Supplementary Material: Global Context and Geometric Priors for Effective Non-Local Self-Attention

Sanghyun Woo[1]
shwoo93@kaist.ac.kr
Dahun KIm[1]
mcahny@kaist.ac.kr
Joon-Young Lee[2]
jolee@adobe.com
In So Kweon[1]
iskweon77@kaist.ac.kr

[1] KAIST
[2] Adobe Research

In this supplementary section, we provide,

1. Implementation details of the proposals and experimental settings,
2. Additional quantitative results,
3. Additional qualitative results.

1 Implementation Details

1.1 Dataset and Evaluation Metrics

1.1.1 Semantic Segmentation

We use the PASCAL Context [19], ADE20K [27], and Cityscapes [5] datasets. The ablation studies are conducted on PASCAL Context [19]. We use the standard metrics of pixel accuracy (pixAcc) and mean Intersection-Over-Union (mIoU). For the scene parsing results on the Pascal Context and ADE20K val sets, we ignore the background pixels in the evaluation, following convention [27]. For the evaluation of semantic segmentation results on the Cityscapes, we use a public evaluation server. We use the SGD optimizer with a "poly" learning rate scheduling $lr = baselr * (1 - \frac{iter}{total\_iter})^{power}$, where power is set to 0.9. The base learning rate is initialized to 0.01 for the ADE20K dataset and 0.001 for others. We use a batch size of 8 with four NVIDIA V100 GPUs. We adopt single-scale testing. To evaluate each baseline module's pure long-term context modeling ability, we do not adopt any auxiliary loss or module (e.g., se loss in EncNet [25] or encoding module in CFNet [26])

Method	Non-local	Context	Rel-Position	Multi-head	Dropout	pixAcc%	mIou%
FCN						75.57	45.78
	✓					76.43	47.25
	✓	✓				77.23	49.11
	✓		✓			77.42	49.33
	✓			✓		77.08	48.98
	✓	✓	✓			78.86	50.74
	✓	✓	✓	✓		78.94	51.18
FCN + Ours	✓	✓	✓	✓	✓	**79.14**	**51.27**

Table 1: Ablation results on major design choices using Pascal Context. We adopt ResNet50 as the FCN backbone.

1.1.2 Detection & Instance Segmentation

We use the COCO Detection dataset [17]. All reported results follow standard COCO-style Average Recall (AR) and Average Precision (AP) metrics. We train the models on train2017 and report the final results on test-dev. We use the SGD optimizer with an initial learning rate of 0.01. The model is trained for a total of 12 epochs, and the learning rate is divided by 10 after 8 and 11 epochs. We use a batch size of 16 with eight NVIDIA V100 GPUs.

1.1.3 Panoptic Segmentation

We use the COCO Panoptic dataset [17]. For a quantitative evaluation, we use the PQ metric, which captures both the recognition and segmentation quality and treats both stuff and thing categories in a unified manner [15]. We use the same training details with the above detection experiments.

2 Quantitative Analysis

2.1 Ablation Studies

We conduct ablation studies to explore how each component of our formulation contributes to the performance gain. We carry out the ablation experiments using the Pascal Context [19] dataset. These results are shown in Table 1.

2.1.1 Non-local Self-Attention

We begin by applying the standard non-local block [23] to the FCN. Here pixAcc and mIoU increase from 75.57 to 76.43 and 45.78 to 47.25, respectively. The incorporation of the non-local block is positive, indicating that long-range relationships are beneficial for the task.

2.1.2 Impact of Contextual Prior

We now introduce the context into the original formulation. Specifically, we modulate the relation computation with the proposed context matrix. In this case, pixAcc and mIoU increase from 76.43 to 77.23 and from 47.25 to 49.11, respectively. These results show that the image-level context indeed provides better relationship learning.

