ECCV'20 ONLINE

23-28 AUGUST 2020

16TH EUROPEAN CONFERENCE ON

COMPUTER VISION

WWW.ECCV2020.EU

Axial-DeepLab: Stand-Alone Axial-Attention for Panoptic Segmentation

Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, Liang-Chieh Chen

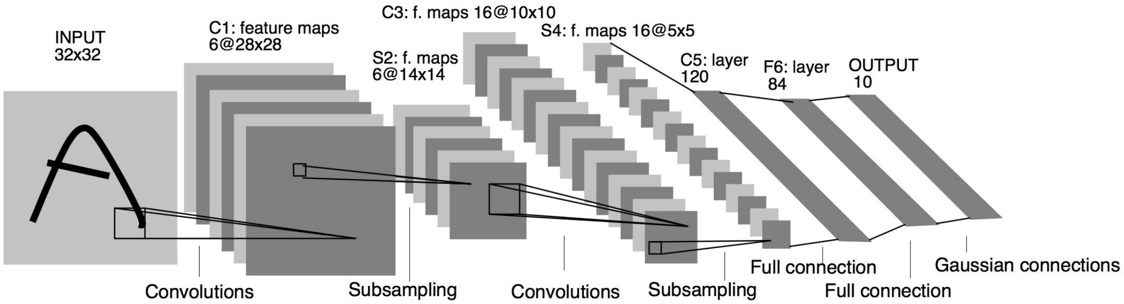
Johns Hopkins University, Google Research

Convolution

- Local square

$$y_o = \sum_{p \in \mathcal{N}_{m \times m}(o)} W_{p-o} x_p$$

Method	Stand-Alone	Long-Range
Convolution	✓	✗



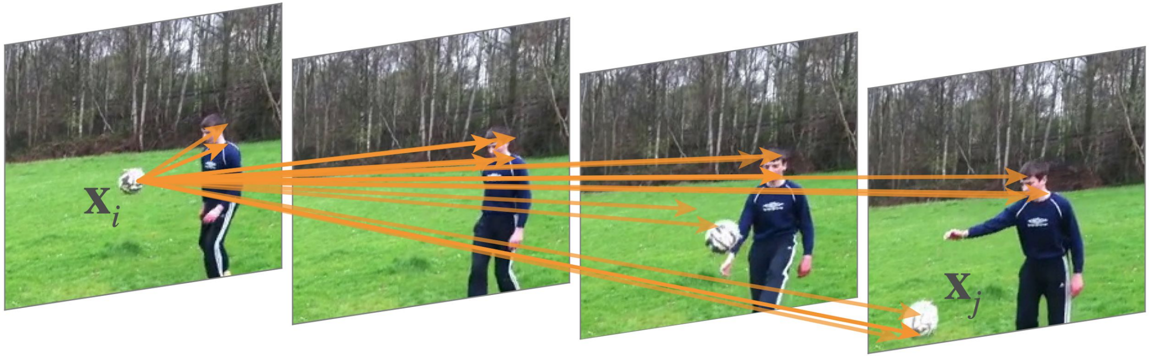
LeCun, Y., et al. Gradient-based learning applied to document recognition. Proceedings of the IEEE. 1998.

Non-Local (a.k.a. self-attention)

- Local square

$$y_o = \sum_{p \in \mathcal{N}_{m \times m}(o)} W_{p-o} x_p$$

Method	Stand-Alone	Long-Range
Convolution	✓	✗
Non-Local	✗	✓



Wang, X., et al. Non-local neural networks. CVPR 2018.
Vaswani, et al. Attention is all you need. NeurIPS 2017.

Non-Local (a.k.a. self-attention)

Method	Stand-Alone	Long-Range
Convolution	✓	✗
Non-Local	✗	✓

- Local square

$$y_o = \sum_{p \in \mathcal{N}_{m \times m}(o)} W_{p-o} x_p$$

- Whole image

$$y_o = \sum_{\boxed{p \in \mathcal{N}}} \mathrm{softmax}_p(q_o^T k_p) v_p$$

Query $q_o = W_Q x_o$

Key $k_p = W_K x_p$

Value $v_p = W_V x_p$

Wang, X., et al. Non-local neural networks. CVPR 2018.
Vaswani, et al. Attention is all you need. NeurIPS 2017.

