Dynamic Multi-scale Filters for Semantic Segmentation

Junjun He [1,2] Zhongying Deng [1] Yu Qiao[1]*
[1]ShenZhen Key Lab of Computer Vision and Pattern Recognition, SIAT-SenseTime Joint Lab,
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China
[2]Shanghai Jiao Tong University
hejunjun@sjtu.edu.cn, {zy.deng1,yu.qiao}@siat.ac.cn

Abstract

Multi-scale representation provides an effective way to address scale variation of objects and stuff in semantic segmentation. Previous works construct multi-scale representation by utilizing different filter sizes, expanding filter sizes with dilated filters or pooling grids, and the parameters of these filters are fixed after training. These methods often suffer from heavy computational cost or have more parameters, and are not adaptive to the input image during inference. To address these problems, this paper proposes a Dynamic Multi-scale Network (DMNet) to adaptively capture multi-scale contents for predicting pixel-level semantic labels. DMNet is composed of multiple Dynamic Convolutional Modules (DCMs) arranged in parallel, each of which exploits context-aware filters to estimate semantic representation for a specific scale. The outputs of multiple DCMs are further integrated for final segmentation. We conduct extensive experiments to evaluate our DMNet on three challenging semantic segmentation and scene parsing datasets, PASCAL VOC 2012, Pascal-Context, and ADE20K. DMNet achieves a new record 84.4% mIoU on PASCAL VOC 2012 test set without MS COCO pre-trained and post-processing, and also obtains state-of-the-art performance on Pascal-Context and ADE20K.

1. Introduction

Semantic segmentation is an important yet challenging task in computer vision which aims at assigning a category label to every pixel in an image. It plays an important role in scene understanding [51, 52], medical image analysis [36], self-driving [40, 45], and many other applications. Recently, approaches based on deep convolutional neural networks (DCNNs) , especially fully convolutional networks (FCNs) [32], have achieved great success in semantic segmentation. However, the existence of objects and stuff with large scale variations often cause difficulty in pixel-level dense prediction, especially for extremely small or large scale objects and stuff. So it is not reasonable to predict the labels of all pixels with a single scale representation. Therefore, multi-scale representations are desired for robust and accurate semantic segmentation [39, 5, 6, 51]. In DCNNs, different scales of objects and stuff can be captured by filters with different receptive fields. The most appropriate receptive field usually corresponds to the sizes of objects and stuff. If the receptive field can only cover a small part of a large scale object or stuff, inconsistent segmentation result may happen. While the receptive field is extremely larger than the object or stuff with small scale, background may dominate the predictions, leading to invisible of small object and stuff. Thus, using multiple receptive fields to capture multi-scale objects and stuff with large variation is critical for dense image prediction in DCNNs.

An intuitive way toward this problem is to utilize multiple kernels with different sizes in parallel. Inception block proposed in [39] adopts multiple branches with different kernel sizes to capture multi-scale information. However, it is inefficient to use large receptive filed for large scale objects, due to the parameters and computational cost increase exponentially to the kernel size. It also increases the risk of over-fitting. Pyramid Pooling Module (PPM) proposed in PSPNet [51], performs pooling operation at different grids which can be seen as a series different non-parametric convolutions with large kernel sizes and strides. It provides an effective method to capture multi-scale context information. But PPM puts equal weights at every position and the fine-detail information may lose in pooling operation, which can hamper the final performance.

Atrous convolution can enlarge the receptive field without adding extra parameters and computational cost, compared with regular convolution with larger filters. [5] proposed Atrous Spatial Pyramid Pooling (ASPP) module to capture multi-scale information by employing multiple dilated filters with different dilation rates which lays out in parallel. It can handle scale variations to some extent, but

[*] Corresponding author

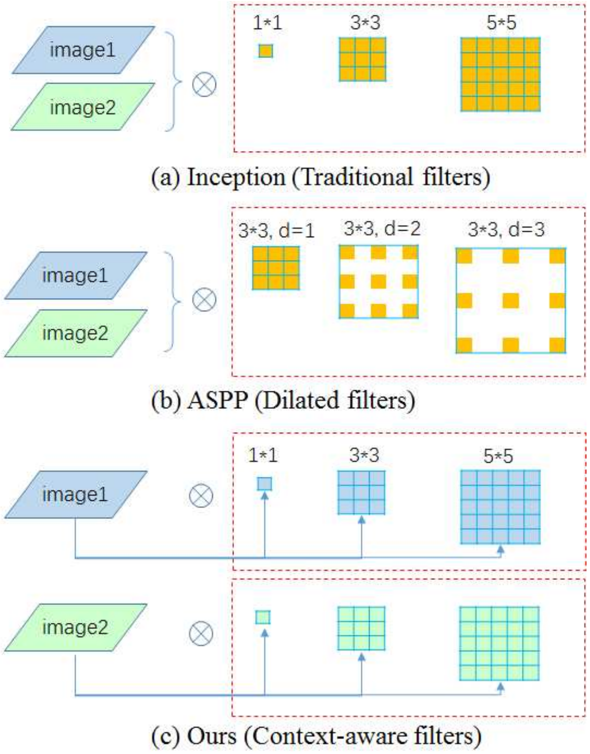
image1
image2
1*1
3*3
5*5
(a) Inception (Traditional filters)
3*3, d=1
3*3, d=2
3*3, d=3
(b) ASPP (Dilated filters)
(c) Ours (Context-aware filters)

