Table 6: Hyper-parameters used during the training of the different models

Parameters	*COCO-standard*			*DIOR*		
	Teacher	Student	Refined	Teacher	Student	Refined
Training steps	20k	180k	10k	20k	90k	10k
Learning rate	0.08	0.08	0.08	0.08	0.08	0.08
Learning rate decay	13k-18k	120k-160k	6k	13k-18k	60k-80k	6k
Batch Size (labeled — pseudo labeled)	64|0	8|56	64|0	32|0	8|24	32|0

Table 7: Comparison between refined student models trained with different view techniques during the generation of candidate labels for the mAP on *COCO-standard*. For the Scale+Flip technique, we use the information of the normal view, the scaled/flipped only view and the scale+flip view. Adding multiple views is a simple yet effective way to improve the quality of candidate labels. We report the mean and standard deviation over 5 randomly sampled dataset.

Transformation	None	Scale	Flip	Scale+Flip
mAP	32.87 ± 0.23	32.76 ± 0.17	32.90 ± 0.19	$\mathbf{32.91 \pm 0.16}$

6. Supplementary Material

6.1. Implementation details.

Networks. We use a pre-trained ResNet-50 [5] as backbone for Faster-RCNN [20] with FPN [13] as object detector.

Training parameters. For the *COCO-standard* setup, the teacher models are warmed-up for 20k steps with a learning rate decay after 13k and 18k steps. Then, our student models are trained for 180k steps, using a global batch size of 64. We apply the same learning rate decay after 120k and 160k steps. We use SGD as optimizer, with an initial learning rate of 0.08 and with default other parameters. The refined student models are trained for 10k steps, using a initial learning rate of 0.0008, which is reduced after 6k steps. For the *DIOR* setup, the student models are trained for 90k steps with a learning rate decay after 60k and 80k steps. The different values are gathered in Table 6.

Data augmentations. For the data augmentations during training, we use some large scale color jittering, such as random changes in brightness, contrast, hue and saturation. We also apply some scale jittering and random horizontal flips.

6.2. Student training.

In Table 7, we show the results of refined student models trained with pseudo-labels generated with different view strategies. The idea of using the four different views (normal, flip, scale and flip+scale). We can see that only scaling up the view gives worse results, but scaling and flipping gives a tiny improvement compared to only flipping the image.

We also study the effect of weighting the loss in Equation (3) during the training of the student. We trained a student by fixing the α term in Equation (3) to 1. On average, the gain of mAP is 0.13 for the model trained with the weighted loss.