Non-Local (a.k.a. self-attention)

Method	Stand-Alone	Long-Range
Convolution	✓	✗
Non-Local	✗	✓

- Local square

$$y_o = \sum_{p \in \mathcal{N}_{m \times m}(o)} W_{p-o} x_p$$

- Whole image

$$y_o = \sum_{p \in \mathcal{N}} \mathrm{softmax}_p(q_o^T k_p) v_p$$

$$\boxed{O(H^2 W^2)}$$

Wang, X., et al. Non-local neural networks. CVPR 2018.
Vaswani, et al. Attention is all you need. NeurIPS 2017.

Stand-Alone Self-Attention

- Local square

$$y_o = \sum_{p \in \mathcal{N}_{m \times m}(o)} W_{p-o} x_p$$

Method	Stand-Alone	Long-Range
Convolution	✓	✗
Non-Local	✗	✓
Stand-Alone	✓	✗

- Whole image

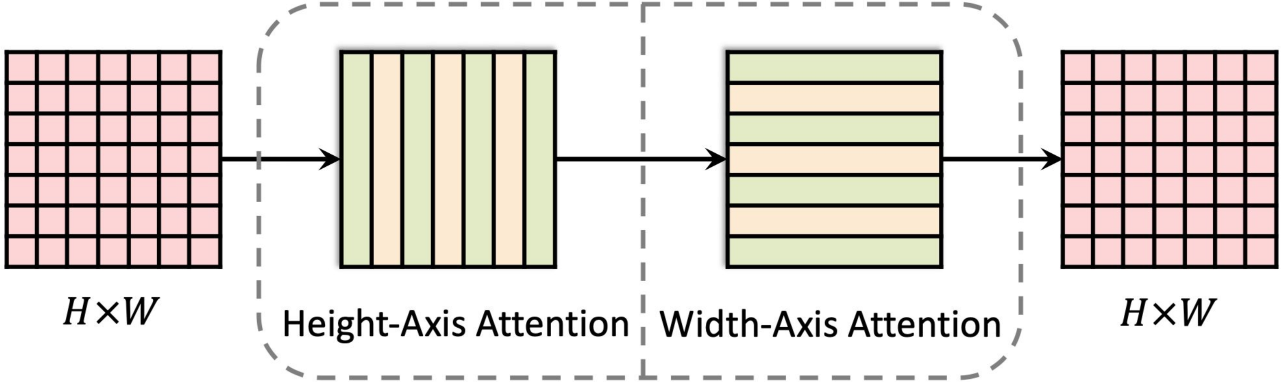
$$y_o = \sum_{p \in \mathcal{N}} \mathrm{softmax}_p(q_o^T k_p) v_p$$

- Local square

$$y_o = \sum_{p \in \mathcal{N}_{m \times m}(o)} \mathrm{softmax}_p(q_o^T k_p) v_p$$

Ramachandran, P., et al. Stand-alone self-attention in vision models. NeurIPS 2019.
Hu, H., et al. Local relation networks for image recognition. ICCV 2019.

