Axial-DeepLab: Stand-Alone Axial-Attention for Panoptic Segmentation (Supplementary Material)

Anonymous ECCV submission

Paper ID 1564

In this supplementary material, we mainly include more visualizations and analyses on COCO validation set. It shows that Axial-DeepLab learns long range attention, and is more robust to occlusion than the previous state-of-the-art. We also show another test server result on Cityscapes. In addition, we include more details of our axial-decoder design, the datasets we use, and some raw data that is used to generate figures in the main paper.

1 COCO Visualization

In Fig. 1, we visualize some panoptic segmentation results on COCO val set. Our Axial-DeepLab-L demonstrates robustness to occlusion, compared with Panoptic-DeepLab (Xception-71).

In Fig. 2 and Fig. 3, we visualize the attention maps of our Axial-DeepLab-L on COCO val set. We visualize a low level block (stage 3 block 2) and a high level block (stage 4 block 3), which are respectively the first block and the last block with resolution 65×65, in the setting of output stride 16. We notice that in our multi-head axial-attention, some heads learn to focus on local details while some others focus on long range context. Additionally, we find that some heads are able to capture positional information and some others learn to correlate with semantic concepts

In Fig. 4, we compare Axial-DeepLab with Panoptic-DeepLab [1], in terms of the three training loss functions, defined in Panoptic-DeepLab [1]. We observe that Axial-DeepLab is able to fit data better, especially on the offset prediction task. This also demonstrates the effectiveness of our position-sensitive attention design, and the long range modeling ability of axial-attention.

2 Cityscapes Test Set

In this section, we report the performance of Axial-DeepLab-L on Cityscapes test set in Tab. 1, without using extra data. In this setting, our Axial-DeepLab-L outperforms Panoptic-DeepLab by 0.4% PQ, without using the trick of output stride 8 [1]. In addition, we directly evaluate the model trained on the training set, without exploiting the validation set. We will probably see more gain if we train on both sets.

