

# Active Pointly-Supervised Instance Segmentation

Chufeng Tang<sup>1</sup>, Lingxi Xie<sup>2</sup>, Gang Zhang<sup>1</sup>, Xiaopeng Zhang<sup>2</sup>,  
Qi Tian<sup>2(✉)</sup>, and Xiaolin Hu<sup>1,3,4(✉)</sup>

<sup>1</sup>Department of Computer Science and Technology, Institute for AI, BNRIst,  
State Key Laboratory of Intelligent Technology and Systems, Tsinghua University

<sup>2</sup>Huawei Inc. <sup>3</sup>Chinese Institute for Brain Research (CIBR)

<sup>4</sup>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

**Abstract.** The requirement of expensive annotations is a major burden for training a well-performed instance segmentation model. In this paper, we present an economic active learning setting, named active pointly-supervised instance segmentation (APIS), which starts with box-level annotations and iteratively samples a point within the box and asks if it falls on the object. The key of APIS is to find the most desirable points to maximize the segmentation accuracy with limited annotation budgets. We formulate this setting and propose several uncertainty-based sampling strategies. The model developed with these strategies yields consistent performance gain on the challenging MS-COCO dataset, compared against other learning strategies. The results suggest that APIS, integrating the advantages of active learning and point-based supervision, is an effective learning paradigm for label-efficient instance segmentation.

**Keywords:** Instance segmentation · Active learning · Point-based supervision · Label-efficient learning

## 1 Introduction

Instance segmentation aims to predict a pixel-wise mask with a category label for each instance in the given image. Despite the rapid development of the instance segmentation methods, the requirement of a massive amount of labeled data is still a heavy burden of training a well-performed instance segmentation model. For example, the annotation of a polygon-based mask for an object in MS-COCO requires 79.2 seconds [34] on average and even higher for the more precise mask annotations in LVIS [18], which is considerably more time-consuming than annotating a bounding box (*e.g.*, 7 seconds via clicking extreme points [39]).

In general, weak supervision and active learning are usually two effective ways to reduce the annotation cost. For the task of instance segmentation, a number of existing works attempted to predict instance masks with weak supervisions, such as category tags [3,41,63], bounding boxes [22,27,29,53], and points [10,30].

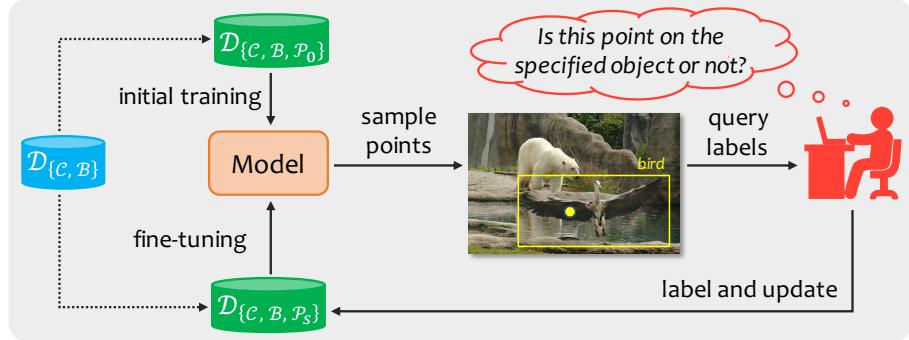


Fig. 1: Overview of the training pipeline for the proposed APIS setting where the annotator is asked for labeling whether the point falls on the specified object (*i.e.*, the bird) or not. The expected label in this case is ‘yes’.  $\mathcal{D}$  is the training data and  $\mathcal{C}, \mathcal{B}, \mathcal{P}$  are the category, bounding box, and point annotations, respectively (see Section 3.1 for details).

However, active learning for instance segmentation has been less investigated. Wang *et al.* [54] first explored the possibility on medical image analysis, but, to the best of our knowledge, no existing work has studied this setting on more complex and larger-scale datasets (*e.g.*, MS-COCO [34]).

In this paper, we present a new setting named **active pointly-supervised instance segmentation (APIS)** to study active learning algorithms for instance segmentation with point supervisions. Under the proposed setting where each image in the data pool has been annotated with category labels and bounding boxes, the goal of the algorithms is selecting the most informative *points* for labeling to maximize the model’s performance. Fig. 1 illustrates the training pipeline of APIS. Compared to the typical active learning settings that the most informative *images* or *instances* are selected and annotated with boxes and masks, 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. As stated in a previous work [10], labeling whether a point falls on the specified object or not takes only 0.9 seconds on average. Note that active learning is a training strategy that point labels are only queried during the training phase of APIS, while no annotations are provided for testing.

APIS raises an important problem that has not been studied before, *i.e.*, *how to estimate the informativeness of a point*, where the informativeness can be roughly defined as the potential gain of segmentation accuracy (*e.g.*, in terms of mAP) if this point is annotated. However, estimating the precise accuracy gain for each point is obviously infeasible. In the literature, uncertainty is widely used to estimate the informativeness of an example. Inspired by this, we designed several metrics to estimate the uncertainty of a point based on the model’s prediction and select the most uncertain point of each instance for annotation.

The labeled points are used together with the category and box annotation to update (*e.g.*, fine-tuning) the instance segmentation model.

Extensive experiments on the challenging MS-COCO dataset demonstrate that the model trained with the actively acquired points performed consistently better than the model trained with randomly sampled points during each active learning step. *Especially, we found that entropy, a conceptually simple metric, works the best among all the proposed metrics for uncertainty estimation.* To understand the results of APIS, we went deep into the point selection process and studied the training dynamics, point distribution and point difficulties. The analyses reveal that the proposed sampling metric provides a good estimation of the point informativeness and therefore leads to higher performance. In addition to the random sampling baseline, we further compared APIS against a setting of active instance segmentation with full supervisions, where the most informative images or instances are selected for mask annotation. Conditioned on the same annotation budget and training time, the model developed under the APIS setting outperformed all other competitors. *The results suggest that active learning cooperates effectively with point supervisions, which can further boost the instance segmentation performance under a limited annotation budget.* We hope the promising results, as well as the comprehensive analysis presented in this work, will draw the attention of the community to APIS and other label-efficient visual recognition techniques.

The contribution of this work is summarized as follows:

- We present a new active instance segmentation setting APIS where the goal is to sample the most informative points to maximize the model’s performance. To our knowledge, this work is the first to explore active learning for instance segmentation with point supervisions.
- We estimate the informativeness with the uncertainty of point predictions, and the model trained with the actively acquired points consistently outperformed the random sampling counterparts on MS-COCO.
- We provide comprehensive comparisons and analyses to understand the results of APIS and concluded that APIS successfully combines the advantages of active learning and point-based supervision to reduce the annotation burden of instance segmentation.

