Sparsely Annotated Semantic Segmentation with Adaptive Gaussian Mixtures (Supplementary Materials)

Linshan Wu[1], Zhun Zhong[2], Leyuan Fang [*1], Xingxin He[1], Qiang Liu[1], Jiayi Ma[3], and Hao Chen[4]

[1]College of Electrical and Information Engineering, Hunan University
[2]Department of Information Engineering and Computer Science, University of Trento
[3]School of Electronic Information, Wuhan University
[4]Department of Computer Science and Engineering, Hong Kong University of Science and Technology

In the supplementary material, we further provide more experimental and visualization results. We analyze the effectiveness of AGMM in Section A. The balance evaluation for multiple losses is presented in Section B. More results on the PASCAL VOC 2012 dataset are discussed in Section C. We further present more details of the Cityscapes dataset in Sections D and E.

A. AGMM analysis

Visualization results. We first present some results of GMM predictions on the PASCAL VOC 2012 dataset, as shown in Figs. 1. The segmentation branch's predictions rely only on the model parameters, while the GMM predictions are generated from the similarity between labeled and unlabeled pixels in the high-dimension feature space. Thus, it can be seen that the semantic discrepancy exists between the predictions of the segmentation branch and GMM branch. Our proposed framework introduces a novel self-supervision loss function to constrain the consistency between these two branches, achieving more robust performance.

Accuracy analysis during training. To get a deeper understanding of our AGMM, we further analyze the accuracy of GMM predictions during the training process, as shown in Fig. 2. It can be seen that compared with the baseline method that uses L_{seg} only, our AGMM can provide more supervision information (GMM predictions) as described in Section 3. By constraining the consistency between segmentation and GMM predictions, the AGMM gains higher accuracy and learns a more robust segmentation model. At the beginning of training, leveraging reliable information from labeled pixels, GMM predictions are more accurate. As the number of epochs increases, the gap between the GMM and segmentation predictions is gradually narrowed while the performances of both are improved. With the constraint of L_{self}, the GMM and segmentation predictions can be consistent. Compared with the baseline method, our proposed AGMM framework can achieve higher accuracy during training.

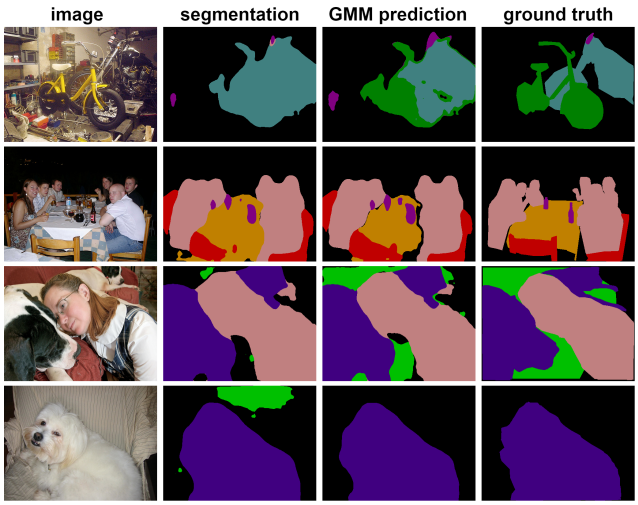

Figure 1. Qualitative results of GMM and segmentation predictions under point-supervised settings in PASCAL VOC 2012 dataset.

B. Balance evaluation for loss functions

We further analyze the balance between L_{seg} and our proposed L_{GMM}. Specifically, we formulate the total loss function L as follows:

$$L = L_{seg} + \lambda L_{GMM}, \tag{1}$$

*Corresponding author

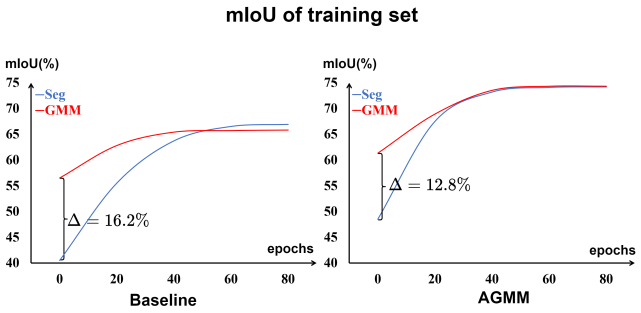

Figure 2. Comparison between segmentation and GMM predictions during training. We report the results on the PASCAL VOC 2012 dataset under point-supervised SASS settings. The baseline method represents using partial cross-entropy loss L_{seg} for supervision only. Δ denotes the difference of mIoU between segmentation and GMM predictions. It can be seen that at the beginning of training, the GMM predictions gain a higher accuracy than the segmentation predictions by a large margin. In our GMM-SASS framework, we utilize the L_{self} to constrain the consistency between segmentation and GMM predictions, achieving better supervision for the segmentation model.

where λ is the factor to balance the L_{seg} and L_{GMM}. The evaluation of λ is illustrated in Table 1. It can be seen that $\lambda = 1$ can yield the best performance.