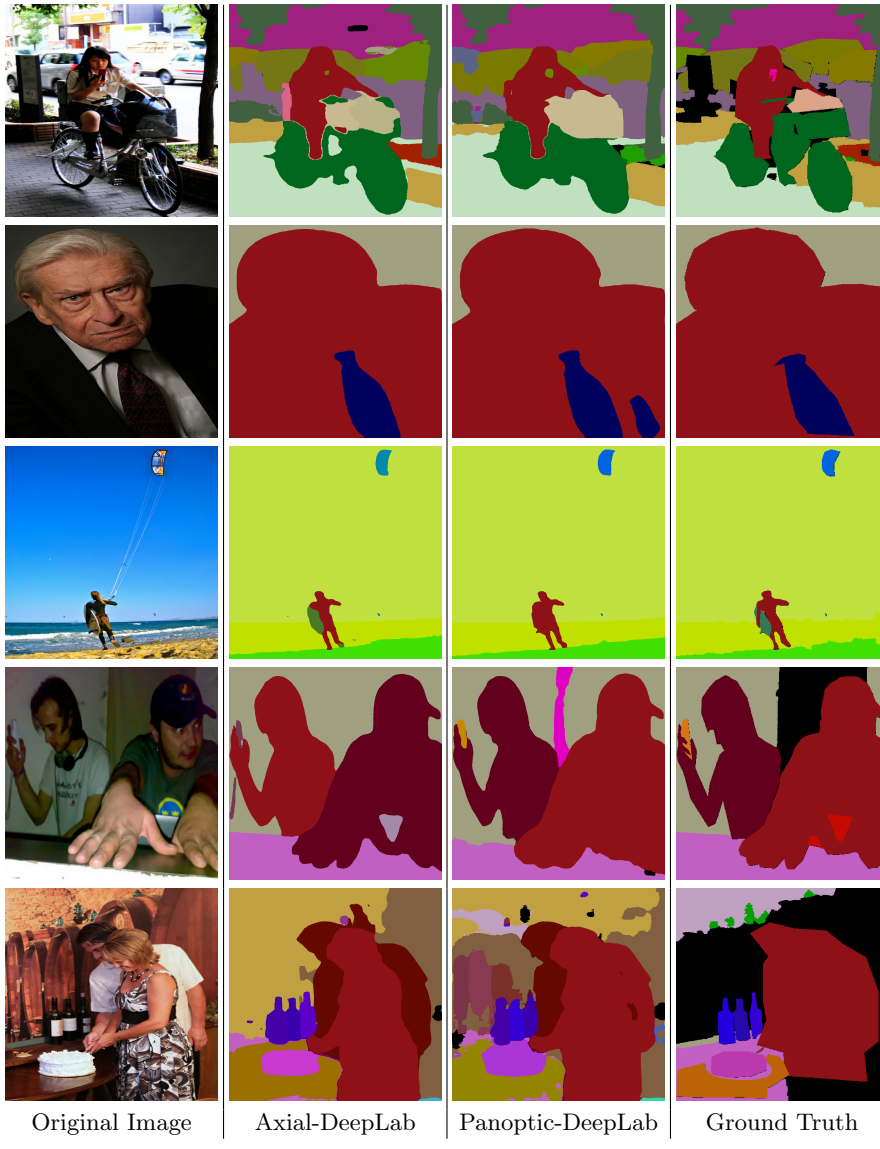

Fig. 1. Visualization on COCO val set. Axial-DeepLab shows robustness to occlusion. In row 1 and row 4, Axial-DeepLab captures the occluded left leg and the remote control cable respectively, which are not even present in ground truth labels. In the last row, Axial-DeepLab distinguishes one person occluding another correctly, whereas the ground truth treats them as one instance

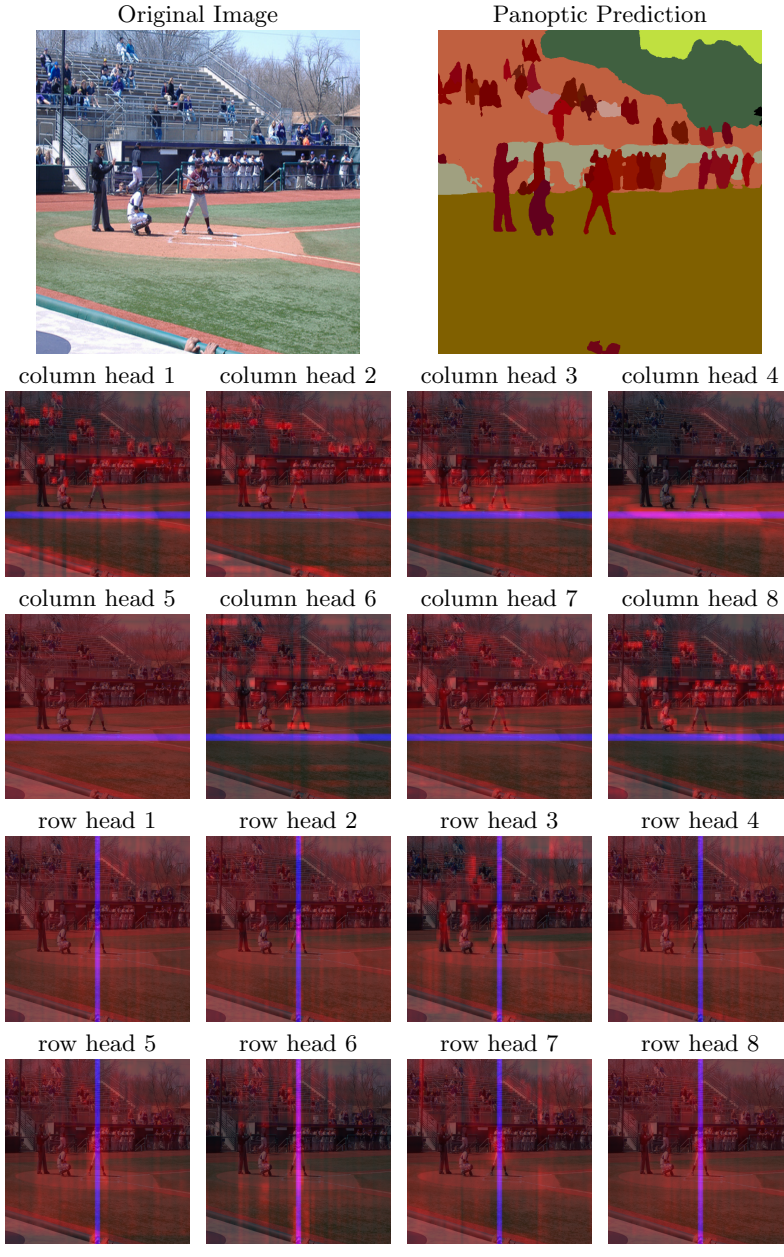

Fig. 2. Attention maps in block 2 of stage 3. We take a row of pixels, and visualize their column (height-axis) attention in all 8 heads. Then, we take a column, and visualize their row attention. Blue pixels are queries that we take, and red pixels indicate the corresponding attention weights. We notice that column head 1 corresponds to human heads, while column head 4 correlates with the field only. Row head 6 focuses on relatively local regions whereas column head 5 pools all over the whole image