Additionally, we conduct experiments to evaluate the influence of the pooling methods when computing the context vector. In particular, we compare our global average pooling

dist type \ enc method	None	Sinusoid
absolute	78.30/50.22 [28]	78.42/50.31
relative	78.65/50.49	**78.86/50.74**

((a)) Exps on distance type and encoding method

emb method	
cont-indep (r) [10, 11]	78.32/50.59
cont-dep ($q^T r$)	**78.86/50.74**

((b)) Exps on embedding method

Number of heads				
H = 1	H = 2	H = 4	H = 8	H = 16
78.86/50.74	78.88/50.84	78.91/51.04	**78.94/51.18**	78.92/51.15

((c)) Exps on the number of heads

Table 2: Detailed ablations on *relative position embedding* and *multi-head* using Pascal Context. Each cell includes segmentation scores (pixAcc%/mIoU%).

with global max pooling. We observe that average pooling (77.23/49.11) outperforms max-pooling (77.09/48.78). Average pooling aggregates the neighboring features with an equal contribution, whereas max-pooling selects a single distinct feature to represent its neighbors. We thus find that average-pooling makes better use of contextual information and promotes effective relation learning. As a result, here we use average-pooled features in the subsequent experiments.

2.1.3 Impact of Geometric Prior

We explore the effectiveness of 2D relative position embeddings. pixAcc and mIoU increase from 77.23 to 78.86 and 49.11 to 50.74, respectively. This indicates that the relative position information further helps relational reasoning and is complementary to the context, which demonstrates the effectiveness of our unified design. Note that previous works are limited to the use of global context [2] or the relative position [1, 21]. The single effect of the relative position is also investigated. pixAcc and mIoU increase from 76.43 to 77.42 and 47.25 to 49.33, respectively. Overall, these results demonstrate the great impact of the relative position information.

Furthermore, in Table 2(a) and Table 2(b), we conduct a detailed analysis of the proposed relative position formulation. First, we experimentally verify that using both the ***relative distance*** and ***sinusoid encoding*** enables finer relation reasoning. We compare a total of four variants in Table 2(a). All variants of position representations are added with $(q^T c)k$ at the same location for a fair comparison. The results show that the relative distance consistently outperforms the absolute representation, with sinusoid encoding further improving the performance. In practice, the relative distance ensures translation-equivariance in the image, and sinusoid encoding allows the model to attend to the relative positions easily. Note that the first variant of absolute distance without sinusoid encoding can be considered as the form presented in earlier work [28], which is clearly inferior to ours (78.30 vs. 78.86, 50.22 vs. 50.74). Second, we investigate the effect of ***content-dependency*** (i.e., r vs q^Tr). To obtain a content-independent result [10, 11], we embed the relative position information without any interaction with the query content. Not surprisingly, we obtain relatively inferior results of 78.32 and 50.59. This implies that content-conditioning is a crucial operation. It allows the model to associate the relative position information with the content of certain object, causing the model to capture high-level, complex motifs. More concisely, we use relative

Method	Dataset	pixAcc% / mIoU%	cosine distance		
			input	output	att
FCN + Non-local [23]	Pascal Context	76.43/47.25	0.254	0.016	0.003
FCN + Ours		**79.14/51.27**	0.271	0.143	0.135
FCN + Non-local [23]	ADE20K Context	79.11/39.32	0.245	0.082	0.019
FCN + Ours		**79.72/40.41**	0.287	0.150	0.101
FCN + Non-local [23]	Cityscapes	95.38/74.19	0.201	0.166	0.193
FCN + Ours		**95.65/75.55**	0.217	0.214	0.197

Table 3: Cosine distance analysis on Pascal Context, ADE20K, and Cityscapes. We adopt ResNet50 as the FCN backbone and compare our formulation with a non-local strategy [23]. We compute the cosine distances of input, output, and attention in the non-local block, following earlier work [2].

distances and encode them using the sinusoid function. The encoded relative distance information is embedded in a content-dependent manner.

2.1.4 Multi-head & Dropout

We further push the performance by employing multi-head and dropout techniques. In this case, pixAcc and mIoU increase from 78.86 to 79.14 and 51.04 to 51.27, respectively. We find that the multi-head strategy brings some extent of the model-ensemble effect. Its individual effect is also investigated, improving the pixAcc and mIoU from 76.43 to 77.08 and 47.25 to 48.98, respectively. In Table 2(c), we conduct an experimental analysis of the head number, finding that increasing the head improves the performance, and that it saturates at H=8. Meanwhile, dropout imposes an information bottleneck during the relation learning step, which encourages general representations.