Figure 1. Three architectures to capture multi-scale feature representations. (a) multiple kernels with different kernel sizes in Inception. (b) multiple kernels with the same kernel sizes but different dilation rates in ASPP (atrous spatial pyramid pooling). (c) multiple context-aware kernels with different kernel sizes whose weights are estimated from the input, in contrast to previous two methods whose weights are fixed after training.

it is hard to achieve a tradeoff between dilation rates and the range of scale variations. Moreover, the sparse sampling method will lose neighbor information and larger dilation rate may cause gridding artifacts. [6] shows that ASPP is sensitive to input image size, and small crop size leads to boundary effect even degeneration. Moreover, the aforementioned methods use fixed parameters after training, which is not adaptive to input images in reference.

In this paper, we advocate an alternative approach for multi-scale feature learning by exploring dynamic multi-scale filters. Specially, we propose a simple yet effective Dynamic Multi-scale Network (DMNet) to handle scale variations of objects and stuff for semantic segmentation. DMNet consists of several Dynamic Convolutional Modules (DCMs), each of which exploit context-aware filters to handle a specific scale related to input. The context-aware filters are generated from input multi-scale features in a dynamic way, which allows them to embed high-level semantics and capture rich contents. Furthermore, our DCM is efficient due to depthwise convolution, which requires less model parameters. Figure 1 shows the key difference of our method from aforementioned methods Inception and ASPP on multi-scale feature representations. The main contributions are as follows.

- We propose DMNet to exploit dynamic multi-scale filters for semantic segmentation in an end-to-end fashion. Compared with previous methods, DMNet are adaptive to image contents.
- We introduce several Dynamic Convolutional Modules to captures multi-scale semantics with context-aware filters. Each DCM can handle a specific scale variation related to the input.
- Our methods achieve state-of-the-art performance on three widely used benchmarks, including PASCAL VOC 2012, Pascal-Context, and ADE20K datasets, and reach a new record of 84.4% mIoU [1] on PASCAL VOC 2012 test set without MS COCO pre-trained and any post-processing.

2. Related Work

Multi-scale features. Due to large scale variations of objects and stuffs in complex scenes [51], it is necessary to adopt multi-scale representations for accurate and effective dense pixel-level image segmentation. Image pyramid is an intuitive way to obtain multi-scale image representations where multiple resized input images are feed to the same model and the results are finally fused [15, 5, 50, 27]. But the increased time in training and inference makes the image pyramid methods inefficient in particular. [36, 2, 7] design the Encoder-Decoder architecture to fuse features from different layers, while [28, 29] predict results from different layers with different receptive fields to capture different scales. The main drawback of these methods is that they ignore the consistency of feature representations across different scales. [5, 6] propose Atrous Spatial Pyramid Pooling (ASPP) module which applies multiple atrous convolutional filters with different dilation rates in parallel, while PSPNet [51] employs Pyramid Pooling Module (PPM) to perform pooling operations at different grid scales. These sparse sampling methods are unable to capture neighbor information and details. [20] utilizes pyramid context to construct multi-scale representations, which are adapted to input images. In contrast to aforementioned methods, we propose an alternative approach to exploit dynamic multi-scale filters for multi-scale features learning. Each branch of our model can capture a specific scale feature related to the input image which is more flexible and adaptive.

[1] Result link on PASCAL VOC 2012 test set without coco pre-trained: http://host.robots.ox.ac.uk:8080/anonymous/GOQMVH.html.

Dynamic filters. Previous works have explored dynamic filters to boost the performance of DCNNs. SAC [50] predicts position-adaptive scale coefficients to relax the fixed-size receptive field, thus it can tackle the problems of invisible small objects and inconsistent large objects segmentation to some extent. [4] explicitly constructs filters bank encoded with objective content which can transform the input image to a specific style. [9, 55] learn offsets for each element of regular convolutional filters to enlarge the sampling field with arbitrary form which can discover geometric-invariant features. DFN [23] exploits dynamic filters to capture different motion patterns within the inputs for video and stereo prediction, and [41] also uses dynamic filters aiming at constructing large receptive fields and receiving local gradients, to produce sharper and more semantic feature maps. [18] generates contrastive convolutional filters conditioned on input image pairs to focus on the distinct characteristics. Whereas, our method generates filters of different kernel sizes from different scale region-based context, which is more effective than previous multi-scale feature learning methods, due to specific scale representation related to input image is captured. Moreover, in our proposed DCM, we adopt depthwise convolution which is more efficient with less parameters.

3. Method

Due to large scale variations of objects and stuff in complex scenes, it is often challenging to make dense pixel-level prediction. Also, it is difficult to represent every pixel with a single scale feature, thus multi-scale representations are necessary for accurate and robust semantic segmentation and scene parsing. In contrast to previous works [5, 51, 39], we propose a simple yet effective Dynamic Multi-scale Network (DMNet), which is an alternative approach for multi-scale feature learning by exploring dynamic multi-scale filters. In the next, we will describe it in detail.

