Appendix

We provide more information here.

A. Additional implementation details

A.1. Evaluation protocol

While evaluating the performance of class-agnostic instance segmentation, we also report the results of an easier protocol AP* which evaluates medium and large objects. Here, only the objects with area greater than 64^2 are considered and their mask AP* with an IoU threshold of 0.5 is computed. AP$^*_M$ and AP$^*_L$ are also reported for medium and large objects, *i.e.*, objects with area in the range of $(64^2, 192^2)$ and those with area greater than 192^2, respectively. The results of MCG and COB are computed using the official segmentation masks.

A.2. Supervised fine-tuning

We evaluate the pre-trained instance segmentation model by fine-tuning it with manual annotations. Specifically, we fine-tune a dynamic SOLO model (aka SOLOv2) on COCO `train2017` and evaluate on COCO `val2017`. Synchronized batch normalization is used in the backbone along with FPN [73] during training. We provide two training settings, *i.e.*, limited fully annotated images, and limited segmentation masks.

Limited images. For the experiments with limited images, we use 5% and 10% images from COCO `train2017`, which corresponds to ~6k and ~12k fully annotated images, respectively. We fine-tune the instance segmenter initialized with the pre-trained model for 20k iterations with an initial learning rate of 0.01, which is then divided by 10 at 12k and 18k iterations.

Limited masks. For the experiments with limited masks, we use 5% and 10% segmentation masks from COCO `train2017`. In this setting, only 5% and 10% of the images have mask annotations, *i.e.*, ~6k and ~12k images, respectively. Specifically, we use all the class labels to supervise the category branch, but only use a part of the annotated masks to supervise the mask branch. The model is trained for 90k iterations with the standard schedule.

A.3. Training details

For the self-supervised pre-trained backbones, we use the official models trained on ImageNet without labels for 200 epochs. For FreeSOLO, we use the images in COCO `train2017` and COCO `unlabeled2017` as the set of unlabeled images, containing a total of ~241k images. We use ResNet-50 as the backbone for all the fine-tuning experiments and ablation study and use ResNet-101 for other results and visualizations. We train for 30k iterations on 8 GPUs with a total of 32 images per mini-batch. The learning rate is set to 0.0025. In the self-training, we repeat the schedule once and train for another 30k iterations.

Copy-paste augmentation. For a pair of images in a batch, we randomly select objects from one image and paste them at random locations on the other image. These objects are not pasted if they have a high overlap (IoU $>= 0.5$) with existing objects.

B. Additional results

We report the results of an easier protocol AP* which evaluates medium and large objects in Table S1. As shown, the gains over MCG and COB are larger, especially for the large objects.

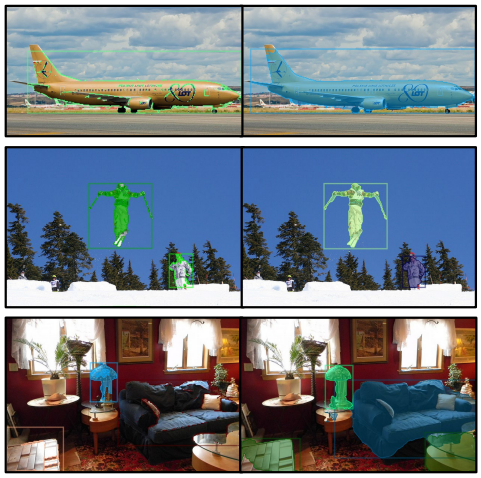
method	AP$_{50}$	AP$_{75}$	AP	AP*	AP$^*_M$	AP$^*_L$
w/ anns:						
MCG [61]	4.6	0.8	1.6	9.4	32.4	7.4
COB [62]	8.8	1.9	3.3	15.6	36.5	11.0
w/o anns:						
FreeSOLO	9.8	2.9	4.0	24.3	21.5	34.3

Table S1. Class-agnostic instance segmentation on COCO `val2017`. Both MCG and COB require annotations.

C. Additional visualizations

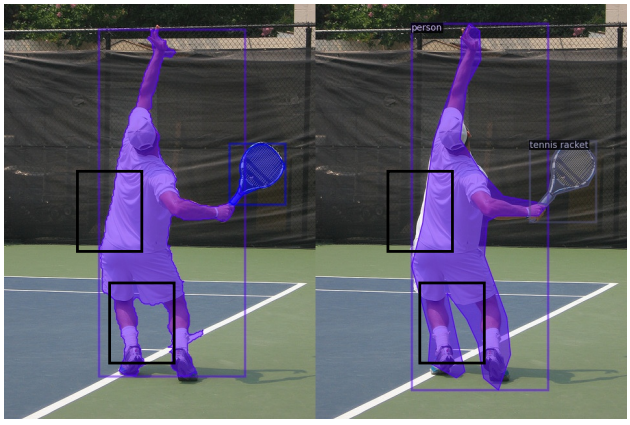
In this section, we provide additional visualizations of FreeSOLO. We show qualitative results of our method for the task of class-agnostic instance segmentation in Figure S1. In Figure S2, we provide more qualitative comparison of FreeSOLO with and without the $\mathcal{L}_{avg\_proj}$. As shown in Figure S3, we further show that FreeSOLO can even produce more precise segmentation results than manual annotations at some object boundaries, which indicates FreeSOLO's great potential for tasks such as auto-labeling.

Figure S1. More qualitative results of FreeSOLO for the task of class-agnostic instance segmentation. The model is trained *without any kind of manual annotations* and can infer at 16 FPS on a V100 GPU. Best viewed on screen.



w/o $\mathcal{L}_{avg\_proj}$ w/ $\mathcal{L}_{avg\_proj}$

Figure S2. Qualitative comparison of FreeSOLO with and without $\mathcal{L}_{avg\_proj}$ when learning from coarse masks. The model trained without $\mathcal{L}_{avg\_proj}$ tends to only segment the contours when trained longer.



FreeSOLO Output COCO Ground Truth

Figure S3. Qualitative comparison of FreeSOLO's predicted masks and ground truth masks. At some object boundaries, FreeSOLO can produce *even more precise* segmentation than manual annotations in some cases.