Method	Dataset	Params	Flops	mAP^{bbox} / AP^{mask}	cosine distance		
					input	output	att
Mask R-CNN [8]	COCO	44.18M	275.58G	37.1/34.1	-	-	-
Mask RCNN [8] + BFPN [23]		44.44M	276.63G	37.9/34.9	0.362	0.080	0.156
Mask RCNN [8] + EBFPN		44.57M	276.24G	**39.1/35.6**	0.355	0.200	0.297

Table 4: Cosine distance analysis on COCO *val*. We use the Mask RCNN model (ResNet50 + FPN backbone) and compare the proposed EBFPN with BFPN [20]. Specifically, we compute the cosine distance of input, output, and attention in the non-local block, following earlier work [2].

2.2 Cosine Distance Analysis of Learned Features

To concretely verify that our attention map is query-specific compared to the original non-local self-attention method, we compute the cosine distances[1] between the input features, output features, and attention maps in the non-local block, following an earlier study [2]. Table 3 and Table 4 show the experimental results. With the original non-local block [23], we can observe the general tendency of the values in 'output' and 'att' columns are one

[1] $avg\_dist = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} dist(x_i, x_j)$, where x_i denotes the feature vector for position i, N indicates the total number of spatial locations, and $dist(x_i, x_j) = (1 - \cos(x_i, x_j))/2$.

Method	Backbone	Params	Flops	mAP	m$AP_{.5}$	m$AP_{.75}$
Mask R-CNN [9]	ResNet101	63.17M	351.65G	39.4(35.7)	60.8(57.6)	43.1(38.2)
Mask R-CNN + BFPN [20]		63.43M	352.70G	40.1(36.4)	62.3(58.8)	43.5(38.6)
Mask R-CNN + EBFPN		63.56M	352.31G	**41.0**(**37.1**)	**63.3**(**59.6**)	**44.7**(**39.5**)
Mask R-CNN [9]	HRNetV2	49.93M	352.92G	40.1(36.3)	61.2(58.0)	43.7(38.7)
Mask R-CNN + BFPN [20]		50.19M	354.97G	40.5(36.5)	62.0(58.8)	44.5(39.3)
Mask R-CNN + EBFPN		50.32M	354.58G	**40.9**(**37.1**)	**62.8**(**59.5**)	**44.5**(**39.7**)

Table 5: Experiments on stronger backbones. Detection/Instance segmentation on COCO *test-dev*.

Method	Params	Flops	mAP	mAP_5	mAP_{75}
Mask R-CNN [9]	44.18M	275.58G	38.0(34.7)	58.9(55.8)	41.4(36.9)
Mask R-CNN + BFPN [20]	44.44M	276.63G	39.2(35.7)	61.2(57.8)	42.3(37.8)
Mask R-CNN + EBFPN	44.57M	276.24G	**40.4**(**36.9**)	**62.9**(**58.9**)	**43.5**(**38.8**)

Table 6: Experiments on a longer training schedule (3×). Detection/Instance segmentation on COCO *test-dev*.

or two orders of magnitude smaller than that of the 'input'. This implies that the globally aggregated output features and computed relations are almost the same for different query positions. On the other hand, our new formulation makes the outputs and relations to be discriminative across different query positions. These are important for the high-level vision tasks, leading to produce significantly better results than in aforementioned study [23].

Despite using the non-local block [23], we note that the cosine distance values of 'att' are relatively high on cityscapes [5]. Recent studies [7, 13, 24] also show similar visualization results. This is possible because the dataset consists only of road driving scenes. Compared to natural scenes [19, 27], driving scenes have a standard perspective geometry [16] and repeating positional patterns [4, 28], allowing the module fairly easily to capture query-specific relationships. However, the output cosine distance values and final segmentation performances are still significantly lower than ours, meaning that our formulation better captures finer relationships.

2.3 Experiments on stronger backbones

To determine whether the proposed method performs well on stronger networks, we adopt ResNet101 and HRNetV2 [22] for the backbone and use Mask-RCNN as the detection head. These results are in Table 5. We find that our method also works well on stronger backbones.

