A Supplementary Material CEREALS – Cost-Effective REgion-based Active Learning for Semantic Segmentation

A.1 Implementation Details

Instead of cropping the annotated regions out of the images, while taking into account their receptive field in input space, we instead mask out all currently unlabeled data in output space, making sure that no loss is computed on unlabeled data when learning the *semantic segmentation model* nor when learning the *cost model*. We then perform an image-based training, from unprocessed input images to spatial label maps. However, our practical implementation of *CEREALS*, which will be made publicly available is supporting both options. For training the utilized models we use Adam as our optimizer with learning rate, alpha and beta set to 0.0001, 0.99 and 0.999 respectively. Furthermore, we claim convergence whenever a model hasn't improved regarding the application loss for at least 10 epochs. We train with the mini-batch size set to 1, such that a gradient step is always being applied w.r.t. one full resolution image of Cityscapes.

Semantic Segmentation Model We do not train the employed model in stages, but directly optimize for *FCN8s*. Regarding the training performed on the full training set of Cityscapes, we report a mean intersection over union (mIoU) of 0.605 which is, as all other results, computed on the full validation dataset of Cityscapes. Note, that the original model achieves a mIoU of 0.65 and that we are able to reproduce this result when the width multiplier is set to 1.0, despite all other changes. Though we utilized this particular model, *CEREALS* can use any model producing semantic segmentation masks as long as it provides probability distributions regarding it's posterior outcome. The *cost model* however, would need to be adapted or made independent of the *semantic segmentation model* in such a case.

Cost Model The only change we made to the original model's architecture is to replace it's softmax activation with a linear activation layer. We trained the model towards minimizing the mean squared error of predicted and ground truth clicks. Since we observed some pixels to have unrealistically many clicks in the ground truth data, we clipped the values to be in $[0, 10]$ range allowing a maximum of 10 ground truth clicks per pixel. As the *semantic segmentation model*, the *cost model* doesn't have any upsampling layer at the end, in order to allow for faster trainings. We instead downscale the provided click data by a factor of 8.

A.2 Results

The evaluation for all processed experiments performed on the modified *FCN8s* architecture are reported in Table 1.

	2048x1024*	**512x512**	**256x256**	**128x128**										
Eq. 1) H (p95)	38.66	16.74	11.71	10.01	**Eq.3)**		**Eq.4)**		**Eq.5** ($\alpha = 0.5$)		**Eq.5** ($\alpha = 0.75$)		**Eq.5** ($\alpha = 0.9$)	
Eq. 2) V (p95)	36.97	19.86	14.85	11.59	**GT**	**Est.**	**GT**	**Est.**	**GT**	**Est.**	**GT**	**Est.**	**GT**	**Est.**
Eq. 1) H (c95)	43.17	33.55	34.05	33.76	23.08	21.42	14.68	**17.07**	10.91	n/a	22.3	20.54	27.79	34.52
Eq. 2) V (c95)	39.02	29.58	28.11	24.81	15.30	18.23	13.97	18.41	n/a	n/a	16.35	19.01	20.11	21.56
Random (p95)	n/a	33.61	35.29	28.57										
Random (c95)	n/a	44.57	38.62	29.86										

Table 1: All results indicate the amount of used annotated pixels (p95) or number of clicks (c95) relative to the amount of all the pixels or clicks within the Cityscapes training set for achieving a mIoU of 95% as compared to the mIoU achieved when learning the employed *FCN8s*-based model on all the data provided by Cityscapes. The results in this table are averaged over five repetitions, as explained in the beginning of the result section in this work. (*Image-based acquisition)

