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 $1024 \times 2048$. 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 specifically, 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 $0.02 \times 0.3$ 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...
**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.

<table>
<thead>
<tr>
<th>Method</th>
<th>OHEM</th>
<th>Multi-grid</th>
<th>MS</th>
<th>mIoU (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCN w/ DGMN</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>79.2</td>
</tr>
<tr>
<td>FCN w/ DGMN</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>79.7</td>
</tr>
<tr>
<td>FCN w/ DGMN</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>80.3</td>
</tr>
<tr>
<td>FCN w/ DGMN</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>81.1</td>
</tr>
</tbody>
</table>

(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 \( \phi \) 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 \( \phi = \{1, 6, 12, 24, 36\} \). For experiments on COCO, where the network stride is 32, we use smaller sampling rates of \( \phi = \{1, 4, 8, 12\} \) 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 \( \text{AP}^{\text{box}} \) metric and by 1.2 points on the \( \text{AP}^{\text{mask}} \) metric. On ResNeXt-101, DGMN (C5) also improves by 1.5 and 0.9 points on the \( \text{AP}^{\text{box}} \) and the \( \text{AP}^{\text{mask}} \), 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.
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.

<table>
<thead>
<tr>
<th>Model</th>
<th>DA</th>
<th>DW</th>
<th>DS</th>
<th>mIoU (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dilated FCN</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>75.0</td>
</tr>
<tr>
<td>+ DGMN (ϕ = {1})</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>76.5</td>
</tr>
<tr>
<td>+ DGMN (ϕ = {1, 1, 1})</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>79.1</td>
</tr>
<tr>
<td>+ DGMN (ϕ = {1, 6, 12})</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>79.2</td>
</tr>
<tr>
<td>+ DGMN (ϕ = {1, 6, 12, 24, 36})</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>80.4</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Model</th>
<th>DA</th>
<th>DW</th>
<th>DS</th>
<th>AP_{box}</th>
<th>AP_{mask}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mask R-CNN baseline</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>37.8</td>
<td>34.4</td>
</tr>
<tr>
<td>+ DGMN (ϕ = {1, 4, 8, 12}, G = 0)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>39.4</td>
<td>35.6</td>
</tr>
<tr>
<td>+ DGMN (ϕ = {1, 4, 8, 12}, G = 0)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>39.9</td>
<td>35.9</td>
</tr>
<tr>
<td>+ DGMN (ϕ = {1, 4, 8}, G = 2)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>39.5</td>
<td>35.6</td>
</tr>
<tr>
<td>+ DGMN (ϕ = {1, 4, 8}, G = 4)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>39.8</td>
<td>35.9</td>
</tr>
<tr>
<td>+ DGMN (ϕ = {1, 4, 8, 12}, G = 4)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>40.2</td>
<td>36.0</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Backbone</th>
<th>AP_{box}</th>
<th>AP_{box}</th>
<th>AP_{box}</th>
<th>AP_{mask}</th>
<th>AP_{mask}</th>
<th>AP_{mask}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mask R-CNN baseline</td>
<td>ResNet-101</td>
<td>40.1</td>
<td>61.7</td>
<td>44.0</td>
<td>36.2</td>
<td>58.1</td>
<td>38.3</td>
</tr>
<tr>
<td>+ DGMN (C5)</td>
<td>ResNet-101</td>
<td>41.7</td>
<td>63.8</td>
<td>45.7</td>
<td>37.4</td>
<td>60.4</td>
<td>39.8</td>
</tr>
<tr>
<td>Mask R-CNN baseline</td>
<td>ResNeXt-101</td>
<td>42.2</td>
<td>63.9</td>
<td>46.1</td>
<td>37.8</td>
<td>60.5</td>
<td>40.2</td>
</tr>
<tr>
<td>+ DGMN (C5)</td>
<td>ResNeXt-101</td>
<td>43.7</td>
<td>65.9</td>
<td>47.8</td>
<td>38.7</td>
<td>62.1</td>
<td>41.3</td>
</tr>
</tbody>
</table>

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


[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