2.4 Experiments on longer training schedule

In general, a longer training schedule improves the performance of the model. To identify whether our method indeed enhances the capability of modeling long-term contexts or merely accelerates training, we conduct experiments on a longer training schedule. Specifically, we extend the total training epoch three times (i.e., 36 epochs). These results are summarized in Table 6 and Table 7. We observe that our method consistently brings further improvements over the baselines for both Mask-RCNN and Panoptic FPN (both with the ResNet50+FPN backbone), which confirms its enhanced capability of modeling long-term contexts.

Method	Params	Flops	PQ	PQ^{Th}	PQ^{St}
Panoptic FPN [9]	45.82M	275.58G	40.5	47.2	29.5
Panoptic FPN + BFPN [20]	46.08M	276.63G	41.2	47.8	31.9
Panoptic FPN + EBFPN	46.21M	276.24G	**42.6**	**48.3**	**33.3**

Table 7: Experiments on a longer training schedule (3×). Panoptic segmentation results on COCO *val*.

Method	Params	Flops	mAP	m$AP_{.5}$	m$AP_{.75}$
M R-CNN [9]	44.18M	275.58G	37.2(34.1)	58.9(55.4)	40.3(36.2)
M R-CNN + BFPN [20]	44.44M	276.63G	38.1(34.8)	60.3(57.1)	41.5(37.1)
M R-CNN + BFPN (Dconv [6])	44.81M	275.74G	38.0(34.8)	60.1(56.8)	41.1(37.3)
M R-CNN + BFPN (GC [3])	44.70M	275.59G	37.8(34.6)	60.4(56.8)	40.7(36.9)
M R-CNN + BFPN (SE [12])	44.44M	276.63G	38.1(34.8)	59.8(56.4)	40.8(36.8)
M R-CNN + EBFPN	44.57M	276.24G	**39.2**(**35.7**)	**61.7**(**57.7**)	**42.7**(**38.1**)

Table 8: Comparison with other widely adopted context aggregation modules [2, 6, 12] on detection models. We report the detection/instance segmentation scores on COCO *test-dev*.

2.5 Comparison with other context aggregating modules

We add more comparisons with other BFPN variants where the NL part is replaced by deformable convolution (DConv), a GC module, and a SE module. As shown in Table 8, the proposed EBFPN significantly outperforms these other methods while using a comparable number of parameters and FLOPs.

2.6 Memory and inference speed comparisons

We report memory (G) and inference speed (fps) here. We use a Tesla V100 GPU, Intel(R) Xeon(R) CPU E5-2686 v4 @ 2.30GHz. Due to the feature relation computation, the non-local and our module slightly decrease the inference speed: Mask-RCNN (1.5 G / 16.8 fps), Mask-RCNN + BFPN (1.7 G / 15.8 fps), and, Mask-RCNN + EBFPN (1.8 G / 15.5 fps). We note that our design is not optimized for speed nor memory, and better speed/accuracy tradeoffs could be achieved by adopting recent optimization techniques (e.g., axial-decomposition, LSH-hashing), which is beyond the scope of this paper but is our future direction.

3 Qualitative Analysis

3.1 Feature visualization

We provide more visualization results of the learned feature relations in Fig. 1. Unlike the non-local case [23], our new formulation learns diverse query-specific relations, including both **intra-class** and **inter-class** relationships. Moreover, in Fig. 2, we visualize how the captured relationships are reflected in the final prediction. Compared to the non-local case [23],

Figure 1: We show learned relations on Pascal Context. Our formulation captures both intra-class and inter-class relationships while the non-local method can only model salient information.

