

Appendix for "SegNeXt: Rethinking Convolutional Attention Design for Semantic Segmentation"

Anonymous Author(s)

Affiliation

Address

email

1 Training Details

We show some training details on different datasets omitted in the main paper in Tab. 1. For different benchmarks, we employ different training settings for fair comparison.

Table 1: Training details on different benchmarks. 80K + 80K means we pretrain 80K iterations on Pascal VOC trainaug set and finetune 80K on its trainval set. 80K + 40K denotes we pretrain 600K iterations on COCO dataset and finetune 40K on its trainval set.

Dataset	Crop Size	Batch Size	Iterations
ADE20K [10]	512×512	16	160K
Cityscapes [2]	$1,024 \times 1,024$	8	160K
COCO-Stuff [1]	512×512	16	80K
Pascal VOC [3]	512×512	16	80K + 80K
Pascal VOC [3] w/ COCO [5]	512×512	16	600K + 40K
Pascal Context [7]	480×480	16	80K
iSAID [9]	896×896	16	160K

2 Ablation about MSCA Head

In addition to using a variant of self-attention as our head, we also used MSCA as our head. Results in Tab. 2 show Ham head [4] achieves a better performance than MSCA head, which demonstrates a CNN-style encoder requires a segmentation head with a global receptive field.

Table 2: Performance of different head in decoder. SegNeXt-T w/ Ham means the MSCAN-T encoder plus the Ham decoder. FLOPs are calculated using the input size of 512×512 . Experiments are conducted on COCO-Stuff dataset.

Architecture	Params. (M)	GFLOPs	mIoU (SS)	mIoU (MS)
SegNeXt-T w/ MSCA	4.4	6.7	38.2	38.6
SegNeXt-T w/ Ham [4]	4.3	6.6	38.7	39.1
SegNeXt-S w/ MSCA	14.0	15.9	42.1	42.4
SegNeXt-S w/ Ham [4]	13.9	15.9	42.2	42.8
SegNeXt-B w/ MSCA	28.0	33.6	45.1	45.5
SegNeXt-B w/ Ham [4]	27.6	34.9	45.8	46.3
SegNeXt-L w/ MSCA	50.1	69.8	45.9	46.4
SegNeXt-L w/ Ham [4]	48.9	70.0	46.5	47.2

8 3 More Qualitative Results

9 In the main paper, we show the qualitative results on Cityscapes dataset. Here, we display qualitative
 10 results on ADE20K dataset in Fig. 1. The figure clearly shows that our method is better at
 11 understanding the details.



Figure 1: Qualitative results on ADE20K dataset. Left: SegFormer-B2. Middle: SegNeXt-B. Right: Ground truth.

12 4 Visualization results

13 We adopt Grad-CAM [8] to conduct visualization. As shown in Fig. 2, we can clearly find our
 14 MSCAN shows better visualization results. In particular, when object occupies most of area in an
 15 image (shown in first three columns) or multiple objects in an image (shown in last three columns),

16 ConvNeXt [6] appears inaccurate, while our MSCAN still works well. It shows the effect of large
 17 receptive field and multi-branch structure.

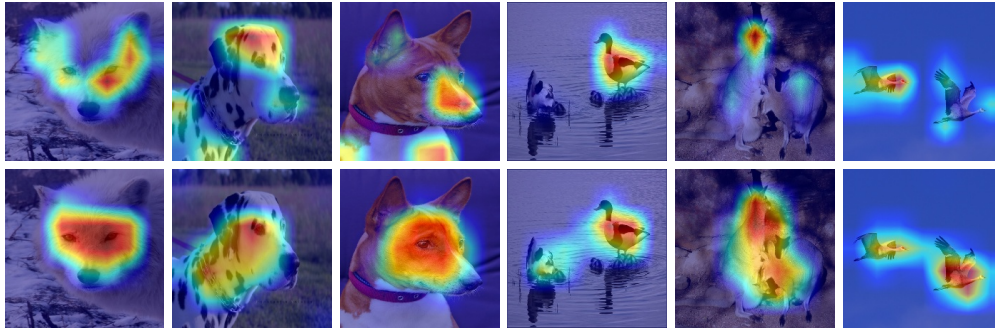


Figure 2: Visualization results by using Grad-CAM [8]. Top: grad-cam figures of ConvNeXt [6]. Bottom: grad-cam figures of our MSCAN.

18 References

- 19 [1] Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: IEEE
 20 Conf. Comput. Vis. Pattern Recog. pp. 1209–1218 (2018)
- 21 [2] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth,
 22 S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: IEEE Conf.
 23 Comput. Vis. Pattern Recog. pp. 3213–3223 (2016)
- 24 [3] Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual
 25 object classes (voc) challenge. *Int. J. Comput. Vis.* **88**(2), 303–338 (2010)
- 26 [4] Geng, Z., Guo, M.H., Chen, H., Li, X., Wei, K., Lin, Z.: Is attention better than matrix
 27 decomposition? In: *Int. Conf. Learn. Represent.* (2021)
- 28 [5] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.:
 29 Microsoft coco: Common objects in context. In: *Eur. Conf. Comput. Vis.* pp. 740–755. Springer
 30 (2014)
- 31 [6] Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s
 32 (2022)
- 33 [7] Mottaghi, R., Chen, X., Liu, X., Cho, N.G., Lee, S.W., Fidler, S., Urtasun, R., Yuille, A.: The
 34 role of context for object detection and semantic segmentation in the wild. In: *IEEE Conf.*
 35 *Comput. Vis. Pattern Recog.* pp. 891–898 (2014)
- 36 [8] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual
 37 explanations from deep networks via gradient-based localization. In: *Int. Conf. Comput. Vis.*
 38 pp. 618–626 (2017)
- 39 [9] Waqas Zamir, S., Arora, A., Gupta, A., Khan, S., Sun, G., Shahbaz Khan, F., Zhu, F., Shao,
 40 L., Xia, G.S., Bai, X.: isaid: A large-scale dataset for instance segmentation in aerial images.
 41 In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*
 42 *Workshops.* pp. 28–37 (2019)
- 43 [10] Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through
 44 ade20k dataset. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 633–641 (2017)