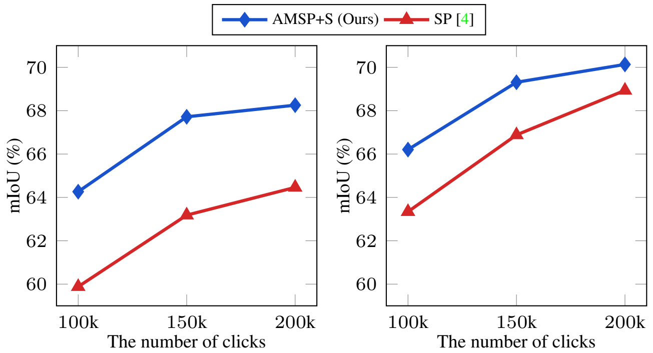
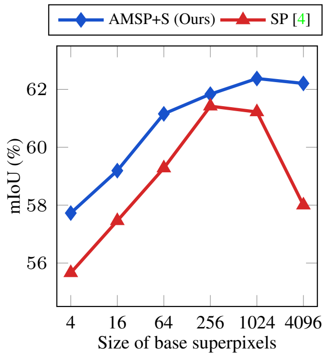
Appendix

A. More gain with other base superpixel sizes



(a) Base superpixel size of 64 (b) Base superpixel size of 256

Figure 7: *Effect of base superpixel size on Cityscapes.* The performance difference is greater when the superpixel size is smaller.



(a) PASCAL

Figure 8: *Effect of base superpixel size on PASCAL.* Our method exhibits robustness to large superpixels, while the baseline is sensitive.

For ease of exposition, Figure 3 presents the gain of our method (compared to *SP* [4]) for a limited set of base superpixel sizes. In this section, we report an additional investigation suggesting further gain with different base superpixels.

Further gain on Cityscapes. In Figure 7a, we additionally provide a comparison between the proposed method (*AMSP+S*) and *SP* [4], where the experiment setup with Cityscapes is identical to that in Figure 3a except that the base superpixel size is 64 (Figure 7a) instead of 256 (Figure 7b). Our adaptive merging method (*AMSP+S*) is especially effective when the superpixel size is small in Figure 7a, thanks to the adaptive merging mechanism. This observation suggests more significant gain of our method with other choices of base superpixel size than that in Figure 3.

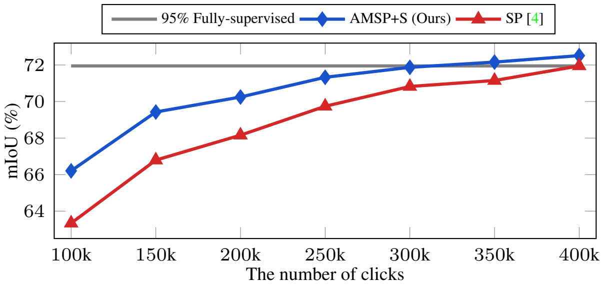


Figure 9: *Additional rounds experiments on Cityscapes.* We extend the experiments in Figure 3a up to a budget of 400k. The performance improvement remains consistent across various additional budgets.

Methods	mIoU
SP [4]	63.77
AMSP+S (bottom 10%)	64.33
AMSP+S (top 10%)	65.99
AMSP+S (complete 100%)	**66.53**

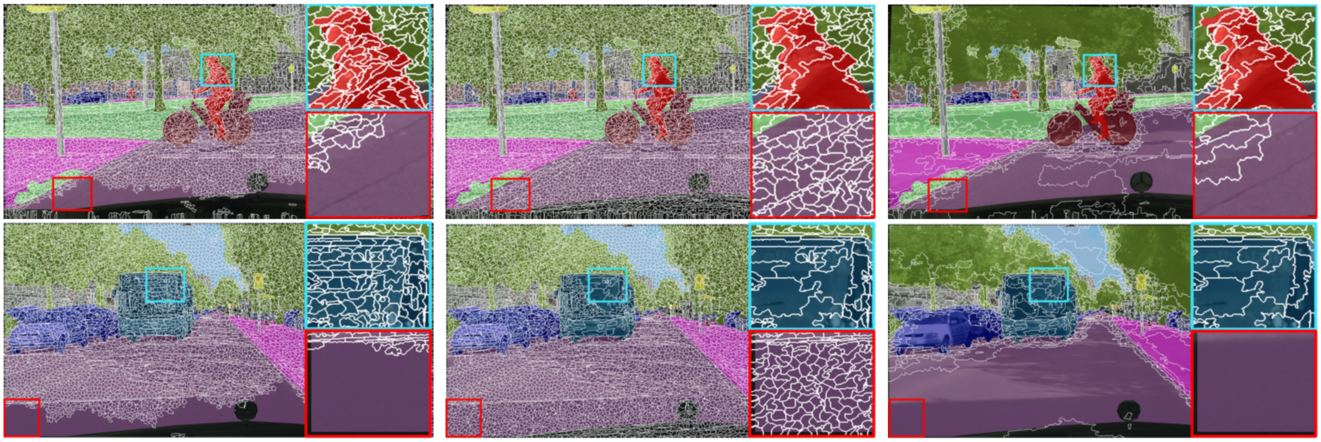
Table 5: *Various levels of partial merging.* Experiments are conducted under the same setting of Figure 3a with 100k clicks (Cityscapes, superpixel size of 256).

