Supplementary Material for "Active Pointly-Supervised Instance Segmentation"

Chufeng Tang[1], Lingxi Xie[2], Gang Zhang[1], Xiaopeng Zhang[2], Qi Tian[2(✉)], and Xiaolin Hu[1,3,4(✉)]

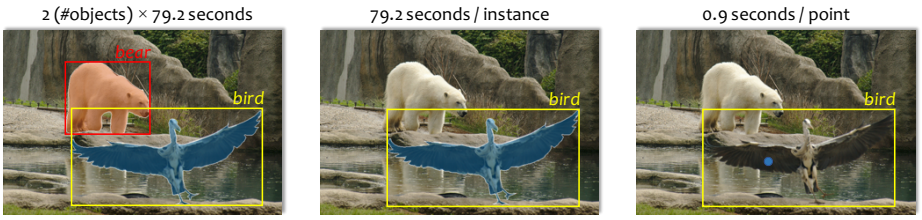
[1]Department of Computer Science and Technology, Institute for AI, BNRist, State Key Laboratory of Intelligent Technology and Systems, Tsinghua University
[2]Huawei Inc. [3]Chinese Institute for Brain Research (CIBR)
[4]IDG/McGovern Institute for Brain Research, Tsinghua University
{tcf18, zhang-g19}@mails.tsinghua.edu.cn, {198808xc, zxphistory}@gmail.com
tian.qi1@huawei.com, xlhu@mail.tsinghua.edu.cn

A. More Details and Results of AFIS

Since there is no existing work can be directly compared to APIS, we established the baseline setting *active fully-supervised instance segmentation (AFIS)*. The mask annotations (image-level or instance-level) are queried by the model during each active learning step, which is conceptually similar to some existing active learning algorithm designed for image classification or object detection. In this section, we provide more description and results of AFIS.

Annotation Schemes. Fig. S1 illustrates different annotation schemes from the perspective of human annotators, as well as the corresponding approximate annotation time. Compared to AFIS, APIS can be studied in a more fine-grained manner because it allocates annotation budgets to pixels, and the annotation of points is considerably faster and cheaper.

Sampling Strategy of AFIS. In Fig. 6 of the main paper, we compared different sampling strategies for the case of *image-level* selection and *instance-level* selection, respectively. **Firstly**, the results of the *Mean Entropy* strategy were unsatisfactory in both cases, even lagging behind the results of random sampling. We diagnosed the problem and found the main reason is that the instances selected under this metric are usually small objects. For example, over 76% of the selected instances at the first step are small (*i.e.*, area$<32^2$, as defined in COCO). For the larger objects, there were usually a lot of low-entropy points (*e.g.*, points on the smooth interior areas or background), which decreased the mean entropy value. It is well-known that the the model's performance is usually poor on small objects even with the mask supervision [6], thus the annotated instances in our cases were not so effective. **Secondly**, we used the detection loss (*e.g.*, GIoU Loss) to measure the *Detection Quality*, *i.e.*, the lower the loss, the higher the quality. As shown, the strategy that selecting with the lowest detection loss (*Min. Det. Loss*) usually produced better results than random sampling for the image-level selection, while for the instance-level selection, the performance was usually on par with random sampling. In addition, we also studied the opposite strategy that selecting with the highest detection



(a) *Image-level*: annotating masks for **all** objects in the image.

(b) *Instance-level*: annotating **one** mask for the specified object.

(c) *Point-level*: answering if **this point** is on the specified object.

Fig. S1: Illustration of different annotation schemes (*image-level*, *instance-level*, and *point-level*) from the perspective of human annotators. The averaged annotation time for these schemes is provided for reference (values adopted from [2]).

Fig. S2: Examples of the selected images under the *Detection Quality* metric for *image-level* AFIS.

loss (*Max. Det. Loss*) and found the results are poor. The results suggest that instance-level cues (*e.g.*, detection quality) are somehow important for AFIS. From these results, we conjecture that instance-level cues (*e.g.*, detection quality) may provide complementary information to assist APIS or a mixed setting that both instance-level and point-level supervision can be chosen. Although the strategy for AFIS has not been thoroughly optimized, we believe that it is sufficient to serve as a reasonable baseline for APIS and provides some useful guidance for future researches in this area.