## 2 Related Work

**Instance Segmentation.** Currently, the fully supervised methods still dominate the popular instance segmentation benchmarks [12,18,34]. Mask R-CNN [21] and its follow-up methods [7,23,35] predict masks based on the region-level features. One-stage methods like CondInst [52] and SOLO [56] directly segment instances at image-level by learning instance-aware kernels. Recently, some query-based methods [9,15,16] further boost the segmentation performance. For the weakly supervised paradigm, category tags [3,41,63] are the simplest supervision but the results are usually uncompetitive. With bounding box annotations, the recently proposed BoxInst [53] and DiscoBox [29] significantly outperformed

previous methods [22,27] on MS-COCO. In addition, PointSup [10] further reduced the gap to the fully-supervised methods by training with boxes and several randomly sampled points for each instance, which is similar the random sampling baseline in APIS where the points are accessed for training step by step. However, there are still great differences that APIS focus on active learning for instance segmentation while PointSup [10] is for weakly-supervised learning.

**Active Learning.** Over the past decades, a great number of active learning algorithms have been proposed but mostly designed for image classification, which can be roughly divided into two categories: uncertainty-based [5,17,25,31,55] and diversity-based [1,19,45,49] algorithms. In recent years, some researchers shifted their attention to the downstream visual recognition tasks such as object detection [2,11,20,26,36,44,60] and semantic segmentation [47,57,59]. However, for the task of instance segmentation, the potential of active learning has been less explored. Only one published work [54] preliminarily studied this problem on the medical image datasets where a triplet uncertainty metric was calculated with the predicted mask IoU scores of Mask Scoring R-CNN [23], which makes the metric relying on a specialized architecture. By contrast, we first studied this problem on the most commonly used MS-COCO dataset and the proposed metrics are model-agnostic. Furthermore, different from all existing settings where images or instances [13] are selected for labeling, the proposed APIS setting performs in a fine-grained and weakly supervised manner, *i.e.*, selecting and labeling points. A few works [14,40] studied weak supervisions for active object detection while they focused on how to decide the annotation scheme (strongly or weakly) for an image, which are complementary to APIS.

**Point-Based or Click-Based Segmentation.** Point-based supervision has been studied in various image segmentation tasks, including semantic segmentation [4,42], instance segmentation [10,30], and panoptic segmentation [32]. In addition, point clicks are widely used in interactive annotation [6,24,33,37,58] methods, which usually requires mask annotations for training the model. During testing (*i.e.*, annotating an unseen image), the user is asked to provide some corrective point clicks iteratively. APIS is intrinsically different to these methods that the model only interacts with users during training and no additional labels are required during testing.

### 3 Method

#### 3.1 Problem Formulation

In this section, we formally define the proposed active pointly-supervised instance segmentation (APIS) setting. Suppose we collect a large training dataset of  $N$  images denoted as  $\mathcal{D} = \{\mathcal{I}_i\}_{i=1}^N$  with annotations  $\{\mathcal{C}, \mathcal{B}\}$ , where  $\mathcal{C} = \{\mathcal{C}_i\}_{i=1}^N$  is the category annotation,  $\mathcal{B} = \{\mathcal{B}_i\}_{i=1}^N$  is the box annotation, and  $Q_i = |\mathcal{B}_i| = |\mathcal{C}_i|$  is the number of instances in the  $i^{\text{th}}$  image. We studied a typical setting of APIS that each instance is sampled with the same number of points and all the points should be located in the corresponding ground-truth bounding box. Ad-

ditionally, we also studied a scenario where *not* all instances were labeled with the same number of points, see Appendix C for details.

Before the active learning cycle starts, we randomly sample and annotate one point for each instance. The initial set of points is denoted as:

$$\mathcal{P}_0 = \{(x_{ij}^0, y_{ij}^0, u_{ij}^0) \mid 1 \leq i \leq N, 1 \leq j \leq Q_i\}, \quad (1)$$

where  $(x_{ij}^0, y_{ij}^0)$  is the coordinates of a point that is located in the  $j^{\text{th}}$  bounding box of the  $i^{\text{th}}$  image, and  $u_{ij}^0 \in \{0, 1\}$  is the point label which indicates whether the point falls on the foreground object or not. The size of  $\mathcal{P}_0$  is  $Q = \sum_{i=1}^N Q_i$  that equals to the total number of instances in  $\mathcal{D}$ . The instance segmentation model  $\mathcal{M}_0$  is initialized by training with the above annotations  $\{\mathcal{C}, \mathcal{B}, \mathcal{P}_0\}$ .

At the  $s^{\text{th}}$  ( $s \geq 1$ ) active learning step, the informativeness of points is estimated based on the predictions of the previous model  $\mathcal{M}_{s-1}$ . We designed several uncertainty-based metrics to estimate the point informativeness, which will be explained in Section 3.2. The most informative point of each instance is selected and the annotators are asked to label it. The newly labeled points are merged with  $\mathcal{P}_{s-1}$  and formed the new point set  $\mathcal{P}_s$ :

$$\mathcal{P}_s = \mathcal{P}_{s-1} \cup \{(x_{ij}^s, y_{ij}^s, u_{ij}^s) \mid 1 \leq i \leq N, 1 \leq j \leq Q_i\}, \quad (2)$$

where the number of points in  $\mathcal{P}_s$  is  $(s+1) \times Q$ . Subsequently, the model  $\mathcal{M}_s$  is fine-tuned from  $\mathcal{M}_{s-1}$  with all available annotations  $\{\mathcal{C}, \mathcal{B}, \mathcal{P}_s\}$ . The above process is repeated multiple steps until the annotation budget has been exhausted or a satisfactory performance has been achieved.

### 3.2 Point Selection for APIS

In this section, we tackle the core problem raised by APIS that how to define a point’s informativeness. Note that informativeness is not totally determined by correctness, *e.g.*, if the model predicts a point as positive with a high confidence while the ground-truth is negative, annotating it may not be an ideal solution (see Sec. 4.3). That said, even provided the mask labels, it is still difficult to determine which point has a potentially large contribution. We refer to some existing active learning methods [5,17,25,31,55] to study uncertainty, and use the uncertainty of points to rank their informativeness.

In modern instance segmentation models, a ground-truth instance is usually assigned as the learning target for multiple predictions during training, known as label assignment [62], which makes NMS (Non-Maximum Supression) being a necessary process during inference. Suppose there are  $K$  mask predictions  $\{\mathbf{m}^k\}_{k=1}^K$  matching to a given instance in  $\mathcal{D}$ , where  $\mathbf{m}^k \in [0, 1]^{H \times W}$  is an image-level probability matrix and  $H \times W$  is the shape of the image. Note that for CondInst [52] the prediction is already image-level while for Mask R-CNN [21] the region-level prediction should be transformed to image-level. Denote the element of  $\mathbf{m}^k$  with coordinates  $(x, y)$  as  $p_{xy}^k$  that indicates the probability of a point located at  $(x, y)$  falling on the object. We designed several metrics to estimate the uncertainty for a point based on its predictive probability of  $p_{xy}^k$ .

