Supplementary Material: Dynamic Graph Message Passing Networks

Li Zhang[1] Dan Xu[1] Anurag Arnab[2*] Philip H.S. Torr[1]

[1]University of Oxford [2]Google Research

{lz, danxu, phst}@robots.ox.ac.uk aarnab@google.com

A. Additional experiments

In this supplementary material, we report additional qualitative results of our approach (Sec. A.1), additional details about the experiments in our paper (Sec. A.2), and also conduct further ablation studies (Sec. A.3).

A.1. Qualitative results

Figure 1 shows qualitative results for semantic segmentation (on Cityscapes) while Figure 2 and 3 show qualitative results for instance segmentation (on COCO).

A.2. Additional experimental details

A.2.1 Datasets

Cityscapes: Cityscapes [6] has densely annotated semantic labels for 19 categories in urban road scenes, and contains a total of 5000 finely annotated images, divided into 2975, 500, and 1525 images for training, validation and testing respectively. We do not use the coarsely annotated data in our experiments. The images of this dataset have a high resolution of 1024×2048. Following the standard evaluation protocol [6], the metric of mean Intersection over Union (mIoU) averaged over all classes is reported.

COCO: COCO 2017 [18] consists of 80 object classes with a training set of 118,000 images, a validation set of 5000 images, and a test set of 2000 images. We follow the standard COCO evaluation metrics [19] to evaluate the performance of object detection and instance segmentation, employing the metric of mean average-precision (mAP) at different box and mask IoUs respectively.

A.2.2 Semantic segmentation on Cityscapes

For the semantic segmentation task on Cityscapes, we follow [31] and use a polynomial learning rate decay with an initial learning rate of 0.01. The momentum and the weight decay are set to 0.9 and 0.0001 respectively. We use 4 Nvidia V100 GPUs, batch size 8 and train for 40000 iterations from an ImageNet-pretrained model. For data augmentation, random cropping with a crop size of 769 and random mirror-flipping are applied on-the-fly during training. Note that following common practice [31, 30, 32, 25] we used synchronised batch normalisation for better estimation of the batch statistics for the experiments on Cityscapes. When predicting dynamic filter weights, we use the grouping parameter $G = 4$. For the experiments on Cityscapes, we use a set of the sampling rates of $\varphi = \{1, 6, 12, 24, 36\}$.

A.2.3 Object detection and instance segmentation on COCO

Our models and all baselines are trained with the typical "1x" training settings from the public Mask R-CNN benchmark [21] for all experiments on COCO. More specificially, the backbone parameters of all the models in the experiments are pretrained on ImageNet classification. The input images are resized such that their shorter side is of 800 pixels and the longer side is limited to 1333 pixels. The batch size is set to 16. The initial learning rate is set to 0.02 with a decrease by a factor of 0.1 after 60000 and 80000 iterations, and finally terminates at 90000 iterations. Following [21, 11], the training warm-up is employed by using a smaller learning rate of 0.02×0.3 for the first 500 iterations of training. All the batch normalisation layers in the backbone are "frozen" during fine-tuning on COCO.

When predicting dynamic filter weights, we use the grouping parameter $G = 4$. For the experiments on COCO, a set of the sampling rates of $\varphi = \{1, 4, 8, 12\}$ is considered. We train models on only the COCO training set and test on the validation set and test-dev set.

A.3. Additional ablation studies

Effectiveness of different training and inference strategies. When evaluating models for the Cityscapes test set, we followed common practice and employed several complementary strategies used to improve performance in semantic segmentation, including Online Hard Example Mining

*Work primarily done at the University of Oxford.

	OHEM	Multi-grid	MS	mIoU (%)
FCN w/ DGMN	✘	✘	✘	79.2
FCN w/ DGMN	✔	✘	✘	79.7
FCN w/ DGMN	✔	✔	✘	80.3
FCN w/ DGMN	✔	✔	✔	**81.1**

Table 1: Ablation studies of different training and inference strategies. Our method (DGMN w/ DA+DW+US) is evaluated under single scale mIoU with ResNet-101 backbone on Cityscapes validation set.