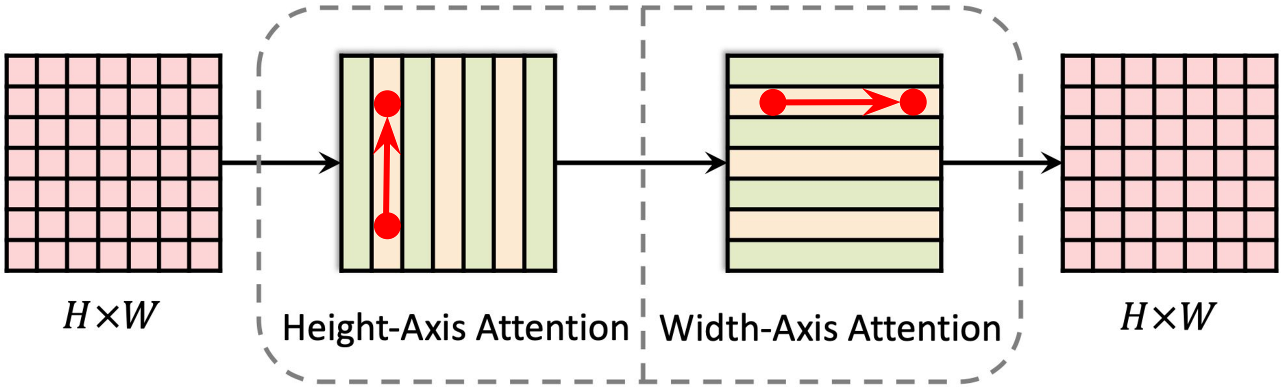
Axial-DeepLab



- Whole image $$y_o = \sum_{p \in \mathcal{N}} \text{softmax}_p(q_o^T k_p) v_p$$
- Whole width-axis $$y_o = \sum_{p \in \mathcal{N}_{1 \times m}(o)} \text{softmax}_p(q_o^T k_p) v_p$$

Ho, J., et al. Axial Attention in Multidimensional Transformers. arXiv 2019.
Huang, Z., et al. Ccnet: Criss-cross attention for semantic segmentation. ICCV 2019.

Axial-DeepLab



- Whole image $$y_o = \sum_{p \in \mathcal{N}} \text{softmax}_p(q_o^T k_p) v_p$$
- Whole width-axis $$y_o = \sum_{p \in \mathcal{N}_{1 \times m}(o)} \text{softmax}_p(q_o^T k_p) v_p$$

Ho, J., et al. Axial Attention in Multidimensional Transformers. arXiv 2019.
Huang, Z., et al. Ccnet: Criss-cross attention for semantic segmentation. ICCV 2019.

Axial-DeepLab

- Local square

$$y_o = \sum_{p \in \mathcal{N}_{m\times m}(o)} W_{p-o} x_p$$

Method	Stand-Alone	Long-Range
Convolution	✓	✗
Non-Local	✗	✓
Stand-Alone	✓	✗
Axial-DeepLab	✓	✓

- Whole image

$$y_o = \sum_{p \in \mathcal{N}} \text{softmax}_p(q_o^T k_p) v_p$$

$\cancel{O(H^2W^2)}$

- Whole width-axis

$$y_o = \sum_{p \in \mathcal{N}_{1\times m}(o)} \text{softmax}_p(q_o^T k_p) v_p$$

$O(HW^2)$

$O(mHW)$

Ho, J., et al. Axial Attention in Multidimensional Transformers. arXiv 2019.
Huang, Z., et al. Ccnet: Criss-cross attention for semantic segmentation. ICCV 2019.

Is this all you need?

$$y_o = \sum_{p \in \mathcal{N}_{1\times m}(o)} \text{softmax}_p(q_o^T k_p) v_p$$

Is this all you need? NO!

$$y_o = \sum_{p \in \mathcal{N}_{1\times m}(o)} \mathrm{softmax}_p(q_o^T k_p) v_p$$

Position Unaware

Method	Position
Convolution	✓
Non-Local	✗

$$y_o = \sum_{p \in \mathcal{N}} W_{\boxed{p-o}} x_p$$

$$y_o = \sum_{p \in \mathcal{N}} \mathrm{softmax}_p(q_o^T k_p) v_p$$

Position Aware

Method	Position
Convolution	✓
Non-Local	✗
Stand-Alone	✓

$$y_o = \sum_{p \in \mathcal{N}} W_{\boxed{p-o}} x_p$$

$$y_o = \sum_{p \in \mathcal{N}} \mathrm{softmax}_p(q_o^T k_p) v_p$$

- Query-dependent positional bias

$$y_o = \sum_{p \in \mathcal{N}} \mathrm{softmax}_p(q_o^T k_p + q_o^T r^q_{\boxed{p-o}}) v_p$$

Ramachandran, P., et al. Stand-alone self-attention in vision models. NeurIPS 2019.
Shaw, P., et al. Self-attention with relative position representations. NAACL 2018.
Dai, J., et al. Deformable convolutional networks. ICCV 2017.