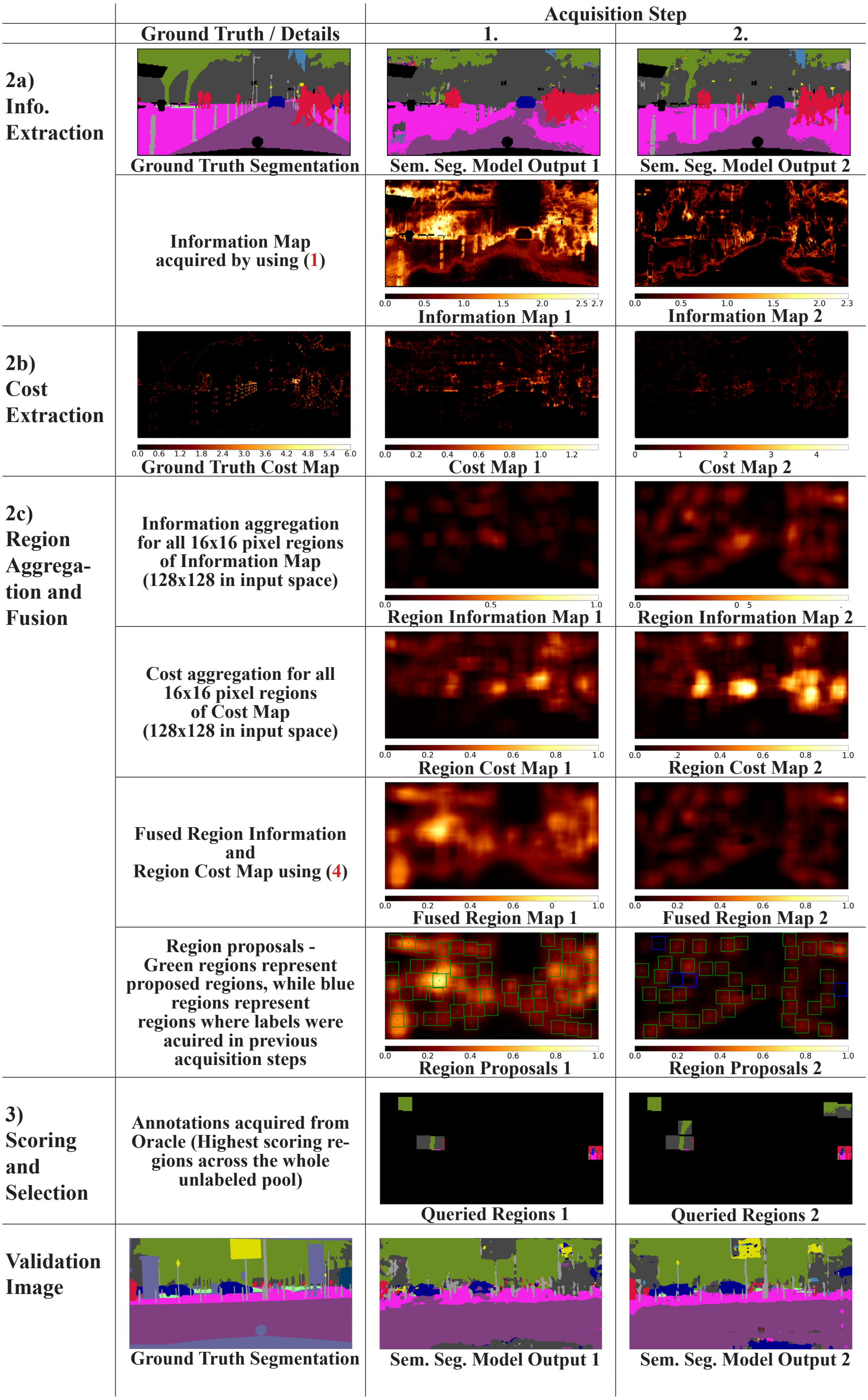

Figure 6: Detailed overview of *CEREALS*. First two acquisition steps.