λ	0.1	0.2	0.5	0.8	1.0	1.5
point-supervised	67.5	68.1	68.6	69.2	**69.6**	69.3
scribble-supervised	74.0	74.8	75.9	76.3	**76.4**	76.2

Table 1. Effectiveness evaluation of λ in Eq. 1. We report the mIoU results on the PASCAL VOC 2012 dataset.

C. Visualization on PASCAL VOC 2012 dataset

To better understand our AGMM, we further give an illustration of visualization of the high-dimension feature space, as shown in Fig. 3. Three t-SNE plots are given respectively on the baseline method, AGMM (stop-gradient), and AGMM. As can be seen, the decision boundaries of features generated by the baseline method and AGMM (stop-gradient) are quite confusing, while our AGMM can learn more discriminative features. This explains why AGMM works from a feature point of view.

We present qualitative segmentation results of AGMM on PASCAL VOC 2012 dataset in Fig. 4. It can be seen that supervised with only sparse labels (points and scribbles), our proposed AGMM can also produce promising segmentation results. Although the sparse labels cannot provide boundary information, our method can still predict complete objects with sharp edges, which demonstrates the effectiveness of our proposed method.

L_{seg}	L_{self}	L_{spar}	L_{con}	MT	clicks 20	clicks 50	clicks 100
✓	-	-	-	-	53.5	60.3	64.2
✓	-	-	-	✓	59.7	65.2	67.8
✓	✓	-	-	-	59.2	66.4	68.3
✓	✓	✓	-	-	60.7	67.1	70.2
✓	✓	✓	✓	-	62.1	68.3	71.6
✓	✓	✓	✓	✓	**66.5**	**71.7**	**73.4**

Table 2. Ablation study for AGMM on the Cityscapes dataset. MT means multi-stage training.

	point-supervised	scribble-supervised
bkg	91.3	93.5
aero	80.3	84.4
bicycle	34.9	39.0
bird	79.3	84.5
boat	68.5	72.8
bottle	62.5	79.1
bus	84.9	94.5
car	79.1	88.2
cat	86.7	91.3
chair	35.8	41.7
cow	86.5	86.9
table	41.0	46.8
dog	81.2	86.0
horse	82.7	87.9
motor	75.8	82.9
person	78.2	83.4
plant	48.0	57.9
sheep	84.4	87.9
sofa	38.3	48.6
train	79.9	86.6
tv	62.1	78.5
mIoU	69.6	76.4

Table 3. IoU results on the val set of PASCAL VOC 2012 dataset.

D. Ablation study for Cityscapes dataset

We conduct thorough experiments on the Cityscapes dataset. Table 2 shows the ablation study for Cityscapes dataset, which demonstrates the effectiveness of our proposed L_{self}, L_{spar}, and L_{con}. In addition, the MT strategy can also gain large improvements on the Cityscapes dataset.

We present the category-wise performance on PASCAL VOC 2012 and Cityscapes datasets in Tables 3 and 4, respectively.

E. Visualization on Cityscapes dataset

We synthesize point annotations on the Cityscapes dataset, as shown in Fig. 5. Given the original images with corresponding full annotations, we randomly select pixels from the original annotations as ground truth for point supervision, while the rest pixels are discarded. We generate the point annotations at 3 levels, including 20 clicks, 50 clicks, and 100 clicks. Each click contains 1×1 pixel. These clicks are sparse but can provide the least category and coarse position information, enabling us to conduct SASS research.

