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.

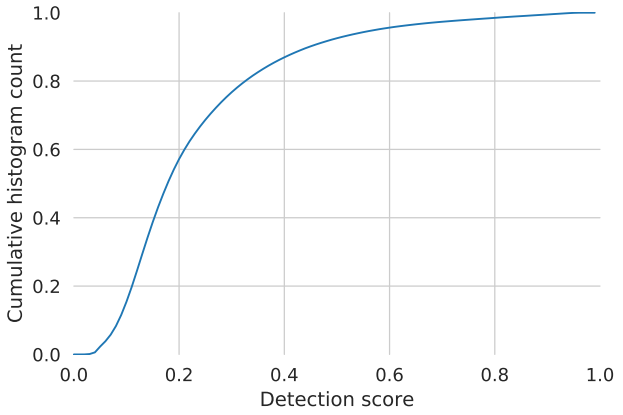

Figure 5: Cumulative histogram of the detection scores from the object detector. Histogram measured on the videos from YT8M, annotated with our detector [1]. In our experiments, we exclude boxes with scores below 0.05, which applies to approximately 3 percent of the objects.

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 ω 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.

LABEL NAME	Videos (%)	Recurrence
STREET LIGHT	14	13.1
FLOWER	15	9.6
CHAIR	15	7.3
LAND VEHICLE	15	5.6
TABLE	20	7.6
TOY	21	13.0
BOTTLE	22	9.7
CAR	24	12.0
BUILDING	28	9.9
WOMAN	30	12.0
WHEEL	30	13.4
POSTER	37	11.1
WINDOW	44	17.9
FOOTWEAR	49	15.0
TREE	50	32.9
MAN	51	14.8
HUMAN FACE	72	20.5
CLOTHING	84	24.1
PERSON	86	30.0

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.

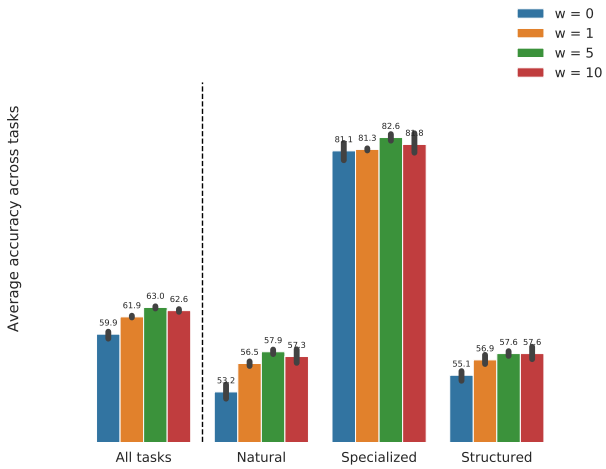

Figure 6: VTAB scores on respective validation sets when changing the weight for the object-level loss. The optimum accuracy occurs at 5, which is the value we use in all experiments. The VTAB scores change away from the optimum, but are relatively stable when comparing to baseline (see Table 1). The error bars indicate bootstrapped 95% confidence intervals.