Qualitative results of image-level AFIS. Fig. S2 shows some examples of the selected images under the *Detection Quality* (*i.e.*, minimum detection loss) metric. As shown, the selected image usually contains fewer objects. Empirically, if the annotation budget, the same as labeling one point for each instance at an active learning step, was allocated to images, we can annotate masks for about 3435 images. Each image contains 2.8 objects on average, which is considerably fewer than the number of annotated objects per image (7.7) in MS-COCO. This observation is contrary to most works on active object detection where the algorithms usually prefer the images with more objects [4] since the annotation costs for different images are considered the same in their experiments. In our

Fig. S3: Examples of the selected instances under the *Detection Quality* metric for *instance-level* AFIS.

setting, the annotation cost is proportional to the number of objects in the image, which is closer to the real-world scenarios.

Qualitative results of instance-level AFIS. Fig. S3 shows some examples of the selected instances under the *Detection Quality* metric. We found that the selected instances usually covered a large area, *e.g.*, over 63% of the selected instances at the first step were large object (*i.e.*, area$>96^2$ pixels, as defined in COCO), which is consistent with the observation that the detection results of larger objects are usually better than the results of smaller objects.

B. More Results of APIS

Relation to Object Scale. We empirically found that the model trained with actively acquired points performed much better on the larger instances (higher ΔAP_L) than random points, as listed in Tabel S1. It indicates that for larger instances, the informativeness of different points is more diverse than smaller one. Besides, the smaller instances are inherently hard to recognize no matter which type of label is given.

APIS on Cityscapes. We additionally reported the results on Cityscapes [3]. The results (red lines in Fig. S4) validate the same conclusion. In this study, we follow the previous work [5] to use MS-COCO pre-training, and the results are slightly unstable due to the small dataset size.

Box-free APIS. In this work, we studied APIS with box-level annotations, but it is also feasible by eliminating bounding box annotations. We validated this through a preliminary solution (*i.e.*, generating pseudo box labels using an off-the-shelf detector to assist APIS) on the Cityscapes dataset. As shown in Fig. S4 (the blue lines), the mask AP dropped by 1%–2% due to the inaccurate box labels. Note that the decrease is moderate since we used a detector that produced high-quality pseudo boxes (only to show the feasibility), and the results might be lower with low-quality pseudo box labels. Additionally, there exist advanced point-based detectors [1] to integrate, which we leave for future work.

Table S1: The mAP improvement of the *Entropy* strategy over random sampling, as well as the results on the small, medium and large instances.

	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$
ΔAP	+0.56	+0.92	+0.80
ΔAP_S	−0.30	+0.37	+0.28
ΔAP_M	+0.69	+0.93	+0.66
ΔAP_L	**+1.17**	**+1.53**	**+1.33**

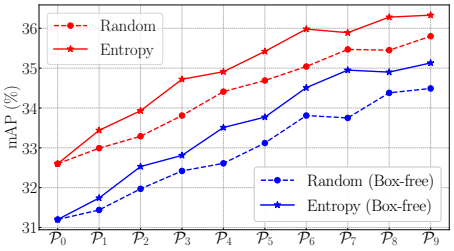


Fig. S4: APIS results on the Cityscapes dataset. Blue lines indicate the box-free APIS results.

More Visualization Results of APIS. We visualized the mask predictions, uncertainty maps, and the selected points for more instances, as shown in Fig. S5 (extension of Fig. 4a in the main paper).

C. APIS with Fewer Labeled Points

For the experiments in the main paper, we labeled one point for each instance at a step, while the difficulty of instance was not considered, *e.g.*, some instances are easier to learn and require fewer (or even zero) points, while others may require more points. We studied this problem by reducing the annotation budget of each step to 100,000 points (8× fewer). The training pipeline keeps unchanged. We explored two different ways to select instances: random sampling and the *Min. Det. Loss* strategy (similar to instance-level AFIS, see Sec. 3.3 in the main paper). As shown in Fig. S6, the actively acquired points still worked better than random points. As for instance selection, sampling instances with higher detection quality (*i.e.*, lowest loss) led to higher performance. With 400k points (32.5%), the model outperformed the previous model trained with 860k points ($\mathcal{P}_0$, 32.0%), but at the cost of longer training time (2×). Both the annotation cost and computational cost should be considered when deciding the number of labeled points at each step, while the former is usually much more expensive in practice.