(OHEM) [26, 23, 15, 30, 28], Multi-Grid [2, 10, 4] and Multi-Scale (MS) ensembling [1, 3, 31, 30, 7]. The contribution of each strategy is reported in Table 1 on the validation set.

Inference time We tested the average run time on the Cityscapes validation set with a Nvidia Tesla V100 GPU. The Dilated FCN baseline and the Non-local model take 0.230s and 0.276s per image, respectively, while our proposed model uses 0.253s, Thus, our proposed method is more efficient than Non-local [27] in execution time, FLOPs and also the number of parameters.

Effectiveness of different sampling rate φ and group of predicted weights G (Section 3.3 and 3.4 in main paper). For our experiments on Cityscapes, where are network has a stride of 8, the sampling rates are set to $\varphi = \{1, 6, 12, 24, 36\}$. For experiments on COCO, where the network stride is 32, we use smaller sampling rates of $\varphi = \{1, 4, 8, 12\}$ in C5. We keep the same sampling rate in C4 when DGMN modules are inserted into C4 as well.

Unless otherwise stated, all the experiments in the main paper and supplementary used $G = 4$ groups as the default. Each group of C/G feature channels shares the same set of filter parameters [5].

The effect of different sampling rates and groups of predicted filter weights are studied in Table 2, for semantic segmentation on Cityscapes, and Table 3, for object detection and instance segmentation on COCO.

Effectiveness of feature learning with DGMN on stronger backbones. Table 4 of the main paper showed that our proposed DGMN module still provided substantial benefits on the more powerful backbones such as ResNet-101 and ResNeXt 101 on the COCO test set. Table 4 shows this for the COCO validation set as well. By inserting DGMN at the convolutional stage C5 of ResNet-101, DGMN (C5) outperforms the Mask R-CNN baseline with 1.6 points on the AP^{box} metricand by 1.2 points on the AP^{mask} metric. On ResNeXt-101, DGMN (C5) also improves by 1.5 and 0.9 points on the AP^{box} and the AP^{mask}, respectively.

A.4. State-of-the-art comparison on COCO

Table 5 shows comparisons to the state-of-the-art on the COCO `test-dev` set. When testing, we process a single scale using a single model. We do not perform any other complementary performance-boosting "tricks". Our DGMN approach outperforms one-stage detectors including the most recent CornerNet [14] by 2.1 points on box Average Precision (AP). DGMN also shows superior performance compared to two-stage detectors including Mask R-CNN [12] and Libra R-CNN [22] using the same ResNeXt-101-FPN backbone.

	DA	DW	DS	mIoU (%)
Dilated FCN	✘	✘	✘	75.0
+ DGMN ($\varphi = \{1\}$)	✔	✘	✘	76.5
+ DGMN ($\varphi = \{1\}$)	✔	✔	✘	79.1
+ DGMN ($\varphi = \{1, 1, 1, 1\}$)	✔	✔	✔	79.2
+ DGMN ($\varphi = \{1, 6, 12\}$)	✔	✔	✔	79.7
+ DGMN ($\varphi = \{1, 6, 12, 24, 36\}$)	✔	✔	✔	80.4

Table 2: Quantitative analysis on different sampling rates of our dynamic sampling strategy in the proposed DGMN model on the Cityscapes validation set. We report the mean IoU and use a ResNet-101 as backbone. All methods are evaluated using a single scale.

	DA	DW	DS	AP^{box}	AP^{mask}
Mask R-CNN baseline	✘	✘	✘	37.8	34.4
+ DGMN ($\varphi = \{1, 4, 8, 12\}, G = 0$)	✔	✘	✘	39.4	35.6
+ DGMN ($\varphi = \{1, 4, 8, 12\}, G = 0$)	✔	✘	✔	39.9	35.9
+ DGMN ($\varphi = \{1, 4, 8\}, G = 2$)	✔	✔	✔	39.5	35.6
+ DGMN ($\varphi = \{1, 4, 8\}, G = 4$)	✔	✔	✔	39.8	35.9
+ DGMN ($\varphi = \{1, 4, 8, 12\}, G = 4$)	✔	✔	✔	**40.2**	**36.0**

Table 3: Quantitative analysis on different numbers of filter groups (G) and sampling rates (φ) for the proposed DGMN model on the COCO 2017 validation set. All methods are based on the Mask R-CNN detection pipeline with a ResNet-50 backbone, and evaluated on the COCO validation set. Modules are inserted after all the 3×3 convolution layers of C5 (*res5*) of ResNet-50.