Further gain on PASCAL. We also demonstrate a larger gap between the proposed method and existing one in PASCAL. In Figure 8, our adaptive merging method (*AMSP+S*) outperforms the baseline (*SP*) for various superpixel sizes as we observed in Figure 3. We stress that the gain of the proposed method is particularly larger than the one reported in Figure 3 when the base superpixel size is 4096, which is much larger than 256 used in Figure 3. This is because the sieving procedure to suppresses the noise from dominant labeling becomes more crucial when querying large superpixels. The experimental setup used in Figure 8 is identical to that of Figure 3d.

Further rounds on Cityscapes. To demonstrate the efficacy of our method across various budgets, we experiment by gradually increasing the budget as illustrated in Figure 9 on Cityscapes. The experimental setting in Figure 9 remains consistent with that of Figure 3a. The advantage of our method over SP [4] is continued in further rounds. We remark that the proposed method nearly achieves the 95% mIoU of the fully supervised model (71.95%) at 300k clicks, whereas SP does at 400k clicks.

B. Rationale for line 3 of Algorithm 2

We explain the rationale behind traversing nodes in the descending order of uncertainty in line 3 of Algorithm 2.

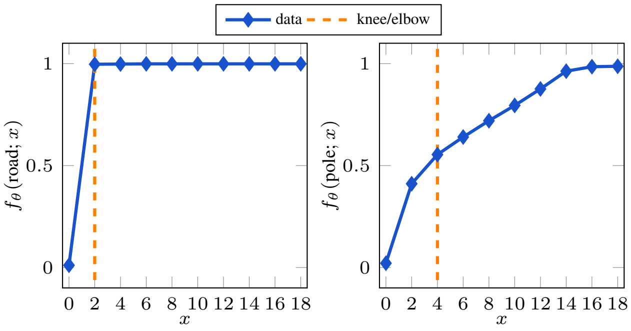


(a) Merging superpixels with low 10% uncertainty (b) Merging superpixels with high 10% uncertainty (c) Merging all superpixels

Figure 10: *Qualitative results for partial merging.* The cyan boxes encompass superpixels exhibiting the highest 10% uncertainty, while the red boxes encompass superpixels with the lowest 10% uncertainty. (b) By merging only a portion of superpixels in the order of high uncertainty, we can reduce time complexity, as it creates similar merged superpixels compared with the cyan box in (c).

Methods	mIoU
SP [4]	63.77
AMSP+S ($\phi(s;\theta) = 0.0$)	<u>65.35</u>
AMSP+S ($\phi(s;\theta) = 0.2$)	61.80
AMSP+S ($\phi(s;\theta) = 0.4$)	57.77
AMSP+S ($\phi(s;\theta) = 0.6$)	45.84
AMSP+S ($\phi(s;\theta) = 0.8$)	38.99
AMSP+S (Kneedle [30])	**66.53**

Table 6: *Various sieving methods.* Experiments are conducted on Cityscapes dataset with an average superpixel size of 256, using 100k costs for two rounds.



(a) Road (b) Pole

Figure 11: *Examples of knee points on Cityscapes.* We obtain (a) a high knee value for the common road class and (b) a low knee value for the rare pole class.

Our merging process requires a linear time complexity proportional to the size of the base superpixels graph. However, due to the advantage of merging in descending uncertainty order, we are able to acquire merged superpixels with considerable uncertainty at the beginning of merging. To reduce merging time complexity, we only merge the top 10% of base superpixels with the highest uncertainty as query candidates. Table 5 shows that it is important to prioritize the merging highly uncertain superpixels, and merging along the ascending order of uncertainty degenerates the performance.

