# Supplementary Material: Revisiting Superpixels for Active Learning in Semantic Segmentation with Realistic Annotation Costs

Lile Cai[1], Xun Xu[1], Jun Hao Liew[2], Chuan Sheng Foo[1]
[1]Institute for Infocomm Research, Singapore
[2]National University of Singapore
{caill, foo_chuan_sheng}@i2r.a-star.edu.sg, alex.xun.xu@gmail.com, liewjunhao@u.nus.edu

In the supplementary material, we first present the effect of different superpixel generation algorithms and show that SEEDS outperforms others in the active learning (AL) tasks. We then present the results of using entropy as uncertainty measure and show that BvSB outperforms entropy for uncertainty sampling. Next we provide the results for the effect of region size on PASCAL VOC and detail the steps of polygon click estimation on this dataset. Finally, we present results on Cityscapes coarse annotation dataset and show that under the same amount of clicks, using fine annotation performs significantly better than coarse annotation.

## 1. Effect of Superpixel Generation Algorithms

The quality of superpixels are usually measured in ASA (Achievable Segmentation Accuracy) and BR (Boundary Recall). Let $\mathcal{G}$ and $\mathcal{S}$ be the set of ground truth segments and superpixel segments for an image, respectively. ASA is the upper bound on the accuracy achievable by any segmentation algorithm using superpixel as a pre-processing step. For the superpixel-based AL method, ASA is essentially the accuracy of the labels provided by dominant labeling. It is computed as:

$$ASA(\mathcal{S}) = \sum_{S_k \in \mathcal{S}} \max_{G_i \in \mathcal{G}} |S_k \cap G_i|/N, \tag{1}$$

where $N$ is the total number of pixels within the image. BR measures how well the superpixel boundaries align with the ground truth boundaries. A pixel on the superpixel boundaries is considered as a true positive (TP) if it is within a local neighborhood of size $(2r+1) \times (2r+1)$ to an arbitrary boundary pixel in $\mathcal{G}$, where $r = 0.0025 \times \sqrt{w^2 + h^2}$, $w$ and $h$ is the image width and height, respectively. BR is then computed as:

$$BR(\mathcal{S}) = \frac{|TP(\mathcal{G}, \mathcal{S})|}{|TP(\mathcal{G}, \mathcal{S})| + |FN(\mathcal{G}, \mathcal{S})|}. \tag{2}$$

The ASA and BR of various superpixel generation algorithms and rectangles (REC) are shown in Table 1. As region size is very small ($32 \times 32$ for image size $1024 \times 2048$), REC can achieve 92.35% ASA, but the BR is low. SLIC significantly improves BR compared to REC and SEEDS obtains higher ASA and BR than SLIC. We also run SEEDS on half-sized images, *i.e.*, input size $512 \times 1024$ for SEEDS_H vs. $1024 \times 2048$ for SEEDS. It can be seen that downsampling the image does not affect ASA much, but degrades BR by 4%.

Table 1: Comparing the quality of different region division schemes on Cityscapes with a region number of 2048 (corresponding to region size $32 \times 32$ for REC).

| | REC | SLIC | SEEDS | SEEDS_H |
|---|---|---|---|---|
| ASA(%) | 92.35 | 93.00 | 94.88 | 94.72 |
| BR(%) | 16.19 | 38.11 | 45.02 | 41.05 |