**(1) Entropy of the Averaged Predictions.** The Shannon Entropy [46] metric is commonly used in existing active learning algorithms [2,17,20,55]. In our case, since mask prediction is a binary classification problem for points, the entropy metric can be defined as:

$$\mathcal{H}(p) = -p \log p - (1-p) \log(1-p), \quad (3)$$

where  $p$  is the probability of any point. We simply average multiple predictions on the same point to calculate the entropy value, and the point with the highest entropy value is selected for the corresponding instance:

$$(\hat{x}, \hat{y}) = \arg \max_{(x,y) \in \Omega} \mathcal{H}(\bar{p}_{xy}), \quad \bar{p}_{xy} = \frac{1}{K} \sum_{k=1}^K p_{xy}^k, \quad (4)$$

where  $(\hat{x}, \hat{y})$  is the coordinates of the actively selected point for the given instance and  $\Omega$  is the spatial constraint for the point candidates. In our setting, the selected points are expected to fall inside the ground-truth bounding boxes. The intuition behind the above sampling strategy is straightforward yet reasonable, *i.e.*, *the more the probability close to 0.5, the higher the entropy, and the more likely the point should be selected*. Designing the constraint  $\Omega$  carefully or averaging multiple predictions with some adaptive weights can make the above strategy more sophisticated, however it is not the focus of this work.

**(2) Disagreement Among Multiple Predictions.** The intuition behind the disagreement metric is that *if multiple predictions for the same point varied significantly, the model should be highly uncertain about that point and we should select that point for labeling*. This idea can be traced back to the classical *query by committee* paradigm [38]. A number of works implicitly or explicitly followed this idea to select samples actively. For example, in the task of active object detection, the offsets between the boxes generated from different SSD layers [44], or the IoU scores between the proposals and final detected boxes [26] were used to measure the disagreement. In our case, we adopt the *variance* across different predictions for a point to measure the disagreement and the point with the largest *variance* is selected for the corresponding instance:

$$(\hat{x}, \hat{y}) = \arg \max_{(x,y) \in \Omega} \mathbb{V}(\mathbf{p}_{xy}), \quad (5)$$

where  $\mathbf{p}_{xy} \in \mathbb{R}^K$  is the probability vector at the location  $(x, y)$  and  $\mathbb{V}(\cdot)$  calculates the variance of it.

In addition to sampling metrics mentioned above, how to form the prediction set  $\{\mathbf{m}^k\}_{k=1}^K$  for a given instance is an important problem. In this work, by considering the properties of the APIS setting, we propose and compare several solutions. (a) *Multiple Anchors*: Since a ground-truth instance is usually assigned to multiple anchors (where anchors are the positive locations for CondInst [52], or the positive proposals for Mask R-CNN [21], etc.) during training, we naturally have multiple predictions for one instance. (b) *Multiple Models*: Another solution is training the same models multiple times with different initializations

to get multiple predictions, which is usually called Deep Ensembles [5,28] in the literature. (c) *Multiple Scales*: Training multiple models is computationally inefficient. An alternative is to forward the same model multiple times under different conditions. In our case, the model is usually trained with multi-scale inputs, thus we can forward the model with an individual scale each time to get multiple predictions. The concept of multi-scale here can be freely replaced by other types of data augmentations (*e.g.*, flipping, rotation).

We calculate the entropy and disagreement metrics on the above prediction sets individually and obtain several different point sampling strategies. Note that the case of multiple anchors also appear in each model or each forward pass, and we simply concatenate them as a single prediction set. For the instance that was not predicted ( $K = 0$ ), we randomly sample a point for it. For the instance that has only one prediction ( $K = 1$ ), we only use entropy as the metric.

### 3.3 Baseline for APIS

Note that there is no existing work can be directly compared to APIS since the problem of active instance segmentation has been less investigated. In addition to the *random sampling* baseline, inspired by the existing works for active image classification or object detection where the full labels are usually queried, we create a baseline setting for comparison, named *active fully-supervised instance segmentation (AFIS)*, where the mask annotations are queried during each active learning step. Note that the category and box annotations  $\{\mathcal{C}, \mathcal{B}\}$  are also provided in advance. We designed several sampling strategies under the AFIS setting, which will serve as the baselines for APIS. A straightforward strategy is selecting the most informative *images* and labeling masks for all instances in the image. An alternative is selecting and annotating the most informative *instances*, which is a fine-grained solution that not all instances in the image are labeled. We call them *image-level* selection and *instance-level* selection for AFIS. See Appendix A for more description of AFIS.

Under the AFIS setting, we propose two metrics to define the informativeness of an image or an instance. (a) *Mean Entropy*: Inspired by the aforementioned uncertainty-based metrics of APIS, we defined the instance uncertainty as the mean entropy of all points inside the ground-truth bounding box, and the uncertainty of an image is defined as the mean uncertainty of all instances in that image. The most uncertain images and instances are selected for the image-level and instance-level cases, respectively. (b) *Detection Quality*: The quality of mask prediction is usually dependent on the detection quality. For an instance that was not accurately detected, the potential of the mask label, if provided, may not be fully utilized. Therefore, we propose to select the instances with the higher detection quality for labeling masks. Since the ground-truth boxes are provided, we can easily use the detection loss (*e.g.*, GIoU Loss [43]) as the metric for instance selection. For the image-level selection, we calculate the average detection loss of all instances for an image, and the images with the lowest loss are selected.

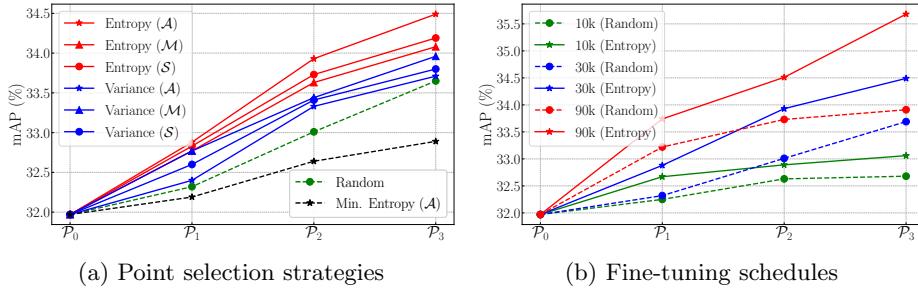


Fig. 2: (a) Comparison of different point selection strategies.  $\mathcal{A}$ ,  $\mathcal{M}$  and  $\mathcal{S}$  indicate the prediction sets constructed from multiple anchors, multiple models, and multiple scales, respectively. (b) Comparison of different fine-tuning schedules.

## 4 Experiments

### 4.1 Experimental Settings

We report results on the MS-COCO [34] dataset. All the active selection strategies were applied on the `train2017` split, including 118k images with 860k instances, where the annotation for points was simulated by adopting the labels of the ground-truth instance masks on the corresponding location. All models were evaluated on the `val2017` split.

