Appendix: A Transformer-based Decoder for Semantic Segmentation with Multi-level Context Mining

Bowen Shi[1†], Dongsheng Jiang[2†], Xiaopeng Zhang[2], Han Li[1], Wenrui Dai[1], Junni Zou[1], Hongkai Xiong[1] and Qi Tian[2⋆]

[1] Shanghai Jiao Tong University
{sjtu_shibowen,qingshi9974,daiwenrui,zoujunni,xionghongkai}@sjtu.edu.cn
[2] Huawei Cloud EI
dongsheng_jiang@outlook.com, zxphistory@gmail.com, tian.qi1@huawei.com

1 Further Discussion about External Tokens

This section discusses the properties of X_{exter} ($\hat{X}_{exter}$). First, we visualize the attention map between X_{inter} and $\hat{x}_{class}$ of $\hat{X}_{exter}$. The qualitative results shown in Fig. 1 demonstrate that each external token is category-specific since it can capture its corresponding class features. We further calculate the mean cosine similarity (mCS) between X_{inter} and $\hat{x}_{info}$ of $\hat{X}_{exter}$. As shown in Tab. 1, our learnable external tokens enjoy lower mean similarity (0.0060) compared to the momentum update type (0.6363). The result reflects that our X_{exter} has obtained features with hierarchical differences, which can play a better complementary role to X_{inter} and lead to better performance.

2 Comparison with Other Decoders

In this section, we replace the initial MLP decoder of SegFormer-B1 with some famous CNN-based decoders and a transformer-based Segmenter decoder and inspect their performance. *Note that we only replace the structure and adapt the input dimension of the decoder, while other training settings are the same as before.* The results shown in Tab. 2 are surprising. We witness performance drops on all CNN-based decoders, probably because the features extracted by the transformer backbone already have a global receptive field, which is totally different from those of CNN. Using Segmenter decoder only achieved 38.88% accuracy, which is 2.09% lower than the baseline. We think this is because we only use the decoder structure of Segmenter but the tricks of Segmenter have a dominant effect on performance, which greatly limits its transferability. In contrast, our method is the only one that achieves a significant performance improvement, and we have demonstrated the excellent transferability of our method to other backbones in previous experiments.

⋆ Corresponding author. † Equal contribution.

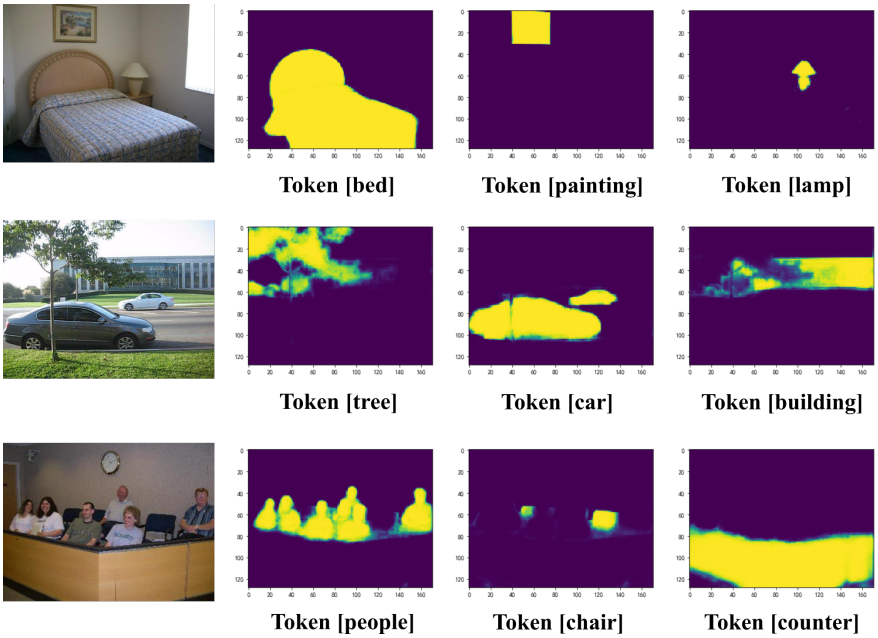

Fig. 1. Visualization of the attention between X_{inter} and $\hat{x}_{class}$ of $\hat{X}_{exter}$.

Table 1. Further studies on the external token based on SegFormer-B1, mCS denotes mean cosine simiarity.

Type of X_{exter}		mCS↓	mIoU↑
Mom.	Learn.		
✓		0.6363	41.74
	✓	0.0060	**42.84**

Table 2. Results of SegFormer-B1 when equipped with different decoders.

Method	mIoU↑
SegFormer-B1	40.97
+ FCN[2]	39.13 (1.84 ↓)
+ PSPNet[5]	38.95 (2.02 ↓)
+ ASPP[1]	39.39 (1.58 ↓)
+ EncNet[4]	39.31 (1.66 ↓)
+ Segmenter[3]	38.88 (2.09 ↓)
+ SegDeformer	**44.05** (3.08 ↑)

References

1. Chen, L., Zhu, Y., Papandreou, G., et. al: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV (2018)
2. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015)
3. Strudel, R., Garcia, R., Laptev, I., et. al: Segmenter: Transformer for semantic segmentation. In: ICCV (2021)
4. Zhang, H., Dana, K., Shi, J., et. al: Context encoding for semantic segmentation. arXiv (2018)
5. Zhao, H., Shi, J., Qi, X., et. al: Pyramid scene parsing network. In: CVPR (2017)