Supplementary Materials for Token Pyramid Transformer

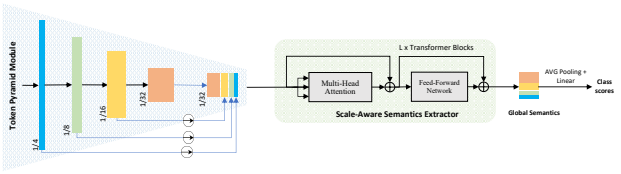


Figure 1. The classification architecture of the proposed Token Pyramid Transformer.

This chapter presents additional materials and results. We give the ImageNet pretraining results in Section A. Then we describe the specific network structure in Section B. Next, we give the performance on Cityscapes. Finally, some visual results are provided.

A. ImageNet Pre-training

For fair comparison, we also use the ImageNet pretrained parameters as initialization. As shown in Figure 1, the classification architecture of the proposed TopFormer appends the average pooling layer and Linear layer on the global semantics for producing class scores. Due to the small resolution (224×224) of input images, we set the target resolution of input tokens of the Semantics Extractor is $\frac{1}{32\times32}$ of input size. The classification results are shown in Table 1. Because our target task is mobile semantic segmentation, we do not explore more technologies, e.g. more epochs and distillation uesd in LeViT, to further improve the accuracy. In the future work, we will continue to improve the classification accuracy.

Model	Params	FLOPs	Top-1 Acc(%)
TopFormer-T	1.50M	126M	66.2
TopFormer-S	3.11M	235M	72.3
TopFormer-B	5.07M	373M	75.3

Table 1. The results on ImageNet classification.

B. Network Structure

The detailed network structures are given in Table 3. Although the Token Pyramid Module have the most layers, as the statistics of the computation and parameters in the paper, the ViT-based Semantics Extractor accounts for the vast majority of parameters.

Methods	Encoder	GFLOPs	mIoU
FCN	MobileNetV2	317.1	61.5
PSPNet	MobileNetV2	423.4	70.2
Segformer	MiT-B0	17.7	71.9
L-ASPP	MobileNetV2	12.6	72.7
LR-ASPP	MobileNetV3-large	9.7	72.4
LR-ASPP	MobileNetV3-small	2.9	68.4
Ours(h)	TopFormer-B	2.7	70.7
Ours(f)	TopFormer-B	11.2	75.0

Table 2. Results on Cityscapes val set. Ours(f) and Ours(h) are denoted as taking a full-resolution input (i.e., 1024×2048) and a half-resolution input (i.e., 512×1024).