**Implementation Details.** We mainly took CondInst [52] with ResNet-50 as the instance segmentation model and adopted the `AdelaiDet` [51] codebase. To train the model with point supervision, we followed the same training protocol as PointSup [10] where the point prediction was sampled from the prediction map using bilinear interpolation and the per-pixel cross-entropy loss was calculated on the labeled points only. The initial model  $\mathcal{M}_0$  was trained for 90k iterations with the initial learning rate being 0.01 and decayed at iteration 60k and 80k, respectively. Other training settings were the same as CondInst. After that, the active learning process was repeated multiple times. At the  $s^{\text{th}}$  active learning step,  $(s + 1)$  points were labeled in total for each instance where  $s$  points were actively acquired, denoted as  $\mathcal{P}_s$  in the following experiments. By default, at each step, the model was fine-tuned for 30k iterations with the initial learning rate being 0.01 and decayed every 10k iterations. All models were trained with the SGD optimizer and multi-scale data augmentation with mini-batch of 16 images on 8 NVIDIA Tesla V100 GPUs. The results were measured by the mask mAP(%) metric of instance segmentation.

### 4.2 Design Principles of APIS

**Point Selection Strategy.** As introduced in Section 3.2, we propose two metrics, named *Entropy* and *Variance* (*i.e.*, the disagreement) here, to estimate the point uncertainty based on the prediction set, and there are three types of prediction sets: *Multiple-Anchors* ( $\mathcal{A}$ ), *Multiple-Models* ( $\mathcal{M}$ ), and *Multiple-Scales* ( $\mathcal{S}$ ).

We composed them into six different point selection strategies, and we observed that all these six strategies performed consistently better than the random sampling baseline, as shown in Fig. 2a. The results demonstrate the effectiveness of active point selection. In addition, the strategies that taking *Entropy* as the metric usually performed better than the *Variance* counterparts. When using *Entropy* as the metric, constructing the prediction set from multiple anchors ( $\mathcal{A}$ ) performed better than from multiple models ( $\mathcal{M}$ ) or scales ( $\mathcal{S}$ ), and the latter two solutions are obviously inefficient in computation. The best-performing strategy is calculating the entropy value for each point based on the predictions from multiple anchors, which surpassed random sampling by 0.56%, 0.92% and 0.80% mAP at the first three steps. Surprisingly, the result at the second step ( $\mathcal{P}_2$ ) even outperformed the random sampling counterparts with larger annotation costs and longer training time ( $\mathcal{P}_3$ ). Unless otherwise specified, we use *Entropy* to denote this best-performing strategy in the rest of this paper.

For the *Entropy* strategy mentioned above, we always chose the most uncertain point of each instance for labeling, and we were also curious about the inverse situation that the most certain points (*i.e.*, point with the lowest entropy value) were selected instead. As shown in Fig. 2a (black dashed line), the points with the lowest entropy even performed worse than the randomly sampled points, which verified that the proposed entropy metric is a simple yet effective way to estimate the point informativeness.

**Fine-Tuning Schedule.** We compared different fine-tuning schedules at each active learning step. For the 10k schedule ( $\sim 1.3$  epoch), the learning rate was fixed at 0.0001. For the 90k schedule (12 epochs), the initial learning rate was 0.01 and reduced by a factor of 10 at iteration 60k and 80k, respectively. As shown in Fig. 2b, we observed that the longer the training time, the larger the gaps between active point selection and random sampling. When adopting the 90k schedule, the gap reached to +1.77% at the third step ( $\mathcal{P}_3$ ). The results suggest that the actively acquired point is indeed more informative, and the longer training time can further release its potential. In this paper, unless otherwise specified, the 30k schedule was used in experiments for convenience.

**Instance Segmentation Model.** In our experiments, we mainly used the CondInst with ResNet-50 backbone as the instance segmentation model. Actually, the proposed APIS setting can be studied with a diverse set of models, and we also studied two alternatives. (a) *Larger backbone*: Using the higher-capacity ResNet-101 as the backbone, the similar observation can be made that active selection works consistently better than random sampling, as shown in Fig. 3a. (b) *Boxly-supervised model*: Recall that  $\mathcal{P}_0$  in above experiments is a set of randomly sampled points which is used to initialize the model to obtain the initial mask predictions. A number of works [22, 27, 29, 53] attempted to predict masks by training with box annotations only. By adopting these models as the initial model in APIS, we can eliminate the need of  $\mathcal{P}_0$ . In Fig. 3b, we adopted BoxInst [53], a dominant method in the boxly-supervised instance segmentation area, as the initial model, and each instance was labeled with  $s$  points in  $\mathcal{P}'_s$  (in comparison,  $s+1$  in  $\mathcal{P}_s$ ). As shown, the proposed strategy also worked in this

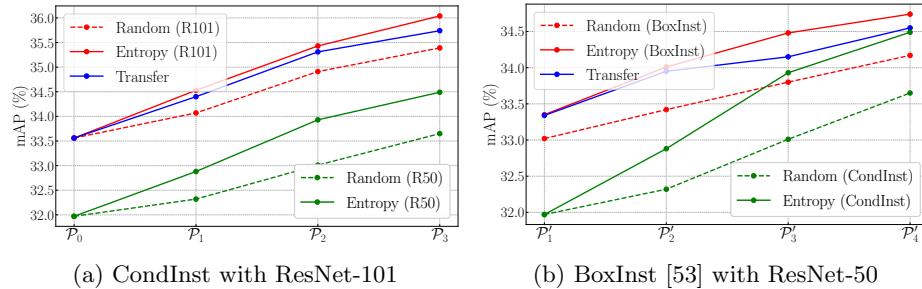


Fig. 3: The results of taking CondInst with ResNet-101 and BoxInst [53] as the instance segmentation model, respectively. *Transfer* indicates that the points are acquired from CondInst with ResNet-50 and transferred to this model.

case. With the power of BoxInst, similar or even better results were achieved with fewer points, *e.g.*, the result of  $\mathcal{P}'_1$  surpassed  $\mathcal{P}_1$  (in Fig. 2a) by 0.47% mAP, although  $\mathcal{P}_1$  has more points.

**Transferability.** We studied the transferability of the actively acquired points. As shown in Fig. 3 (blue lines), we transferred the points acquired from one model (CondInst with ResNet-50) to other models, CondInst with ResNet-101 and BoxInst, respectively. Specifically, the point set  $\mathcal{P}_s$  of the former model served as the supervision for the latter two models in the  $s^{\text{th}}$  step. As shown, the results of the transferred points are slightly lower than the results of selecting points for those models from scratch, but still higher than random sampling.