Alternatives

Method	Position
Convolution	✓
Non-Local	✗
Stand-Alone	✓

- Query-dependent positional bias
- Key-dependent positional bias

$$y_o = \sum_{p \in \mathcal{N}} W_{p-o} x_p$$

$$y_o = \sum_{p \in \mathcal{N}} \mathrm{softmax}_p(q_o^T k_p) v_p$$

$$y_o = \sum_{p \in \mathcal{N}} \mathrm{softmax}_p(q_o^T k_p + q_o^T r_{p-o}^q) v_p$$

$$y_o = \sum_{p \in \mathcal{N}} \mathrm{softmax}_p(q_o^T k_p + k_p^T r_{p-o}^k) v_p$$

Alternatives

Method	Position
Convolution	✓
Non-Local	✗
Stand-Alone	✓

- Query-dependent positional bias
- Key-dependent positional bias
- Content-based position retrieval

$$y_o = \sum_{p \in \mathcal{N}} W_{p-o} x_p$$

$$y_o = \sum_{p \in \mathcal{N}} \mathrm{softmax}_p(q_o^T k_p) v_p$$

$$y_o = \sum_{p \in \mathcal{N}} \mathrm{softmax}_p(q_o^T k_p + q_o^T r_{p-o}^q) v_p$$

$$y_o = \sum_{p \in \mathcal{N}} \mathrm{softmax}_p(q_o^T k_p + k_p^T r_{p-o}^k) v_p$$

$$y_o = \sum_{p \in \mathcal{N}} \mathrm{softmax}_p(q_o^T k_p)(v_p + r_{p-o}^v)$$

Position Sensitive

Method	Position
Convolution	✓
Non-Local	✗
Stand-Alone	✓
Axial-DeepLab	✓✓✓

$$y_o = \sum_{p \in \mathcal{N}} W_{p-o} x_p$$

$$y_o = \sum_{p \in \mathcal{N}} \text{softmax}_p(q_o^T k_p) v_p$$

$$y_o = \sum_{p \in \mathcal{N}} \text{softmax}_p(q_o^T k_p + q_o^T r_{p-o}^q) v_p$$

$$y_o = \sum_{p \in \mathcal{N}} \text{softmax}_p(q_o^T k_p + q_o^T r_{p-o}^q + k_p^T r_{p-o}^k)(v_p + r_{p-o}^v)$$

Summary

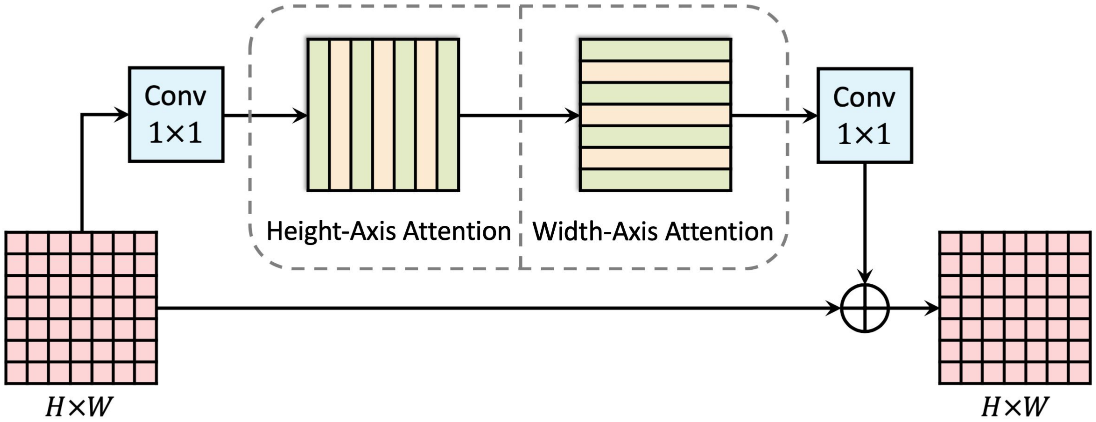
Method	Stand-Alone	Long-Range	Position
Convolution	✓	✗	✓
Non-Local	✗	✓	✗
Stand-Alone	✓	✗	✓
Axial-DeepLab	✓	✓	✓✓✓

