A. Statistics on the annotated YT8M

This section shows statistics on the YT8M, annotated with the object detector [1]. We annotate each frame of YT8M with the object detector and store the five objects with highest detection scores. Our method relies on objects recurring multiple times in a video. The method works better when objects occur multiple times in the selected frames. Therefore, Table 6 displays statistics for objects that occur in most videos. For each object, we count how often the object recurs in the 32 frames sampled with the strategy from [68]. For example, in 49 percent of videos, an object with class Footwear occurs. Each of those videos has, on average, 15 instances of the Footwear class.

We discard objects with a low detection score. Figure 5 shows the fraction of boxes below a certain threshold. All methods in this work use a threshold of 0.05, which discards about 3 percent of the objects. We experimented with higher thresholds, but this resulted in worse VTAB scores.

Table 6: Recurrence of objects within the 32 frames sampled for learning from one video. For example, on average, 86% of the videos contain an object labeled PERSON. In each video where a PERSON occurs, the detector annotated an average of 30 instances. We show averages over ten thousand videos that we randomly sampled from the training set.

B. Sensitivity to hyperparameters

Our experiments use three important hyperparameters. We used the validations sets from the VTAB benchmark to set the hyperparameters. This section shows the sweeps we make so one can judge the sensitivity for each hyperparameter. Figure 6 shows the search for hyperparameter \( w \) from Equation (3). Figure 7 shows the search for a positive coefficient to include the cross entropy loss in the experiment for Table 1, row Also predict cross entropy. Figure 8 shows the search for a positive coefficient for the cross entropy loss when learning from the soft labels from ImageNet for Table 1, row Distilling from ImageNet.
Table 7: VTAB accuracies for each method and dataset considered in our work. Each number represents the accuracy after transferring the model learned with the method to the specific dataset. Each dataset has only 1000 labeled samples. We follow the transfer protocol from [60]

![Figure 7](vtab_scores_natural.png)  
Figure 7: VTAB scores on respective validation sets when changing the weight for the additional supervised loss on the objects. The optimum accuracy occurs at 0.1, which is the value we use in the ablation experiment. The error bars indicate bootstrapped 95% confidence intervals.

![Figure 8](vtab_scores_stratified.png)  
Figure 8: VTAB scores on respective validation sets when changing the weight for cross entropy loss on the soft labels. This corresponds to row Distilling from IMAGENET reported in Table 1. The optimum accuracy occurs at 1.0, which is the value we use for the experiment. The error bars indicate bootstrapped 95% confidence intervals.