MetaSeg: MetaFormer-based Global Contexts-aware Network for Efficient Semantic Segmentation

Beoungwoo Kang*, Seunghun Moon*, Yubin Cho*, Hyunwoo Yu*, and Suk-Ju Kang
Sogang University, Republic of Korea
{beoungwoo, moonsh97, dbqls1219, hyunwoo137, sjkang}@sogang.ac.kr

Appendix

- In Section A, we provide additional experiments on adopting Transformer-based encoder as a backbone in the proposed MetaSeg network.
- In Section B, we provide a comparison of our MetaSeg with other segmentation networks.
- In Section C, we provide additional results on the inference speed (FPS).
- In Section D, we provide additional qualitative results compared with the proposed and previous model on ADE20K, Cityscapes, COCO-Stuff and Synapse datasets.
- Code and README file were submitted as a zip file for reproducibility.

A. Effectiveness of Our MetaSeg for Various Transformer-based Backbone

In Table 1, we conducted the experiment on using the Transformer-based encoder as a backbone of our MetaSeg. Previously, Mix Transformer (MiT) [8] and Lite Vision Transformer (LVT) [9] backbones adopt SegFormer [8] as its semantic segmentation decoder. Compared to SegFormer [8], our MetaSeg remarkably reduces the computational costs (GFLOPs) by 53.6 % and 43.4 % with mIoU improvements of 1.9% and 0.4% for MiT [8] and LVT [9], respectively. These results indicate that the Transformer-based backbones as well as the CNN-based backbones can effectively leverage our MetaSeg for efficient semantic segmentation task.

B. Comparison of Our MetaSeg with Other Semantic Segmentation Networks

In Table 2, we compared our method with other segmentation networks to demonstrate the power of our MetaSeg. We experimented with the same backbone for a fair comparison. Compared to other networks, our model showed significant computational reduction with higher mIoU performance. This result indicates that our MetaSeg is a powerful and efficient segmentation network by leveraging the MetaFormer block that uses our efficient CRA module as a token mixer.

*These authors contributed equally.

Backbone	Method	Params (M)	ADE20K	
			GFLOPs ↓	mIoU (%) ↑
MiT [8]	SegFormer [8]	3.8	8.4	37.4
	MetaSeg (Ours)	**4.1**	**3.9**	**39.3**
LVT [9]	SegFormer [8]	3.9	10.6	39.3
	MetaSeg (Ours)	**4.2**	**6.0**	**39.7**

Table 1. Effectiveness of our MetaSeg for Transformer-based backbones on ADE20K validation set.

Method	Params (M)	GFLOPs ↓	mIoU (%) ↑
FCN [5]	9.8	39.6	19.7
PSPNet [10]	13.7	53.0	29.7
DeepLabV3 [1]	18.7	75.4	34.1
DeepLabV3+ [2]	15.4	69.5	34.0
MetaSeg (Ours)	**3.4**	**4.6**	**34.7**

Table 2. Comparison of our MetaSeg with other segmentation Networks on ADE20K dataset. For a fair comparison, we use the same backbone, MobileNetV2 [6].

Method	Params(M)	Cityscapes		
		GFLOPs ↓	mIoU (%) ↑	FPS (img/s)↑
SegFormer-B0 [8]	3.8	125.5	76.2	12.52
FeedFormer-B0 [7]	4.5	107.4	77.9	17.33
SegNeXt-T [3]	4.3	50.5	79.8	22.73
MetaSeg-T (Ours)	4.7	**47.9**	**80.1**	**23.46**

Table 3. Comparison of our MetaSeg with previous state-of-the-art methods on Cityscapes.

