Supplementary Material. ISD: Self-Supervised Learning by Iterative Similarity Distillation

Transfer evaluation training details: we freeze the backbone and forward train set images without augmentation (resize shorter side to 256, take a center crop of size 224, and normalize with ImageNet statistics). Then we train a linear layer on top of extracted features. We split each dataset to train, validation, and test set. We search for best lr in 10 log spaced values between -3 and 0 and weight decay in 9 log spaced values between -10 and -2, then we train linear layer with best parameters on train+validation set and evaluate it on test set.

Dataset	Classes	Train samples	Val samples	Test samples	Accuracy measure	Test provided
Food101 [6]	101	68175	7575	25250	Top-1 accuracy	-
CIFAR-10 [25]	10	49500	500	10000	Top-1 accuracy	-
CIFAR-100 [25]	100	45000	5000	10000	Top-1 accuracy	-
Sun397 (split 1) [45]	397	15880	3970	19850	Top-1 accuracy	-
Cars [24]	196	6509	1635	8041	Top-1 accuracy	-
DTD (split 1) [13]	47	1880	1880	1880	Top-1 accuracy	Yes
Pets [33]	37	2940	740	3669	Mean per-class accuracy	-
Caltech-101 [15]	101	2550	510	6084	Mean per-class accuracy	-
Flowers [28]	102	1020	1020	6149	Mean per-class accuracy	Yes

Table A1: The train, val, and test split sizes for transfer datasets are listed above. **Test split:** For DTD and Flowers datasets, we use the provided test sets. Otherwise, in case of Sun397, Cars, CIFAR-10, CIFAR-100, Food101, and Pets datasets, the val set provided in the dataset is used as the hold-out test set. Also, for Caltech-101, the hold-out test set is created by randomly sampling 30 images/class from the train set. **Val split:** For DTD and Flowers datasets, we use the provided val sets. Otherwise, the val set is created by a randomly sampled subset of the train set is used as the val set. We report the strategy for splitting val sets for different datasets: 5 samples/class for Caltech-101, 20% samples/class for Cars, 50 samples/class for CIFAR-100, 50 samples/class for CIFAR-10. 75 samples/class for Food101, 20 samples/class for Pets, 10 samples/class for Sun397. We attempt to be as close to the details provided in BYOL [18] as possible.

	LR	m	Aug.	Proj.	τ_t	τ_s	NN	20-NN
1	step	0.999	s/s	✗	0.02	0.02	41.5	46.6
2	step	0.999	w/s	✗	0.02	0.02	41.7	46.7
3	step	0.99	s/s	✗	0.02	0.02	40.2	45.2
4	cosine	0.999	s/s	✗	0.02	0.02	40.9	45.7
5	cosine	0.99	w/s	✓	0.02	0.02	34.6	38.3
6	cosine	0.99	w/s	✓	0.02	0.2	39.4	44.4
7	cosine	0.99	w/s	✓	0.01	0.1	31.7	37.3
8	step	0.999	w/s	✗	0.02	0.2	6.0	8.0
9	step	0.999	w/s	✗	0.02	0.5	5.5	7.0
10	step	0.999	w/s	✗	0.03	0.3	2.9	3.9

Table A2: We explore various hyper-parameters for ResNet-18 model. "s/s" refers to the setting where both views use strong augmentation while in "w/s" the teacher view uses weak augmentation while the student view uses strong augmentation. The projection layer is a 2-layer MLP with 1024 as hidden dim and 128 as the output dim. There is a BatchNorm layer followed by ReLU between the two layers. In step learning rate decay, the LR is reduced by a factor of 0.2 at 140 and 180 epochs. The training happens for 200 epochs. The 1st row uses the same setting as the ResNet-18 model in the main paper. The 6th row uses the same settings as the ResNet-50 model in the main paper.