3.1. Overview of Dynamic Multi-scale Network

The architecture of DMNet is shown in Figure 2. DMNet consists of a backbone convolutional neural network (CNN) and several Dynamic Convolutional Modules (DCMs). The backbone CNN for feature extraction can be VGG, ResNet, DenseNet or Xception. While each DCM can capture a specific scale feature representation related with the input image, different DCMs of DMNet can obtain multiple scale feature representations, which is more flexible and adaptive. We arrange DCMs in parallel, thus not sacrifice the consistency of feature representation power across different scales. The outputs of each specific feature representation are aggregated with the original feature extracted by backbone CNN to form the final feature representations for every pixel. Then, these aggregated robust multi-scale feature representations are feed to pixel-level predictor for dense image segmentation.

3.2. Dynamic Convolutional Module

The goal of DCM is to capture a specific scale representation for the input image adaptively. In DCNNs, different receptive fields are suitable for obtaining different scale representations, so we explore multi-scale filters for multi-scale feature learning. We name these filters as context-aware filters which are dynamically generated from region context conditioned on the input image. The context-aware filters are embedded with rich contents and high-level semantics intrinsically, in spite of dense small kernel sizes. It is adaptive to the input image, and more flexible than traditional filters, thus it can capture internal variation of the input image. Next, we describe the context-aware filters and DCM in detail.

Given a feature map $x \in \mathbb{R}^{h \times w \times c}$ extracted by backbone CNN as the input of DCM, we first apply feature reduction f_k to it and get the reduced feature map $f_k(x) \in \mathbb{R}^{h \times w \times c'}$, where h, w, c are the height, width and number of channels respectively. c' is the number of channels of the reduced input feature map ($c' < c$) and k is the kernel size of context-aware filters. Denote g_k as kernel generator with kernel size k, and the generated filters $g_k(x) \in \mathbb{R}^{k \times k \times c'}$ are referred as context-aware filters. And then the reduced feature map is convolved with the generated context-aware filters with depthwise convolution to obtain a specific scale representation.

$$h_k = f_k(x) \otimes g_k(x), \tag{1}$$

where symbol $\otimes$ is the depthwise convolution. Then h_k is processed with 1×1 convolution to fuse the channel information, as depthwise separable convolution. The final output of the Dynamic Convolutional Module (DCM) is $O_k \in \mathbb{R}^{h \times w \times c'}$.

In our implementation, the feature reduction f_k is a 1×1 convolution operation and the filter generator g_k consists of one adaptive average pooling operation (AAP) and one 1×1 convolution operation. For a specific DCM with kernel size k, $k \times k$ region-based context is first obtained by applying AAP to the input of DCM, and convolved with the 1×1 convolution to generate the $k \times k$ context-aware filters. Then, the generated context-aware filters convolve with the reduced feature map $f_k(x)$ to get h_k. Finally, a 1×1 convolution is applied to fuse channel information, and the specific scale representation O_k is obtained as the output of DCM. As mentioned above, we can obtain arbitrary size of context-aware filters with only one 1×1 convolution layer. Therefore, we can generate context-aware filters with different kernel sizes for different DCMs to capture multi-scale contents.

