Supplementary: Dynamic Token Pruning in Plain Vision Transformers for Semantic Segmentation

Quan Tang[1] Bowen Zhang[2] Jiajun Liu[3] Fagui Liu[1] Yifan Liu[2]

[1]South China University of Technology [2]The University of Adelaide [3]CSIRO

1. Visualized Results

The computation is unevenly allocated among different images when applying the proposed DToP, which attributes computation cost to dissimilar recognition difficulties. We present visualized examples for a simple illustration in Figure 1. We see that the reduction of computation cost in GFLOPs can be as high as 57.7% in simple-scene images, such as the example in the first row that contains only the building and sky. For complex-scene images where the object number increases and the scale varies, fewer tokens trigger the early exit, and less GFLOPs reduction is obtained. Even though the computation cost fluctuates among images, the segmentation accuracy remains stable compared with the baseline results.

2. Downsampling Methods

DToP serves as an unsymmetrical downsampling operator by making an early exit of easy tokens. We compare several commonly used symmetrical downsampling operators, including stride convolution, average pooling and nearest sampling. We apply these operators at the end of the 9th layer of ViT-Base [1] backbone to ensure an approximate computation overhead. The shrunk architecture in SegViT [8] is also presented for comparison, which is considered an unsymmetrical downsampling method. Results with the ADE20K dataset [9] are shown in Table 1. The proposed DToP outperforms all symmetrical methods by a large margin.

3. Comparison with Expedite-ViT

Expedite-ViT [3] proposes a token clustering layer to merge similar tokens in the middle network to reduce computations. And it adopts a corresponding token reconstruction layer to rebuild the original tokens before the final prediction. The conceptual difference between our DToP and Expedite-ViT is that DToP makes early predictions on easy tokens by assessing all token recognition difficulties. Here we summarize their experimental comparisons in Table 2 using the Segmenter [6] framework. Our DToP achieves superior performance on the Segmenter baseline.

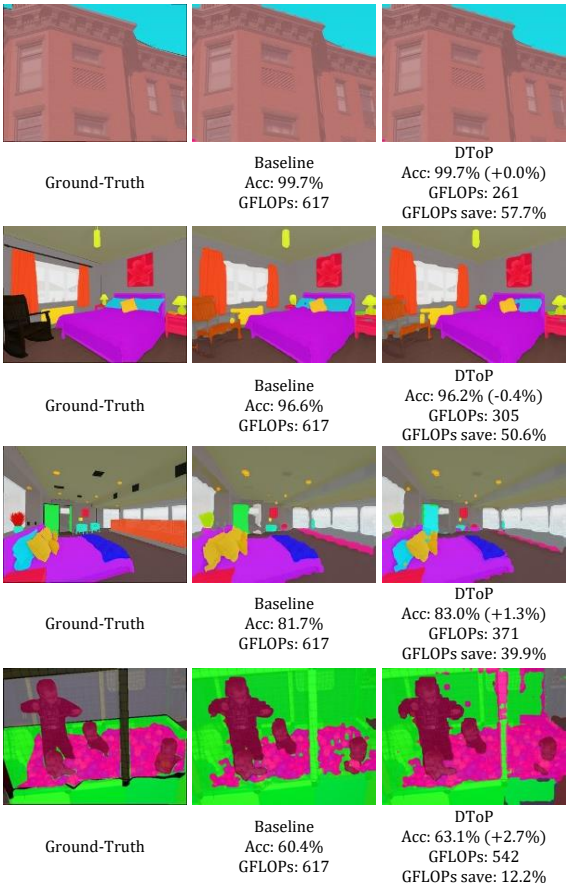

Figure 1: Visualised examples using ADE20K dataset. As computing the intersection over union (IoU) is unreasonable within a single image, we use the category-agnostic pixel accuracy (Acc) instead. GFLOPs means float-point operations in Giga. Best viewed in color.

