

Supplemental Material

S1 Parameter Settings

S1.1 Pre-Training Approaches

We provide parameter settings for each pre-training approach below.

- **Supervised:** Following [29], systems were trained for 40 epochs with a minibatch size of 256, an initial learning rate of 0.1 with a decrease of $10\times$ every 15 epochs, momentum of 0.9, and weight decay of $1e-4$. Standard random crop and horizontal flips were used for augmentation, unless noted otherwise. The network with best performance on the validation set is picked.
- **MoCo-V2:** Following [16], we train MoCo-V2 models for 800 epochs using a learning rate of 0.015 and minibatch size of 128.
- **Barlow Twins:** Following [69], we train Barlow Twins models for 300 epochs with a minibatch size of 1024, learning rate of 0.2, and a trade-off parameter λ of 0.0051.
- **SwAV:** Following [9], SwAV models are trained for 400 epochs with an initial learning rate of 0.6, final learning rate of 0.0006, epsilon of 0.03, and minibatch size of 64. We set the number of prototypes to $10|\mathcal{Y}|$ in a queue of length 384.

S1.2 Offline Linear Evaluations

To linearly evaluate the pre-trained features on the ImageNet dataset, the classifier was trained for 100 epochs with a minibatch size of 256. For supervised features with SwAV and Barlow Twins, we used a learning rate of 0.1 which we decay by a factor of 10 at 60 and 80 epochs and L2 weight decay of $1e-5$. These settings did not work well for MoCo-V2, so we used the settings for linear evaluation from [32] instead, which were a learning rate of 30 that we decay by a factor of 10 at epochs 60 and 80 with no weight decay.

For linear evaluation on the Places-365 dataset, we trained the softmax classifier for 28 epochs with a minibatch size of 256. We used the same learning rates and L2 weight decays from the ImageNet linear evaluation and reduced the learning rate by $10\times$ at epochs 10 and 18.

S1.3 Online Continual Learning Methods

We provide parameter settings for each continual learning method below.

- **SLDA:** Following [27], we use a plastic covariance matrix and shrinkage of $1e-4$.
- **Online Softmax with Replay:** We use a learning rate of 0.1. The buffer contains a maximum of 735K feature vectors each with 512 dimensions (1.5 GB). This buffer size was chosen to match the size of the buffer for REMIND in GB.
- **REMIND:** We follow the parameter settings from [29], i.e., starting learning rate of 0.1, 32 codebooks each of size 256, 50 randomly selected replay samples, a buffer size of 959,665 (equal to 1.5 GB), manifold mixup and random resize crop data augmentation, and extracting mid-level features such that two convolutional layers and the final classification layer remain plastic during online learning.

S2 Additional Results

S2.1 Relative Performance Improvements

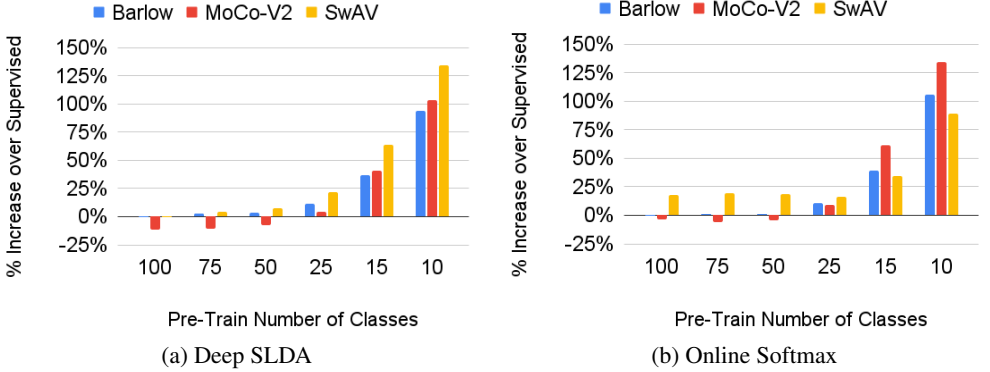


Figure S1: Relative performance improvements over the supervised baseline using MoCo-V2, Barlow Twins, and SwAV features for online continual learning with (a) Deep SLDA and (b) Online Softmax on ImageNet.

Fig. S1 shows the relative performance improvements exhibited by the Deep SLDA and Online Softmax methods when performing online continual learning on the ImageNet dataset using MoCo-V2, Barlow Twins, and SwAV features.