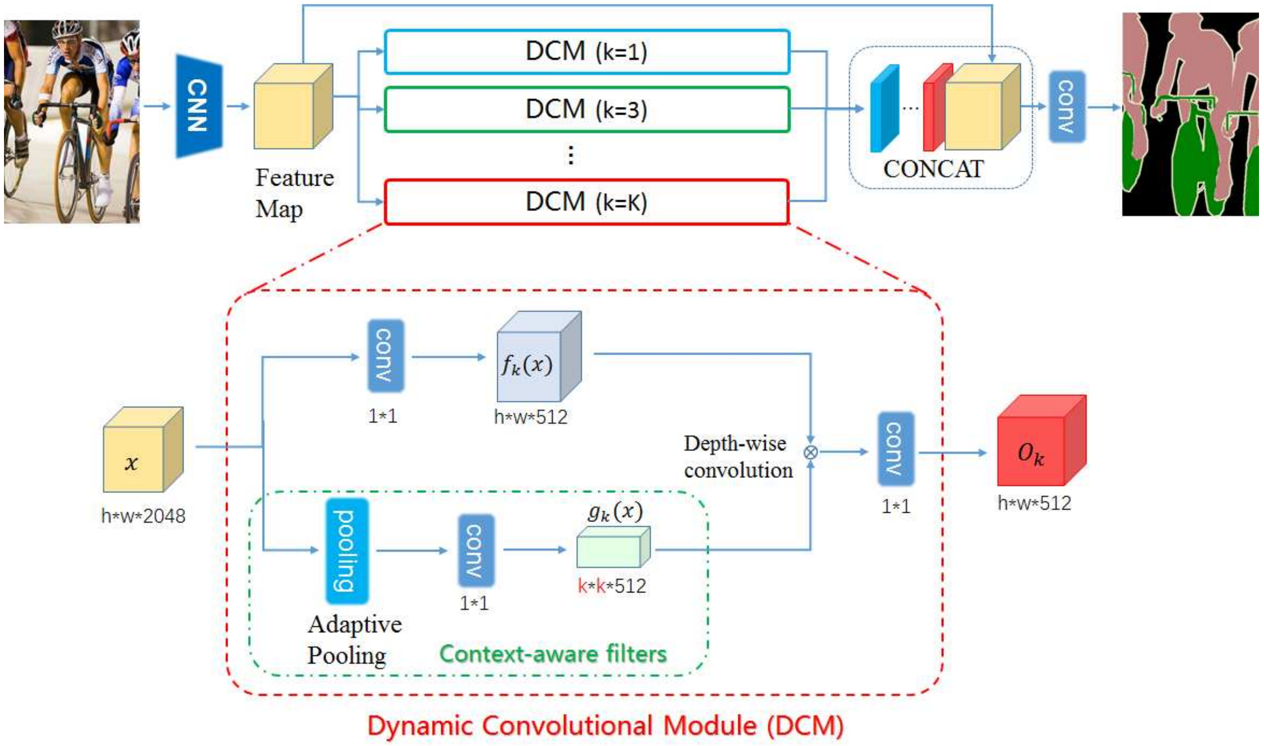

Figure 2. The pipeline of Dynamic Multi-scale Network (DMNet). DMNet consists of a backbone convolutional neural network (CNN) and several Dynamic Convolutional Modules (DCMs). The backbone CNN for feature extraction can be VGG, ResNet, DenseNet or Xception. While each DCM can capture a specific scale feature representation related with the input image, different DCMs of DMNet can obtain multiple scale feature representations. The outputs of each specific feature representations are aggregated with the original feature extracted by backbone CNN to form the final feature representations for every pixel. Finally, these aggregated robust multi-scale feature representations are feed to pixel-level predictor for dense image segmentation.

3.3. Discussion

Comparison with other network architectures. Many previous works [5, 6, 7, 51] use fixed filters which are not adapted to the input image and only capture a series of preset scale feature representations. ASPP [5] and PPM [51] which adopt predefined atrous convolution with different dilation rates and pooling operations with different grids respectively. These two methods are not only sensitive to input image size, but also sensitive to the scale difference between images in training and inference phrase. The fixed weights, preset dilation rates and pooling grids cannot capture the internal scale variations of input images with arbitrary scales and sizes.

Comparison with other kinds of filters. Previous works lie in enlarging the kernel size to obtain large receptive field. Inception [39] expands the kernel size in a dense way while ASPP expands the kernel size in a sparse way. Therefore, multi-scale representations can be obtained by different kernels with different expansion rates. However, with receptive field increases larger, the parameters of the former one is exploded and suffers overfitting and heavy computational cost. Thought the latter one can arbitrary enlarge the receptive field, the sparse sampling way may lose fine-detail information and cause gridding artifacts. We can obtain dense context-aware filters with arbitrary size adding only one 1×1 convolution layer.

4. Experiments

To show the effectiveness of our proposed DMNet, extensive experiments are conducted on PASCAL VOC 2012 [13], Pascal-Context [33], and ADE20K [54] datasets.

4.1. Implementation Details

We use ImageNet [37] to pre-train our backbone CNN, i.e. ResNet [21]. The stride of the backbone CNN is removed and the dilation rate is set to 2 and 4 for its last two stages, which is the same as [48, 5, 49]. This modification makes the size of output feature map to be 1/8 of the input image. At training stage, the crop size of input image is 512×512 on PASCAL VOC 2012 [13] and Pascal-Context [33] dataset. For ADE20K [54], since its average image size is larger than other two datasets [5, 51, 49], we set larger crop size of 576×576 accordingly. In addition, the input image is randomly flipped and scaled to perform data augmentation. The scaling rate ranges from 0.5 to 2.0. When evaluating the CNN model, we also flip and resize the testing images to multiple scales. To predict semantic labels of each pixel, bilinear interpolation is applied to the output predictions to obtain target size. Finally, we use averaged predictions of different scales as final prediction [30, 51, 38, 46, 17, 10, 20, 12, 11] and adopt mean of class-wise intersection over union (mIoU) as our evaluation metric .

We set the initial learning rate to 0.01 for PASCAL VOC 2012 [13] and ADE20K dataset [54], 0.005 for Pascal-Context dataset [33]. The initial learning rate is multiplied by $(1-\frac{iter}{total\_iter})^{power}$ at different training iteration, where the power is 0.9 [49]. We adopt SGD [3] with momentum of 0.9 and weight decay of 0.0001 as the optimizer. It takes about 80 epochs for our CNN model to converge on PASCAL VOC 2012 [13] and Pascal-Context dataset [33], and 120 epochs on ADE20K dataset [54]. We implement all experiments based on PyTorch [35] with synchronized batch normalization [22].

4.2. PASCAL VOC 2012

PASCAL VOC 2012 [13] is a widely used benchmark dataset for semantic segmentation which contains 21 object classes (including a background class). Since there are only 1,464 training images in original PASCAL VOC 2012, most works [5, 49, 7] use an augmented training set of PASCAL VOC 2012 [19] whose training set includes 10,582 images. In our experiments, we also take the same augmented training set as these works.

Firstly, we conduct the ablation study to illustrate the effectiveness of our proposed modules, and then compare our DMNet with the state-of-the-art methods. We implement our method with different sets of kernel sizes including $\{1\}$, $\{1, 3\}$, $\{1, 3, 5\}$, $\{1, 3, 5, 7\}$ and the backbone is ResNet50 pre-trained on ImageNet. In each set, the number of elements equals to the number of DCM, and the element corresponds to the kernel size of context-aware filters in DCM.