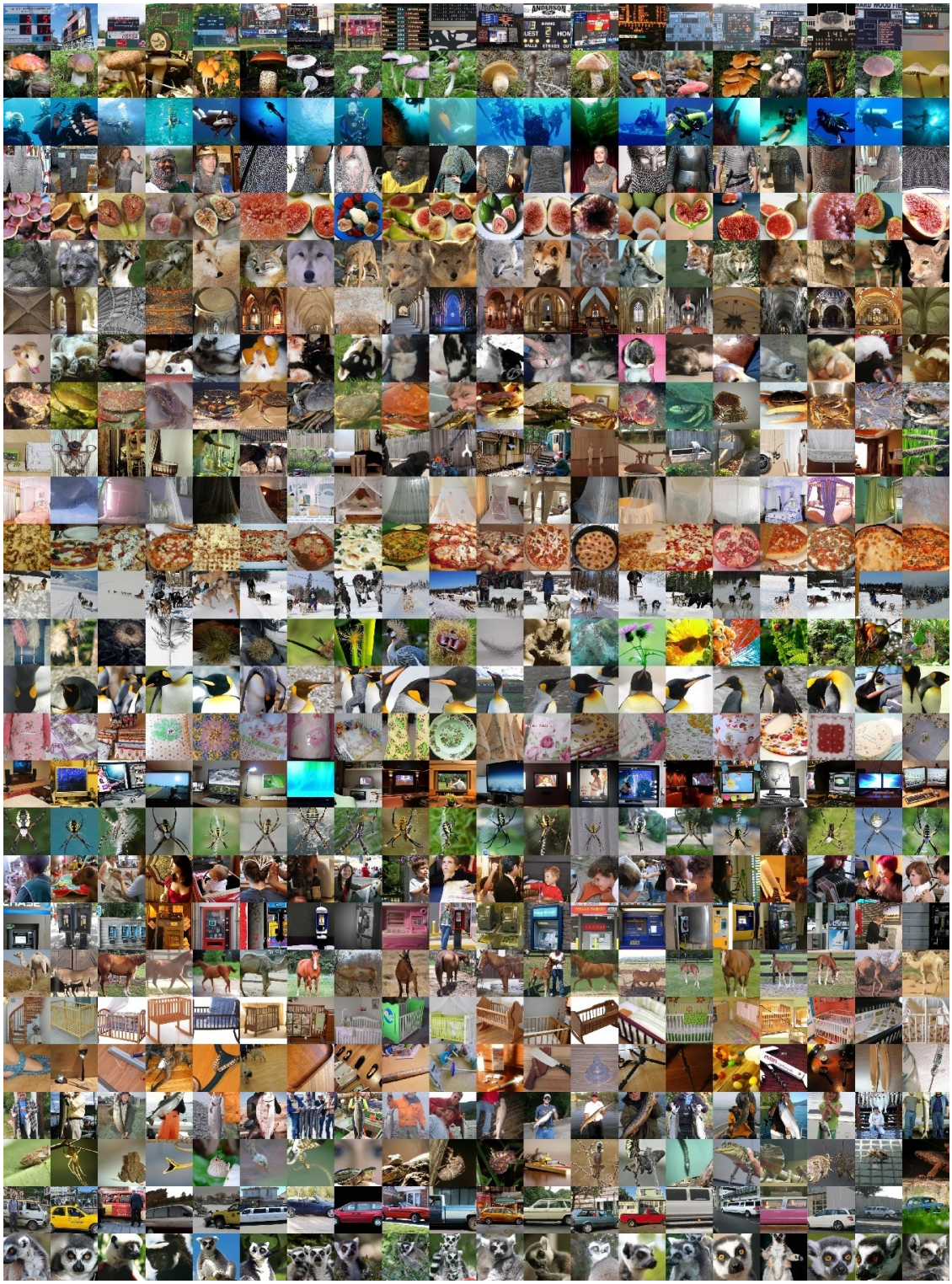

Figure A1: **Random Clusters:** We cluster ImageNet dataset into 1000 clusters using k-means and show random samples from random clusters. We have no cherry-picking for this visualization. Interestingly, images from each row(each cluster) are semantically similar.