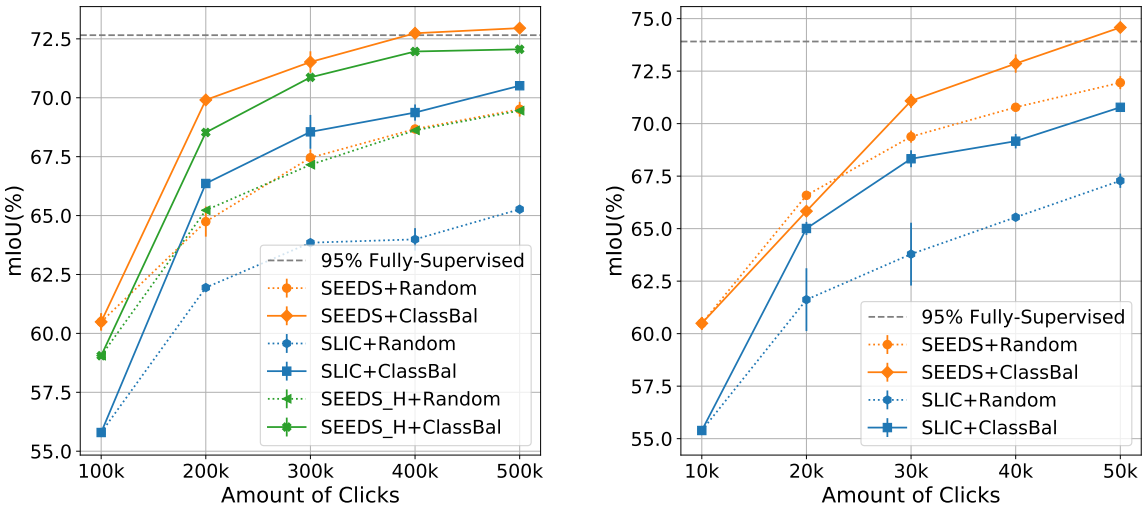
Figure 1 shows the AL results using different superpixel algorithms. It can be seen that SEEDS leads to better results than SLIC on both Cityscapes and PASCAL VOC 2012. Comparing SEEDS and SEEDS_H, it can be seen that the two perform closely for Random sampling, but SEEDS performs better when using ClassBal sampling.

## 2. Effect of Uncertainty Measures

Entropy is a commonly used measure for uncertainty. It is defined as:

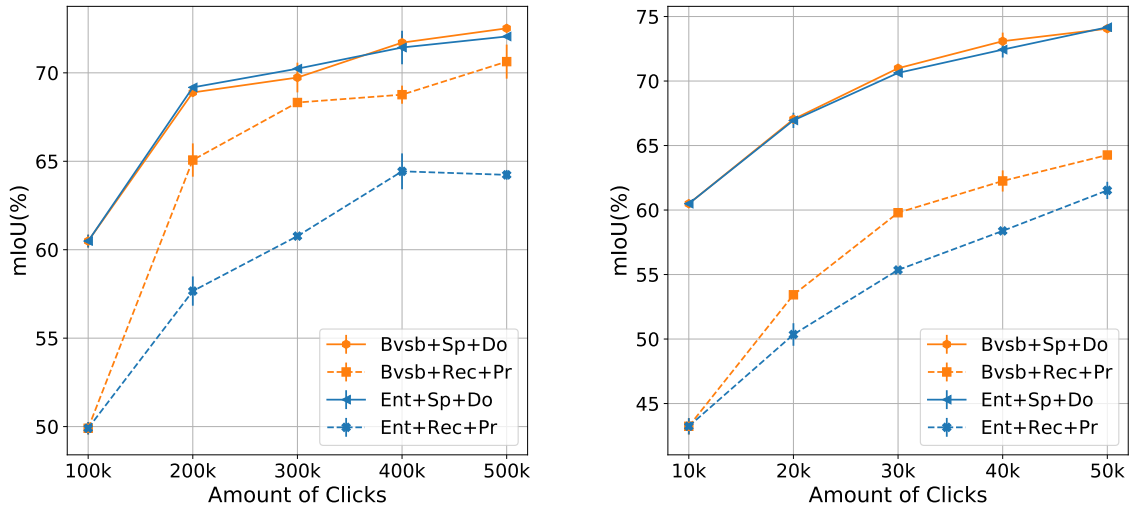
$$H(y|x) = -\sum_{c} p(y = c|x) \log(p(y = c|x)), \tag{3}$$

where $p(y = c|x)$ is the class posterior given by the model. Figure 2 shows the AL results with uncertainty sampling using different uncertainty measures on Cityscapes and PASCAL VOC 2012. It can be seen that BvSB performs closely to Entropy for the superpixel-based approach, but performs much better than Entropy for the traditional Rectangle+Polygon-based approach.




(a) (b)

Figure 1: Comparing AL results for different superpixel algorithms. (a) Results on Cityscapes. (b) Results on PASCAL VOC 2012.




