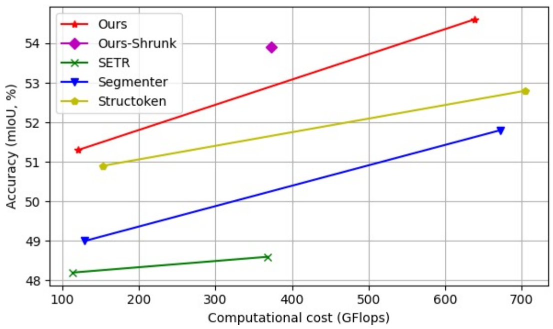

Figure 1: Accuracy vs. computational cost on the ADE20K `val.` split. All the performances and computational costs are measured at single scale inference. Our proposed SegViT structure can achieve a better trade-off and the best performance among methods using ViT backbone.

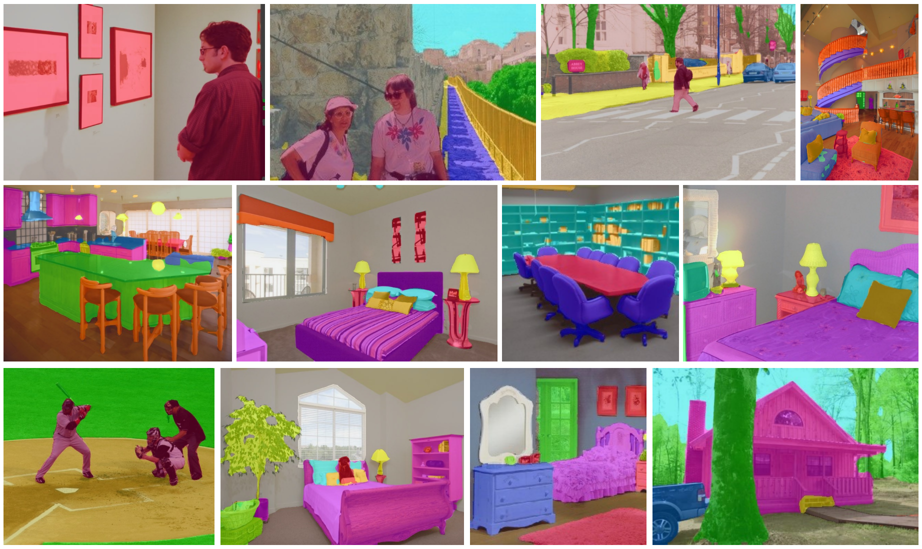

Figure 2: Competitive segmentation results on the ADE20K `val.` split.

A Appendix

In this section, we show more evaluation results to demonstrate the performance of the proposed SegViT structure.

A.1 Illustration of the accuracy vs. computational cost

As shown in Fig. 1, we achieve the best performance with a better trade-off between computational cost and accuracy compared to other methods that use ViT backbone. Also, for our SegViT *Shrunk* version, we dramatically reduced the computation while still retaining competitive performance.

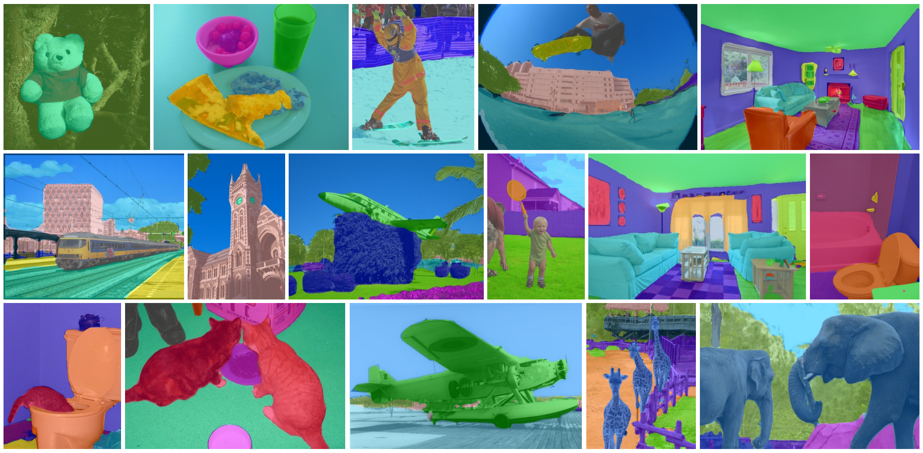

Figure 3: Competitive segmentation results on the COCO-stuff-10K dataset.