$$y_o = \sum_{p \in \mathcal{N}_{1 \times m}(o)} \text{softmax}_p(q_o^T k_p + q_o^T r_{p-o}^q + k_p^T r_{p-o}^k)(v_p + r_{p-o}^v)$$

Is this all you need?

$$y_o = \sum_{p \in \mathcal{N}_{1\times m}(o)} \text{softmax}_p(q_o^T k_p + q_o^T r_{p-o}^q + k_p^T r_{p-o}^k)(v_p + r_{p-o}^v)$$

Stand-Alone Axial Block



$$y_o = \sum_{p \in \mathcal{N}_{1 \times m}(o)} \operatorname{softmax}_p(q_o^T k_p + q_o^T r_{p-o}^q + k_p^T r_{p-o}^k)(v_p + r_{p-o}^v)$$

He, K., et al. Deep residual learning for image recognition. CVPR 2016.

Results: ImageNet Classification

Method	Params	M-Adds	Top-1
ResNet-50	25.6M	4.1B	76.9
Stand-Alone Self-Attention	18.0M	3.6B	77.6
Position-Sensitive Axial-Attention	**12.5M**	**3.3B**	**78.1**

$$y_o = \sum_{p \in \mathcal{N}_{1 \times m}(o)} \text{softmax}_p(q_o^T k_p + q_o^T r_{p-o}^q + k_p^T r_{p-o}^k)(v_p + r_{p-o}^v)$$

Russakovsky, O., et al. Imagenet large scale visual recognition challenge. IJCV 2015.
He, K., et al. Deep residual learning for image recognition. CVPR 2016.
Ramachandran, P., et al. Stand-alone self-attention in vision models. NeurIPS 2019.

Is this all you need? YES!

$$y_o = \sum_{p \in \mathcal{N}_{1\times m}(o)} \mathrm{softmax}_p(q_o^T k_p + q_o^T r_{p-o}^q + k_p^T r_{p-o}^k)(v_p + r_{p-o}^v)$$

Results: Cityscapes

Backbone	ASPP	PS	Params	M-Adds	PQ	AP	mIoU
ResNet-50			24.8M	374.8B	58.1	30.0	73.3
ResNet-50	✓		30.0M	390.0B	59.8	32.6	77.8

Kirillov, A., et al. Panoptic segmentation. CVPR 2019.
Cordts, M., et al. The cityscapes dataset for semantic urban scene understanding. CVPR 2016.
Cheng, B., et al. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. CVPR 2020.

Results: Cityscapes

Backbone	ASPP	PS	Params	M-Adds	PQ	AP	mIoU
ResNet-50			24.8M	374.8B	58.1	30.0	73.3
ResNet-50	✓		30.0M	390.0B	59.8	32.6	77.8
Stand-Alone			17.3M	317.7B	58.7	31.9	75.8
Stand-Alone	✓		22.5M	332.9B	60.9	30.0	78.2
Stand-Alone		✓	17.3M	326.7B	59.9	32.2	76.3
Stand-Alone	✓	✓	22.5M	341.9B	**61.5**	**33.1**	**79.1**

Kirillov, A., et al. Panoptic segmentation. CVPR 2019.
Cordts, M., et al. The cityscapes dataset for semantic urban scene understanding. CVPR 2016.
Cheng, B., et al. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. CVPR 2020.

