Supplementary Material: Mean Shift for Self-Supervised Learning

Trasfer evaluation training details: we use the pre-trained models as frozen feature extractors and train a single linear layer on top of the features. The following pre-processing is applied to all images: resize shorter side to 256, take a center crop of size 224, and normalize with ImageNet statistics. No training time augmentations are used. We use the LBFGS optimizer with options: `max_iter=20` and `history_size=10`. We perform a grid search for learning rate and weight decay on the val set, and retrain on the combined train+val set with the best hyperparameters. learning rate is searched in a set of 10 log spaced values between -3 and 0, and weight decay is searched in a set of 9 log spaced values between -10 and -2.

Dataset	Classes	Train samples	Val samples	Test samples	Accuracy measure	Test provided
Food101 [8]	101	68175	7575	25250	Top-1 accuracy	-
CIFAR-10 [29]	10	49500	500	10000	Top-1 accuracy	-
CIFAR-100 [29]	100	45000	5000	10000	Top-1 accuracy	-
Sun397 (split 1) [48]	397	15880	3970	19850	Top-1 accuracy	-
Cars [28]	196	6509	1635	8041	Top-1 accuracy	-
Aircraft [30]	100	5367	1300	3333	Mean per-class accuracy	Yes
DTD (split 1) [15]	47	1880	1880	1880	Top-1 accuracy	Yes
Pets [36]	37	2940	740	3669	Mean per-class accuracy	-
Caltech-101 [20]	101	2550	510	6084	Mean per-class accuracy	-
Flowers [32]	102	1020	1020	6149	Mean per-class accuracy	Yes

Table A1: We list the sizes of train, val, and test splits of the transfer datasets. **Test split:** We use the provided test sets for Aircraft, DTD, and Flowers datasets. In case of Sun397, Cars, CIFAR-10, CIFAR-100, Food101, and Pets datasets, we use the provided val set as the hold-out test set. In case of Caltech-101, we use a random split of 30 images per category as the hold-out test set. **Val split:** We use the provided val sets for the datasets DTD and Flowers. For all other datasets, the val set is created by randomly sampling a subset of the train set. In order to be as close to BYOL [22] transfer setup as possible, we use the following val set splitting strategies for each dataset. Aircraft: 20% samples/class. Caltech-101: 5 samples/class. Cars: 20% samples/class. CIFAR-100: 50 samples/class. CIFAR-10: 50 samples/class. Food101: 75 samples/class. Pets: 20 samples/class. Sun397: 10 samples/class.