We present qualitative segmentation results of AGMM

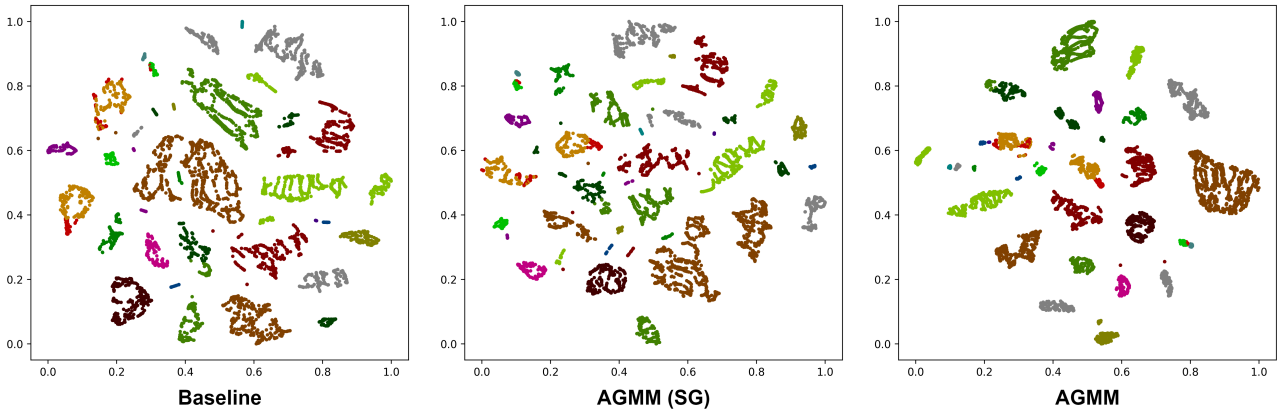

Figure 3. Visualization of the high-level feature space learned by baseline, AGMM (Stop-Gradient), and AGMM, respectively. We use t-SNE to visual the features. The results are presented under the point-supervised SASS setting in the PASCAL VOC 2012 dataset.

	20 clicks	50 clicks	100 clicks
road	94.7	96.5	97.5
sidewalk	73.3	78.2	80.6
building	84.5	88.0	89.9
wall	28.4	37.2	47.6
fence	43.3	48.7	54.3
pole	47.0	53.3	56.0
light	49.1	57.5	62.7
sign	61.8	67.9	72.4
vege	88.2	89.7	91.2
ter. class	51.2	57.5	58.4
sky	89.2	91.8	93.2
person	71.5	75.1	77.5
rider	48.7	52.0	55.1
car	90.1	92.4	93.3
truck	37.7	63.9	60.2
bus	67.4	76.2	80.9
train	46.7	48.4	63.8
motor	39.9	53.3	52.4
bicycle	66.6	69.9	72.5
mIoU	62.1	68.3	71.6

Table 4. IoU results on the val set of Cityscapes dataset.

on Cityscapes dataset in Fig. 6. As shown in Fig. 6, with more complex scenes and clutter backgrounds in the Cityscapes dataset, our AGMM can also predict complete objects and sharp boundaries, which demonstrates the effectiveness of our proposed framework.

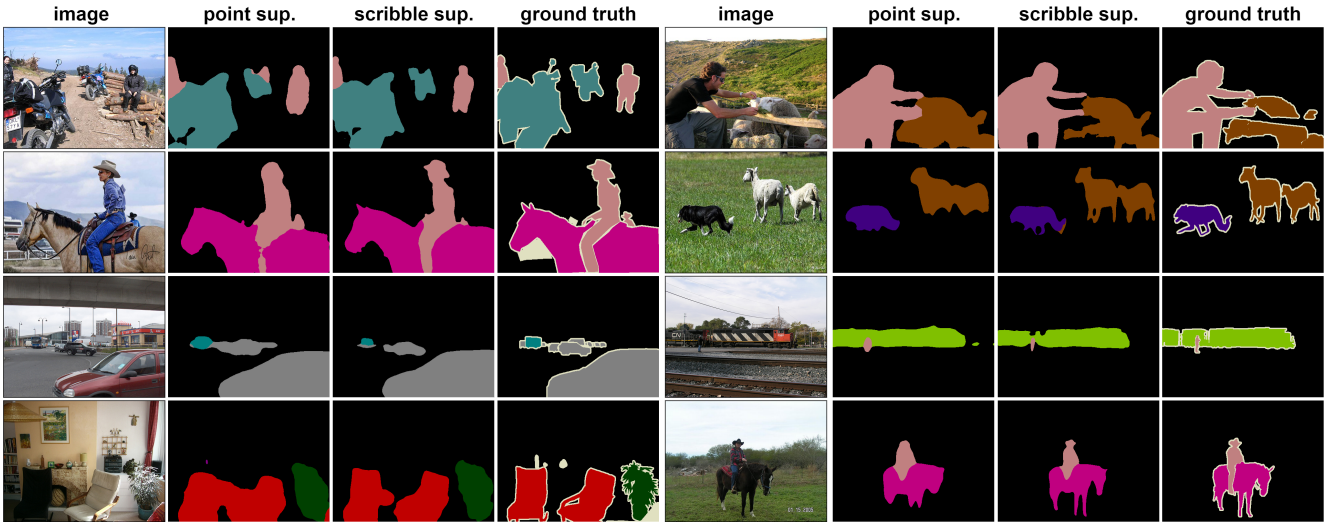

Figure 4. Qualitative segmentation results of AGMM under point- and scribble-supervised settings on PASCAL VOC 2012 dataset.