Results: Cityscapes

Backbone	ASPP	PS	Params	M-Adds	PQ	AP	mIoU
ResNet-50			24.8M	374.8B	58.1	30.0	73.3
ResNet-50	✓		30.0M	390.0B	59.8	32.6	77.8
Stand-Alone			17.3M	317.7B	58.7	31.9	75.8
Stand-Alone	✓		22.5M	332.9B	60.9	30.0	78.2
Stand-Alone		✓	17.3M	326.7B	59.9	32.2	76.3
Stand-Alone	✓	✓	22.5M	341.9B	**61.5**	**33.1**	**79.1**
Axial-DeepLab-S		✓	**12.1M**	**220.8B**	**62.6**	**34.9**	**80.5**

Results: Cityscapes

Backbone	ASPP	PS	Params	M-Adds	PQ	AP	mIoU
ResNet-50			24.8M	374.8B	58.1	30.0	73.3
ResNet-50	✓		30.0M	390.0B	59.8	32.6	77.8
Stand-Alone			17.3M	317.7B	58.7	31.9	75.8
Stand-Alone	✓		22.5M	332.9B	60.9	30.0	78.2
Stand-Alone		✓	17.3M	326.7B	59.9	32.2	76.3
Stand-Alone	✓	✓	22.5M	341.9B	**61.5**	**33.1**	**79.1**
Axial-DeepLab-S		✓	**12.1M**	**220.8B**	**62.6**	**34.9**	**80.5**
Axial-DeepLab-M		✓	25.9M	419.6B	63.1	35.6	80.3
Axial-DeepLab-L		✓	44.9M	687.4B	63.9	35.8	81.0
Axial-DeepLab-XL		✓	173.0M	2446.8B	64.4	36.7	80.6

Long-Range helps

$$y_o = \sum_{p \in \mathcal{N}_{1\times \boldsymbol{m}}(o)} \text{softmax}_p(q_o^T k_p + q_o^T r_{p-o}^q + k_p^T r_{p-o}^k)(v_p + r_{p-o}^v)$$

Backbone	Span m	Params	M-Adds	PQ	AP	mIoU
ResNet-101	-	43.8M	530.0B	59.9	31.9	74.6
Axial-ResNet-L	5 × 5	44.9M	617.4B	59.1	31.3	74.5
Axial-ResNet-L	9 × 9	44.9M	622.1B	61.2	31.1	77.6
Axial-ResNet-L	17 × 17	44.9M	631.5B	62.8	34.0	79.5
Axial-ResNet-L	33 × 33	44.9M	650.2B	63.8	35.9	80.2
Axial-ResNet-L	65 × 65	44.9M	687.4B	**64.2**	**36.3**	**80.6**

More Results

Dataset	Split	Metric	SOTA	Axial-DeepLab
Cityscapes	test	PQ	65.5	66.6 (+1.1)
COCO (bottom-up)	test	PQ	41.4	44.2 (+2.8)
Mapillary Vistas	val	PQ	40.3	41.1 (+0.8)
Mapillary Vistas	val	mIoU	57.6	58.4 (+0.8)

Lin, T.Y., et al. Microsoft coco: Common objects in context. ECCV 2014.
Neuhold, G., et al. The mapillary vistas dataset for semantic understanding of street scenes. ICCV 2017.
Liu, C., et al. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. CVPR 2019.

