U-MixFormer: UNet-like Transformer with Mix-Attention for Efficient Semantic Segmentation

Seul-Ki Yeom
Nota AI GmbH
Mariendorfer Damm 1, 12099 Berlin, Germany
skyeom@nota.ai

Julian von Klitzing
Nota AI GmbH
Mariendorfer Damm 1, 12099 Berlin, Germany
julian.von.klitzing@campus.tu-berlin.de

A. Experimental Results

A.1. Training and Evaluation Setting

We followed the default settings provided by mmsegmentation. Models were trained on a server with 2 NVIDIA A100 GPUs, using pre-trained encoders from the ImageNet-1K dataset. Training included augmentations like random resizing (with a ratio between 0.5 and 2), random horizontal flipping, and cropping — to dimensions of 512×512 for the ADE20K dataset and 1024×1024 for Cityscapes. Following [1], for our largest MiT encoder, B5, we adjusted the cropping size to 640×640 on ADE20K. AdamW optimizer was employed across 160K iterations for both datasets. We set the batch sizes to 16 for ADE20K and 8 for Cityscapes. We initialized the learning rate at 6e-5 and adopted a polynomial learning rate decay schedule with a default factor of 1.0. For the loss function, we employed a standard cross-entropy loss with a weight of 1.0, ensuring robust training stability and balanced class representation.

Evaluations were conducted on the ADE20K and Cityscapes *valid*. Particularly Cityscapes was used for a sliding window, cropping into windows size of 1024×1024. The semantic segmentation results regarding the mean Intersection over Union (mIoU) based on a single-scale inference paradigm are presented.

A.2. Additional Qualitative Results

In addition to the qualitative results presented in Figure 5 for U-MixFormer, SegFormer, and FeedFormer, Supplementary Figure 1 shows further examples of U-MixFormer's superiority in accurately segmenting object boundaries.

A.3. Effectiveness of Decoding Head with the same MiT Encoder

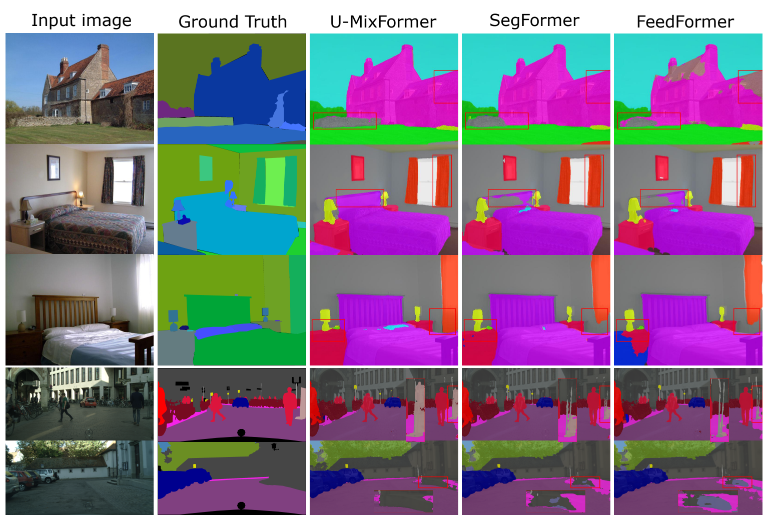
For convenience, Supplementary Table 1 summarizes the results of Tables 1 and 2 for U-MixFormer, SegFormer, and FeedFormer, which share the same encoder MiT in different sizes from the smallest B0 to the largest B5. Considering the same size for MiT, U-MixFormer achieves higher performance (mIoU) while maintaining lower computational cost (GFLOPs).

References

[1] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, October 2021. 1

Supplementary Table 1. Performance and efficacy comparison of three transformer decoders, using the MiT encoder in varying size (B0 - B5) on ADE20K.

Method	GFLOPs ↓						mIoU ↑					
	B0	B1	B2	B3	B4	B5	B0	B1	B2	B3	B4	B5
SegFormer	8.4	15.9	62.4	79.0	95.7	183.3	37.4	42.2	46.5	49.4	50.3	51.0
FeedFormer	7.8	-	42.7	-	-	-	39.2	-	48.0	-	-	-
U-MixFormer	**6.1**	**17.8**	**40.0**	**56.8**	**74.5**	**152.5**	**41.2**	**45.2**	**48.2**	**49.8**	**50.4**	**52.0**



Supplementary Figure 1. Qualitative analysis on ADE20K and Cityscapes datasets for U-MixFormer, SegFormer, and FeedFormer. All methods utilize the same encoder MiT-B0. U-MixFormer's superior object boundaries segmentation: first row (house/wall), second row (bed/wall), third row (box/lamp), fourth row (group of "pole" segments), fifth row (group of "building" segments)