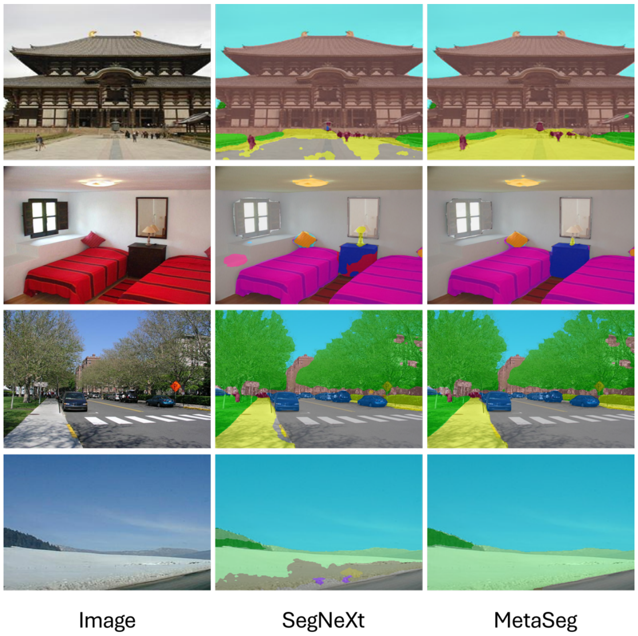

Figure 1. Qualitative results on ADE20K. Compared to the previous state-or-the-art method, our MetaSeg generates more accurate segmentation maps across various categories.

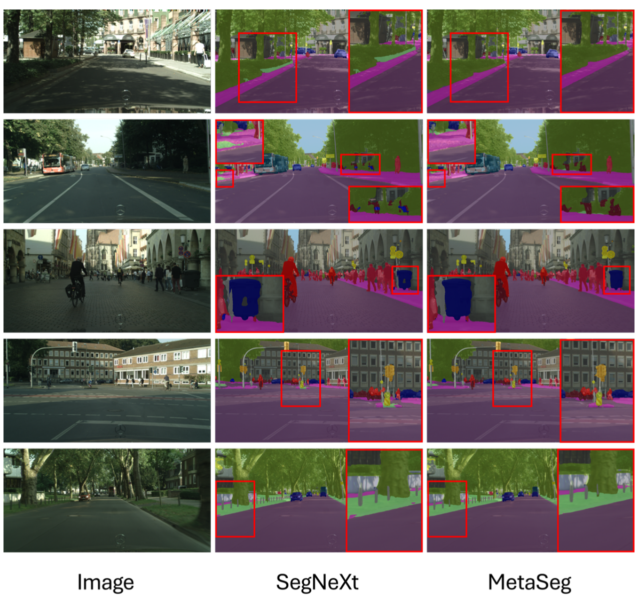

Figure 2. Qualitative results on Cityscapes. For multiple categories, our MetaSeg provides more precise predictions than SegNeXt [3].

C. Inference Speed Comparison

In Table 3, we represent the inference speed comparisons under the mmsegmentation code base without any additional accelerating techniques. We tested Frame Per Second (FPS) of a single image of 1024×2048 on Cityscapes test dataset using a single RTX3090 GPU. The results show that our MetaSeg is fastest compared with other lightweight semantic segmentation models [3, 7, 8], while achieving the highest mIoU performance.

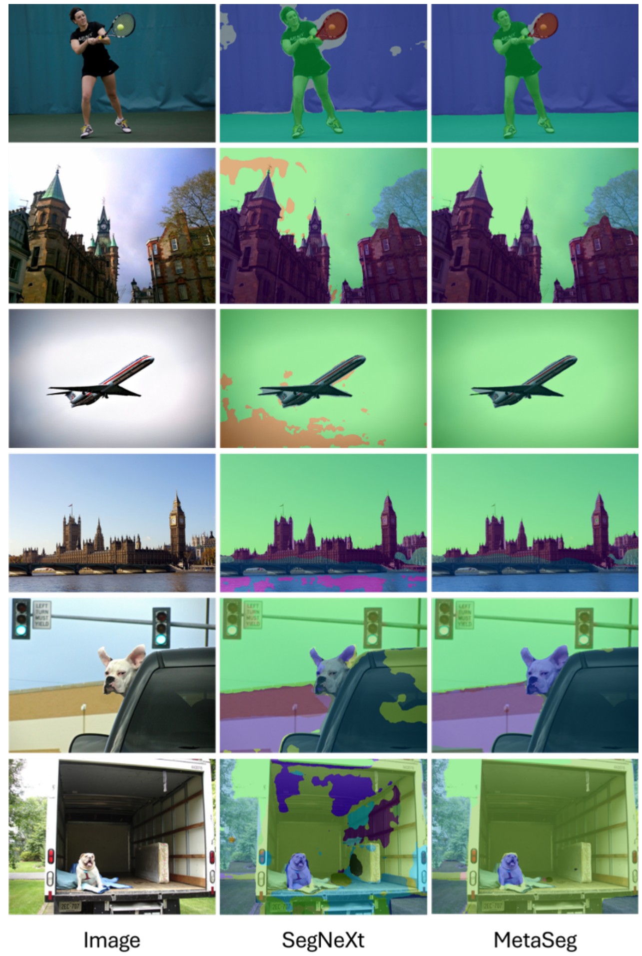

Figure 3. Qualitative results on COCO-Stuff. Compared to SegNeXt [3], our MetaSeg provides more detailed segmentation predictions.