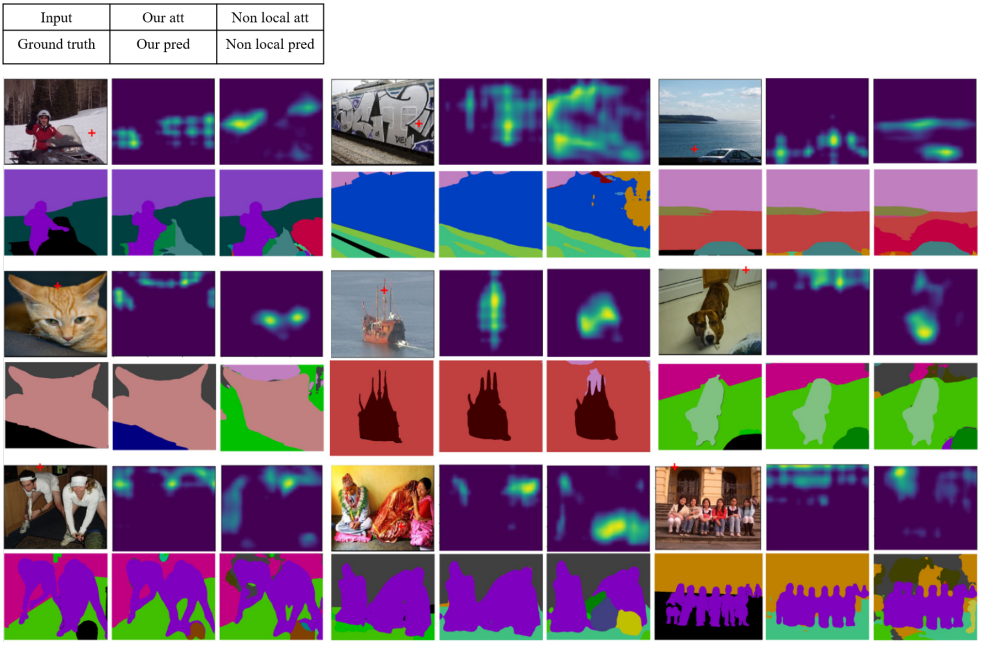

Figure 2: We show how the captured relations (i.e., attention) of the '+' sign marked in the input image are reflected in the final prediction. We use Pascal Context.

we observe that our formulation better exploits necessary neighboring features in different scenarios. We find this could support the model for more accurate predictions in ambiguous regions.

3.2 Output visualization

To assess the qualitative impact of our two instantiations further, we provide visual results on both semantic segmentation and panoptic segmentation. We demonstrate that our new formulation significantly improves the baselines on various datasets [5, 17, 19, 27].

3.2.1 Semantic segmentation

We use ResNet50-FCN [8] as a baseline. We remove the last two down-sampling operations and adopt the multi-grid dilated convolutions [3]. We append our module at the end of FCN. We use Pascal Context [19], Ade20K [27], and Cityscapes [5] to obtain visual results, which are shown in Fig. 3, Fig. 4 and Fig. 5, respectively. It is readily apparent that our predictions are more semantically accurate and spatially consistent compared to the baseline, demonstrating the efficacy of the non-locally aggregated context and relative position information.

3.2.2 Panoptic segmentation

We adopt PanopticFPN [14] (with ResNet50 + FPN backbone) as a baseline. We then apply the proposed EBFPN. We use COCO *val* [17] for visualization. In Fig. 6, we observe significant visual improvement when the EBFPN approach is applied to the baseline.