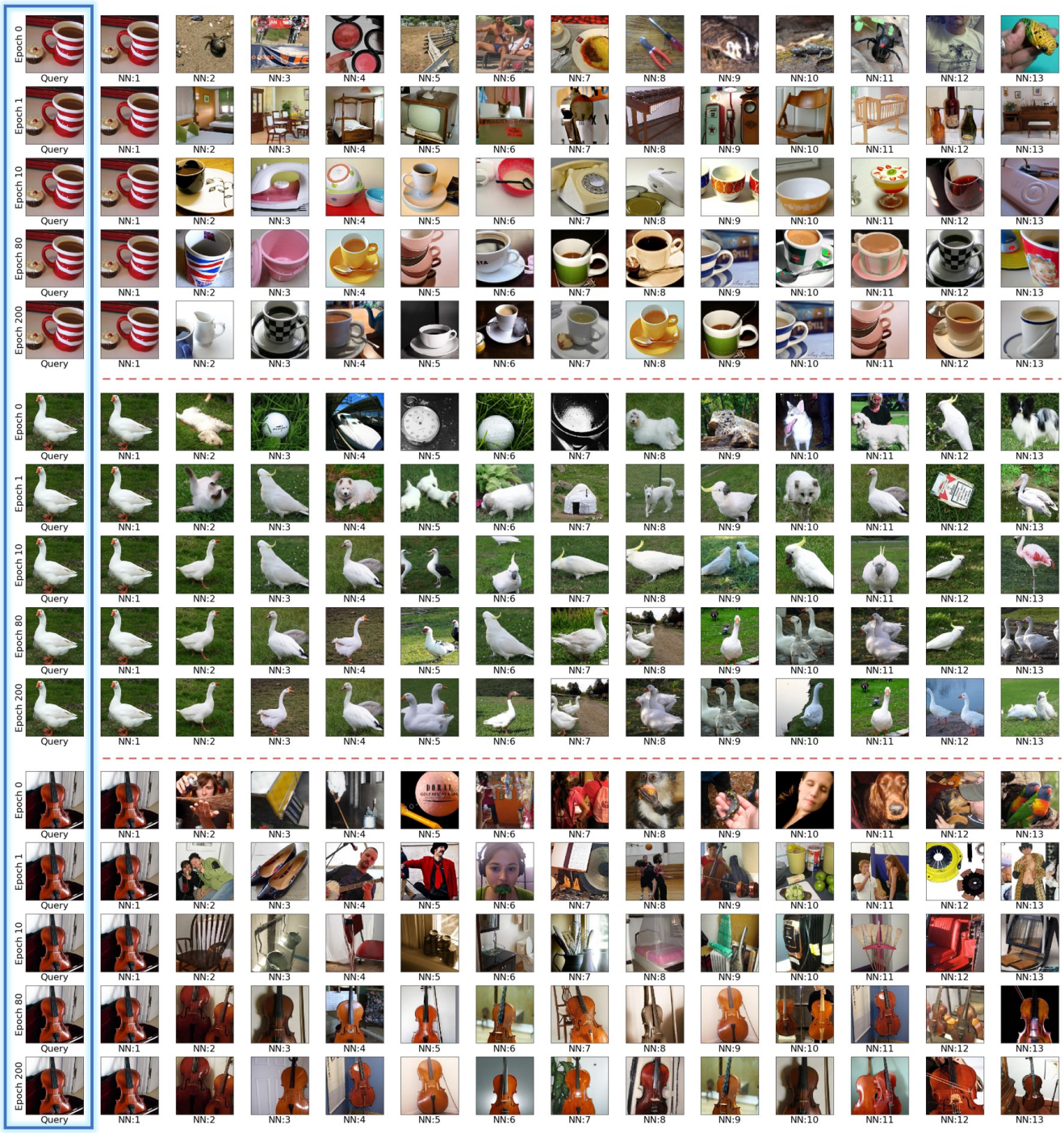

Figure A1: **Nearest neighbors (NN) of the model at each epoch:** Similar to Figure 2.

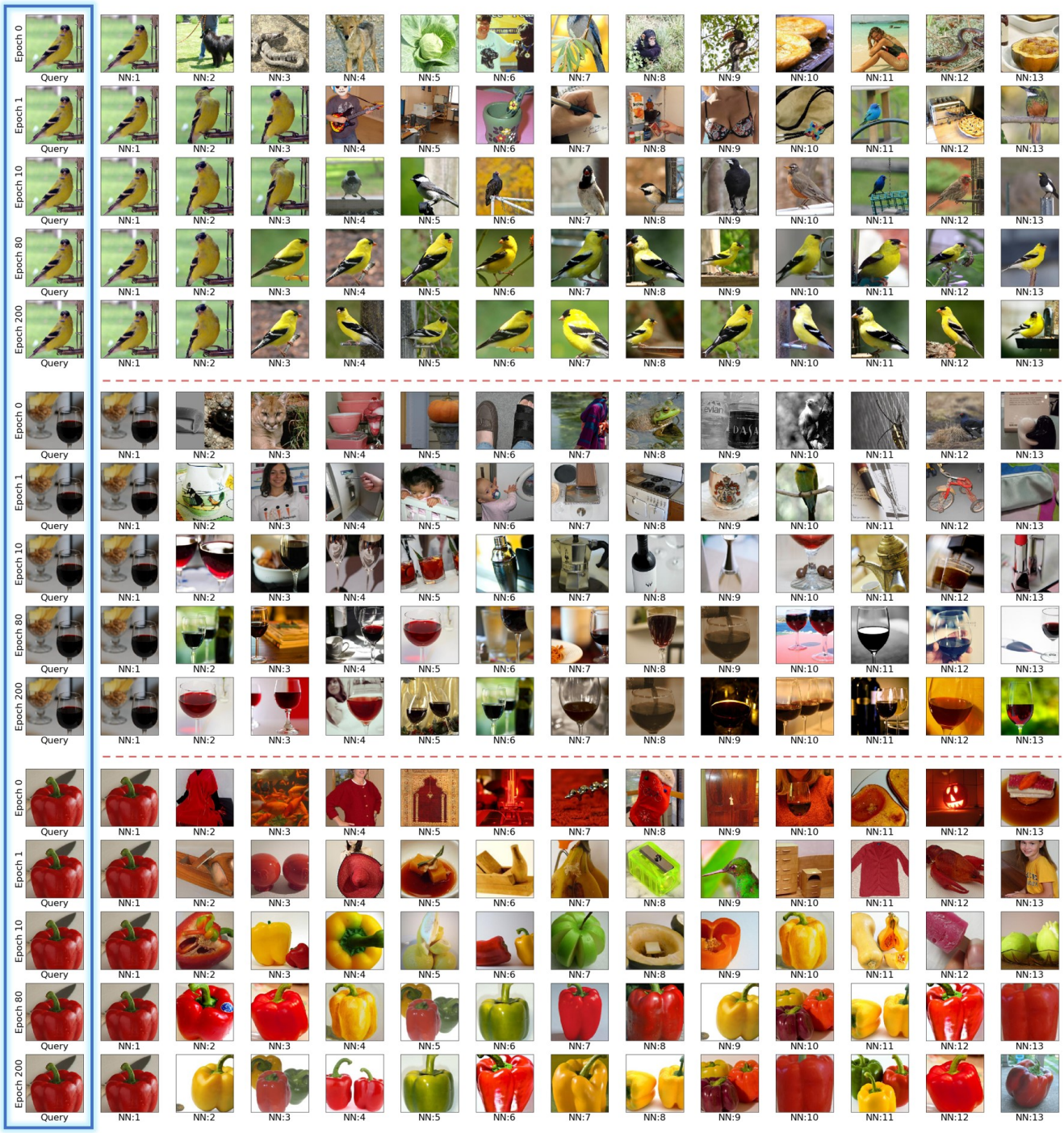

Figure A2: **Nearest neighbors (NN) of the model at each epoch:** Similar to Figure 2.

Figure A3: **Random Clusters:** We forward ImageNet training set through our ResNet-50 model and cluster them into 1000 clusters using k-means. We select 30 clusters randomly and show 20 randomly sampled images from each cluster without cherry-picking. Each row corresponds to a cluster. Note that semantically similar images are clustered together.