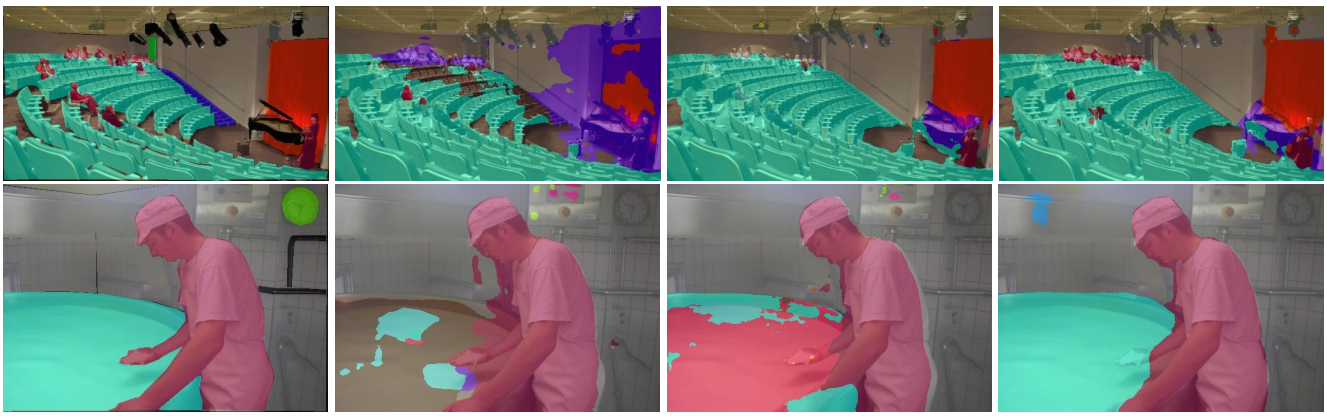
C. The Performance on Cityscapes

Training Settings Our implementation is based on MMSegmentation and Pytorch. We perform 80K iterations. The initial learning rate is 0.0003 and weight decay is 0.01. A poly learning rare scheduled with factor 1.0 is used. For full-resolution version, the training images are randomly scaling and then cropping to fixed size of 1024×1024. As for the half-resolution version, the training images are resized to 1024×512 and randomly scaling, the crop size is 1024×512. We follow the data augmentation strategy of Segformer for fair comparison.

Experimental results To validate the performance of the proposed method, we directly fed a full-resolution input and a half-resolution input into the trained segmentation models for testing, respectively. As shown in Table 2, the proposed method with a full-resolution, denoted as Ours(f), achieves about 2.6% higher accuracy in mIoU than L-ASPP based on MobileNetV2 with lower computation. The experimental results demonstrate that TopFormer could achieve good trade-off between accuracy and computation even if the input image is with large resolution.

Stage	Output size	Tiny	Small	Base
Token Pyramid Module	256 × 256	Conv,3 × 3, 16, 2 MB, 3, 1, 16, 1		
	128 × 128	MB, 3, 4, 16, 2 MB, 3, 3, 16, 1	MB, 3, 4, 24, 2 MB, 3, 3, 24, 1	MB, 3, 4, 32, 2 MB, 3, 3, 32, 1
	64 × 64	MB, 5, 3, 32, 2 MB, 5, 3, 32, 1	MB, 5, 3, 48, 2 MB, 5, 3, 48, 1	MB, 5, 3, 64, 2 MB, 5, 3, 64, 1
	32 × 32	MB, 3, 3, 64, 2 MB, 3, 3, 64, 1	MB, 3, 3, 96, 2 MB, 3, 3, 96, 1	MB, 3, 3, 128, 2 MB, 3, 3, 128, 1
	16 × 16	MB, 5, 6, 96, 2 MB, 5, 6, 96, 1	MB, 5, 6, 128, 2 MB, 5, 6, 128, 1 MB, 3, 6, 128, 1	MB, 5, 6, 160, 2 MB, 5, 6, 160, 1 MB, 3, 6, 160, 1
Semantics Extractor	8 × 8	L=4,H=4	L=4,H=6	L=4,H=8
Semantics Injection Module	16 × 16,32 × 32,64 × 64	M=128	M=192	M=256
FLOPs		0.6G	1.2G	1.8G

Table 3. Detailed architecture configs of the proposed method. The input is with resolution 512 × 512. For Token Pyramid Module, Conv refers to regular convolution layer, [MB, 5, 3, 48, 1] refers to MobileNetV2 block with kernel size=5, expand ratio=3, output channels=48 and stride=1. For Semantics Extractor, L is the number of Transformer Blocks. H is the number of heads in a multi-head self-attention block.



(a) GT **(b)** MBV2-DeepLabV3+ **(c)** Segformer-B0 **(d)** Ours

Figure 2. The visualization of the Ground Truth, MBV2-Deeplabv3+, SegFormer-B0 and the proposed TopFormer on ADE20K val set. We use TopFormer-B to conduct visualization.

D. Visualization

We present some visualization comparisons among the proposed TopFormer and other CNNs- and ViT-based methods on the ADE20K validation (val) set in Figure 2. Here, we choose deeplabv3+ based on mobilenetV2 as a representative of CNNs-based methods and Segformer as a representative of ViT-based methods. These two methods both have larger model size and computational cost. As shown in Figure 2, the proposed method could achieve better segmentation results than these two methods.