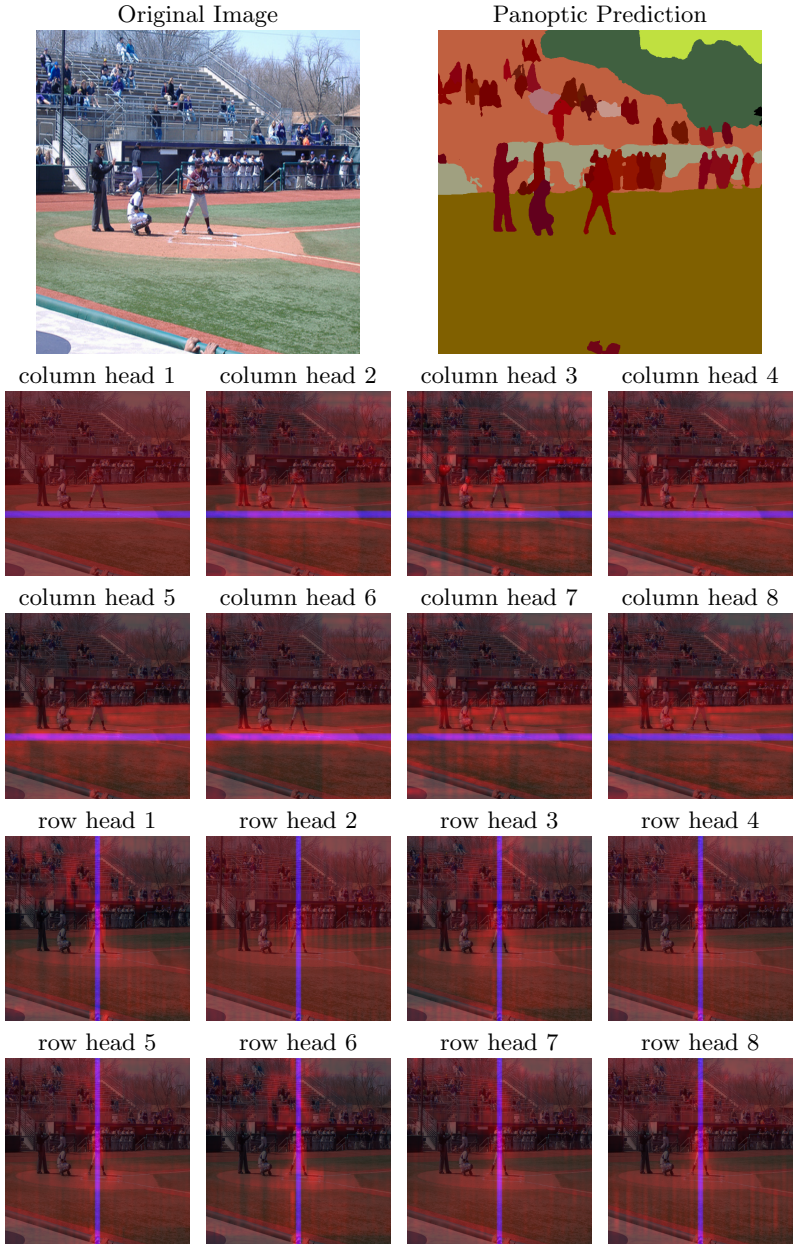

Fig. 3. Attention maps in block 3 of stage 4. They focus more on long range context than those in Fig. 2, although all of them have a global receptive field