### 4.3 Analysis for APIS

**Visualization Analysis.** To better understand how the point is selected and how it works, we visualized the mask predictions, uncertainty maps, and the selected points for some instances, as shown in Fig. 4a. We observed that the highlighted regions in uncertainty maps often corresponded to the two types of mistakes in the mask predictions. One is the typical over-segmented or under-segmented regions (*e.g.*, holes on the object, or patches on the background). Another one is the error around the boundaries of the predicted masks (*i.e.*, imprecise boundaries). Therefore, the points selected from these misclassified regions can provide a valuable feedback about how the model performs and it is possible to correct the mistakes by fine-tuning with the selected points in the subsequent active learning steps (*e.g.*, the last column). Interestingly, some prediction errors were still corrected even without sampling the points from those regions (*e.g.*, dashed boxes in the 3<sup>rd</sup> row). The reason might be that the patterns of these error regions also appeared in other instances that have the desired annotations. On the other side, there were also some failure cases where the prediction around the selected point even got worse after fine-tuning (*e.g.*, the 4<sup>th</sup> row). More visualized results are included in Appendix B.

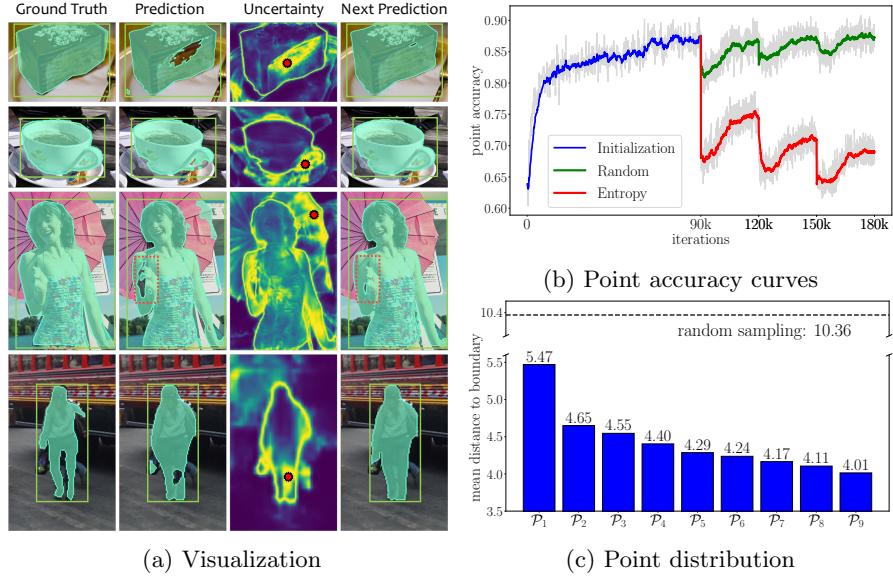


Fig. 4: (a) Visualization of (from left to right): ground-truth masks, mask predictions, uncertainty maps, and mask predictions after fine-tuning with the selected points (red spots) for some instances. (b) Accuracy curves of the actively acquired points and random points during training. (c) The mean distances to the object boundaries of the actively acquired points for each step. The mean distance for random points is provided for reference (dash line).

**Point Accuracy Curves.** In addition to the case studies above, the accuracy (*i.e.*, accuracy of binary classification for points during training) curves of the actively acquired points and randomly sampled points are plotted in Fig. 4b. As shown, the accuracy of the actively acquired points dropped dramatically at the beginning of each step (iteration 90k, 120k and 150k, respectively), which shows that the selected points were often misclassified at the previous step and it is possible to correct the errors by fine-tuning the model with their labels (*e.g.*, the first three rows of Fig. 4a). In contrast, the accuracy of random points dropped slightly at each time and always stayed at a high level, which indicates that most of these points were already handled by the model and of course less informative. In addition, the accuracy improvements usually decreased with more steps. It suggests that, with more steps, the mask prediction gets gradually better and the selected point usually gets harder.

**Point Distribution.** We calculated the distance between the actively acquired points and the ground-truth instance boundaries, and show the mean distances at each step in Fig. 4c. As shown, the selected points were usually closer to boundaries than the random points and the mean distance became smaller with more steps. This is as expected — as the model gets gradually better (*e.g.*,

Table 1: Sampling points with different difficulties.  $\star$  indicates that mask labels were used for selection.

Strategy	$\mathcal{P}_0$	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$
Random	31.97	32.32	33.01	33.69
Entropy	31.97	<b>32.88</b>	<b>33.93</b>	<b>34.49</b>
*Max Error	31.97	7.95	12.45	14.24
*Least Error	31.97	32.16	32.50	32.95

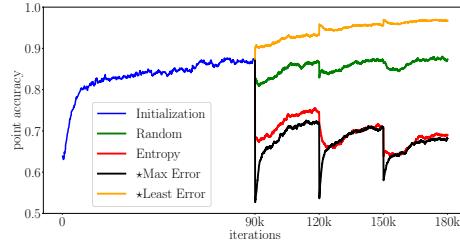


Fig. 5: The point accuracy (measured on `train2017`) curves of sampling points with different difficulties. At the first step ( $\mathcal{P}_1$ ), the misclassification ratio was 82% and 7% for *Max Error* and *Least Error*, respectively.

over/under-segmented regions has been corrected), the predicted boundaries get closer to the actual object boundaries, thus the remaining high-entropy points are mostly located around object boundaries (see the 2<sup>nd</sup> column of Fig. 4a). However, the points around object boundaries are inherently hard to classify even with full supervisions, as studied in previous works [50,61], and their labels might be noisy due to the coarse polygon-based mask annotations in MS-COCO. In summary, with more steps, the algorithm tends to select points around object boundaries, yet these annotations, though being difficult, often bring marginal performance gain, which confirmed the observations in above analysis.

**Point Difficulty.** From the accuracy decrements (after adding new points) in Fig. 4b we can calculate that about 51% of the actively acquired points ( $\mathcal{P}_1$ ) were misclassified by the previous model ( $\mathcal{M}_0$ ), while the ratio is 23% for random sampling, which suggests that the actively acquired points are more difficult for the model to learn. To study the influence of point difficulty, two experiments were conducted where we selected the points with maximum error or with minimum error at each step. The results were compared in terms of both mask mAP (Table 1) and point accuracy (Fig. 5). For *Least Error*, although the point accuracy (orange line) always stayed at a high level, the results of mAP still lagged behind random sampling, which suggests that these well-classified points were too easy for the model and of course less informative. For *Max Error*, the results of mAP were extremely poor while the point accuracy (black line) still increased during each training step. The reason might be that these points were mostly the hard cases and training with them directly made the model overfitted. Two conclusions can be drawn from the above results: (a) Point difficulty can heavily impact the performance of active learning, where neither the easiest nor the most difficult points should be selected. The proposed *Entropy* metric achieves a reasonable balance in this factor. (b) APIS is a challenging but non-trivial problem. It is still difficult to determine which point has a potentially large contribution even provided the mask labels. The results and analyses suggest that uncertainty is a better solution towards the right direction than correctness.