Context-aware filters v.s. traditional filters. We replace context-aware filters in each DCM with traditional filters of the same kernel size whose paremeters is fixed after training. Depthwise convolution is adopted in DCM, so we conduct both regular convolution and depthwise convolution in traditional filters for thorough studies. From Table 1, we can see that the performance of traditional filters with depthwise convolution is slightly worse than regular convolution, since depthwise convolution has less parameters. Our method outperforms traditional filters with regular convolution and depthwise convolution in setting with a large margin. Our context-aware filters improve the results of traditional filters by about 7% for kernel size $\{1\}$ and $\{1,3\}$ setting with mIoU 77.75% and 78.52%, respectively. It is noting that our methods use depthwise convolution with less parameters, but clearly outperform traditional filters with regular convolution. We owe the superior performance to the fact that context-aware filters of DCM can capture a specific scale representation related to input image. These results indicate that DCM is effective and adaptive.

Kernel size	$\{1\}$	$\{1,3\}$	$\{1,3,5\}$	$\{1,3,5,7\}$
Regular conv	70.76	71.41	**72.65**	72.37
Depthwise conv	70.59	71.12	**72.24**	71.88
Ours	77.75	78.52	**78.76**	78.21

Table 1. Comparison between Dynamic convolution in DCM with context-aware filters and traditional convolution (regular convolution and depthwise convolution), i.e. replacing context-aware filters with traditional filters. Depthwise convolution is adopted in our DCM. Four sets of kernel sizes are implemented. In each set, the number of elements equals to the number of branches of models, and the element corresponds to the kernel size. And the number of elements is also the number of DCMs.

Context-aware filters v.s. dilated filters. Following DeepLab [5] and PSPNet [51], we adopt four branches in our model to learning multi-scale features for fair comparison. Here, the context-aware filters are replaced by dilated filters with the kernel size unchanged (3×3) and dilation rates varied in each branch. We also conduct both regular convolution and depthwise convolution with dilated filters. To comprehensively study the effectiveness of DMNet, we set a series of dilation rates. The dilation rates are set as $\{2, 4, 8, 12\}$ and $\{6, 12, 18, 24\}$ corresponding to ASPP-S and ASPP-L in DeeplabV2 [5] respectively. But there are some differences: we insert these dilation filters in the inter-media convolution before the score map to capture multi-scale feature representations, while DeepLabV2 [5] inserts them in the last convolutional layer used for predicting final multi-scale score maps.

From Table 2, we can observe that the performances are highly related with dilation rates. As the dilation rates increase, the performance of atrous convolution increases, but it begins to decline with too large dilation rates due to degeneration of artous convolution. Note that the performance of dilation rates $\{2, 4, 8, 12\}$ (ASPP-S) is better than $\{6, 12, 18, 24\}$ (ASPP-L), which is different from the observation in DeepLabV2 [5]. The difference may be because atrous convolution is sensitive to insert positions. Our four-branch model with kernel sizes $\{1,3,5,7\}$ of context-aware filters outperforms all atrous convolution with different dilation rates, which exhibits that the context-aware filters are more effective to capture rich contents for segmentation.

Parameters and flops of different methods. Compared with traditional filters based method Inception (filter sizes $\{1,3,5,7\}$) and dilated filters based method ASPP (filter sizes $\{3,3,3,3\}$, dilation rates $\{6,12,24,36\}$), the proposed context-aware filters based method DMNet has less parameters and flops, as shown in Table 3 (without backbone and classifier, input size 512*512).

Filter types	Dilation rates	Regular	Depthwise
Dilated	{1,2,3,4}	73.37	72.98
	{1,3,5,7}	76.22	74.29
	{2,4,8,12}	77.18	77.30
	{6,12,18,24}	76.94	76.26
	{6,12,24,36}	**77.60**	**77.34**
	{12,24,36,48}	77.17	77.10
	{18,30,42,54}	76.88	76.92
Context-aware	-		**78.21**

Table 2. Performance of baselines with different dilation rates. The kernel size of each branch is 3, yet with different dilation rates. In each set, the number of elements equals to the number of branches of models, and the element corresponds to the dilation rate. 'Regular' means regular convolution and 'depthwise' means depthwise convolution.

Methods	Inception	ASPP	Ours
Parameters	26M	14M	9M
Flops	104G	56G	20G

Table 3. Parameters and flops of different methods.

Context from different stages. We also explore how context from different stages (to generate context-aware filters) influence the context-aware filters and the final segmentation results. Feature maps from different stages of ResNet50 are adopted to generate context-aware filters. It shows in Table 4 that the higher layer feature maps are utilized, the better performance is achieved. Indeed, the feature maps of higher layers usually contain higher level semantics and richer contents which can be intrinsically inherited by context-aware filters. Therefore, high-level context features are more effective for generating context-aware filters.

Different stage	mIoU(%)
Stage5	**78.21**
Stage4	75.85
Stage3	73.27
Stage2	73.13

Table 4. Comparison of context-aware filters generated by context from different stages of ResNet.

