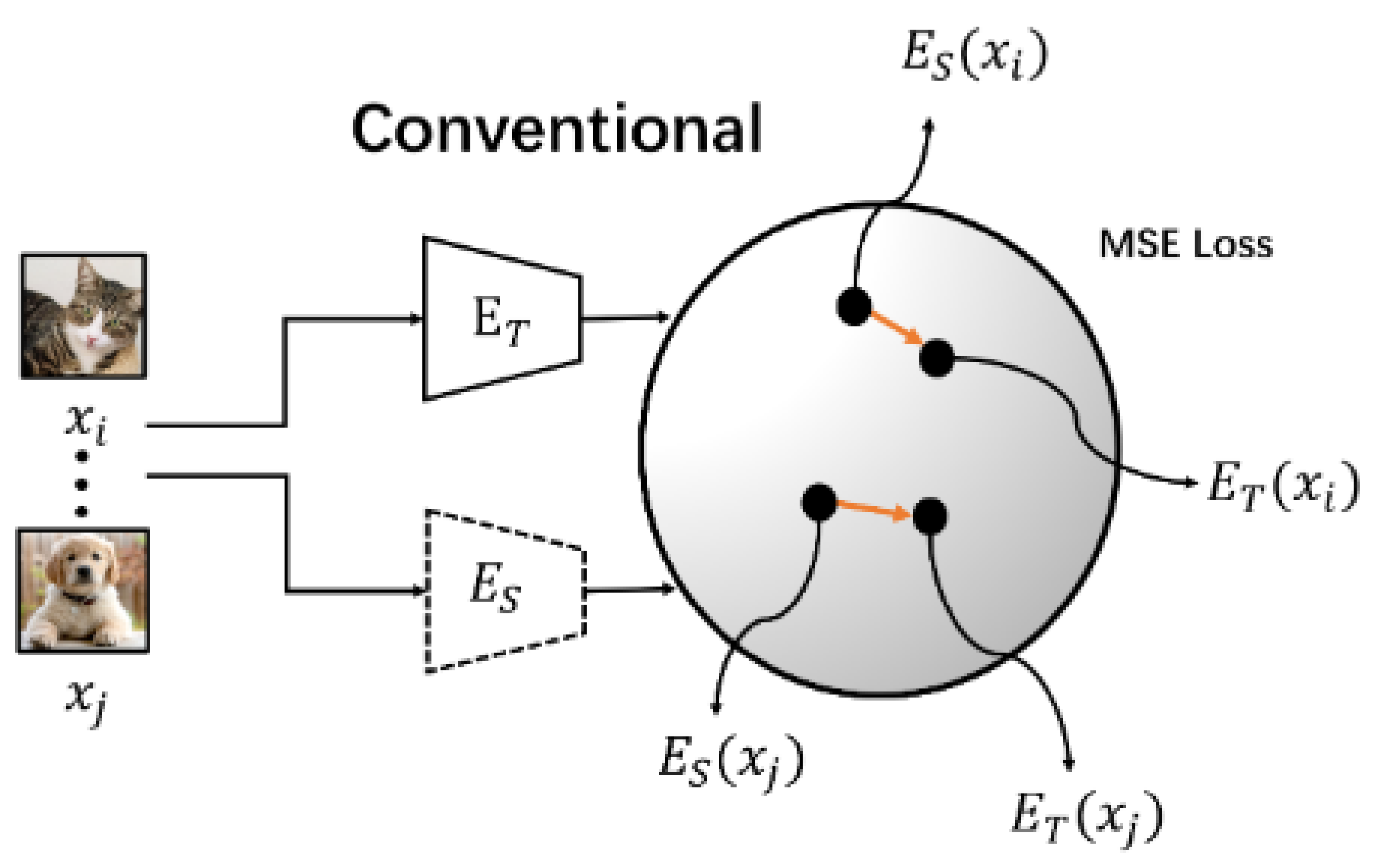
Adversary’s Goal:

- **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.