In Figure 10, we exemplify the merged superpixels from the partial merging in the ascending or descending order of uncertainty, and the full merging, where the cyan boxes contain higher values of acquisition function than the red boxes. The partial merging with the ascending order of uncertainty regrettably merges the superpixels that would not be selected in AL, while that with the ascending order efficiently combines the base superpixels of which selection is highly like. This difference indeed results in a huge gap in the final performance as shown in Table 5.

C. Rationale for the adaptive threshold $\phi(s;\theta)$ in the sieving

We provide the reason for introducing the threshold function $\phi(s)$ personalized for each superpixel s, described in Section 3.3. We obtain the dominant label $\mathrm{D}(s)$ for a queried superpixel s, however, we only propagate the label to pixels $x \in s$ that are predicted to have a positive impact

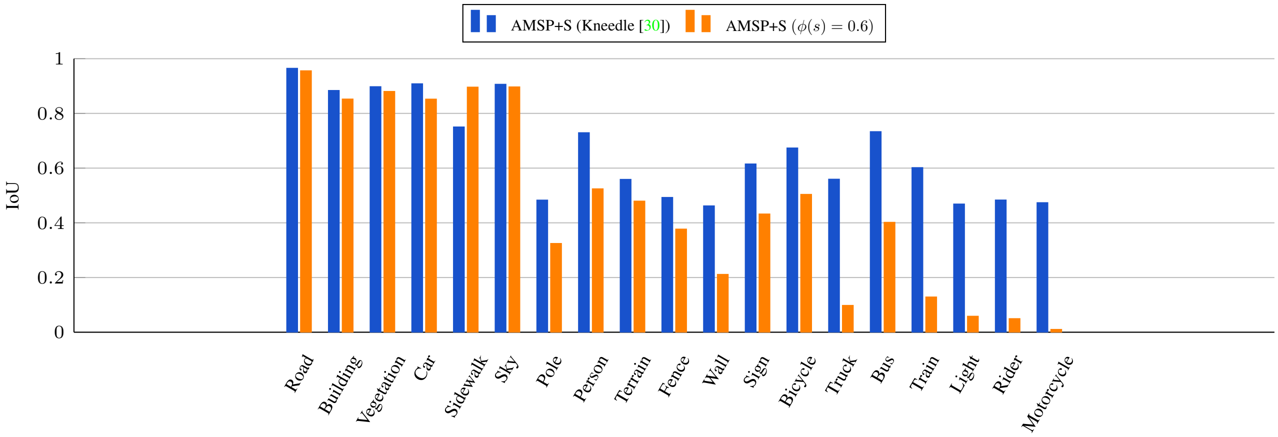
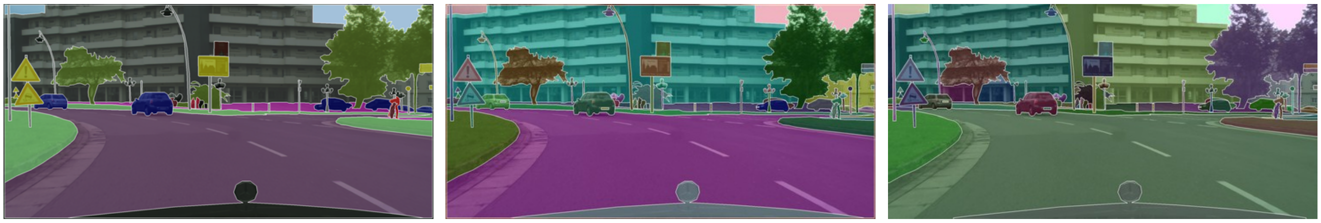


Figure 12: *Class-wise IoU according to* $\phi(s;\theta)$. Applying the same $\phi(s)$ of 0.6 to all pixels results in excessive sieving for relatively rare classes, leading to decreased performance for these classes (*e.g.* Light, Rider, and Motorcycle). Based on the ground-truth, class labels are organized in order of the total pixel count for each class.