Different backbones. With the kernel sizes {1,3,5,7} of context-aware filters in our DMNet, we show the mIoU(%) of DMNet with different backbones, i.e. ResNet50 and ResNet101, in Table 5. It is obvious that the mIoU of our DMNet can be improved by using deeper backbones. To further show the significance of our multiple DCMs, we remove these DCMs from the DMNet and get the FCN. The mIoU of FCN is shown in the second column of Table 5. Without DCMs, the performance decreases by 8.38% and 3.58% for ResNet50 and ResNet101 respectively, which verifies the effectiveness of our multi-scale DCMs. Moreover, the performance of FCN decreases by 6.41% when its backbone of ResNet101 is replaced by ResNet50 (76.24% v.s. 69.83%), which means FCN is more dependent on the stronger backbone to achieve good performance. For DCMs, its performance does not change much with different backbone.

Backbone	**FCN**	**DMNet**
ResNet50	69.83	78.21
ResNet101	76.24	80.82

Table 5. Influence of different backbones. Note that we get the FCN by removing all DCMs from DMNet.

Training and evaluation strategies. As shown in Table 6, we further explore the influence of different training and evaluation strategies. Table 6 shows that introducing deep supervision to ResNet101 can improve the performance. We argue that this is because deep supervision optimizes the learning process. Moreover, horizontally flipping or scaling the image for evaluation also contributes to the improvement of mIoU. Fine-tuning the trained model with original training set can also improve the result. The final result on PASCAL VOC 2012 validation set without MS COCO pre-trained is 82.82% mIoU .

Backbone	**DS**	**Flip**	**MS**	**FT**	**mIoU%**
ResNet101					80.87
ResNet101	✓				81.08
ResNet101	✓	✓			81.50
ResNet101	✓	✓	✓		82.15
ResNet101	✓	✓	✓	✓	**82.82**

Table 6. Influence of different strategies in training and evaluation. The kernel sizes of context-aware filters in different branches are {1,3,5,7}. DS: deep supervision [51], Flip: horizontally flipping input image for evaluation, MS: multi-scale evaluation. FT: fine tune the trained model on PASCAL VOC 2012 original training set.The results are evaluated on the validation set of PASCAL VOC 2012 dataset.

Comparison with state-of-the-arts. To further demonstrate the effectiveness of our DMNet, we compare it with the state-of-the-art methods on the test set of PASCAL VOC 2012. For evaluation, we set kernel sizes to {1,3,5,7} of context-aware filters for the four branches of DMNet and adopt the deep supervision, flip, and multi-scale strategies. DMNet takes ResNet101 pre-trained on ImageNet as its backbone. We first train DMNet on the augmented training set, and then fine-tune on original training and validation set. From Table 7, we can observe that our DMNet outperforms other methods on most categories of PASCAL VOC 2012. Especially for the objects which are relatively small, e.g. bike and motorbike, our DMNet can capture more details of these objects partly because context-aware filters densely sample over the feature maps. Furthermore, our DMNet achieves the state-of-the-art performance without MS COCO pre-trained, i.e. 84.4% mIoU, since it can capture more details and multi-scale semantics. With MS COCO pretrained, our proposed method also achieves the best performance of 87.06% mIoU among the methods based on backbone ResNet101.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mIoU%
FCN [32]	76.8	34.2	68.9	49.4	60.3	75.3	74.7	77.6	21.4	62.5	46.8	71.8	63.9	76.5	73.9	45.2	72.4	37.4	70.9	55.1	62.2
DeepLabv2 [5]	84.4	54.5	81.5	63.6	65.9	85.1	79.1	83.4	30.7	74.1	59.8	79	76.1	83.2	80.8	59.7	82.2	50.4	73.1	63.7	71.6
CRF-RNN [53]	87.5	39.0	79.7	64.2	68.3	87.6	80.8	84.4	30.4	78.2	60.4	80.5	77.8	83.1	80.6	59.5	82.8	47.8	78.3	67.1	72.0
DeconvNet [34]	89.9	39.3	79.7	63.9	68.2	87.4	81.2	86.1	28.5	77.0	62.0	79.0	80.3	83.6	80.2	58.8	83.4	54.3	80.7	65.0	72.5
DPN [31]	87.7	59.4	78.4	64.9	70.3	89.3	83.5	86.1	31.7	79.9	62.6	81.9	80.0	83.5	82.3	60.5	83.2	53.4	77.9	65.0	74.1
Piecewise [27]	90.6	37.6	80.0	67.8	74.4	92	85.2	86.2	39.1	81.2	58.9	83.8	83.9	84.3	84.8	62.1	83.2	58.2	80.8	72.3	75.3
ResNet38 [43]	94.4	72.9	94.9	68.8	78.4	90.6	90.0	92.1	**40.1**	90.4	71.7	89.9	93.7	91.0	89.1	71.3	90.7	61.3	**87.7**	**78.1**	82.5
PSPNet [51]	91.8	71.9	94.7	71.2	75.8	95.2	89.9	95.9	39.3	90.7	71.7	90.5	94.5	88.8	89.6	72.8	89.6	**64.0**	85.1	76.3	82.6
EncNet [49]	94.1	69.2	**96.3**	**76.7**	**86.2**	96.3	90.7	94.2	38.8	90.7	73.3	90.0	92.5	88.8	87.9	68.7	92.6	59.0	86.4	73.4	82.9
Ours	**96.1**	**77.3**	94.1	72.8	78.1	**97.1**	**92.7**	**96.4**	39.8	**91.4**	**75.5**	**92.7**	**95.8**	**91.0**	**90.3**	**76.6**	**94.1**	62.1	85.5	77.6	**84.4**

Table 7. Results of each categories on PASCAL VOC 2012 test set. Our DMNet gets 84.4% without MS COCO dataset pre-trained.