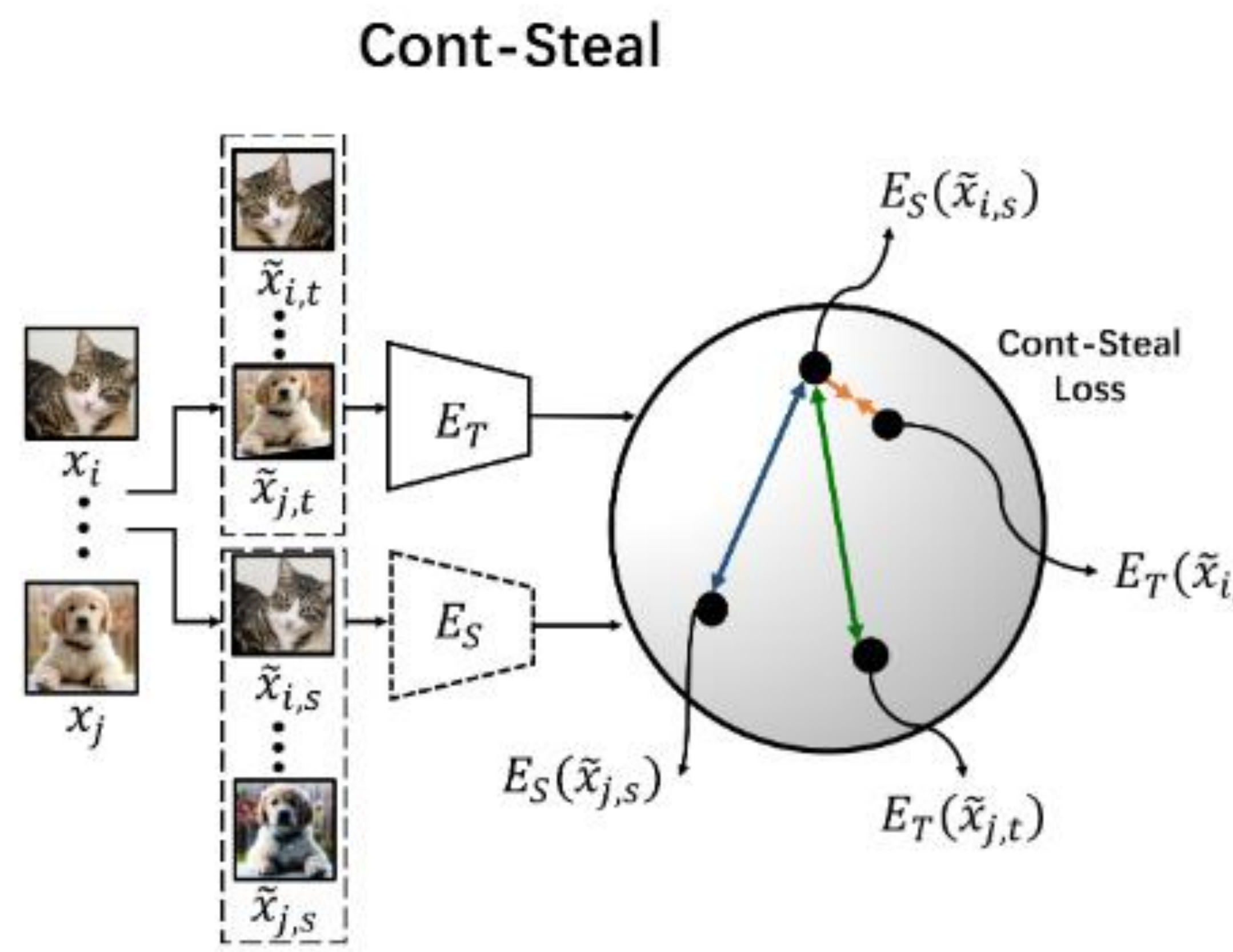
Method



Conventional attacks will use MSE loss to optimize the surrogate encoder. The loss function can be formulated as:

$$L_{MS} = \sum_{k=1}^{N} \mathrm{l}\left(\bar{E}_T\left(1^k\right)/E_S\left(x_k\right)\right).$$