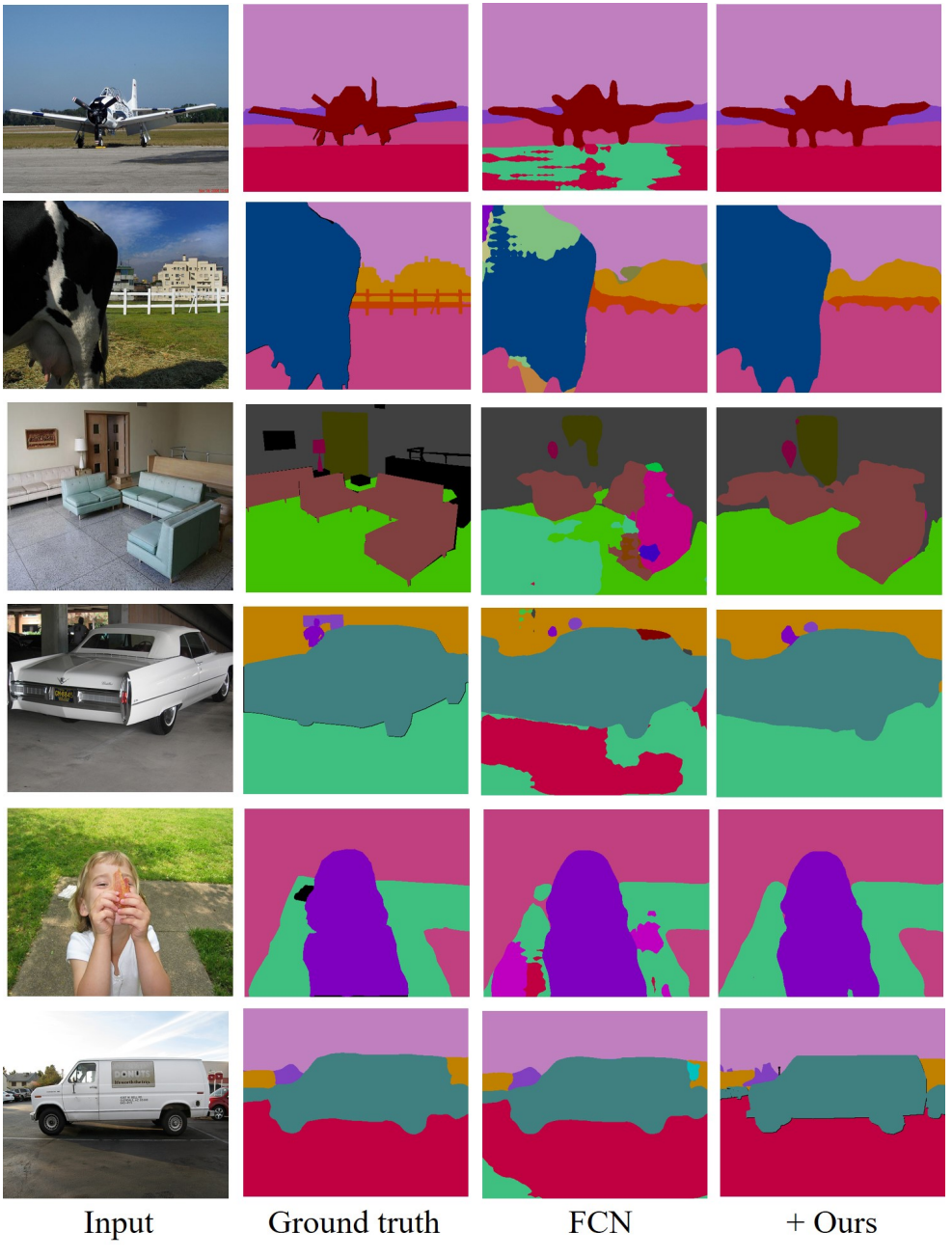

Figure 3: **Qualitative improvement by integrating our module into the FCN [8].** We use Pascal Context.