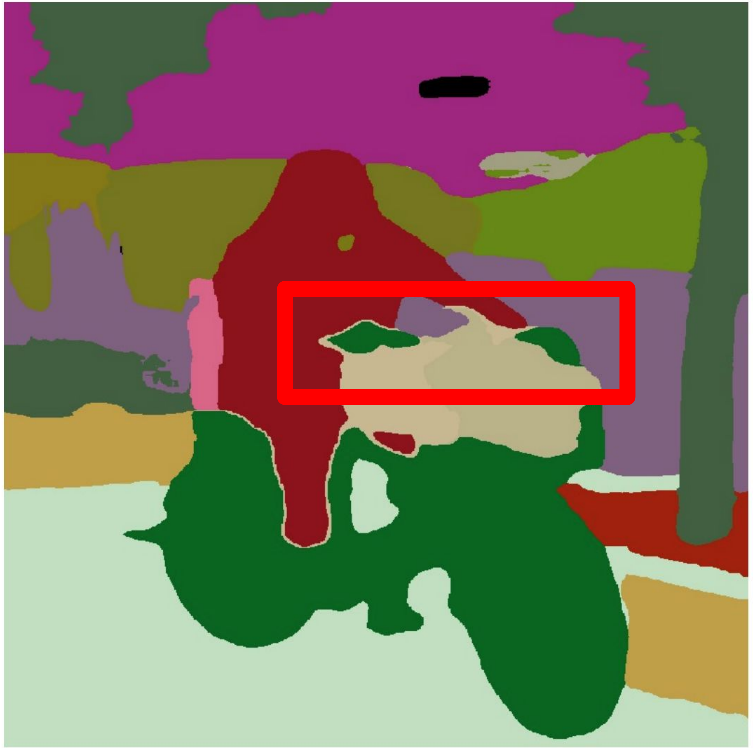
More Results

Dataset	Split	Metric	SOTA	Axial-DeepLab
Cityscapes	test	PQ	65.5	66.6 (+1.1)
COCO (bottom-up)	test	PQ	41.4	44.2 (+2.8)
Mapillary Vistas	val	PQ	40.3	41.1 (+0.8)
Mapillary Vistas	val	mIoU	57.6	58.4 (+0.8)

Auto-DeepLab-XL++

Lin, T.Y., et al. Microsoft coco: Common objects in context. ECCV 2014.
Neuhold, G., et al. The mapillary vistas dataset for semantic understanding of street scenes. ICCV 2017.
Liu, C., et al. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. CVPR 2019.