4. Per-Category Results

We present in Figure 2 per-category scores on the Pascal Context [5] dataset as an example. We observe that DToP

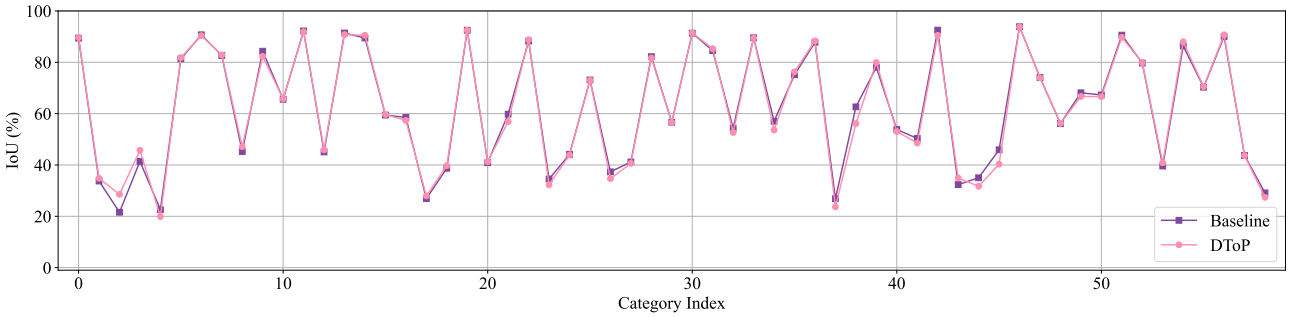

Figure 2: Per-category scores on Pascal Context dataset with 59 classes excluding *background*.

Methods	GFLOPs	mIoU (%)
Baseline	109.9	49.7
$Conv$, $stride = 2$	88.4	44.8
2×2 average pool	87.8	44.4
2×2 nearest sampling	87.8	46.1
SegViT shrunk [8]	97.1	50.0
DToP (ours)	86.8	49.8

Table 1: Comparisons to standard symmetrical downsampling methods under similar computation budget. All methods except baseline follow the @Finetune training scheme.

Method	GFLOPs	mIoU(%)
Segmenter [6]	129.6	49.6
+ Expedite-ViT [3]	100.5	48.9
+ ours ($p_0 = 0.90$)	**98.8**	49.9
+ ours ($p_0 = 0.95$)	106.5	**50.3**

Table 2: Comparisons with Expedite-ViT using ADE20K based on Segmenter. ViT-Base is adopted as the backbone.

yields negligible performance impact on each category as it *finalizes* easy tokens' predictions instead of *discarding* then *rebuilding* them and making predictions at the final layer.

5. Why Plain ViT

We focus our pruning method for semantic segmentation on the plain ViT backbone [1] as it offers several advantages over pyramid structures [4] or efficient transformers. The plain ViT structure has the potential to unify multiple dense prediction tasks and can be improved with more flexible self-supervised methods [2, 7]. Additionally, it is capable of connecting visual and language inputs, allowing zero-/few-shot and continuous learning for dense prediction tasks. The proposed dynamic pruning method enables a new paradigm for using ViT in the future. This approach allows for a foundation ViT to be trained on large-scale datasets yet still be applied to various local datasets with flexible computational reduction.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. Int. Conf. Learn. Representations*, 2021. 1, 2

[2] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 16000–16009, 2022. 2

[3] Weicong Liang, Yuhui Yuan, Henghui Ding, Xiao Luo, Weihong Lin, Ding Jia, Zheng Zhang, Chao Zhang, and Han Hu. Expediting large-scale vision transformer for dense prediction without fine-tuning. *Advances in Neural Information Processing Systems*, 35:35462–35477, 2022. 1, 2

[4] 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 *Proc. IEEE Int. Conf. Comp. Vis.*, pages 10012–10022, 2021. 2

[5] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 891–898, 2014. 1

[6] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 7262–7272, 2021. 1, 2

[7] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 9653–9663, 2022. 2

[8] Bowen Zhang, Zhi Tian, Quan Tang, Xiangxiang Chu, Xiaolin Wei, Chunhua Shen, et al. Segvit: Semantic segmentation with plain vision transformers. *Advances in Neural Information Processing Systems*, 35:4971–4982, 2022. 1, 2

[9] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 633–641, 2017. 1