Model	Backbone	AP^{box}	AP^{box}_{50}	AP^{box}_{75}	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}
Mask R-CNN baseline	ResNet-101	40.1	61.7	44.0	36.2	58.1	38.3
+ DGMN (C5)		**41.7**	**63.8**	**45.7**	**37.4**	**60.4**	**39.8**
Mask R-CNN baseline	ResNeXt-101	42.2	63.9	46.1	37.8	60.5	40.2
+ DGMN (C5)		**43.7**	**65.9**	**47.8**	**38.7**	**62.1**	**41.3**

Table 4: Quantitative results via applying the proposed DGMN module into different strong backbone networks for object detection and instance segmentation on the COCO 2017 validation set.

	Backbone	AP^{box}	AP^{box}_{50}	AP^{box}_{75}	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}
One-stage detectors							
YOLOv3 [24]	Darknet-53	33.0	57.9	34.4	-	-	-
SSD513 [20]	ResNet-101-SSD	31.2	50.4	33.3	-	-	-
DSSD513 [9]	ResNet-101-DSSD	33.2	53.3	35.2	-	-	-
RetinaNet [17]	ResNeXt-101-FPN	40.8	61.1	44.1			
CornerNet [14]	Hourglass-104	42.2	57.8	45.2			
Two-stage detectors							
Faster R-CNN+++ [13]	ResNet-101-C4	34.9	55.7	37.4	-	-	-
Faster R-CNN w FPN [16]	ResNet-101-FPN	36.2	59.1	39.0	-	-	-
R-FCN [8]	ResNet-101	29.9	51.9	-	-	-	-
Mask R-CNN [12]	ResNet-101-FPN	40.2	61.9	44.0	36.2	58.6	38.4
Mask R-CNN [12]	ResNeXt-101-FPN	42.6	64.9	46.6	38.3	61.6	40.8
Libra R-CNN [22]	ResNetX-101-FPN	43.0	64.0	47.0	-	-	-
DGMN (ours)	ResNeXt-101-FPN	**44.3**	**66.8**	**48.4**	**39.5**	**63.3**	**42.1**

Table 5: Object detection and instance segmentation performance using a *single-model* on the COCO `test-dev` set. We use *single scale* testing.