Deep SLDA (S1a) shows a maximum relative improvement of 134.31% for SwAV features, 93.91% for Barlow Twins features, and 103.25% for MoCo-V2 features when only 10 classes are used for pre-training. Deep SLDA follows the same trend as REMIND from Fig. 1, with SwAV outperforming Barlow Twins and MoCo-V2 for all pre-train sizes. MoCo-V2 shows negative relative improvement for 50, 75, and 100 pre-training classes, while SwAV and Barlow Twins show small negative relative improvements of -0.69% and -0.56% respectively for 100 classes. These few cases of negative relative performances occur when there are a large number of pre-training classes, which is less desirable for pre-training as it requires more data.

Online Softmax (S1b) shows a maximum relative improvement of 88.96% for SwAV features, 105.67% for Barlow Twins features, and 133.94% for MoCo-V2 features when only 10 classes are used for pre-training. Surprisingly, MoCo-V2 outperforms SwAV and Barlow Twins for 10 and 15 pre-train classes, which differs from the results for REMIND and Deep SLDA. We see a small negative relative improvement with MoCo-V2 for 50, 75 and 100 pre-train classes, while Barlow Twins shows a small negative relative improvement for 100 pre-train classes only. SwAV shows positive relative improvement for all pre-train sizes. Overall, these results demonstrate that self-supervised features are superior to supervised features for continual learning when less data is used during pre-training.

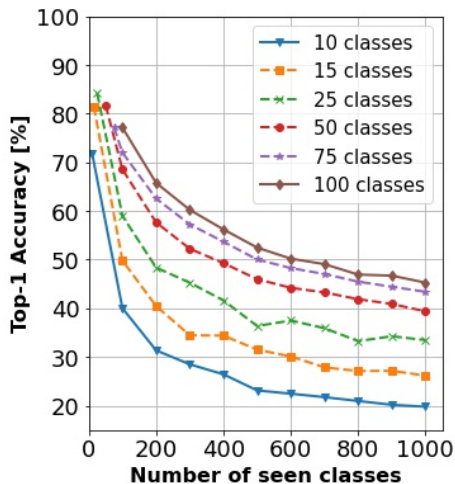
S2.2 Learning Curves

Learning curves for online continual learning on ImageNet using REMIND, Deep SLDA and Online Softmax with various features are in Fig. S2, Fig. S3 and Fig. S4, respectively. These curves show the top-1 accuracy every time a multiple of 100 classes has been seen by the model, including the performance right after pre-training.

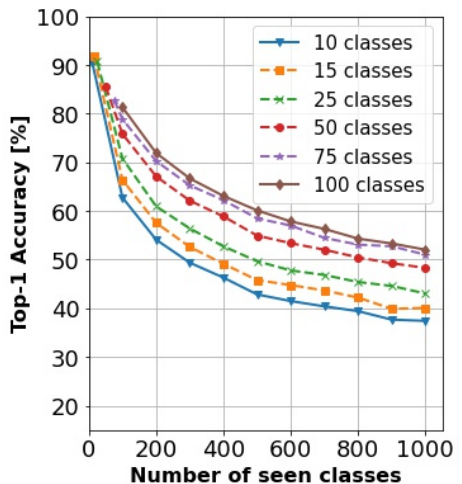
Learning curves for REMIND using supervised, MoCo-V2, Barlow Twins, and SwAV features for various pre-training set sizes show that adding more classes during pre-training consistently improves REMIND's performance for all features used. Furthermore, using SwAV, Barlow Twins, or MoCo-V2 consistently improves performance over supervised pre-training, which can be seen by the vertical shift of the learning curves across different features.

Similar to REMIND, adding more classes during pre-training improves Deep SLDA's performance for all features used. The vertical shift of the curves upwards across features shows that MoCo-V2, Barlow Twins, and SwAV consistently outperform supervised features for 10, 15 and 25 pre-train classes. Even though supervised features show the best performance for 100 classes, MoCo-V2, Barlow Twins, and SwAV show competitive results.