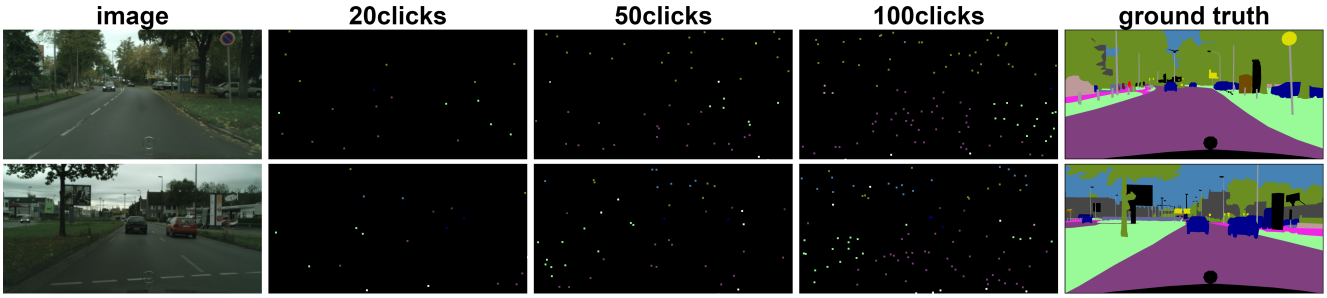

Figure 5. The randomly selected point annotations (20 clicks, 50 clicks, and 100 clicks) for the Cityscapes dataset. We enlarge the clicks from 1×1 size to 40×40 size for better visualization.

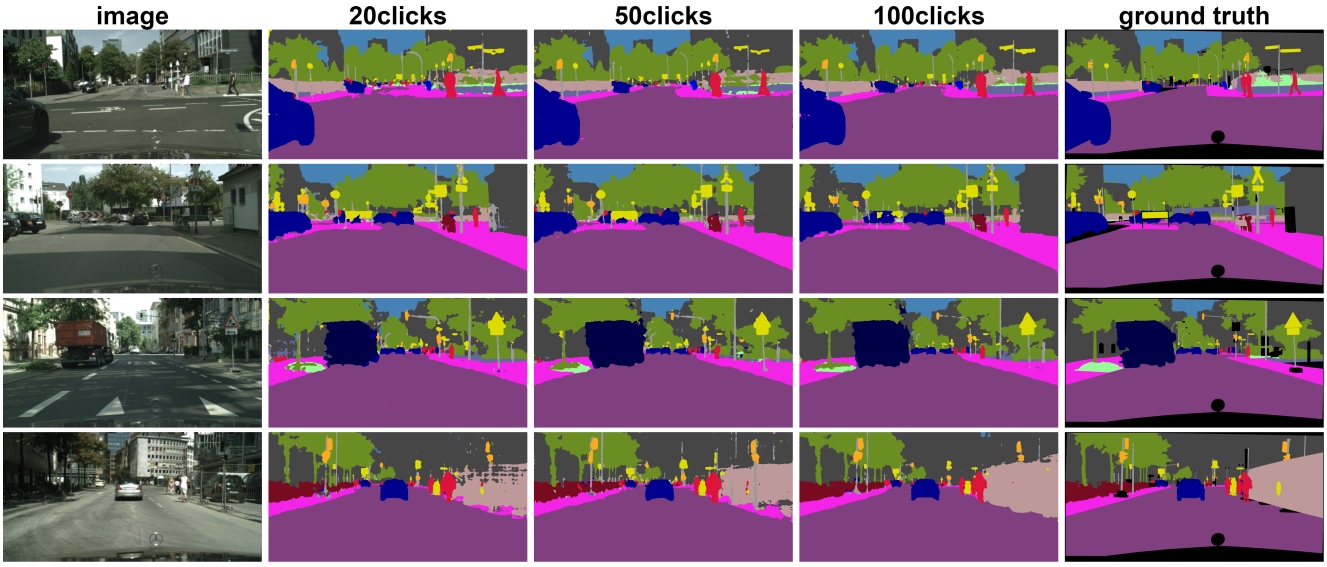

Figure 6. Qualitative segmentation results of AGMM under point-supervised settings (20 clicks, 50 clicks, and 100 clicks per image 2048×1024) on Cityscapes dataset.