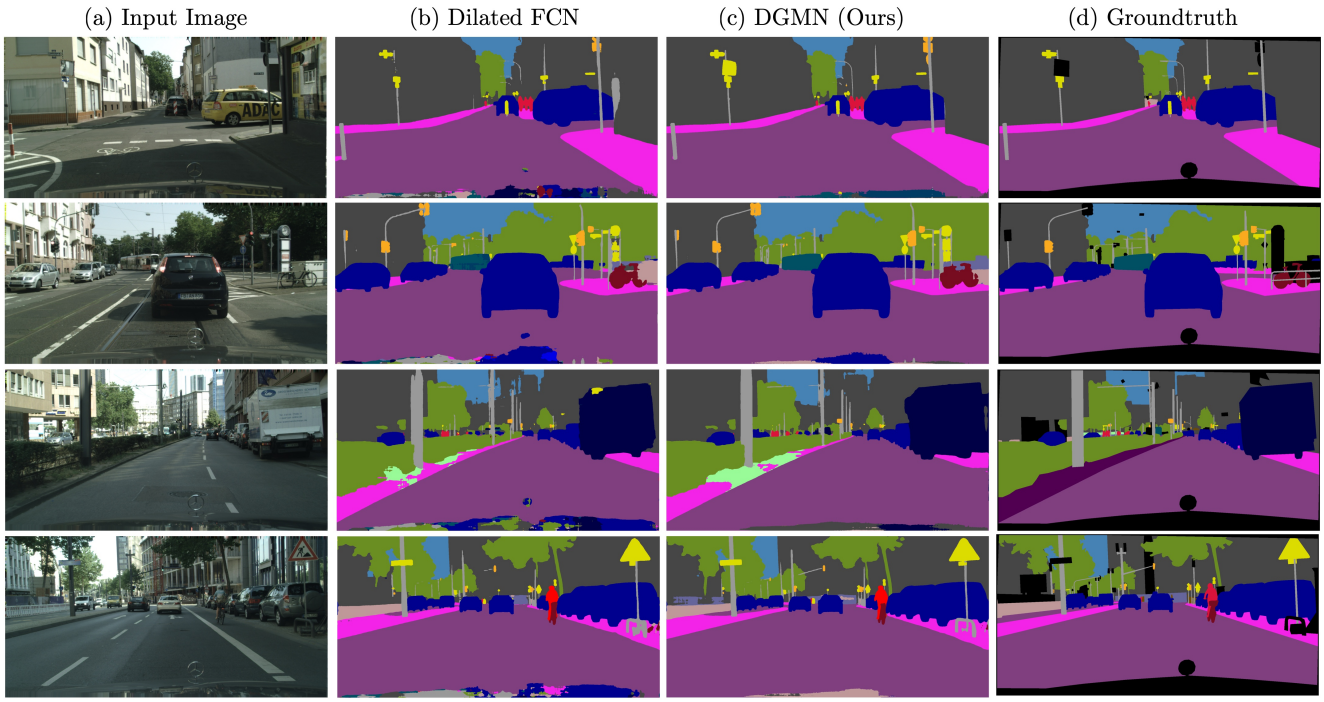

Figure 1: Qualitative results of the Dilated FCN baseline [29] and our proposed DGMN model on the Cityscapes dataset.

Figure 2: Qualitative examples of the instance segmentation task on the COCO validation dataset. The odd rows are the results from the Mask R-CNN baseline [21, 12]. The even rows are the results from our DGMN approach. Note how our approach often produces better segmentations and fewer false-positive and false-negative detections.

