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.
Figure 1. Qualitative results on ADE20K. Compared to the previous state-of-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 $1024 \times 2048$ 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.

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
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


