Supplementary Material

Table S1. Model stealing performance for COVID-19 classification task. We report total accuracy (Total), class-wise accuracies for all 3 classes, and agreement (Agr.). Query budget is 5000. Proposed method achieves thief accuracy close to the baselines, while having the best agreement value.

Arch	Method	Total	COVID-19	Pneumonia	Regular	Agr.
Victim	-	89.91	83.43	95.40	92.24	-
ResNet-50[13]	Random [21]	65.97	40.13	74.71	80.17	70.59
	k-Center [23]	65.55	57.32	68.97	69.83	68.49
	Random+QW	63.87	33.12	72.41	81.47	71.22

Table S2. Model extraction performance on general vision tasks for natural images with 5000 queries. The proposed method is implemented with two different anchor models: Random+QW and k-Center+QW. The best method for each dataset is depicted in **bold**, and the next best is underlined. Proposed method outperforms the baselines in terms of both accuracy and agreement for all datasets. Note: Dual Students[7] is a data-free method that uses synthetically generated data instead of a proxy dataset, but requires millions of queries. We implement [7] with a budget of 500K queries for the smaller datasets, yet it fails to match the performance of the other methods operating at 5000 queries.

Method	Venue	MNIST		SVHN		CIFAR10		Caltech256		CUBS200		Indoor67	
		Acc	Agr	Acc	Agr	Acc	Agr	Acc	Agr	Acc	Agr	Acc	Agr
Random[21]	CVPR'19	80.55	80.59	67.54	67.84	65.89	66.66	39.77	40.32	14.74	15.75	33.36	36.22
Entropy[23]	AAAI'20	80.55	80.59	41.02	41.11	47.53	48.16	38.88	39.78	13.87	15.06	35.82	39.33
k-Center[23]	AAAI'20	71.92	71.99	72.78	73.25	68.02	68.71	44.66	45.16	18.57	20.14	40.37	42.91
BBD + Random [30]	ECCV'22	27.49	27.51	58.87	59.04	47.04	47.59	40.56	41.16	16.00	16.64	34.40	38.06
BBD + k-Center [30]	ECCV'22	58.53	58.52	43.99	44.18	40.59	40.58	41.72	42.05	16.48	17.41	30.07	33.88
Dual Students [7]	ICLR'23	18.62	19.24	6.69	10.89	12.86	10.16	-	-	-	-	-	-
Random + QW		80.00	80.08	70.83	71.20	73.07	73.62	45.19	45.36	14.67	15.43	36.12	38.63
k-Center + QW		**85.08**	**86.01**	**76.58**	**76.92**	**74.85**	**74.78**	**50.48**	**50.00**	**20.21**	**21.32**	**42.24**	**43.73**

Table S3. Thief model accuracy under SOTA model stealing defenses. We evaluate three MS attacks on GBC malignancy classification victim model, under three defense techniques. Note that the RadFormer victim model is non-differentiable, rendering it infeasible for the defenses to compute gradients. Hence, for this experiment, we use a differentiable version of RadFormer, containing only the global branch. As can be observed, there is no significant impact (lowering of thief accuracy) that is consistent across all MS attacks. The paper advocates more research in this topic to prevent stealing of proprietary information through this route of MS attacks.

Method	No Defense	MAD [22]	AM [16]	GRAD^2 [18]
Victim model	89.34	80.32	88.52	86.88
Random	75.40	78.68	62.29	69.67
Entropy [23]	74.59	64.75	65.57	75.40
k-Center [23]	70.49	72.13	72.95	79.50

Table S4. Training hyperparameters for anchor and student models, corresponding to the two victim models. B_l and B_u are mini-batch sizes for labeled and unlabeled data respectively. Input image pre-processing for ViT, DeiT and Inception-v3 includes random horizontal flip and random augmentation; for ResNet-50 includes random crop, jitter, and random horizontal flip. For student model training, we use cosine learning rate decay with warmup.

		GBC				COVID-19
		ResNet50	Inception-v3	ViT	DeiT	ResNet50
Anchor training	learning rate	0.005	0.005	0.005	0.005	0.01
	momentum	0.9	0.9	0.9	0.9	0.9
	epochs	100	100	100	100	100
	batch size	16	16	16	16	128
	weight decay	0.0005	0.0005	0.0005	0.0005	0.0005
Student training	learning rate	0.02	0.01	0.01	0.01	0.02
	momentum	0.9	0.9	0.9	0.9	0.9
	weight decay	0.0005	0.0005	0.0005	0.005	0.005
	epochs	100	100	100	100	100
	warmup epochs	10	10	10	10	10
	B_l	16	16	16	16	16
	B_u	112	112	48	48	112