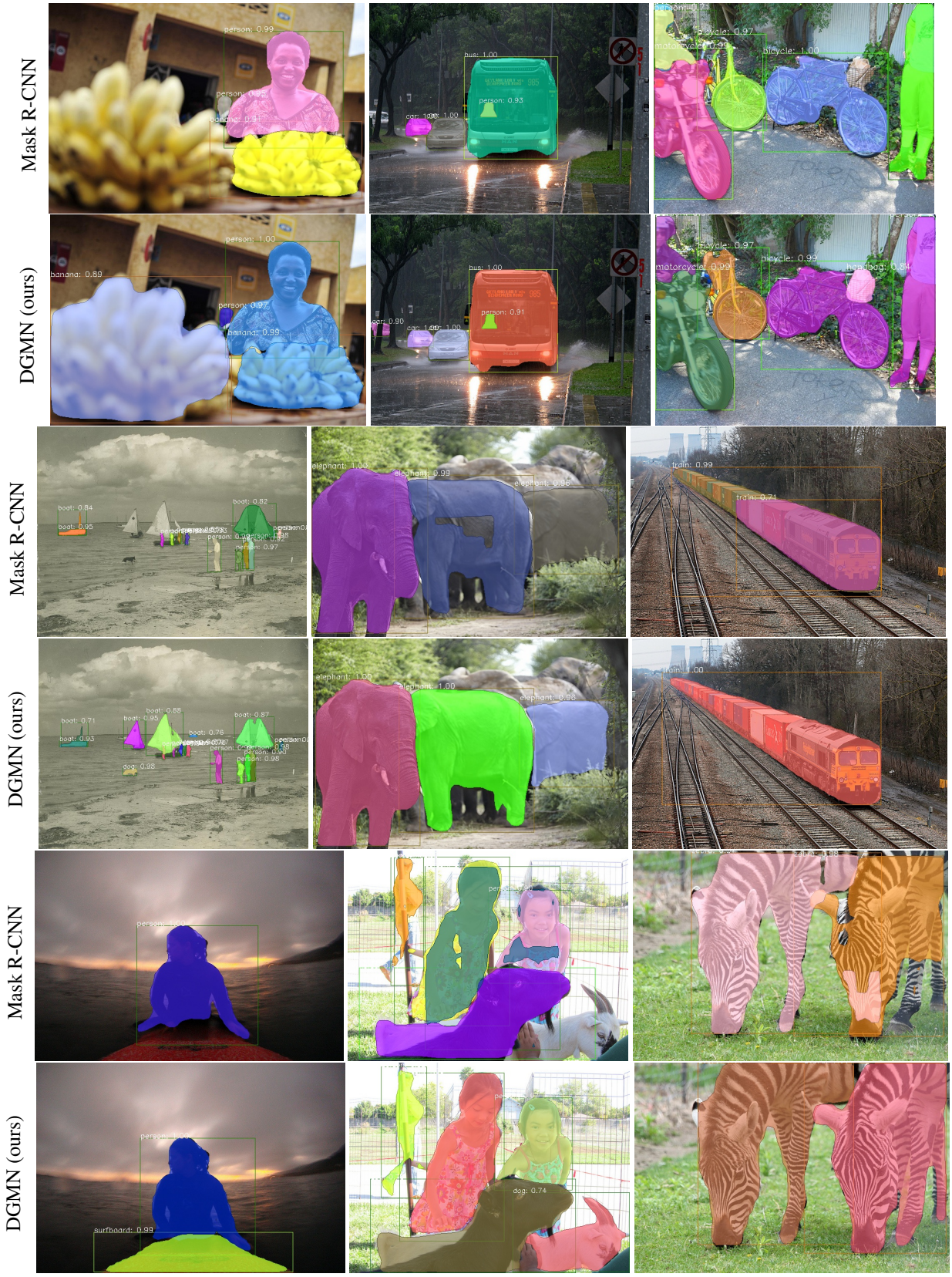

Figure 3: More qualitative examples of the instance segmentation task on the COCO validation dataset. The odd rows are the results from the Mask R-CNN baseline [21, 12]. The even rows are the detection results from our DGMN approach. Note how our approach often produces better segmentations and fewer false-positive and false-negative detections.

References

[1] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected CRFs. In *ICLR*, 2015. 2

[2] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 2

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 2

[4] Yunpeng Chen, Marcus Rohrbach, Zhicheng Yan, Shuicheng Yan, Jiashi Feng, and Yannis Kalantidis. Graph-based global reasoning networks. In *CVPR*, 2019. 2

[5] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, 2017. 2

[6] 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 *CVPR*, 2016. 1

[7] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *ICCV*, 2015. 2

[8] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *NeurIPS*, 2016. 3

[9] Cheng-Yang Fu, Wei Liu, Ananth Ranga, Ambrish Tyagi, and Alexander C Berg. Dssd: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*, 2017. 3

[10] Jun Fu, Jing Liu, Haijie Tian, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, 2019. 2

[11] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. In *arXiv preprint arXiv:1706.02677*, 2017. 1

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 2, 3, 5, 6

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[14] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *ECCV*, 2018. 2, 3

[15] Qizhu Li, Anurag Arnab, and Philip HS Torr. Holistic, instance-level human parsing. In *BMVC*, 2017. 2

[16] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 3

[17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 3

[18] 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 *ECCV*, 2014. 1

[19] 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 *ECCV*, 2014. 1

[20] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016. 3

[21] Francisco Massa and Ross Girshick. maskrcnn-benchmark: Fast, modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch. https://github.com/facebookresearch/maskrcnn-benchmark, 2018. 1, 5, 6

[22] Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, and Dahua Lin. Libra r-cnn: Towards balanced learning for object detection. In *CVPR*, 2019. 2, 3

[23] Tobias Pohlen, Alexander Hermans, Markus Mathias, and Bastian Leibe. Full-resolution residual networks for semantic segmentation in street scenes. In *CVPR*, 2017. 2

[24] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 3

[25] Samuel Rota Bulò, Lorenzo Porzi, and Peter Kontschieder. In-place activated batchnorm for memory-optimized training of dnns. In *CVPR*, 2018. 1

[26] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *CVPR*, 2016. 2

[27] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 2

[28] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *ECCV*, 2018. 2

[29] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016. 4

[30] Yuhui Yuan and Jingdong Wang. Ocnet: Object context network for scene parsing. *arXiv preprint arXiv:1809.00916*, 2018. 1, 2

[31] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 1, 2

[32] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *ECCV*, 2018. 1