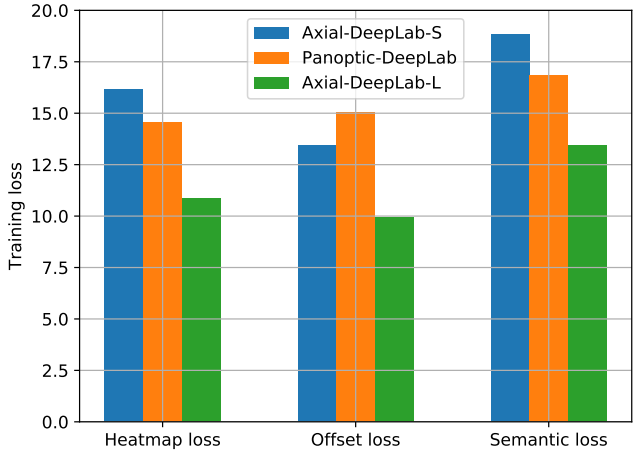

Fig. 4. Training loss on COCO. Equipped with position-sensitive axial-attention, our Axial-DeepLab fits data distribution better than Panoptic-DeepLab [1], especially on the task of predicting the offset to the object center, which requires precise and long range positional information

3 Axial-Decoder Design

In Fig. 5, we show our design of axial-decoders. This is an example decoder in Axial-DeepLab-L from output stride 8 to output stride 4. We apply three such blocks, analogous to the three 5×5 convolutions in Panoptic-DeepLab [1].

4 Datasets

ImageNet: The ImageNet-1K dataset [12] contains 1.2M training images and 50K validation images.

COCO: COCO dataset [8] contains 118K, 5K, and 20K images for training, validation, and testing, respectively. It consists of 80 *thing* and 53 *stuff* classes.

Mapillary Vistas: The street-view Mapillary Vistas [10] consists of 18K, 2K, and 5K images for training, validation and testing, respectively. There are 37 *thing* classes and 28 *stuff* classes in a variety of image resolutions, ranging from 1024 × 768 to more than 4000 × 6000.

Cityscapes: It has 2975, 500, and 1525 traffic-related images for training, validation, and testing, respectively. It contains 8 *thing* and 11 *stuff* classes [2].

5 Raw Data

In companion to Fig. 3 of the main paper where we compare parameters and M-Adds against accuracy on ImageNet classification, we also show the performance of our models in Tab. 2.

Table 1. Cityscapes test set. **C:** Cityscapes coarse annotation. **V:** Cityscapes video. **MV:** Mapillary Vistas

Method	Extra Data	PQ	AP	mIoU
GFF-Net [6]		-	-	82.3
Zhu *et al.* [15]	C, V, MV	-	-	83.5
AdaptIS [13]		-	32.5	-
UPSNet [14]	COCO	-	33.0	-
PANet [9]	COCO	-	36.4	-
PolyTransform [7]	COCO	-	40.1	
SSAP [3]		58.9	32.7	-
Li *et al.* [5]		61.0	-	-
Panoptic-DeepLab [1]		62.3	34.6	79.4
TASCNet [4]	COCO	60.7	-	-
Seamless [11]	MV	62.6	-	-
Li *et al.* [5]	COCO	63.3	-	-
Panoptic-DeepLab [1]	MV	65.5	39.0	84.2
Axial-DeepLab-L		**62.7**	**33.3**	**79.5**
Axial-DeepLab-XL		62.8	34.0	79.9
Axial-DeepLab-L	MV	65.6	38.1	83.1

Table 2. ImageNet validation set results. **Width:** the width multiplier that scales the models up. **Full:** Stand-alone self-attention models without spatial convolutions

Method	Width	Full	Params	M-Adds	Top-1
Conv-Stem + PS-Attention	0.5		5.1M	1.2B	75.5
Conv-Stem + PS-Attention	0.75		10.5M	2.3B	77.4
Conv-Stem + PS-Attention	1.0		18.0M	3.7B	78.1
Conv-Stem + PS-Attention	1.25		27.5M	5.6B	78.5
Conv-Stem + PS-Attention	1.5		39.0M	7.8B	79.0
Conv-Stem + Axial-Attention	0.375		7.4M	1.8B	76.4
Conv-Stem + Axial-Attention	0.5		12.4M	2.8B	77.5
Conv-Stem + Axial-Attention	0.75		26.4M	5.7B	78.6
Conv-Stem + Axial-Attention	1.0.		45.6M	9.6B	79.0
Full Axial-Attention	0.5	✓	**12.5M**	**3.3B**	**78.1**
Full Axial-Attention	0.75	✓	**26.5M**	**6.8B**	**79.2**
Full Axial-Attention	1.0	✓	**45.8M**	**11.6B**	**79.3**

In companion to Fig. 4 of the main paper where we demonstrate the relative improvements of Axial-DeepLab-L over Panoptic-DeepLab (Xception-71) in our scale stress test on COCO, we also show the raw performance of both models in Fig. 6.