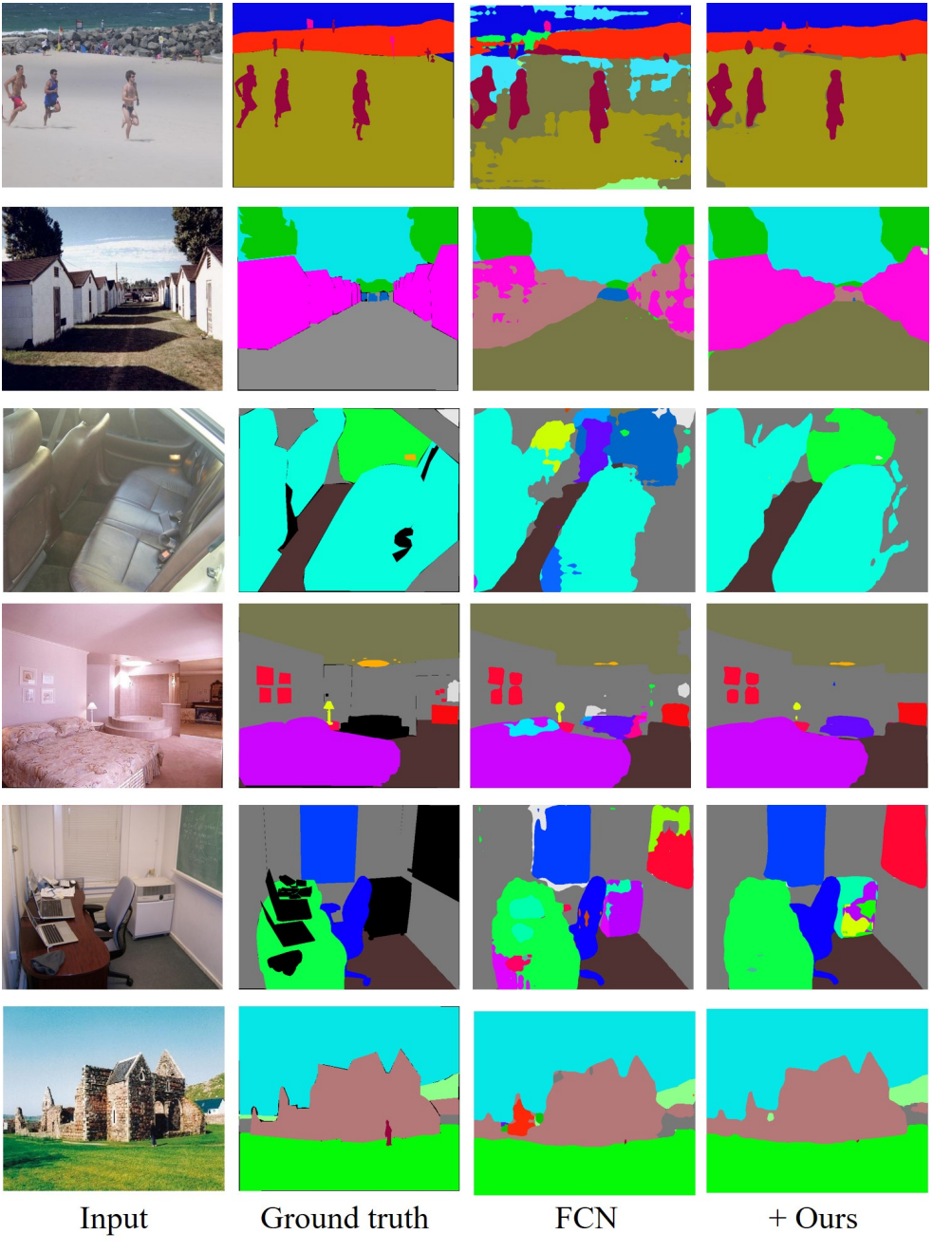

Figure 4: **Qualitative improvement by integrating our module into the FCN [5].** We use ADE20K.