(a) Semantic segmentation

(b) Panoptic segmentation

(c) Oracle superpixels (ours)

Figure 13: *Difference between conventional segmentations and oracle superpixels.* (a) When sharing the same class label, they are depicted as identical superpixels (*i.e.* green color on separate trees). (b) Although a building is divided by a pole, it is represented as a single superpixel (*i.e.* cyan color). (c) We consider a building as two distinct superpixels (*i.e.* cyan and light yellow colors).

on the training of model θ as:

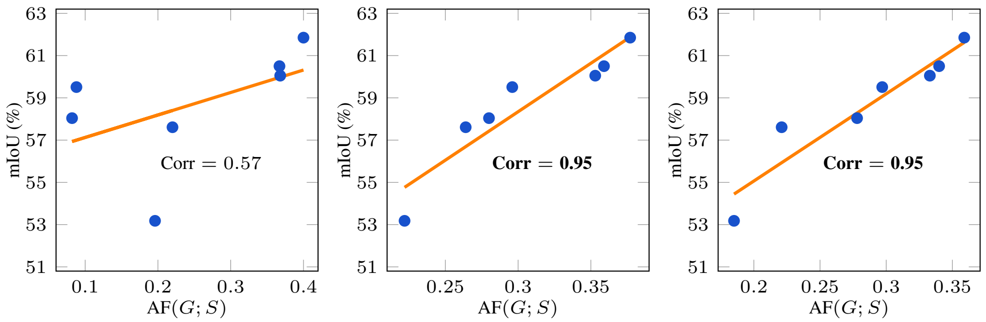
$$h(s;\theta) := \{x \in s : f_\theta\big(\mathrm{D}(s);x\big) \geq \phi(s;\theta)\}\ , \tag{16}$$

where $f_\theta\left(\mathrm{D}(s);x\right)$ implies the confidence of pixel x to dominant label $\mathrm{D}(s)$ given θ and $\phi(s;\theta)$ determines the degree of sieving. In Table 6, we study the effect of various $\phi(s;\theta)$. When the same $\phi(s;\theta)$ is applied to all pixels, it causes class imbalance by leaving relatively easy classes as described in Figure 12. To avoid this issue, we utilize the Kneedle algorithm [30] to obtain different $\phi(s;\theta)$ for each superpixel s. Specifically, $\phi(s;\theta)$ is a knee point of the cumulative distribution function of values of $f_\theta\big(\mathrm{D}(s);x\big)$ in superpixel $x \in s$. However, for the Kneedle algorithm to work accurately, the curve of cumulative distribution must be either convex or concave. In addition, the algorithm may provide inaccurate knee points on very smooth curves. To address this issue, we use a subset of uniformly sampled values based on $f_\theta(\mathrm{D}(s);x)$, instead of using the distribution for all pixels. We sample 20 and 5 pixels for Cityscapes and PASCAL datasets, respectively. In Figure 11, different knee points are detected according to the dominant class of superpixels.

Effect of sieving. Our sieving method exhibits a significant effect on larger superpixels, as illustrated in Figure 3c and Figure 8. Especially, in Figure 8 with a large base superpixel size of 4096, the first sieving excises 45.87% of the mislabeled pixels that disagree with their dominant labels. Furthermore, we observe that the sieving is progressively refined round by round. For instance, in Figure 3a, the portion of the mislabeled labels removed by the sieving increases over four rounds as follows: 3.58%, 8.54%, 10.46%, and 12.43%. Our sieving technique enhances label quality by retaining only high-confidence labels and continuously improves through multiple rounds.

D. Further discussion on the oracle superpixels

In Section 4.1, we introduce the oracle superpixels, which we believe is an achievable optimal set of superpixels for active learning. For clarification, we provide the detailed



(a) Semantic segmentation (b) Panoptic segmentation (c) Oracle superpixels

Figure 14: *Relationship between AF*$(G; S)$ *and mIoU varying* G*.* AF$(G; S)$ and mIoU exhibit a high correlation when ground-truth G is represented by the panoptic segmentation and oracle superpixels in Figure 13. For the correlation calculation, *Oracle* in Table 1 is excluded.