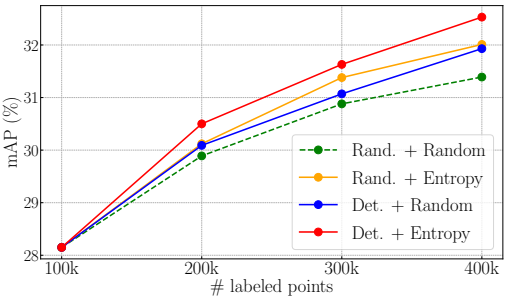


Fig. S6: The results of APIS with fewer labeled points. Rand. and Det. indicate random sampling and the *Min. Det. Loss* strategy, respectively.

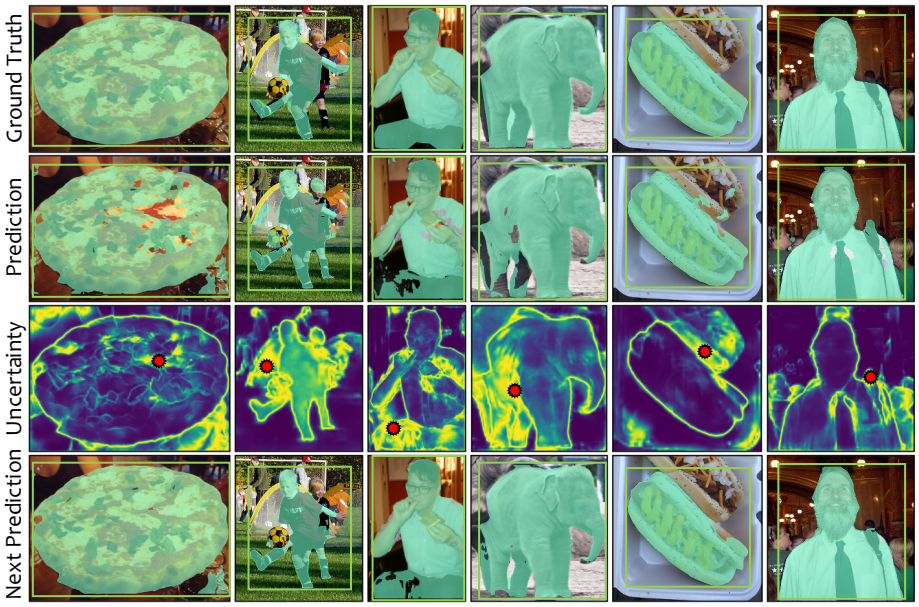


Fig. S5: Visualization of (from top to bottom): ground-truth masks, mask predictions (averaged over multiple predictions), uncertainty maps (the brighter the more uncertain), and mask predictions after fine-tuning with the selected points for some instances (extension of Fig. 4a in the main paper). The red spots indicate the selected points. Best viewed in colour.

References

1. Chen, L., Yang, T., Zhang, X., Zhang, W., Sun, J.: Points as queries: Weakly semi-supervised object detection by points. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 8823–8832 (2021)
2. Cheng, B., Parkhi, O., Kirillov, A.: Pointly-supervised instance segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2617–2626 (2022)
3. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3213–3223 (2016)
4. Haussmann, E., Fenzi, M., Chitta, K., Ivanecky, J., Xu, H., Roy, D., Mittel, A., Koumchatzky, N., Farabet, C., Alvarez, J.M.: Scalable active learning for object detection. In: 2020 IEEE Intelligent Vehicles Symposium (IV). pp. 1430–1435 (2020)
5. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Int. Conf. Comput. Vis. pp. 2961–2969 (2017)
6. Singh, B., Davis, L.S.: An analysis of scale invariance in object detection snip. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3578–3587 (2018)