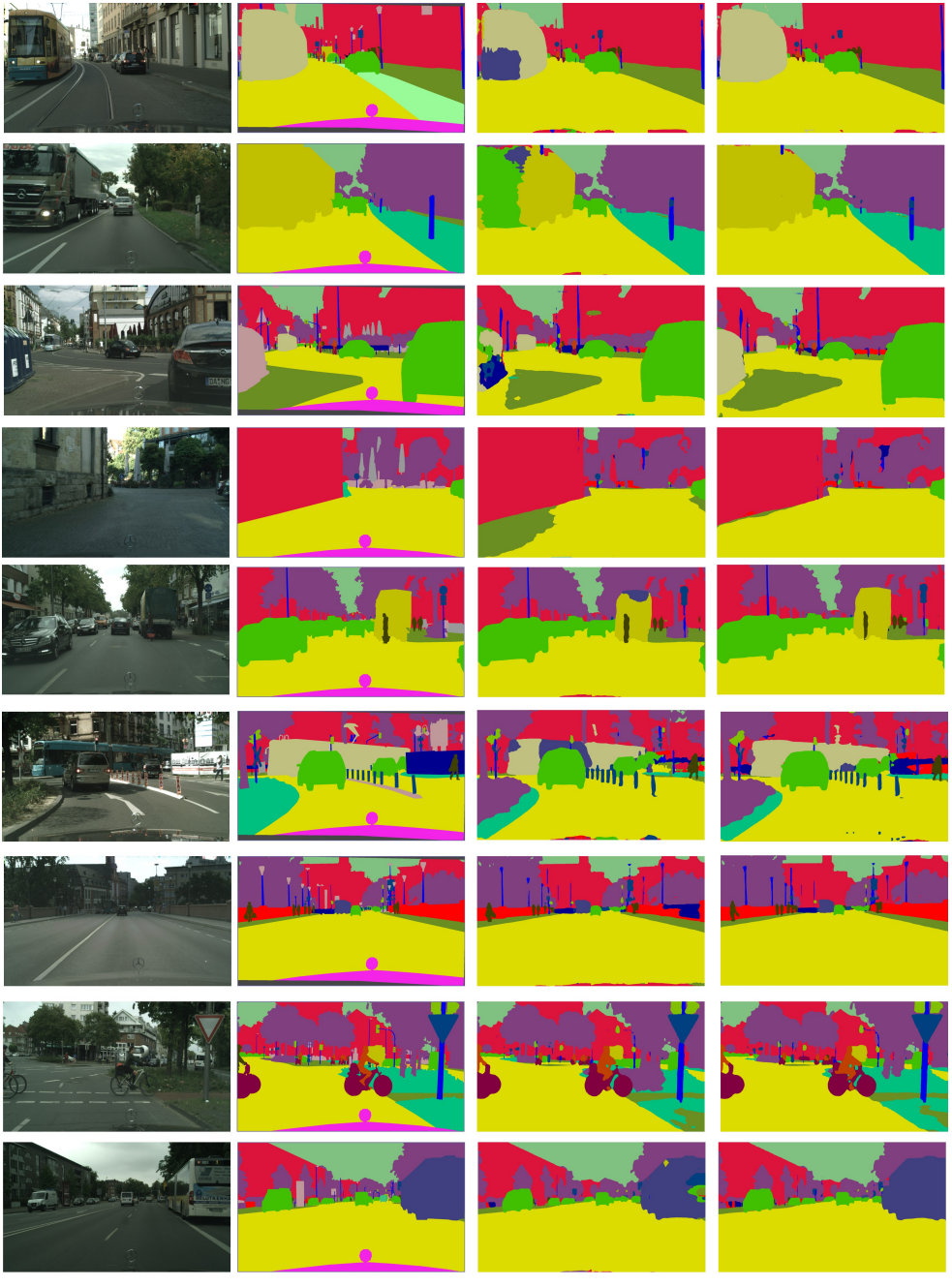

Figure 5: **Qualitative improvement by integrating our module into the FCN [8].** We use Cityscapes.

Figure 6: **Qualitative improvement by integrating EBFPN into the PanopticFPN [14].** We use COCO.

References

[1] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3286–3295, 2019.

[2] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[3] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[4] Yuhua Chen, Wen Li, and Luc Van Gool. Road: Reality oriented adaptation for semantic segmentation of urban scenes. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 7892–7901, 2018.

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016.

[6] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, pages 764–773, 2017.

[7] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 3146–3154, 2019.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, pages 2961–2969, 2017.

[10] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 3588–3597, 2018.

[11] Han Hu, Zheng Zhang, Zhenda Xie, and Stephen Lin. Local relation networks for image recognition. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, October 2019.

[12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018.

[13] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, pages 603–612, 2019.

[14] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 6399–6408, 2019.

[15] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 9404–9413, 2019.

[16] Xin Li, Zequn Jie, Wei Wang, Changsong Liu, Jimei Yang, Xiaohui Shen, Zhe Lin, Qiang Chen, Shuicheng Yan, and Jiashi Feng. Foveanet: Perspective-aware urban scene parsing. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, pages 784–792, 2017.

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proc. of European Conf. on Computer Vision (ECCV)*, pages 740–755. Springer, 2014.

[18] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. In *Proc. of Neural Information Processing Systems (NeurIPS)*, pages 9605–9616, 2018.

[19] 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. of Computer Vision and Pattern Recognition (CVPR)*, pages 891–898, 2014.

[20] Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, and Dahua Lin. Libra r-cnn: Towards balanced learning for object detection. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 821–830, 2019.

[21] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018.

[22] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[23] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 7794–7803, 2018.

[24] Fan Zhang, Yanqin Chen, Zhihang Li, Zhibin Hong, Jingtuo Liu, Feifei Ma, Junyu Han, and Errui Ding. Acfnet: Attentional class feature network for semantic segmentation. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, pages 6798–6807, 2019.

[25] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 7151–7160, 2018.

[26] Hang Zhang, Han Zhang, Chenguang Wang, and Junyuan Xie. Co-occurrent features in semantic segmentation. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 548–557, 2019.

[27] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 633–641, 2017.

[28] Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proc. of European Conf. on Computer Vision (ECCV)*, pages 289–305, 2018.