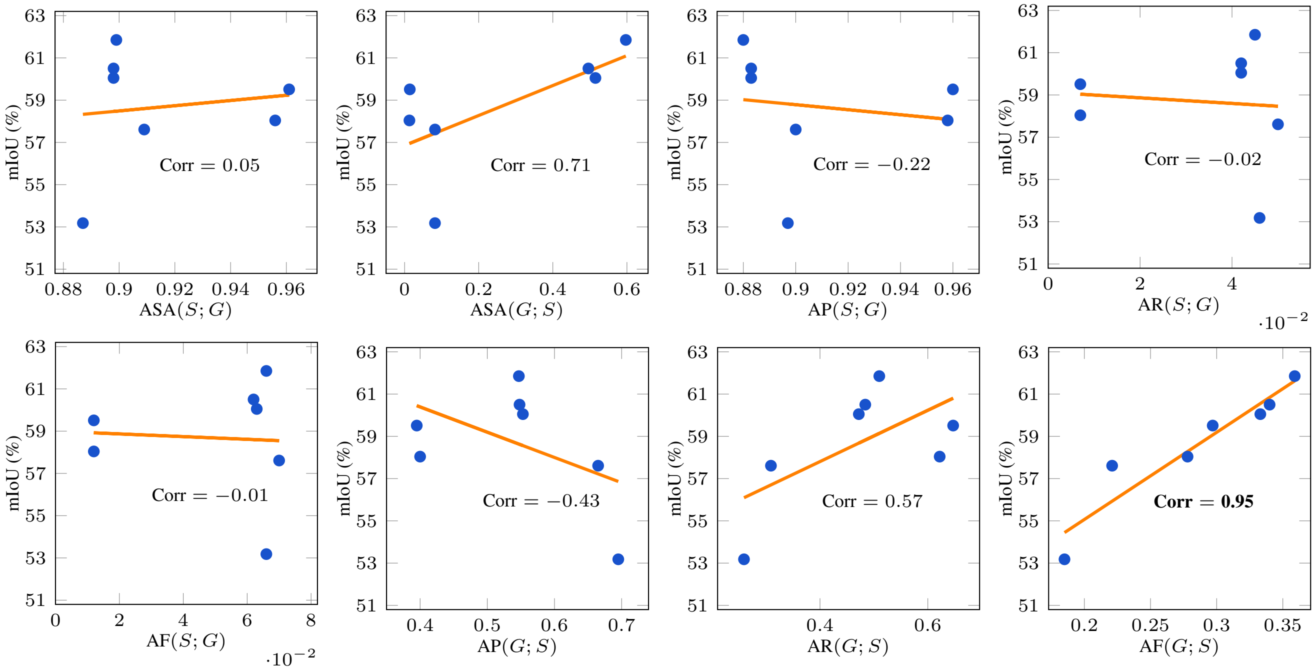
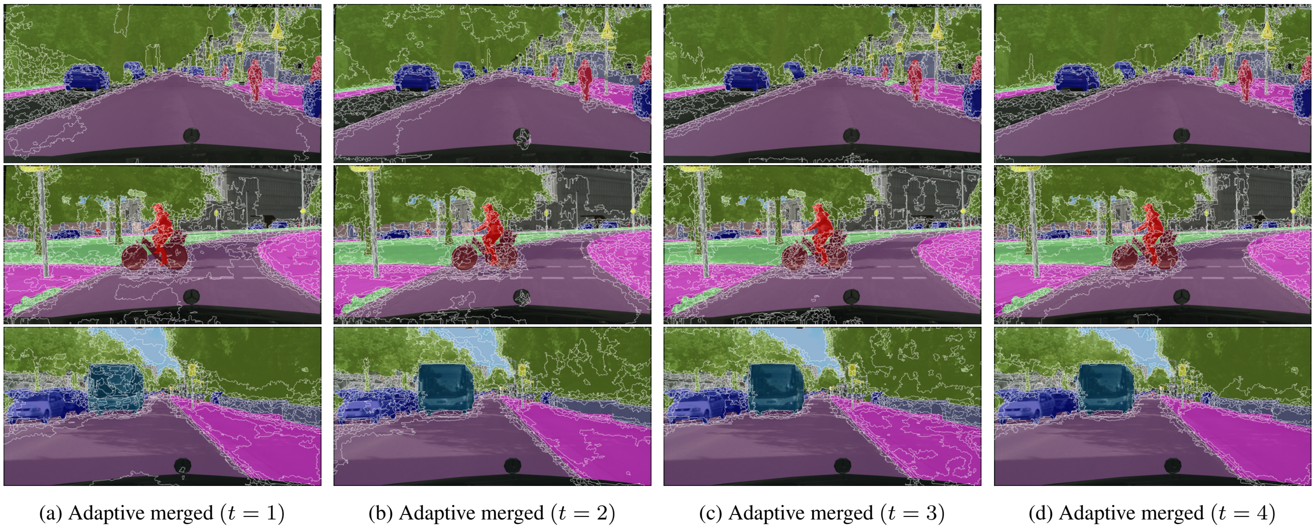


Figure 15: *Relationship between metrics and mIoU.* The correlation between AF$(G; S)$ and mIoU is especially high. For the correlation calculation, *Oracle* in Table 1 is excluded.