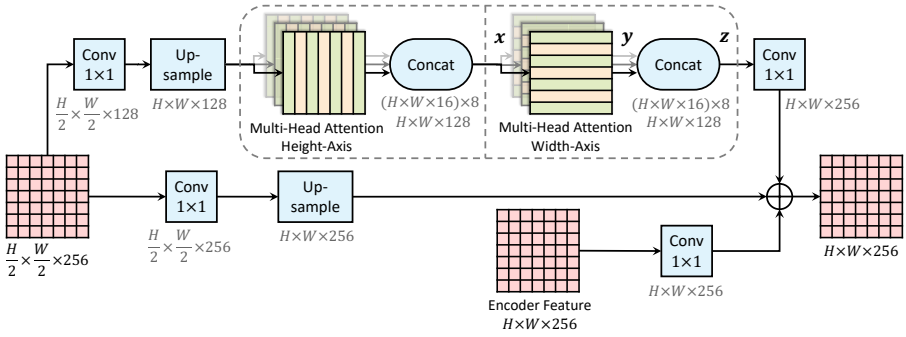

Fig. 5. An axial-decoder block. We augment an axial-attention block with up-samplings, and encoder features

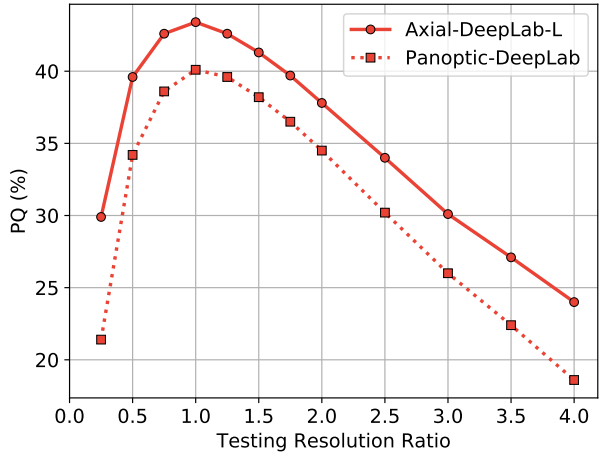

Fig. 6. Scale stress test on COCO val set

References

1. Cheng, B., Collins, M.D., Zhu, Y., Liu, T., Huang, T.S., Adam, H., Chen, L.C.: Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. arXiv:1911.10194 (2019) 1, 5, 6
2. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016) 5
3. Gao, N., Shan, Y., Wang, Y., Zhao, X., Yu, Y., Yang, M., Huang, K.: Ssap: Single-shot instance segmentation with affinity pyramid. In: ICCV (2019) 6
4. Li, J., Raventos, A., Bhargava, A., Tagawa, T., Gaidon, A.: Learning to fuse things and stuff. arXiv:1812.01192 (2018) 6
5. Li, Q., Qi, X., Torr, P.H.: Unifying training and inference for panoptic segmentation. arXiv:2001.04982 (2020) 6

6. Li, X., Zhao, H., Han, L., Tong, Y., Yang, K.: Gff: Gated fully fusion for semantic segmentation. arXiv:1904.01803 (2019) 6
7. Liang, J., Homayounfar, N., Ma, W.C., Xiong, Y., Hu, R., Urtasun, R.: Polytransform: Deep polygon transformer for instance segmentation. arXiv:1912.02801 (2019) 6
8. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014) 5
9. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation. In: CVPR (2018) 6
10. Neuhold, G., Ollmann, T., Rota Bulo, S., Kontschieder, P.: The mapillary vistas dataset for semantic understanding of street scenes. In: ICCV (2017) 5
11. Porzi, L., Bulò, S.R., Colovic, A., Kontschieder, P.: Seamless scene segmentation. In: CVPR (2019) 6
12. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. IJCV **115**, 211–252 (2015) 5
13. Sofiiuk, K., Barinova, O., Konushin, A.: Adaptis: Adaptive instance selection network. In: ICCV (2019) 6
14. Xiong, Y., Liao, R., Zhao, H., Hu, R., Bai, M., Yumer, E., Urtasun, R.: Upsnet: A unified panoptic segmentation network. In: CVPR (2019) 6
15. Zhu, Y., Sapra, K., Reda, F.A., Shih, K.J., Newsam, S., Tao, A., Catanzaro, B.: Improving semantic segmentation via video propagation and label relaxation. In: CVPR (2019) 6