Table 5: Ablation Study for Attention Mechanism. ChannelAttention means SENet Block, SpatialAttention means Residual Attention Block. "sep" denotes the support branch and the query branch adopt separate attention. "share" represents the support branch and the query branch share the same attention.

Method	1-shot	#params(M)
MCG	63.3	87.2
MCG-ChannelAttention-sep[1]	63.6	89.6
MCG-ChannelAttention-share[2]	61.7	89.8
MCG-SpatialAttention-sep	63.3	93.3
MCG-SpatialAttention-share	65.8	89.5

[1] For fair comparison with SpatialAttention method, we change the fusion width to 428 to make #param nearly the same.

[2] For fair comparison with SpatialAttention method, we change the fusion width to 480 to make #param nearly the same.

Table 6: Ablation Study for loss function in Conv-LSTM. Baseline is our A-MCG module, we mainly compare the difference between 1-loss Conv-LSTM and 5-loss LSTM. The experiment is conducted on PASCAL-$\boldsymbol{i^5}$ sub-dataset 0.

Method	1-shot	5-shot	#params(M)
baseline	65.8	66.2	89.5
1-loss Conv-LSTM	65.1	67.5	90.8
5-loss Conv-LSTM	66.1	67.9	90.8

shot learning. Interestingly, both the 1-shot and 5-shot result on 5-loss LSTM outperform our baseline, which sufficiently validates our motivation.

Furthermore, we also conduct k-shot learning where k ranges from 1 to 10 in Fig. 6. k-loss Conv-LSTM fully surpasses the traditional logical or method in all shot number range. When $k \leq 4$, the performance of 1-loss Conv-LSTM is less than 5-loss Conv-LSTM, while partially larger than our baseline. This proves that our 5-loss Conv-LSTM better integrates multi-shot support features than traditional method.

Result in PASCAL VOC. As shown in Table 7, our A-MCG architecture could outperform nearly 61.2% in 1-shot mIoU, 62.2% in 5-shot mIoU. Based on our baseline, we continue applying the MCG, attention mechanism, Conv-LSTM, reach a new state-of-the-art result on PASCAL-i^5 dataset in the end.

Table 7: Result on PASCAL-$\boldsymbol{i^5}$ Dataset. All results are computed by taking the average of the 5 sub-datasets in PASCAL-$\boldsymbol{i^5}$. The 5-shot result is obtained by logic or fusion except the method with Conv-LSTM.

Method	1-shot	5-shot	#params(M)
OSLSM (Shaban et al. 2017)	40.8	43.9	276.7
co-FCN(Multi-class) (Rakelly et al. 2018)	50.9	50.9	–
co-FCN(Overall) (Rakelly et al. 2018)	60.1	60.8	–
Baseline	53.0	54.8	85.1
MCG	55.3	56.5	87.2
A-MCG	57.3	57.8	89.5
A-MCG-Conv-LSTM	**61.2**	**62.2**	90.8

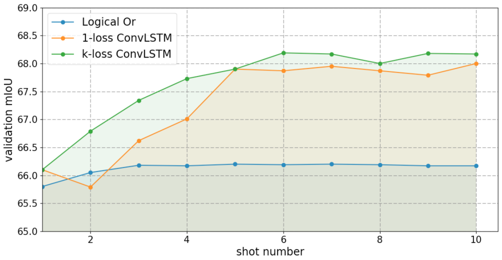

Figure 6: The relationship between shot number and validation mIoU, we mainly compare among three multi-shot learning fusion strategies: (1). Logical Or. (2). 1-loss Conv-LSTM. (3). k-loss Conv-LSTM(k=5 in our experiment). The experiment is conducted on PASCAL-$\boldsymbol{i^5}$ sub-dataset 0.

Table 8: COCO Dataset result.

method	1-shot	5-shot	#params(M)
Baseline	49.98	51.2	85.1
A-MCG-Conv-LSTM	52	54.7	90.8

Result in COCO Dataset. To evaluate our algorithm in more complex dataset, we evaluate our algorithm in COCO dataset. For the COCO dataset evaluation, we divide the 80 classes into 4 sub-dataset, thus every sub-dataset is comprised of 20 classes. We cross-validate the performance of our algorithm and the result is shown in Table 8. As COCO dataset owns much more classes compared with Pascal VOC (80 vs 20). The complexity of this dataset makes our performance much less obvious in COCO than in PASCAL VOC. However, the result in Table 8 demonstrates that our A-MCG-Conv-LSTM model persistently improve our baseline about 3% mIoU both in 1-shot and 5-shot result.

Conclusion

We propose an Attention-based Multi-Context Guiding network (A-MCG) which incorporates multi-level concentrated context. The benefits of our network are three folds: (1). The shallow part of our network generates low-level semantic features, meanwhile deep part of our network captures high-level semantic features. Context features in equal level are fused by our MCG module, which highly facilitates the support branch to globally "support" the query branch. (2). Spatial Attention is employed along with the whole MCG branch, which makes our network focus on different scales of context information. (3). The import of Conv-LSTM enables the network to better integrate the feature from the support set in multi-shot semantic segmentation. The performance of our model surpasses state-of-the-art in few-shot semantic segmentation. In the future, we will exploit few-shot learning in multi-class segmentation at one time.