Appendix

A. COCO experiments with fewer labels

We conducted additional experiments on the COCO dataset using fewer labeled training samples compared to earlier work: from 0.4% down to 0.1%. We report results for a supervised training baseline and our semi-supervised approach in Tab. 8. In this extremely challenging, very sparsely labeled data regime, our approach shows marked improvements over the supervised baseline, *e.g.*, +3.4 points for 0.1% labeled images using the DINOv2 backbone. Moreover, we find that in this regime the DINOv2 backbone is more effective than the Swin-L (IN-21k) backbone, *e.g.*, in the 0.4% labeled data case, corresponding to less than 500 of image annotations: the DINOv2 backbone improves mask-AP by 8 points from 23.0 to 31.0, reaching similar performance as the previous state-of-the-art approach Polite Teacher with 25 times less annotations (30.8 mask AP at 10% labeled data, see Tab. 3d. This is in line with our observation on the 5% labeled data case for Cityscapes in Tab. 3c, and 1% labeled data case for COCO in Tab. 3d. See Fig. 7 for an illustration of segmentations obtained with this model.

Amount of labeled data used	0.1%	0.2%	0.4%
	Supervised models		
Mask2Former - Swin-L (IN-21k)	5.3	10.2	15.9
Mask2Former - ViT-L (DINOv2)	**10.2**	**19.4**	**25.7**
	Semi-supervised models		
Ours - Swin-L (IN-21k)	5.8	16.1	23.0
Ours - ViT-L (DINOv2)	**13.6**	**24.9**	**31.0**

Table 8. Evaluation of supervised and semi-supervised models on COCO using extra small labeled training sets.

B. Details on models and training efficiency

Training efficiency. Our approach can be seen as a two-stage knowledge distillation method where the teacher and student share the same architecture. Compared to PT [9], we pretrain the teacher network using the available labeled samples only and use its predictions during the student's burn-in stage. Compared to NB [28], we generate pseudo-ground truths in an online manner instead of doing it offline prior to the student's training. These changes, although computationally demanding, are justified by the large performance improvements over the previous baselines. For example, the student training with ResNet-50 backbone on COCO consumes 89% more GPU memory and the iterations take approximately twice longer compared to the teacher pre-training. Peak performance is usually achieved after a few thousand iterations in sparse regimes. Therefore, in practice, it only takes a few dozen hours to train the models end-to-end.

Backbone	Parameters (M)	GFLOPS
R50	44	224.8 ± 24.6
Swin-B	107	464.2 ± 48.7
Swin-L	216	864.7 ± 90.2
ViT-B (DINOv2)	108	944.5 ± 93.4
ViT-L (DINOv2)	326	1285.3 ± 127.1
ViT-B (Deit)	108	692.8 ± 98.6

Table 9. FLOP and parameter count for the different models used in our project.

Recent works in KD have explored alternative strategies where no separate teacher network is required, see *e.g.* [15], with applications to image classification. Such alternatives present interesting directions of future work to improve the efficiency of semi-supervised methods.

Model characteristics. We present the different characteristics of the models used. We report both the number of parameters and the FLOP count for each architecture used in our project in Tab. 9 as measured using `count_flops` function in the Detectron2 library.

Comparing training protocols. Section 4.3 provides ablation for the training protocol, which compares the main changes in the distillation strategy with respect to PT while keeping the overall model architecture constant. Particularly, we :

1. Isolate the effect of our revisited burn-in stage.
2. Isolate the effect of student data augmentations (the teacher augmentations are the same between PT and ours).

In Tab. 4, we can see that using the standard burn-in stage as in PT reduces the AP by 3.7 points. In Fig. 6, we track the mask-AP evolution when using different data augmentations, we can see that our data augmentation yields better performance and more stable convergence that PT which additionally uses random cutout. Hence, both changes show improved performance with respect to the SOTA protocol while using the same underlying meta architecture, backbone, training epochs etc. This is evidence that our distillation protocol and revisited burn-in stage are both important factors for the improved performance, beyond the backbone and meta-architecture choices.

Estimation of carbon footprint. On COCO, it took 25 hours to train our ViT-L (DINOv2) model using 1% of annotations, and about 2 days to train a Swin-L (IN-21k) model using 10% of annotations. Trainings are approximately 2.5 times faster on Cityscapes. Given the same for-

Figure 7. Illustration of predictions obtained with model trained on COCO with only 0.4% of labels. The model uses a DINOv2 [22] backbone and achieves an AP of 31.0, which is superior to what the previous SOTA achieved using 25x more labels (PT achieves 30.8 AP with 10% of labels).

mula used in [22], a Thermal Design Power (TDP) of V100-32G equal to 250W, a Power Usage Effectiveness (PUE) of 1.1, a carbon intensity factor of 0.385 kg CO_2 per KWh, a time of 2 days $\times$ 24 hours $\times$ 16 GPUs = 768 GPU hours to train our approach with a SWIN-L, it leads to 211 kWh, an equivalent CO_2 footprint of 211 $\times$ 0.385 = 81.2 kg.