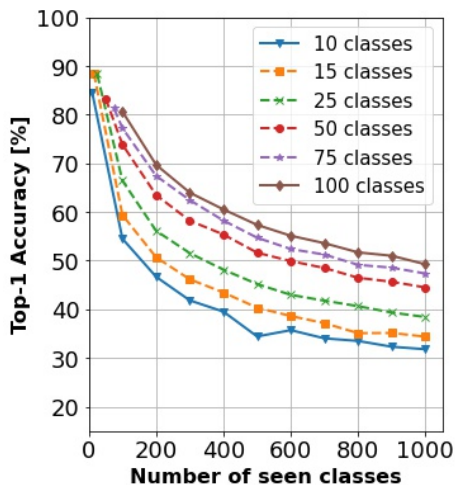
Online Softmax also has better performance when using more pre-train classes across different features. Surprisingly, MoCo-V2 outperforms Barlow Twins, SwAV, and supervised features for 10 and 15 pre-train classes. SwAV features outperform supervised features across all pre-train sizes. It is worth noting that when 100 pre-train classes are used, MoCo-V2, SwAV, and supervised features start around the same top-1 accuracy, but after the model has finished learning all 1000 classes, SwAV achieves the highest performance. This behaviour means that SwAV features obtained during pre-training on 100 classes are more useful when continually learning all 1000 classes.



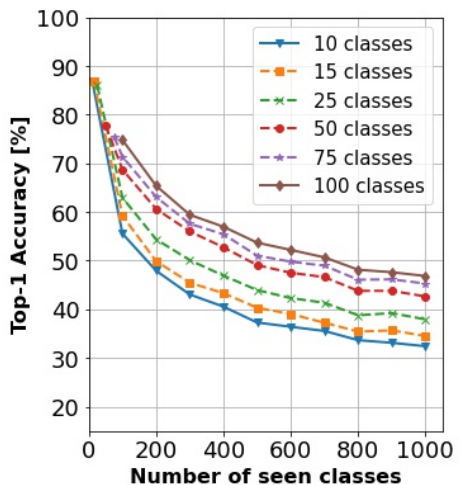
(a) Supervised



(b) SwAV

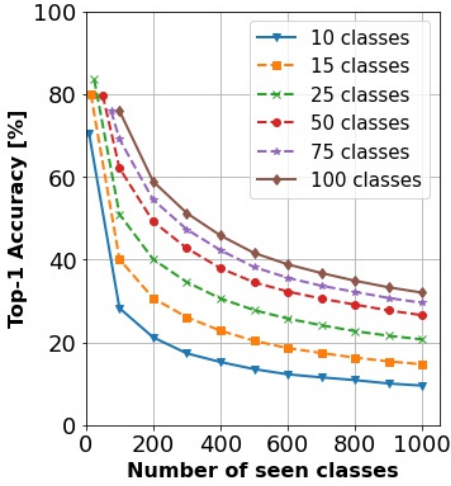


(c) MoCo-V2

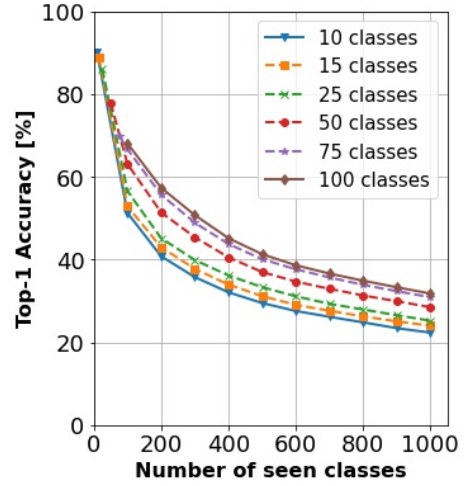


(d) Barlow Twins

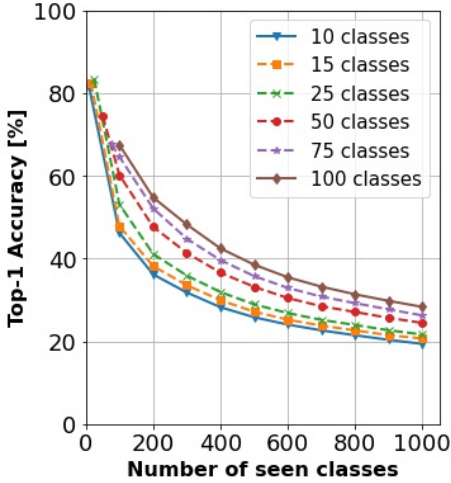
Figure S2: Learning curves on ImageNet for each pre-train size with REMIND using (a) supervised, (b) SwAV, (c) MoCo-V2, and (d) Barlow Twins features.