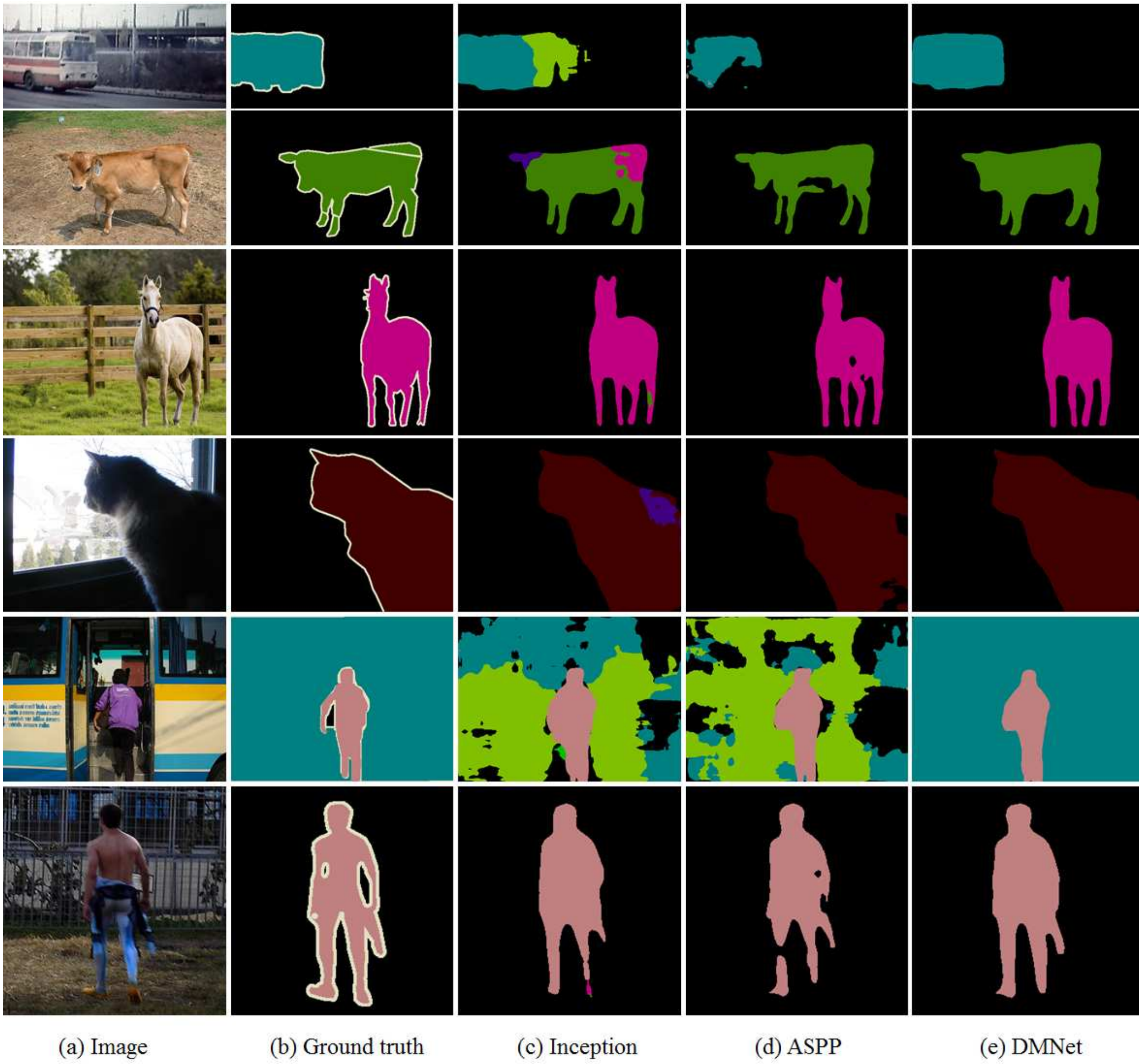

Figure 3. Visualization of segmentation results on PASCAL VOC 2012 with different multi-scale feature leaning architectures.

4.3. Pascal-Context

Pascal-Context dataset [33] provides additional annotations of the whole scene for PASCAL VOC 2010 [14]. It includes 4,998 images for training and 5,105 images for testing. We adopt the same training and test protocol as [49, 26] and compare different methods in Table 8. Several observations can be obtained from Table 8. Firstly, our DMNet shows better performance than EncNet [49] and DANet [16] whose backbone model is the same as ours. Secondly, our DMNet even surpass these methods which either use

Method	Backbone	mIoU%
FCN-8S [32]		37.8
CRF-RNN [53]		39.3
ParseNet [30]		40.4
BoxSup [8]		40.5
HO_CRF [1]		41.3
Piecewise [27]		43.3
VeryDeep [42]		44.5
DeepLab-v2 [5]	ResNet101-COCO	45.7
RefineNet [26]	ResNet152	47.3
MSCI [25]	ResNet152	50.3
EncNet [49]	ResNet101	51.7
DANet [16]	ResNet101	52.6
Ours	ResNet101	**54.4**

Table 8. Segmentation results of state-of-the-art methods on PASCAL-Context dataset.

deeper backbone model [26, 25] or utilize additional MS COCO dataset to pre-train their model [5]. The superior result (54.4%) over other state-of-the-art methods again manifests the effectiveness of our DMNet.