(a) (b)

Figure 2: BvSB vs Entropy for uncertainty sampling. (a) Results on Cityscapes. (b) Results on PASCAL VOC 2012.

## 3. Effect of Region Size on PASCAL VOC

Table 2 lists the number of clicks for different region sizes on PASCAL VOC 2012 training set and Fig. 3 shows the performance of different region sizes at a fixed budget of 20k clicks. Similar to Cityscapes, smaller region sizes incur more class clicks and intersection clicks, yet both Sp+Do and Rec+Pr benefit from smaller region sizes. Also, Sp+Do outperforms Rec+Pr from region size $16 \times 16$ to $64 \times 64$.

## 4. Polygon Click Estimation for PASCAL VOC

PASCAL VOC does not release the polygon click information and we estimate the polygon clicks by first running `alphashape` algorithm[1] to fit a concave hull to the segment mask and then applying the RDP algorithm[2] to reduce the number of points on the hull. The quality of fitting is controlled by parameter alpha (for alphashape) and epsilon (for rdp). The quality of the fitted polygon can be measured by the IoU (Intersection over Union) between the fit-

[1] https://pypi.org/project/alphashape/
[2] https://rdp.readthedocs.io/en/latest/

| | Polygon($c_p$) | Class ($c_c$) | Intersection ($c_i$) | Total |
|---|---|---|---|---|
| Rec+Pr | | | | |
| $32 \times 32$ | 138,921 | 456,583 | 332,800 | 928,304 |
| $128 \times 128$ | 153,132 | 79,607 | 85,489 | 318,228 |
| Image | 157,389 | 29,516 | 0 | 186,905 |
| Sp+Do | | | | |
| 256 | 0 | 231,573 | 0 | 231,573 |
| 16 | 0 | 8,345 | 0 | 8,345 |

Table 2: The number of clicks for different region sizes on PASCAL VOC 2012 training set. $32 \times 32$ corresponds to 256 regions and $128 \times 128$ corresponds to 16 regions per image. Note that $c_p$ is slightly decreasing for smaller region size as we do not count the box corners and polygon clicks on the region boundary have been counted into $c_i$.

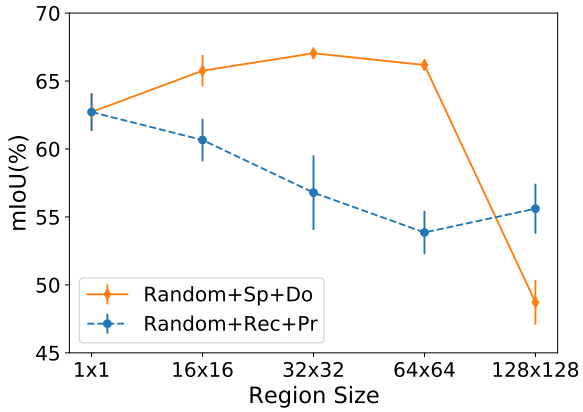



Figure 3: Effect of region size on PASCAL VOC 2012 val subset using the Random baseline.

ted polygon mask and the original object mask. Table 3 lists the mIoU, average number of points per polygon, average running time per image for different combinations of alpha and epsilon values. Setting alpha=0 is equivalent to fitting a convex hull to the mask. It is fast and the number of points per polygon is small, but the mIoU is low. Increasing alpha to 0.5 boosts the mIoU to 90.16%, but results in hundreds of points per polygon, which does not reflect the realistic annotation cost. The rdp algorithm helps to reduce the number of points while maintaining reasonably good mIoU. We choose alpha=0.5 and epsilon=1 in our experiments.

## 5. Benchmark with Cityscapes Coarse Annotation

Cityscapes provides both fine and coarse pixel-wise annotation for the 2975 training and 500 validation images. The fine annotation takes more than 1.5 hours on average for a single image, while the coarse takes less than 7 mins per image but the accuracy of object boundaries is compromised. In Section 4 of the paper, we conduct experiments using fine annotation. Here we add the experiments on coarse annotation.

Table 3: Effect of alpha and epsilon values on the quality of the fitted polygons. The results are reported on a random subset of 100 images (262 polygons) from PACSAL VOC 2012 training set.