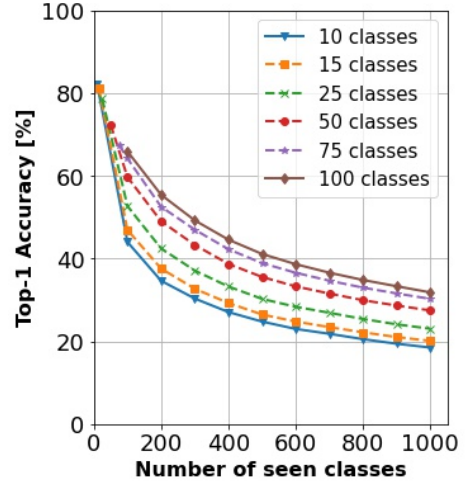
(a) Supervised



(b) SwAV

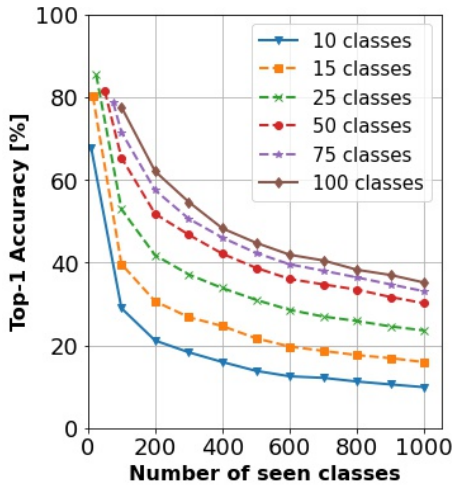


(c) MoCo-V2

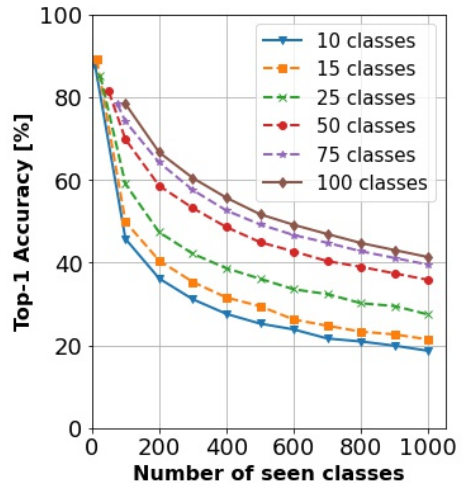


(d) Barlow Twins

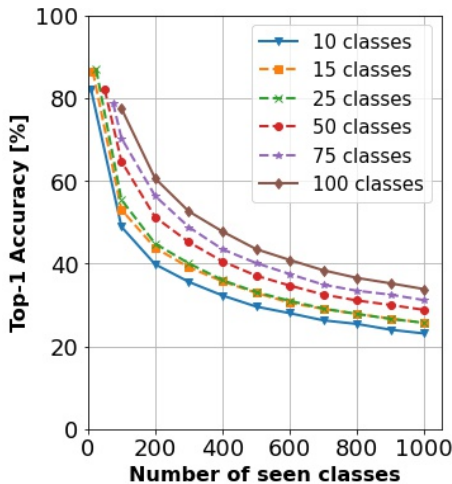
Figure S3: Learning curves on ImageNet for each pre-train size with Deep SLDA using (a) supervised, (b) SwAV, (c) MoCo-V2, and (d) Barlow Twins features.



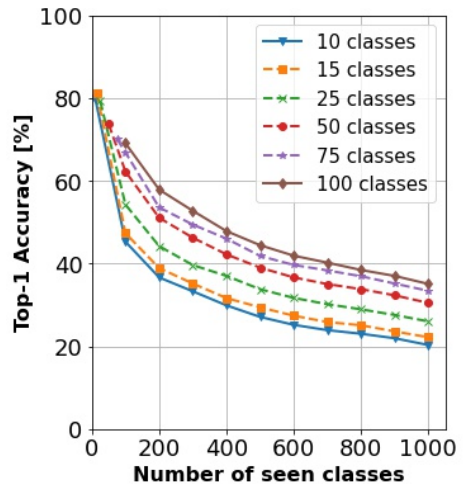
(a) Supervised



(b) SwAV



(c) MoCo-V2



(d) Barlow Twins

Figure S4: Learning curves on ImageNet for each pre-train size with Online Softmax using (a) supervised, (b) SwAV, (c) MoCo-V2, and (d) Barlow Twins features.