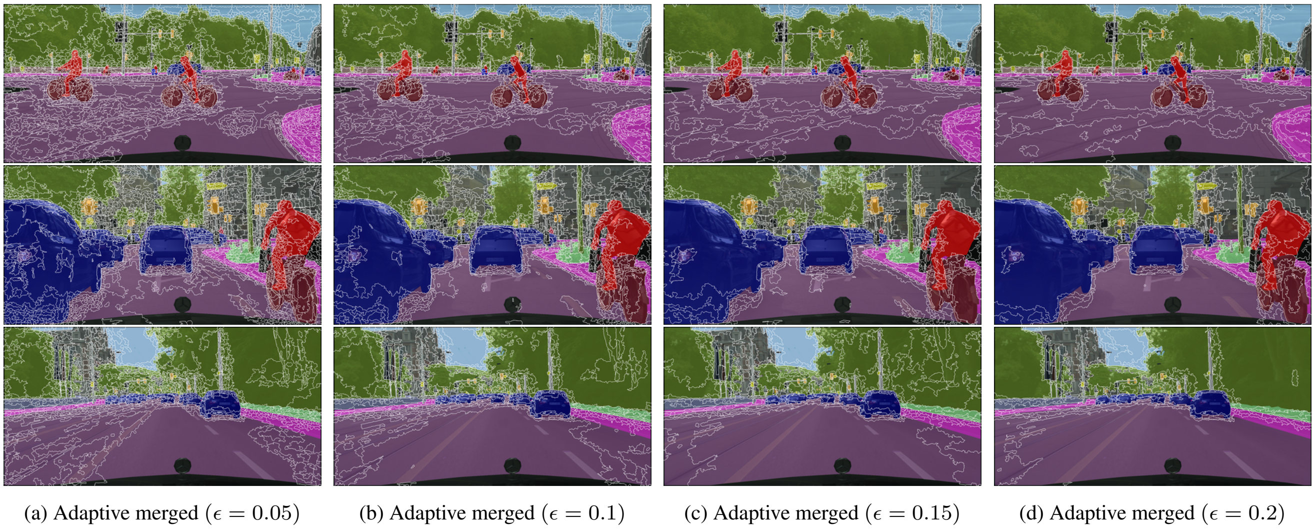
process of generating the proposed oracle superpixels. In addition, we provide further insights into the achievable notion of optimal superpixels.

The Cityscapes dataset is equipped with the ground-truth annotations for semantic segmentation, represented by dense pixel-wise labels: *i.e.*, each pixel in an annotated image is assigned an ID that represents a ground-truth semantic category (Figure 13a). In such annotation, each group of pixels that share the same ID aligns perfectly with the boundary of semantic objects. However, each such group is not guaranteed to be a single-connected component of pixels. For example, different cars in Figure 13a are assigned the same blue color despite being physically separated, and a car divided into two parts due to an obstructing pole is still colored blue. This is opposed to what we hope to achieve by merging two adjacent superpixels repeatedly. To address this issue, we subdivide each superpixel as necessary to ensure that every pixel within a superpixel is adjacent to each other. We utilize OpenCV [3] and Shapely [14] to identify the maximal connected component of pixels sharing the same semantic. We apply the same procedure to annotated images in the PASCAL dataset Figure 13 illustrates the distinction between conventional semantic and panoptic segmentation and our oracle superpixels.



(a) Adaptive merged ($t = 1$) (b) Adaptive merged ($t = 2$) (c) Adaptive merged ($t = 3$) (d) Adaptive merged ($t = 4$)

Figure 16: *Qualitative results with varying round.* (a-d) Superpixels generated with proposed adaptive merging at rounds 1 to 4. Thanks to the improved model, we observe that the merging becomes more accurate as the round increases. We use the model reported in Figure 3a.