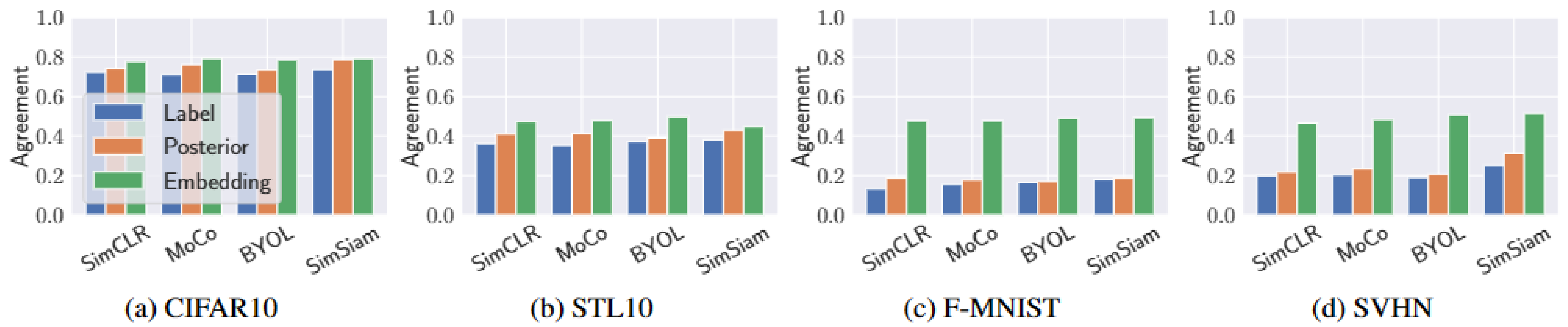
The loss function will make different embeddings of same images closer to each other. In this way, the surrogate model will behave similarly as target model.



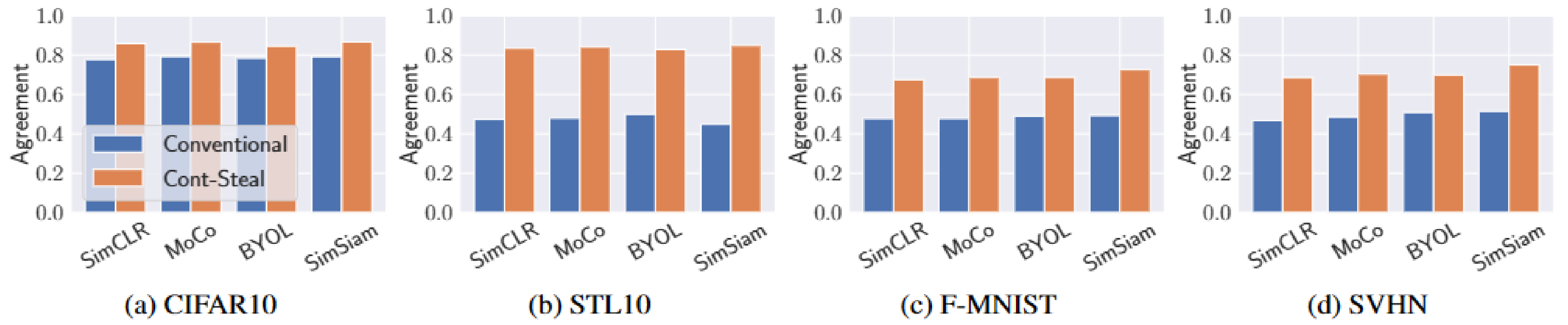
Cont-Steal attacks will first generate different views of given images using RandAug. Then, the cont-steal loss will try to enforce the surrogate embedding of an image close to its target embedding (defined as a positive pair) and also push away embeddings of different images irrespective of being generated by the target or the surrogate encoders (defined as negative pairs).

Apply The Surrogate Encoder to Downstream Tasks: To evaluate the effectiveness of model stealing attacks against the encoder, the adversary can leverage the same downstream task to both the target and surrogate encoders. Then, the adversary quantifies the attack effectiveness by measuring the performance of the target/surrogate classifier on the downstream tasks.

Results



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



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

Our results demonstrate that the conventional attacks **are more effective against encoders** than against downstream classifiers. Cont-Steal **outperforms** the conventional model stealing attacks to a large extent. Also, when the surrogate dataset is totally different from target dataset, Cont-Steal can have better performance.

Conclusion

We introduce the first model stealing attacks against image encoders:

- We pioneer the investigation of the vulnerability of unsupervised image encoders against model stealing attacks. We discover that encoders are more vulnerable than classifiers;
- We propose Cont-Steal, the first contrastive learning-based stealing attack against encoders that outperforms the conventional attacks to a large extent;
- 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.