D. Additional Qualitative Results

In Fig.1, 2 and 3, we visualized additional qualitative results of our MetaSeg and the previous state-of-the-art method on ADE20K, Cityscapes, COCO-Stuff datasets, respectively. Compared to SegNeXt [3], our MetaSeg showed more accurate predictions for large regions. Our MetaSeg also predicted more detailed for the object boundaries than SegNeXt [3]. In addition, we visualized more qualitative results of our MetaSeg and HiFormer [4] on Synapse dataset. As shown in Fig.4, our MetaSeg predicted more precisely than HiFormer [4] for various categories. These

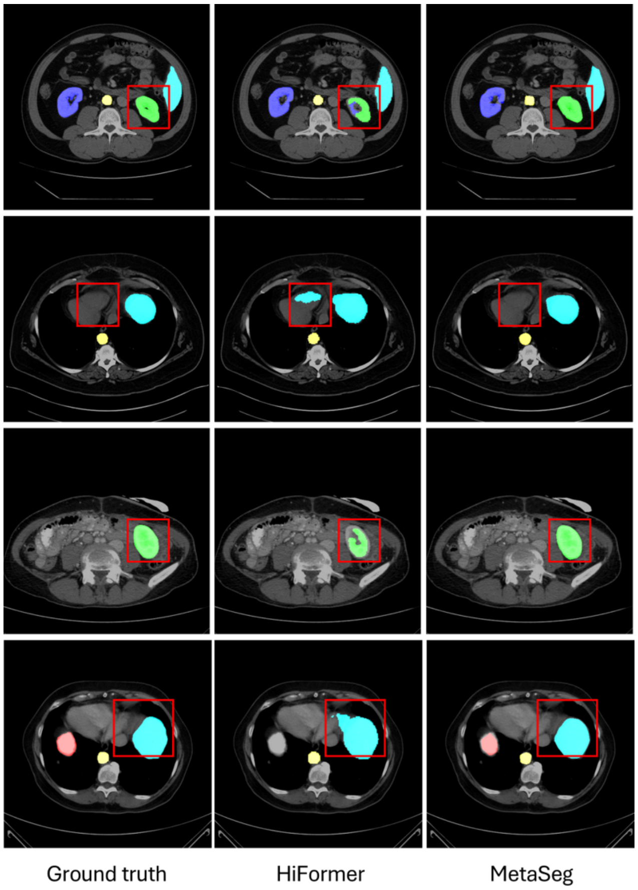

Figure 4. Qualitative results on Synapse. Our MetaSeg predicts more precisely than HiFormer [4] across various categories.

results indicate that our MetaSeg can effectively capture the local to global information by extensively leveraging the MetaFormer architecture from the encoder to the decoder.

References

[1] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 1

[2] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 1

[3] Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-Min Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. *arXiv preprint arXiv:2209.08575*, 2022. 1, 2

[4] Moein Heidari, Amirhossein Kazerouni, Milad Soltany, Reza Azad, Ehsan Khodapanah Aghdam, Julien Cohen-Adad, and Dorit Merhof. Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6202–6212, 2023. 2, 3

[5] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1

[6] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 1

[7] Jae-hun Shim, Hyunwoo Yu, Kyeongbo Kong, and Suk-ju Kang. Feedformer: Revisiting transformer decoder for efficient semantic segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2263–2271, 2023. 1, 2

[8] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 1, 2

[9] Chenglin Yang, Yilin Wang, Jianming Zhang, He Zhang, Zijun Wei, Zhe Lin, and Alan Yuille. Lite vision transformer with enhanced self-attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11998–12008, 2022. 1

[10] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 1