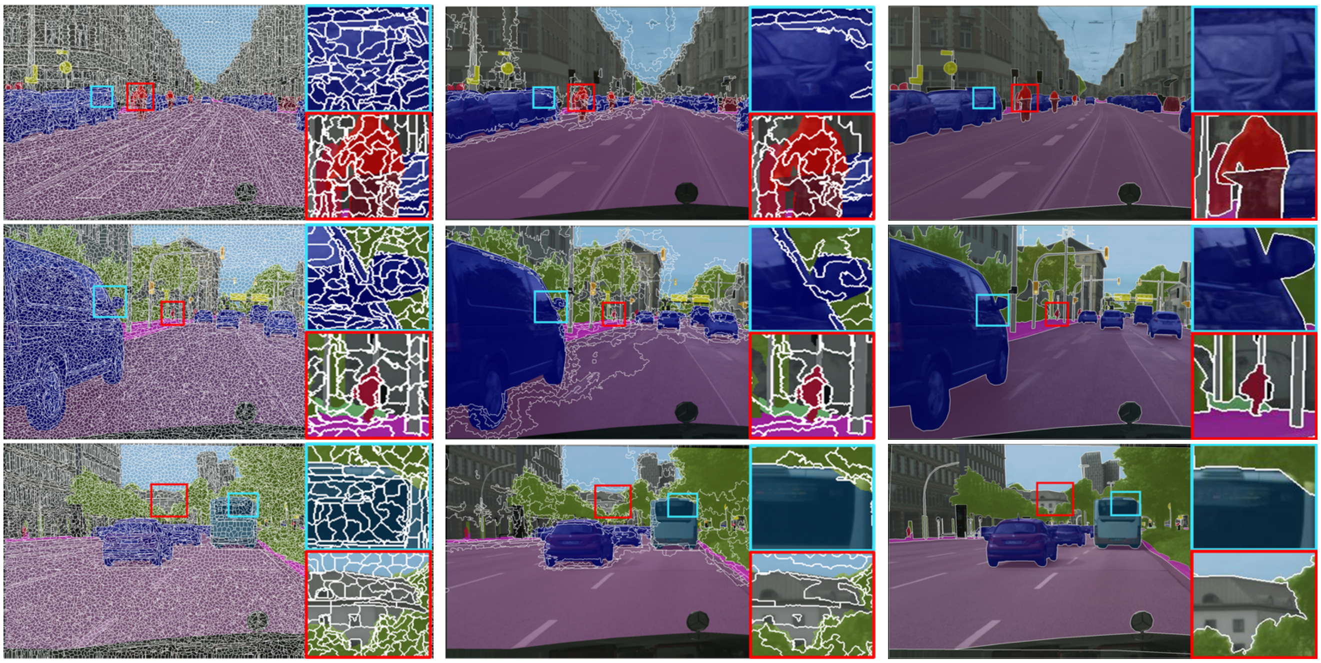
(a) Adaptive merged ($\epsilon = 0.05$) (b) Adaptive merged ($\epsilon = 0.1$) (c) Adaptive merged ($\epsilon = 0.15$) (d) Adaptive merged ($\epsilon = 0.2$)

Figure 17: *Qualitative results with varying ϵ.* (a-d) Superpixels are generated with proposed adaptive merging with ϵ: 0.05, 0.1, 0.15, 0.2. We observe that an increase in ϵ gives more aggressive merging. Merging is conducted on Cityscapes with a base superpixel size of 256.

The Cityscapes and PASCAL datasets are divided into 327k and 16k oracle superpixels, respectively. It is worth noting that the PASCAL has a lower number of oracle superpixels due to the smaller number of classes per image. In other words, only a few objects are of interest in each image, and the rest are simply treated as the background.

E. Further analysis on the achievable metrics

In Table 1, we evaluate various superpixels using eight metrics with oracle superpixels as ground-truth G. Figure 15 shows the correlation between each metric and mIoU. We observe that our AF$(G; S)$ can be utilized to look-ahead a model's performance in active learning without training. In addition, we examine how different ground-truth G impacts AF$(G; S)$. In the field of semantic segmentation, two conventional segmentations, semantic and panoptic segmentations in Figure 13, are widely used as ground-truth. Figure 14 indicates that using panoptic segmentation and oracle superpixels for G results in higher correlation between AF$(G; S)$ and mIoU than semantic segmentation. However,



(a) Base superpixels [35]

(b) Merged superpixels (Ours)

(c) Oracle superpixels

Figure 18: *Qualitative results of adaptive superpixels.* (a) Base superpixel generated by SEEDS [35] with size 256. (b) Superpixels generated with proposed adaptive merging at round 4. (c) Oracle superpixels generated from the ground truth.