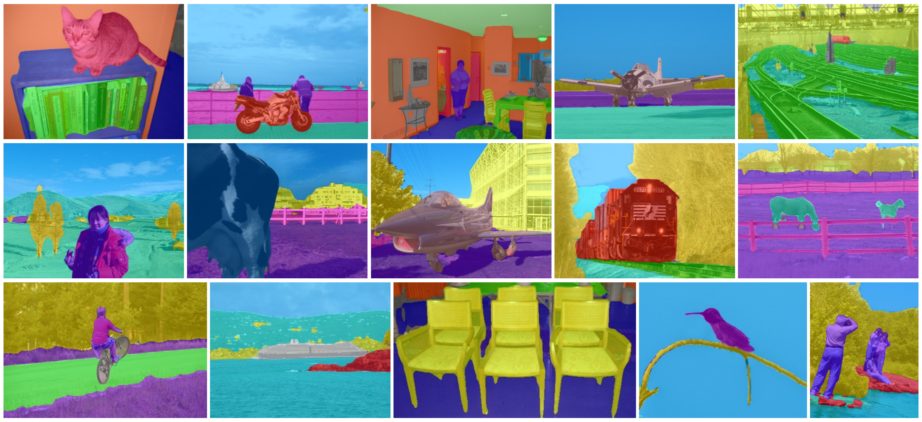

Figure 4: Competitive segmentation results on the PASCAL-Context dataset with 60 classes.

A.2 More Visualization Results

Competitive segmentation results on three datasets. Fig. 2, Fig. 3 and Fig. 4 illustrates the model performance on ADE20K, COCO-Stuff-10K and PASCAL-Context datasets respectively using single-scale inference. We can see that in various indoor and outdoor scenes, SegViT can produce satisfactory segmentation results.

A.3 More Ablation Study Results

Different decoder methods have their corresponding feature merge types and loss types. We compare the difference on a plain ViT base backbone shown in Table 1. For hierarchical backbones, such as Swin, since the resolution of the

Table 1: Ablation results of different decoder methods with their corresponding feature merge types and loss types. ViT-Base is employed as the backbone for all the variants.

	Multi-level Features		Loss Types			
Decoder	FPN	Token Merge	Pixel level	Dot product	Attention Mask	mIoU (ss)
SETR-MLA [3]	✓		✓			48.2
Segmenter [4]			✓			49.0
MaskFormer [2]	✓			✓		46.7
Ours-Variant 1					✓	49.6
Ours-Variant 2		✓		✓		50.6
Ours		✓			✓	51.2

feature maps of each stage is reduced, to get feature maps with large resolution and rich semantic information, FPN is necessary. However, in Table 1, FPN structure can not work well with plain vision transformers. For non-hierarchical backbones, such as ViT, the resolution is maintained and the last layer feature map contains the richest semantic information. Thus, our proposed method that uses the tokens to merge features of different levels got better performance. By simply changing the FPN structure with out ATM based Token merge, we improve the performance from 46.7% to 50.6%. For the loss type, pixel level loss indicates the regular cross-entropy loss applied to the feature map. The dot product loss indicates the loss used in [1] and [2]. Attention mask loss means the mask supervision is directly applied to the similarity map generated by the ATM during attention calculation. We can see that the loss supervised on the attention mask further improve the result with 0.6%.

References

[1] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comp. Vis.*, pp. 213–229, Springer, 2020.

[2] B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," *Proc. Advances in Neural Inf. Process. Syst.*, vol. 34, 2021.

[3] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr, *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 6881–6890, 2021.

[4] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE Int. Conf. Comp. Vis.*, pp. 7262–7272, 2021.