	Method	All tasks	Natural	Specialized	Structured	Caltech101	Cifar-100	DTD	Flowers102	Pets	SUN397	SVHN	Retinopathy	EuroSAT	Resics45	Camelyon	CLEVR-dist	CLEVR-count	DMLab	dSPRITES-orient	dSPRITES-pos	sNORB-azimuth	sNORB-elevation	KITTI
Comparison	RESNET50 from scratch	42.1	26.9	65.8	43.6	37.7	11.0	23.0	40.2	13.3	3.9	59.3	63.1	84.8	41.6	73.5	54.8	38.5	35.8	37.3	87.9	20.9	36.9	36.9
	Transitive Invariance [72]	44.2	35.0	61.8	43.3	54.9	7.1	38.3	28.2	32.3	7.4	77.0	63.1	84.1	50.0	50.0	61.7	12.7	35.0	59.3	86.1	21.1	29.2	41.6
	MS [13]	47.2	33.4	68.4	47.9	52.3	12.7	37.3	32.6	15.8	6.8	81.8	57.3	89.7	49.7	76.8	55.7	43.2	38.4	46.4	81.2	34.8	35.1	48.4
	MT [13]	59.2	51.9	78.9	55.8	76.2	26.2	49.3	63.5	48.5	10.6	89.1	71.7	93.3	70.2	80.3	62.1	55.6	44.3	43.2	86.6	39.1	38.9	76.3
	MobileNetV2	65.9	69.5	81.9	54.8	88.5	45.4	59.1	87.3	86.7	32.0	87.3	71.1	94.3	80.5	81.6	55.8	44.8	46.6	51.6	90.0	37.5	38.7	73.4
	IMAGENET supervised	68.5	71.3	83.0	58.9	84.7	60.0	68.2	87.3	90.3	36.1	72.2	75.4	95.2	81.4	80.0	57.7	73.9	45.6	59.7	88.5	29.1	34.2	82.3
	IMAGENET supervised (3x)	69.5	72.6	83.8	59.5	85.6	61.0	69.6	88.8	90.9	37.4	75.0	78.0	95.7	82.5	78.9	61.4	64.6	45.3	60.5	93.2	32.9	36.6	81.5
	Detector backbone [1]	61.6	60.0	80.4	53.5	84.3	38.2	48.4	77.4	58.6	25.2	88.1	70.6	94.0	73.4	83.5	58.2	42.8	47.8	46.4	73.4	39.4	42.9	77.4
	BigBiGAN [15]	59.1	56.6	79.1	51.3	80.8	39.2	56.6	77.9	44.4	20.3	76.8	69.3	95.6	74.0	77.4	55.6	53.9	38.7	46.7	70.6	27.2	46.3	71.4
Video only	VIVI [68]	60.8	55.1	80.0	56.3	74.8	29.2	48.6	76.9	54.8	13.6	87.6	71.4	94.4	74.1	80.1	59.0	54.0	47.1	50.9	91.7	37.0	42.4	68.2
	OURS	64.0	58.9	81.8	59.7	81.5	35.9	51.6	76.9	60.1	17.1	89.3	72.7	94.7	76.9	82.7	62.6	61.5	50.8	53.3	92.2	41.5	39.0	76.6
	Rand boxes and labels	60.3	55.1	80.0	54.9	75.7	28.2	49.6	76.7	53.1	14.7	88.0	71.3	93.8	74.0	80.8	60.8	55.7	34.5	50.7	94.0	37.2	37.0	69.5
	Rand boxes	63.4	57.5	81.1	59.7	79.4	31.4	51.3	77.0	58.8	16.2	88.5	72.2	94.3	74.5	83.5	61.3	60.0	48.0	52.2	95.0	40.6	42.1	78.2
	Distilling from ImNet	63.1	59.6	81.5	56.9	81.3	35.4	58.4	75.5	54.3	24.9	87.5	73.0	95.4	75.8	82.0	61.0	50.0	47.0	50.7	89.3	36.5	41.8	79.3
	Include CE loss	64.9	60.5	81.3	60.5	83.9	38.9	55.2	76.2	59.2	20.7	89.3	71.2	94.7	77.0	82.3	63.7	65.6	50.8	52.8	94.2	34.7	40.4	81.7
	Distilling detector	57.1	52.2	77.2	51.3	78.2	29.6	49.1	56.9	47.2	21.0	83.5	70.0	91.8	66.3	80.5	58.1	44.9	41.4	44.9	77.5	30.9	32.4	80.2
	Filter half of detections	61.9	57.6	80.3	56.4	79.7	31.8	50.4	78.3	59.2	14.5	89.2	71.0	94.3	74.9	81.0	58.4	50.6	48.4	52.2	90.9	34.9	41.6	74.4
Cotraining	VIVI [68]	69.0	70.0	83.5	60.8	87.2	51.5	64.6	88.2	85.9	32.3	79.9	72.5	95.5	81.0	84.9	61.2	74.6	44.7	61.9	90.6	29.4	43.7	80.5
	OURS	69.5	70.7	83.2	61.5	88.0	53.1	64.9	88.1	86.3	33.5	81.0	72.2	94.5	80.5	85.6	56.5	79.3	46.6	60.5	92.7	28.6	45.2	82.3
	VIVI (3x) [68]	70.5	72.6	83.8	62.0	88.0	54.3	69.4	89.6	87.9	34.6	84.2	72.9	95.3	82.3	84.9	58.3	74.5	46.3	67.8	92.1	33.1	44.1	80.2
	OURS (3x)	70.8	72.2	83.4	63.4	87.1	55.3	67.8	90.0	87.7	35.6	81.6	72.0	95.0	82.4	84.1	63.2	80.5	47.3	66.0	87.8	33.9	46.0	82.6

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]

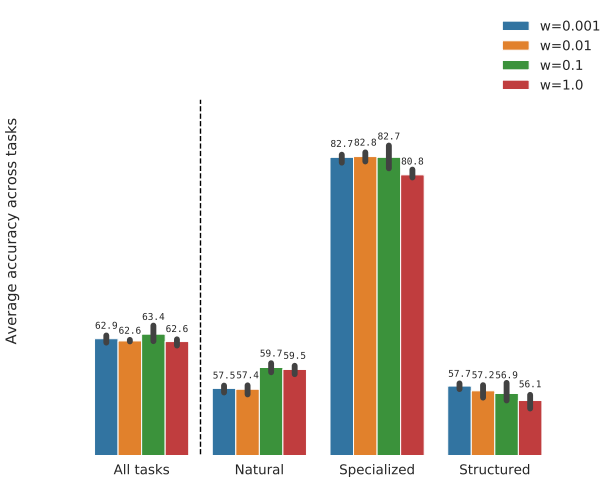

Figure 7: VTAB scores on respective validation sets when changing the weight for the additional supervised loss on the objects. The optimum accuracy occurs at 0.1, which is the value we use in the ablation experiment. The error bars indicate bootstrapped 95% confidence intervals.

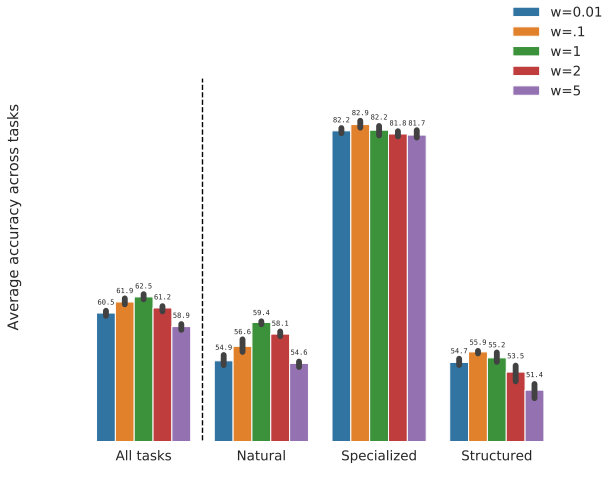

Figure 8: VTAB scores on respective validation sets when changing the weight for cross entropy loss on the soft labels. This corresponds to row *Distilling from* IMAGENET reported in Table 1. The optimum accuracy occurs at 1.0, which is the value we use for the experiment . The error bars indicate bootstrapped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