Examples

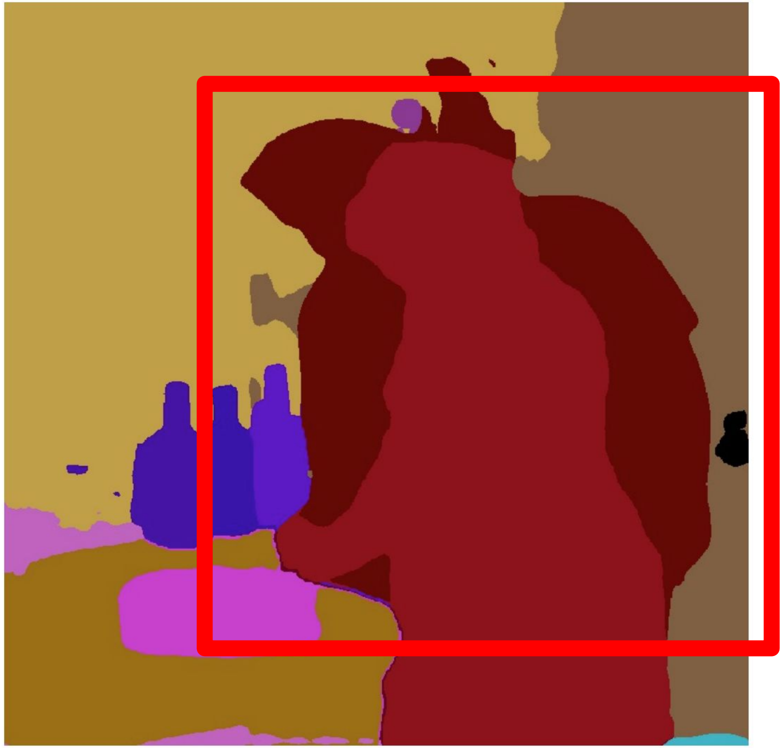



http://farm5.staticflickr.com/4134/4782858440_3885462451_z.jpg

Examples

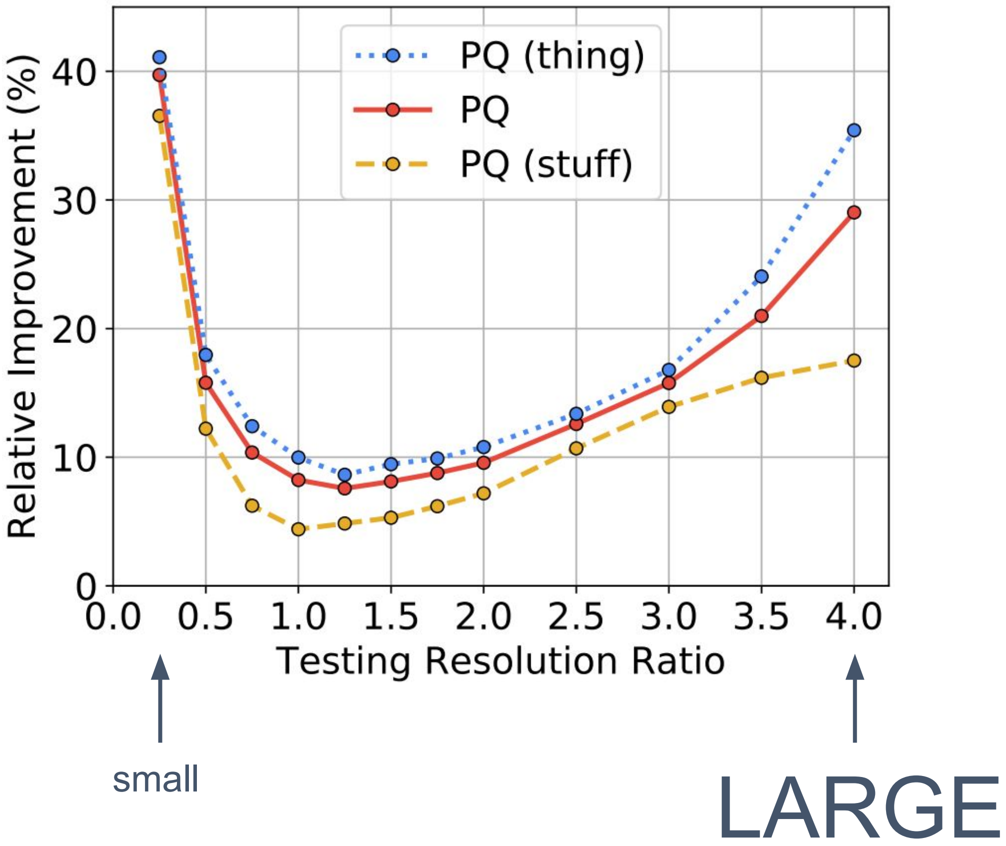
http://farm4.staticflickr.com/3189/2947274789_a1a35b33c3_z.jpg

Examples

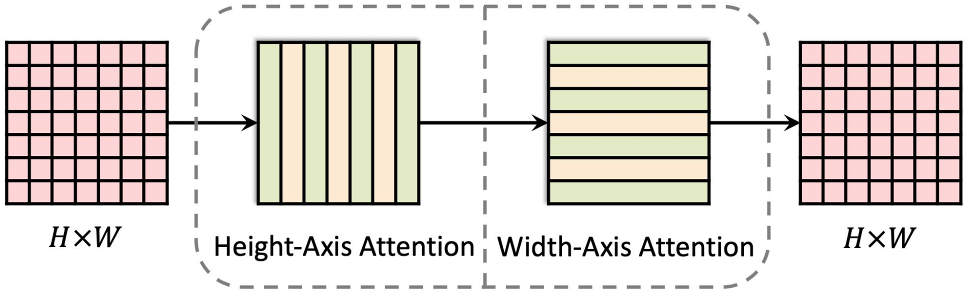


Scale stress test

- Robust to out-of-distribution scales (both ***small*** and ***large***)



Conclusion



Method	Stand-Alone	Long-Range	Position
Convolution	✓	✗	✓
Non-Local	✗	✓	✗
Stand-Alone	✓	✗	✓
Axial-DeepLab	✓	✓	✓✓✓