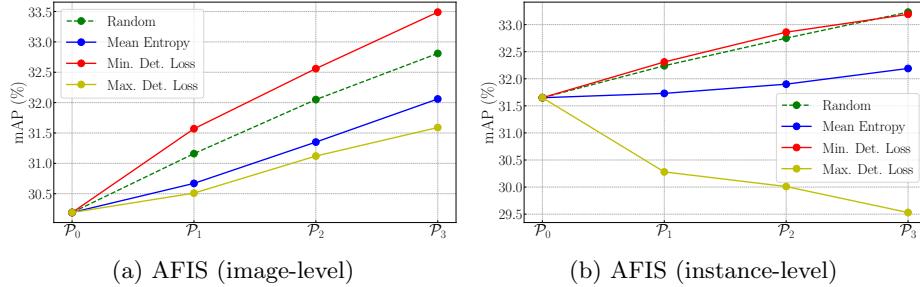


Fig. 6: Comparison of different image-level and instance-level selection strategies designed for AFIS. See Sec. 3.3 and Appendix A for details.

#### 4.4 Comparison to Other Learning Strategies

**Comparison of APIS and AFIS.** As introduced in Sec. 3.3, a baseline setting AFIS was established for comparison. For fair comparison, the annotation budget and training time should be the same with APIS during each step. As stated in previous works [10,34], it takes 0.9 seconds on average to label a point and 79.2 seconds to create a polygon-based instance mask in MS-COCO. In above experiments, one point was labeled for each instance at each step, thus 860000 points in total. If the same budget is allocated to instances, we can annotate masks for  $860000/(79.2/0.9) \approx 9773$  instances. Unlike previous works that the annotation cost for different images are treated identical, in our case the cost is proportional to the number of instances in the image, which varies across images. Therefore, given a fixed annotation budget, the number of annotated images is dependent on the sampling strategy.

We compared different sampling strategies for the case of *image-level* selection and *instance-level* selection, respectively.  $\mathcal{P}_s$  here indicates that the cost for labeling images or instances is exactly the same as  $\mathcal{P}_s$  in APIS. Similarly, some images or instances were randomly selected (with the same budget as  $\mathcal{P}_0$ ) for model initialization. As shown in Fig. 6, the results of the *Mean Entropy* strategy were unsatisfactory in both cases, even lagging behind random sampling. On the other hand, the strategy that selecting instances with the lowest detection loss (*Min. Det. Loss*) usually produced on par or even better results than other competitors. Since AFIS is not the focus of this paper, we provided more results and analyses of AFIS in Appendix A.

The best-performing strategies found above were adopted for comparison in Fig. 7. Three conclusions can be drawn from the results: (1) Point-based supervision is an effective way for training instance segmentation models. Even with random points, the model can still outperformed those trained with mask supervisions under the same annotation costs and training time. (2) Active learning is a label-efficient training strategy for instance segmentation, especially when the annotation budget is limited. As shown, the active selection strategy usually performed better than random sampling in all settings. (3) The proposed

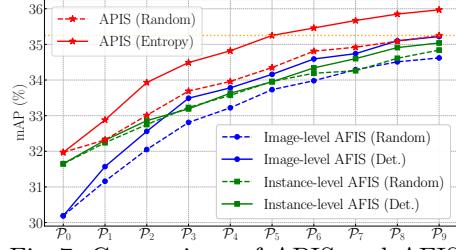


Fig. 7: Comparison of APIS and AFIS with the same annotation budget and training time. Det. indicates the *Min. Det. Loss* strategy.

Table 2: Comparison of APIS and weakly-supervised instance segmentation methods (with ResNet-50 backbone). †: our impl. (no point aug.).

Method	Anno.	Iter.	mAP
CondInst [52]	fully sup.	270k	37.5
DiscoBox [29]	$\{\mathcal{C}, \mathcal{B}\}$	270k	31.4
BoxInst [53]	$\{\mathcal{C}, \mathcal{B}\}$	270k	31.8
PointSup <sup>†</sup> [10]	$\{\mathcal{C}, \mathcal{B}, \mathcal{P}_{10}\}$	270k	35.4
APIS (ours)	$\{\mathcal{C}, \mathcal{B}, \mathcal{P}_7\}$	270k	35.4
APIS (ours)	$\{\mathcal{C}, \mathcal{B}, \mathcal{P}_{10}\}$	360k	36.0

APIS setting, combining active learning and point-based supervision, is a more powerful yet economic choice to train instance segmentation models under limited annotation budgets. The model developed under this setting consistently outperformed all other competitors at each active learning step. For example, at the 5<sup>th</sup> step ( $\mathcal{P}_5$ ), our model achieved on par or even better results than other models with more points and longer training time (e.g.,  $\mathcal{P}_9$ ).

**Comparison to Weakly-Supervised Methods.** In Table 2, we compared APIS with some existing weakly-supervised instance segmentation methods. Compared to boxly-supervised methods [29,53], labeling points additionally usually leads to considerable improvements. Compared to PointSup [10] (based on CondInst with ResNet50) trained with 10 randomly sampled points, APIS achieved the same results with 3 fewer points per instance. We argue that random sampling not considered the informativeness of different points, thus leading to sub-optimal results. In spite of this, we have to realize that APIS and PointSup varied significantly in motivation, which optimized for active learning and weakly-supervised learning, respectively.

## 5 Conclusion

In this paper, we propose APIS, a new active learning setting for instance segmentation where the most informative points are selected for annotation. We formulate this setting and propose several sampling strategies. Extensive experiments and detailed analysis on MS-COCO demonstrate that APIS is a powerful but economic learning strategy for training instance segmentation models with limited annotation budgets. We hope this work could inspire future researches on related topics, e.g., point-based supervision and label-efficient learning.

**Acknowledgements.** This work was supported in part by the National Natural Science Foundation of China (Nos. U19B2034, 62061136001 and 61836014).

## References

1. Agarwal, S., Arora, H., Anand, S., Arora, C.: Contextual diversity for active learning. In: Eur. Conf. Comput. Vis. pp. 137–153 (2020)
2. Aghdam, H.H., Gonzalez-Garcia, A., Weijer, J.v.d., López, A.M.: Active learning for deep detection neural networks. In: Int. Conf. Comput. Vis. pp. 3672–3680 (2019)
3. Arun, A., Jawahar, C., Kumar, M.P.: Weakly supervised instance segmentation by learning annotation consistent instances. In: Eur. Conf. Comput. Vis. pp. 254–270 (2020)
4. Bearman, A., Russakovsky, O., Ferrari, V., Fei-Fei, L.: What’s the point: Semantic segmentation with point supervision. In: Eur. Conf. Comput. Vis. pp. 549–565 (2016)
5. Beluch, W.H., Genewein, T., Nürnberg, A., Köhler, J.M.: The power of ensembles for active learning in image classification. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 9368–9377 (2018)
6. Benenson, R., Popov, S., Ferrari, V.: Large-scale interactive object segmentation with human annotators. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 11700–11709 (2019)
7. Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Shi, J., Ouyang, W., Loy, C.C., Lin, D.: Hybrid task cascade for instance segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4974–4983 (2019)
8. 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)
9. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1290–1299 (2022)
10. Cheng, B., Parkhi, O., Kirillov, A.: Pointly-supervised instance segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2617–2626 (2022)
11. Choi, J., Elezi, I., Lee, H.J., Farabet, C., Alvarez, J.M.: Active learning for deep object detection via probabilistic modeling. In: Int. Conf. Comput. Vis. (2021)
12. 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)
13. Desai, S.V., Balasubramanian, V.N.: Towards fine-grained sampling for active learning in object detection. In: IEEE Conf. Comput. Vis. Pattern Recog. Worksh. pp. 924–925 (2020)
14. Desai, S.V., Chandra, A.L., Guo, W., Ninomiya, S., Balasubramanian, V.N.: An adaptive supervision framework for active learning in object detection. In: Brit. Mach. Vis. Conf. (2019)
15. Dong, B., Zeng, F., Wang, T., Zhang, X., Wei, Y.: Solq: Segmenting objects by learning queries. Adv. Neural Inform. Process. Syst. (2021)
16. Fang, Y., Yang, S., Wang, X., Li, Y., Fang, C., Shan, Y., Feng, B., Liu, W.: Instances as queries. In: Int. Conf. Comput. Vis. pp. 6910–6919 (2021)
17. Gal, Y., Islam, R., Ghahramani, Z.: Deep bayesian active learning with image data. In: Int. Conf. Machine Learning. pp. 1183–1192 (2017)
18. Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5356–5364 (2019)