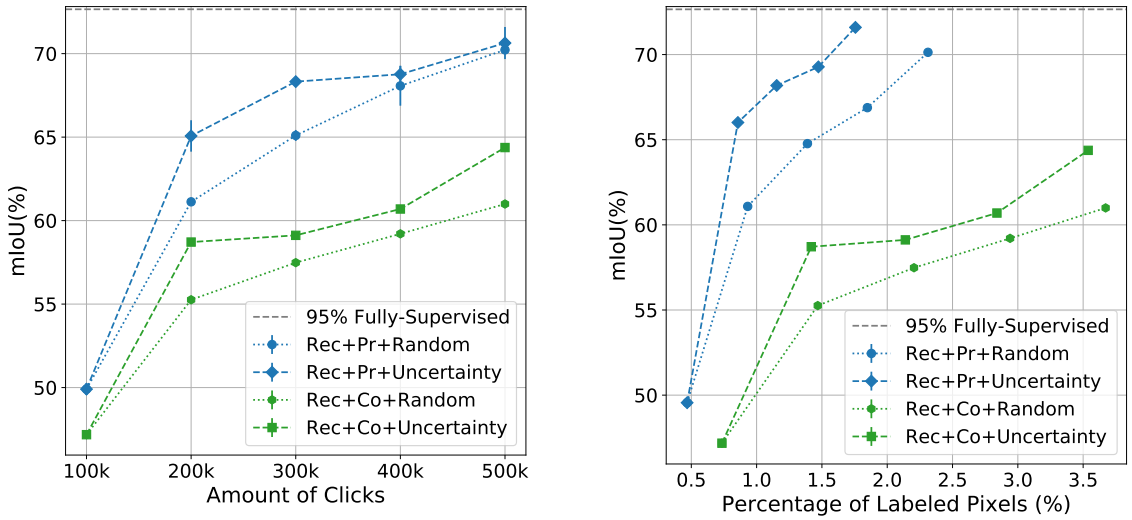
| alpha | epsilon | mIoU (%) | Avg #of Points | Avg Time (sec) |
|---|---|---|---|---|
| 0 | \ | 68.05 | 18 | 0.37 |
| 0.1 | \ | 84.94 | 309 | 38.27 |
| 0.5 | \ | 90.16 | 457 | 37.59 |
| 0.5 | 0 | 90.16 | 159 | 38.18 |
| 0.5 | 1 | 88.44 | 43 | 44.82 |
| 0.5 | 2 | 85.96 | 28 | 43.46 |
| 0.5 | 3 | 83.50 | 22 | 41.42 |

Table 4 lists the number of clicks for fine (Rec+Pr) vs. coarse (Rec+Co) annotation on Cityscapes training set. It can be seen that coarse annotation requires only 25% polygon clicks of fine annotation, but when region size is reduced to $32 \times 32$, the number of class clicks and intersection clicks are close to fine annotation.

Table 4: The number of clicks for fine (Rec+Pr) vs. coarse (Rec+Co) annotation on Cityscapes training set.

| | Polygon($c_p$) | Class ($c_c$) | Intersection ($c_i$) | Total |
|---|---|---|---|---|
| Rec+Pr | | | | |
| $32 \times 32$ | 4,170,635 | 8,952,624 | 4,987,933 | 18,111,192 |
| Image | 4,756,857 | 338,302 | 0 | 5,095,159 |
| Rec+Co | | | | |
| $32 \times 32$ | 1,075,895 | 8,283,616 | 4,258,278 | 13,617,789 |
| Image | 1,223,030 | 86,409 | 0 | 1,309,439 |

Figure 4 presents the AL curves with coarse annotation (Rec+Co). It can be seen that though Rec+Co is able to annotate more regions (pixels) than Rec+Pr using the same amount of clicks, Rec+Pr still outperforms Rec+Co by a large margin.




(a) (b)

Figure 4: Active learning results on Cityscapes fine vs. coarse annotation datasets. (a) Benchmarking at fixed amount of annotation budget measured in clicks. (b) Plot the same results with annotation cost measured in the percentage of labeled pixels.