Supplementary Material

A Dataset Details

The CIFAR-10 dataset contains images of resolution 32 × 32. The training set contains 50,000 images (5,000 from each class), and the test set contains 10,000 images. The Indoor-67 dataset contains 15,620 images, with a minimum resolution of 200 pixels in the smaller axis. The dataset is split into 14,280 training and 1,340 test images. Caltech-256 contains images of varying sizes spanning 256 object classes. Since no official train-test split has been reported, we keep aside 25 images from each class for testing, resulting in a training set of 23,380 images and test set of 6,400 images. Figure S1 shows sample images from each of the three victim datasets.

Figure S1: Sample images from victim datasets.

B Additional Results

B.1 Same Victim and Thief Architectures

We had assumed so far that the thief does not have knowledge of the victim model's architecture, and therefore uses a large pre-trained model that is publicly available. Another scenario usually considered in the model stealing literature [28, 29] is when both victim and thief model are based on the same neural network architecture. We study this scenario in this section, where the thief has knowledge of the victim's architecture, and therefore, fine-tunes from the same pre-trained model as the victim. We report the accuracies of the victim and thief models, along with the agreement between victim and thief in Table S1, and plot the agreements in Figure S2 for better visualization. In addition to the seven victim models considered in Section 5.1, we include two self-supervised foundation models: DINO [5] and CLIP [33], pre-trained on the ImageNet-1K and LAION-2B datasets respectively.

We observe that for the high capacity ViT models with stronger pre-training, both accuracy and agreement of the thief models increase. Moreover, self-supervised foundation models (CLIP and DINO) are also more vulnerable to model stealing compared to ResNets. This reiterates our finding from Section 5.1 that even though stronger foundation models increase the accuracy of victim models, it increases the risk of model stealing too.

	Victim arch	ResNet-18	ResNet-34	ResNet-50	ResNet-101	ViT-S/16	ViT-B/16	ViT-B/16 DINO	ViT-B/16 CLIP	ViT-L/16
	Params (M)	11.18	21.29	23.53	42.52	21.66	85.80	85.80	85.80	303.31
	Pretraining dataset	IN-1K	IN-1K	IN-1K	IN-1K	IN-21K	IN-21K	IN-1K	LAION-2B	IN-21K
CIFAR	Victim accuracy	80.74	81.96	80.24	84.64	86.38	86.93	95.06	96.26	97.61
	Thief accuracy	67.02	72.12	68.15	77.97	90.70	86.42	79.63	89.79	96.55
	Thief agreement	69.42	72.42	65.53	75.51	84.63	80.96	80.96	90.74	94.68
Indoor	Victim accuracy	54.70	58.06	45.97	51.49	78.51	82.76	80.22	83.58	87.39
	Thief accuracy	28.06	31.27	22.09	26.04	57.46	49.70	52.84	63.58	57.39
	Thief agreement	41.94	45.22	42.69	39.79	58.73	51.49	57.91	69.70	58.51
Caltech	Victim accuracy	67.75	74.36	58.47	74.78	84.78	87.67	87.66	91.98	94.13
	Thief accuracy	18.83	37.17	14.88	28.61	63.02	63.23	59.78	68.30	61.30
	Thief agreement	25.27	44.89	34.56	38.25	61.23	61.14	62.19	71.39	60.95

Table S1: Model stealing results when thief model architecture is same as victim model architecture. Both victim and thief models are linear probed. IN stands for ImageNet.

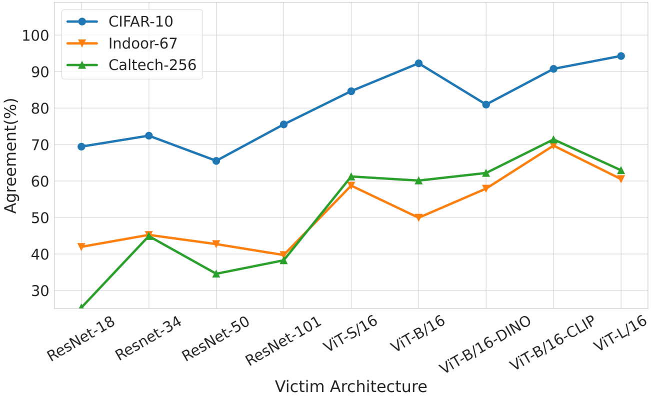

Figure S2: Agreement between victim and thief models. Both victim and thief share the same architecture.

C Extensions of Existing Results

C.1 Ablation study from Main Paper

We provide here detailed results to support the ablation studies reported in Section 5.4. Figure S3 and Figure S4 show the impact of varying the query budget and sample selection technique respectively, for victims trained on the Indoor-67 dataset using the linear probing method. The thief is a ViT-B/16 model (chosen because of faster training time) and is trained using linear probing. We observe a similar trend in both cases: the ViT models with their rich feature representations are stolen with higher agreements compared to ResNet models.