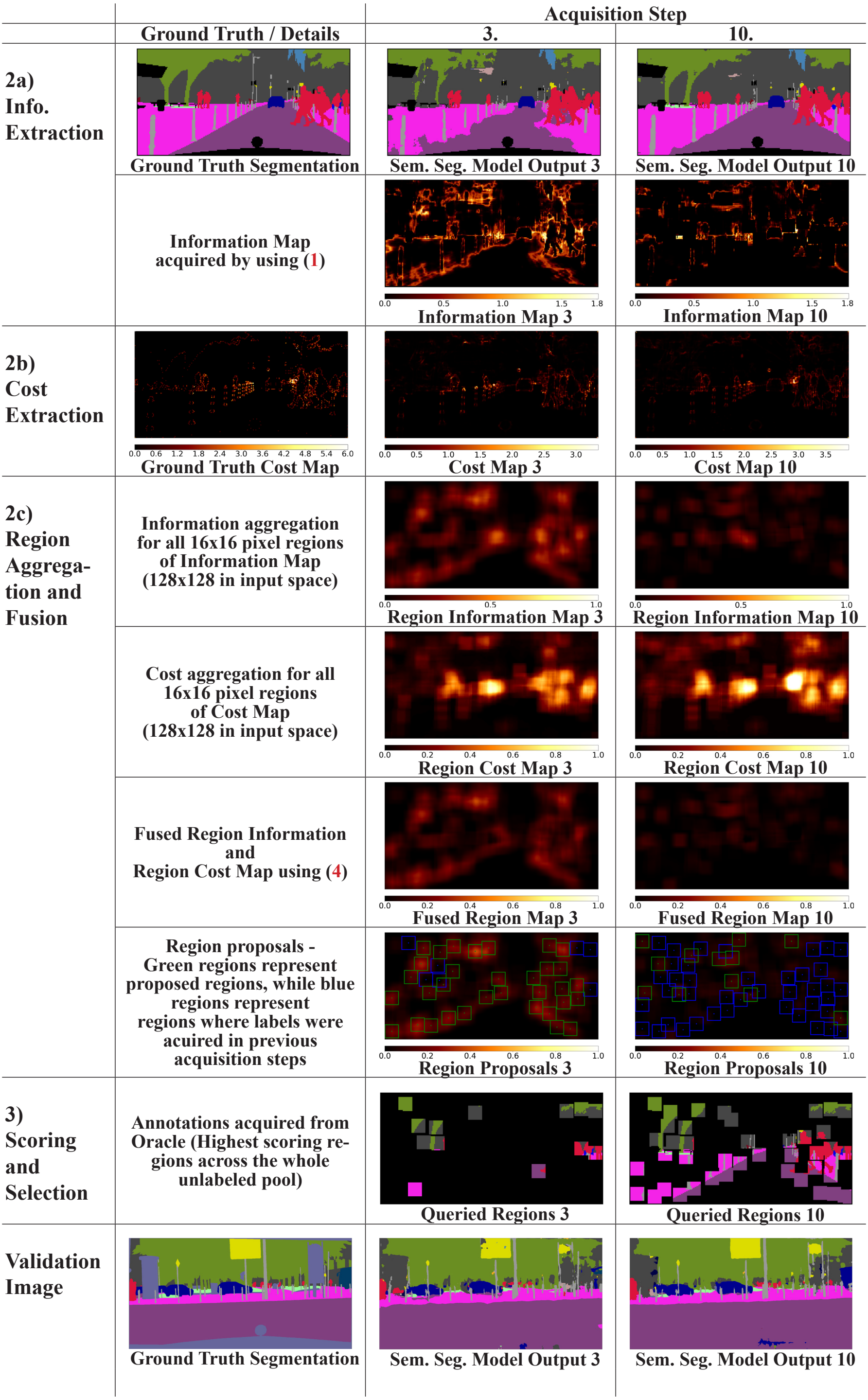
Acquisition Step
Ground Truth / Details
3.
10.
2a) Info. Extraction
Ground Truth Segmentation
Sem. Seg. Model Output 3
Sem. Seg. Model Output 10
Information Map acquired by using (1)
0.0 0.5 1.0 1.5 1.8
Information Map 3
Information Map 10
2b) Cost Extraction
0.0 0.6 1.2 1.8 2.4 3.0 3.6 4.2 4.8 5.4 6.0
Ground Truth Cost Map
0.0 0.5 1.0 1.5 2.0 2.5 3.0
Cost Map 3
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5
Cost Map 10
2c) Region Aggregation and Fusion
Information aggregation for all 16x16 pixel regions of Information Map (128x128 in input space)
0.0 0.5 1.0
Region Information Map 3
Region Information Map 10
Cost aggregation for all 16x16 pixel regions of Cost Map (128x128 in input space)
0.0 0.2 0.4 0.6 0.8 1.0
Region Cost Map 3
Region Cost Map 10
Fused Region Information and Region Cost Map using (4)
Fused Region Map 3
Fused Region Map 10
Region proposals - Green regions represent proposed regions, while blue regions represent regions where labels were acuired in previous acquisition steps
Region Proposals 3
Region Proposals 10
3) Scoring and Selection
Annotations acquired from Oracle (Highest scoring regions across the whole unlabeled pool)
Queried Regions 3
Queried Regions 10
Validation Image
Ground Truth Segmentation
Sem. Seg. Model Output 3
Sem. Seg. Model Output 10

Figure 7: Detailed overview of *CEREALS*. Acquisition steps three and ten.