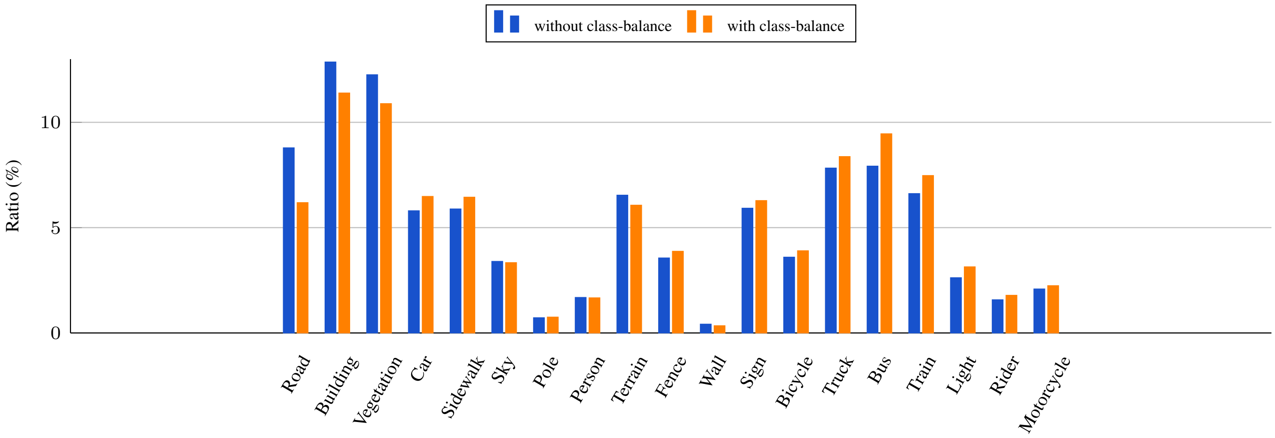


Figure 19: *Effect of class-balanced acquisition function.* According to the ground-truth, class labels are arranged based on the total pixel count for each class, *i.e.* classes become rarer in images as you move from left to right along the x-axis. We observe that classes on the left are selected less with the class-balanced term, while classes on the right are selected more.

obtaining panoptic segmentation requires more costs than semantic segmentation since it utilizes additional instance information. It is worth noting that our oracle superpixels (Figure 13c) can be easily generated even in cost-limited practical situations as they are produced from semantic segmentation (Figure 13a).

F. Additional qualitative adaptive superpixels

To facilitate comprehension of the merged superpixels, we display superpixels generated across diverse settings. The appearance of merged superpixels is mainly determined by the model's performance and ϵ. Figure 16 highlights that as the round progresses, the model's performance improves, leading to more accurate merging. With the model fixed at round 4, Figure 17 shows the impact of adjusting ϵ. As ϵ

Notations	Description
$\mathcal{I}$	the set of unlabeled images
$\mathcal{C}$	the set of class labels
t	a round
x	a pixel
s	a superpixel
$S_t(i)$	the set of superpixels in an image i in round t
$\mathcal{S}_t$	the set of superpixels in all images in round t, $\mathcal{S}_t := \bigcup_{i\in\mathcal{I}} S_t(i)$
B	the query budget per round
$\mathcal{B}_t$	the set of B selected superpixels in round t, $B_t \subset \mathcal{S}_t, \lvert B_t\rvert = B$
θ_t	the model at the end of round t
$y_\theta(x)$	the estimated dominant label of pixel x given θ
$\mathrm{D}(s)$	the true dominant label of superpixel s
$\mathrm{D}_\theta(s)$	the estimated dominant label of superpixel s given θ
$\mathcal{G}(S) := (S, \mathcal{E}(S))$	the graph consisting of the superpixels in S as nodes and the edge set $\mathcal{E}(S)$ such that $(s, n) \in \mathcal{E}(S)$ for each pair of adjacent superpixels $s, n \in S$.
ϵ	the hyperparameter for merging in (2)

Table 7: *Notations.* The notations used in the paper are defined.

grows, the merging process intensifies, ultimately decreasing the overall number of superpixels. In addition, Figure 18 shows further examples of our merged superpixels.

G. Class-balanced sampling

To observe the impact of the class-balanced acquisition function in (7), we analyze the class distribution of selected superpixels both with and without the class-balanced term. In Figure 19, where class labels are sorted such that the left (road) and right (motorcycle) ends represent the most and least popular classes, it is evident that the class-balanced term results in a higher selection of rarer classes, as intended.