C.2 Qualitative Results

We extend the qualitative results reported in Section 5.3 for the CIFAR-10 dataset to other datasets. Figure S5 and Figure S6 show the t-SNE visualizations for the Indoor-67 and Caltech-256 datasets, respectively. The ViT backbones are able to form well-separated clusters on both the datasets, leading to better performance of the respective victim models, and also easier stealing.

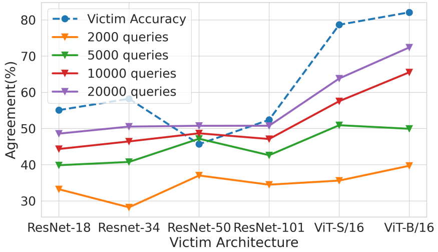

Figure S3: Impact of varying the query budget, for 'random' sample selection. Indoor-67 dataset, ViT-B/16 thief.

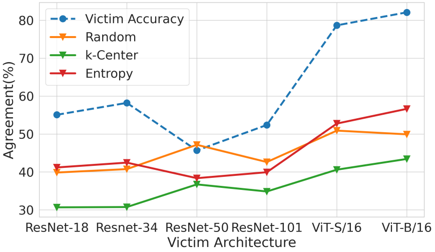

Figure S4: Impact of varying sample selection method, for a query budget of 5000. Indoor-67 dataset, ViT-B/16 thief.

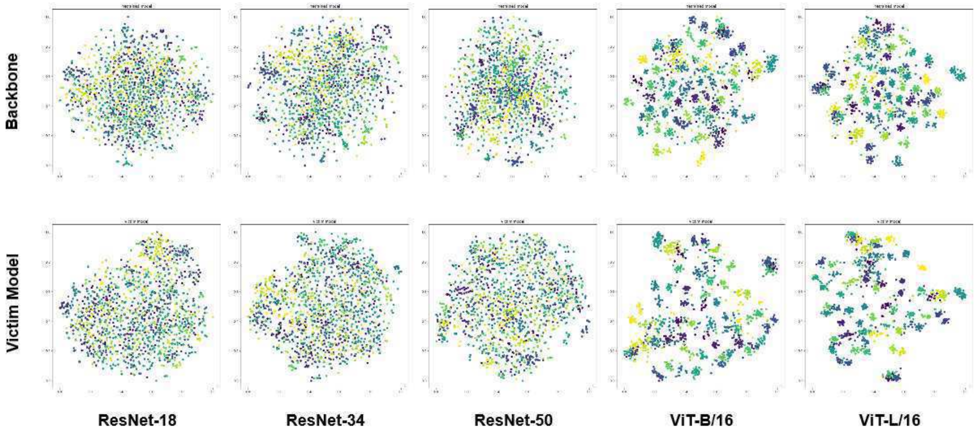

Figure S5: t-SNE visualizations of embeddings for backbone models (top row), and corresponding victim models (bottom row) trained on Indoor-67 dataset, for various model architectures.

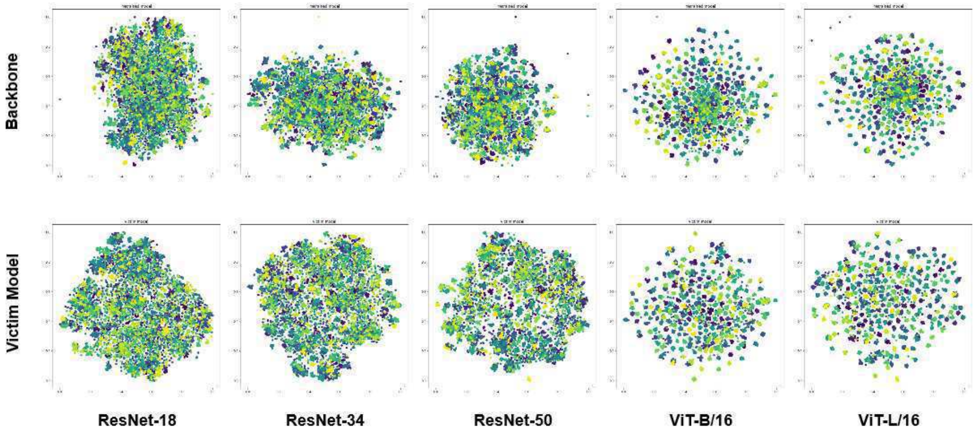

Figure S6: t-SNE visualizations of embeddings for backbone models (top row), and corresponding victim models (bottom row) trained on Caltech-256 dataset, for various model architectures.