19. Hasan, M., Roy-Chowdhury, A.K.: Context aware active learning of activity recognition models. In: Int. Conf. Comput. Vis. pp. 4543–4551 (2015)
20. 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)
21. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Int. Conf. Comput. Vis. pp. 2961–2969 (2017)
22. Hsu, C.C., Hsu, K.J., Tsai, C.C., Lin, Y.Y., Chuang, Y.Y.: Weakly supervised instance segmentation using the bounding box tightness prior. In: Adv. Neural Inform. Process. Syst. pp. 6586–6597 (2019)
23. Huang, Z., Huang, L., Gong, Y., Huang, C., Wang, X.: Mask scoring r-cnn. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 6409–6418 (2019)
24. Jang, W.D., Kim, C.S.: Interactive image segmentation via backpropagating refinement scheme. IEEE Conf. Comput. Vis. Pattern Recog. pp. 5292–5301 (2019)
25. Joshi, A.J., Porikli, F., Papanikolopoulos, N.: Multi-class active learning for image classification. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2372–2379 (2009)
26. Kao, C.C., Lee, T.Y., Sen, P., Liu, M.Y.: Localization-aware active learning for object detection. In: Asian Conference on Computer Vision. pp. 506–522 (2018)
27. Khoreva, A., Benenson, R., Hosang, J., Hein, M., Schiele, B.: Simple does it: Weakly supervised instance and semantic segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 876–885 (2017)
28. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Adv. Neural Inform. Process. Syst. (2017)
29. Lan, S., Yu, Z., Choy, C., Radhakrishnan, S., Liu, G., Zhu, Y., Davis, L.S., Anandkumar, A.: Discobox: Weakly supervised instance segmentation and semantic correspondence from box supervision. In: Int. Conf. Comput. Vis. (2021)
30. Laradji, I.H., Rostamzadeh, N., Pinheiro, P.O., Vazquez, D., Schmidt, M.: Proposal-based instance segmentation with point supervision. In: IEEE Int. Conf. Image Process. pp. 2126–2130 (2020)
31. Lewis, D.D., Catlett, J.: Heterogeneous uncertainty sampling for supervised learning. In: Int. Conf. Machine Learning (1994)
32. Li, Y., Zhao, H., Qi, X., Chen, Y., Qi, L., Wang, L., Li, Z., Sun, J., Jia, J.: Fully convolutional networks for panoptic segmentation with point-based supervision. arXiv preprint arXiv:2108.07682 (2021)
33. Li, Z., Chen, Q., Koltun, V.: Interactive image segmentation with latent diversity. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 577–585 (2018)
34. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Eur. Conf. Comput. Vis. pp. 740–755 (2014)
35. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 8759–8768 (2018)
36. Liu, Z., Ding, H., Zhong, H., Li, W., Dai, J., He, C.: Influence selection for active learning. In: Int. Conf. Comput. Vis. pp. 9274–9283 (2021)
37. Maninis, K.K., Caelles, S., Pont-Tuset, J., Van Gool, L.: Deep extreme cut: From extreme points to object segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 616–625 (2018)
38. Melville, P., Mooney, R.J.: Diverse ensembles for active learning. In: Int. Conf. Machine Learning (2004)
39. Papadopoulos, D.P., Uijlings, J.R., Keller, F., Ferrari, V.: Extreme clicking for efficient object annotation. In: Int. Conf. Comput. Vis. pp. 4930–4939 (2017)

40. Pardo, A., Xu, M., Thabet, A., Arbelaez, P., Ghanem, B.: Baod: budget-aware object detection. In: IEEE Conf. Comput. Vis. Pattern Recog. Worksh. pp. 1247–1256 (2021)
41. Pont-Tuset, J., Arbelaez, P., Barron, J.T., Marques, F., Malik, J.: Multiscale combinatorial grouping for image segmentation and object proposal generation. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(1), 128–140 (2016)
42. Qian, R., Wei, Y., Shi, H., Li, J., Liu, J., Huang, T.: Weakly supervised scene parsing with point-based distance metric learning. In: AAAI. pp. 8843–8850 (2019)
43. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 658–666 (2019)
44. Roy, S., Unmesh, A., Namboodiri, V.P.: Deep active learning for object detection. In: Brit. Mach. Vis. Conf. (2018)
45. Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. arXiv preprint arXiv:1708.00489 (2017)
46. Shannon, C.E.: A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review* **5**(1), 3–55 (2001)
47. Shin, G., Xie, W., Albanie, S.: All you need are a few pixels: Semantic segmentation with pixelpick. In: Int. Conf. Comput. Vis. Worksh. pp. 1687–1697 (2021)
48. 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)
49. Sinha, S., Ebrahimi, S., Darrell, T.: Variational adversarial active learning. In: Int. Conf. Comput. Vis. pp. 5972–5981 (2019)
50. Tang, C., Chen, H., Li, X., Li, J., Zhang, Z., Hu, X.: Look closer to segment better: Boundary patch refinement for instance segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 13926–13935 (2021)
51. Tian, Z., Chen, H., Wang, X., Liu, Y., Shen, C.: AdelaiDet: A toolbox for instance-level recognition tasks. <https://git.io/adelaidet> (2019)
52. Tian, Z., Shen, C., Chen, H.: Conditional convolutions for instance segmentation. In: Eur. Conf. Comput. Vis. pp. 282–298 (2020)
53. Tian, Z., Shen, C., Wang, X., Chen, H.: Boxinst: High-performance instance segmentation with box annotations. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5443–5452 (2021)
54. Wang, J., Wen, S., Chen, K., Yu, J., Zhou, X., Gao, P., Li, C., Xie, G.: Semi-supervised active learning for instance segmentation via scoring predictions. In: Brit. Mach. Vis. Conf. (2020)
55. Wang, K., Zhang, D., Li, Y., Zhang, R., Lin, L.: Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology* **27**(12), 2591–2600 (2016)
56. Wang, X., Zhang, R., Shen, C., Kong, T., Li, L.: Solo: A simple framework for instance segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* (2021)
57. Wu, T.H., Liu, Y.C., Huang, Y.K., Lee, H.Y., Su, H.T., Huang, P.C., Hsu, W.H.: Redal: Region-based and diversity-aware active learning for point cloud semantic segmentation. In: Int. Conf. Comput. Vis. pp. 15510–15519 (2021)
58. Xu, N., Price, B., Cohen, S., Yang, J., Huang, T.S.: Deep interactive object selection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 373–381 (2016)
59. Yang, L., Zhang, Y., Chen, J., Zhang, S., Chen, D.Z.: Suggestive annotation: A deep active learning framework for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 399–407 (2017)

60. Yuan, T., Wan, F., Fu, M., Liu, J., Xu, S., Ji, X., Ye, Q.: Multiple instance active learning for object detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5330–5339 (2021)
61. Zhang, G., Lu, X., Tan, J., Li, J., Zhang, Z., Li, Q., Hu, X.: Refinemask: Towards high-quality instance segmentation with fine-grained features. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 6861–6869 (2021)
62. Zhu, B., Wang, J., Jiang, Z., Zong, F., Liu, S., Li, Z., Sun, J.: Autoassign: Differentiable label assignment for dense object detection. arXiv preprint arXiv:2007.03496 (2020)
63. Zhu, Y., Zhou, Y., Xu, H., Ye, Q., Doermann, D., Jiao, J.: Learning instance activation maps for weakly supervised instance segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3116–3125 (2019)

## Appendix

### 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, 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<sup>2</sup>, 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 [48], 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

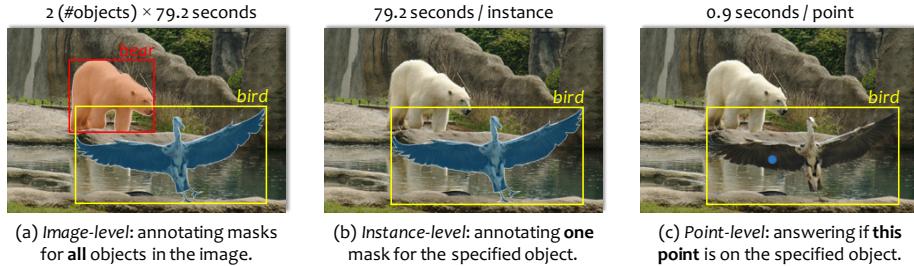


Fig. S1: Illustration of different annotation schemes (*image-level*, *instance-level*, and *point-level*) from the perspective of human annotators. The averaged annotation time for these schemes is provided for reference (values adopted from [10]).



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

the highest detection 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 [20] since the annotation costs for different images are considered the same in their experiments. In our setting, the annotation cost is proportional to the number of objects in the image, which is closer to the real-world scenarios.



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

**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  $\Delta AP_L$ ) 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 [12]. The results (red lines in Fig. S4) validate the same conclusion. In this study, we follow the previous work [21] 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 [8] to integrate, which we leave for future work.

**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).

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

	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$
$\Delta AP$	+0.56	+0.92	+0.80
$\Delta AP_S$	-0.30	+0.37	+0.28
$\Delta AP_M$	+0.69	+0.93	+0.66
$\Delta AP_L$	<b>+1.17</b>	<b>+1.53</b>	<b>+1.33</b>

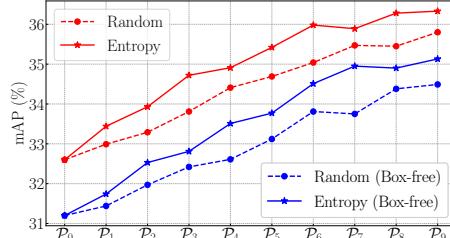


Fig. S4: APIS results on the Cityscapes dataset. Blue lines indicate the box-free APIS results.

## C APIS with Fewer Labeled Points

In above experiments, 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 ( $8\times$  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).

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 (32.0% for  $\mathcal{P}_0$ ), but at the cost of longer training time ( $2\times$ ). 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.

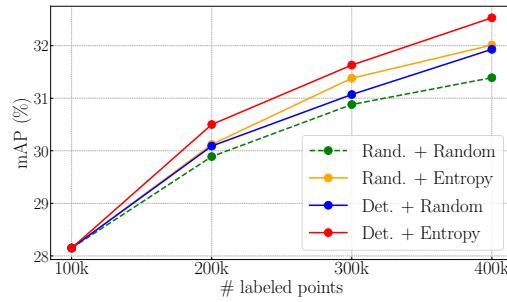


Fig. S6: The results of APIS with fewer labeled points. Rand. and Det. indicate random sampling and the *Min. Det. Loss* strategy, respectively.

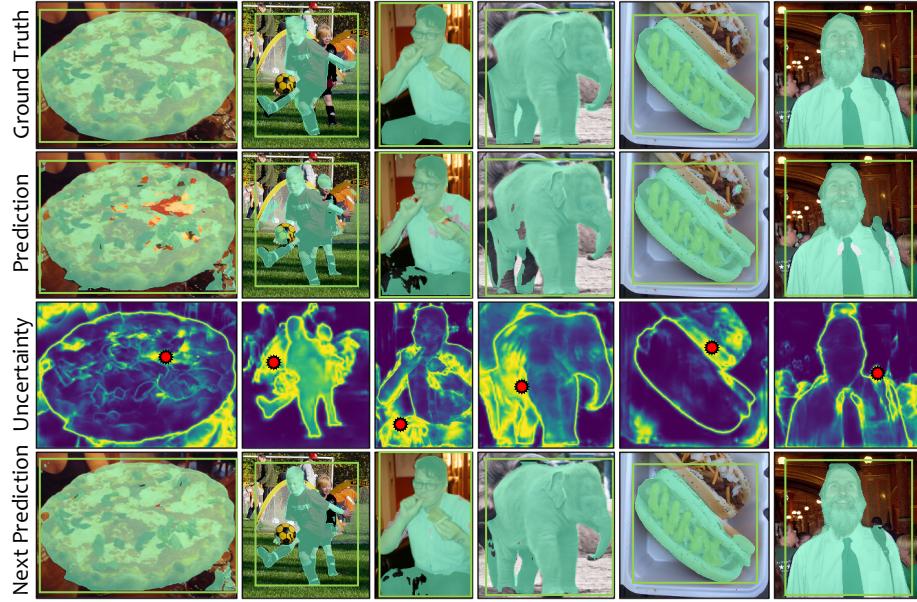


Fig. S5: Visualization of (from top to bottom): ground-truth masks, mask predictions (averaged over multiple predictions), uncertainty maps (the brighter the more uncertain), and mask predictions after fine-tuning with the selected points for some instances (extension of Fig. 4a). The red spots indicate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