4.4. ADE20K

Furthermore, we evaluate our DMNet on a more challenging dataset, ADE20K [54], to show its effectiveness. There are 20K training samples, 2K validation and 3K test images in ADE20K dataset, totally 150 classes with dense labels. This dataset is more challenging because the scene in this dataset is more diverse and complex. Table 9 summarizes the mIoU of several state-of-the-art methods. Our DMNet significantly surpasses these methods with deeper backbone models, e.g. RefineNet [26] and PSPNet [51]. With the same backbone of ResNet101, our DMNet also outperforms other state-of-the-art methods, which proves that DMNet is an effective and efficient method for semantic segmentation.

Method	Backbone	mIoU%
FCN [32]		29.39
SegNet [2]		21.64
DilatedNet [47]		32.31
CascadeNet [54]		34.90
RefineNet [26]	ResNet152	40.7
PSPNet [51]	ResNet101	43.29
PSPNet [51]	ResNet269	44.94
EncNet [49]	ResNet101	44.65
SAC [50]	ResNet101	44.30
PSANet [52]	ResNet101	43.77
UperNet [44]	ResNet101	42.66
DSSPN [24]	ResNet101	43.68
APCNet [20]	ResNet101	45.38
Ours	ResNet101	**45.50**

Table 9. Segmentation results of state-of-the-art methods on ADE20K validation set.

4.5. Visualization

To better demonstrate the effectiveness of our DMNet, we visualize the segmentation results of different multi-scale feature learning methods, including Inception, ASPP, and our DMNet in Figure 3. The detailed settings of Inception and DMNet are just the same as 1st row and 3rd row with kernel size of {1,3,5,7} in Table 1. While for the ASPP, we adopt the network with dilation rate of {6,12,24,36} which achieves 77.6% mIoU in Table 2. From Figure 3, we can see that the Inception may segment the whole object with large size into several different categories, e.g. the bus in the 1st row, the leg of horse (3rd row) and human (last row) are classified into other objects. The inconsistency of the segmentation results probably comes from its small receptive field. For ASPP, it ignores the details of a certain object and leads to the 'hole' in the segmentation result. For example, ASPP mis-classifies the belly of the cow (2nd row) and horse (3rd row), and the back of the cat into background class. This is because the atrous convolution of ASPP convolves in a sparse sampling method and ASPP itself is sensitive to input image size. Different from Inception and ASPP, our DMNet can capture more details and utilize high-level semantics to achieve consistent segmentation results.

5. Conclusion

In this paper, we introduce a novel DMNet to extract multi-scale features for semantic segmentation. The DMNet incorporates multiple Dynamic Convolutional Modules (DCMs) which exploit of context-aware filters to handle the scale variations of objects. The context-aware filters are dynamically generated from input image features and embedded with high-level semantics, which makes them capable to capture more details. Extensive experiments on PASCAL VOC 2012, Pascal-Context and ADE20K show that our DMNet can not only capture more details but also adapt to objects of different scales. The state-of-the-art performance on these datasets further illustrates the effectiveness of our DMNet.

Acknowledgements. This work is partially supported by National Natural Science Foundation of China (U1813218U1613211), Shenzhen Research Program (JCYJ20170818164704758, CXB201104220032A, JSGG20180507182100698), the Joint Lab of CAS-HK.

References

[1] Anurag Arnab, Sadeep Jayasumana, Shuai Zheng, and Philip HS Torr. Higher order conditional random fields in deep neural networks. In *European Conference on Computer Vision*, pages 524–540. Springer, 2016. 8

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*, 2015. 2, 8

[3] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010. 5

[4] Dongdong Chen, Lu Yuan, Jing Liao, Nenghai Yu, and Gang Hua. Stylebank: An explicit representation for neural image style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1897–1906, 2017. 3

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018. 1, 2, 3, 4, 5, 7, 8

[6] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 1, 2, 4

[7] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *arXiv preprint arXiv:1802.02611*, 2018. 2, 4, 5

[8] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1635–1643, 2015. 8

[9] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 3

[10] Henghui Ding, Xudong Jiang, Ai Qun Liu, Nadia Magnenat Thalmann, and Gang Wang. Boundary-aware feature propagation for scene segmentation. In *Proceedings of the International Conference on Computer Vision*, 2019. 5

[11] Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Semantic correlation promoted shape-variant context for segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8885–8894, 2019. 5

[12] Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Context contrasted feature and gated multi-scale aggregation for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2393–2402, 2018. 5

[13] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 4, 5

[14] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. 7

[15] Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. Learning hierarchical features for scene labeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1915–1929, 2013. 2

[16] Jun Fu, Jing Liu, Haijie Tian, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. *arXiv preprint arXiv:1809.02983*, 2018. 7, 8

[17] Jun Fu, Jing Liu, Yuhang Wang, Yong Li, Yongjun Bao, Jinhui Tang, and Hanqing Lu. Adaptive context network for scene parsing. In *Proceedings of the IEEE international conference on computer vision*, 2019. 5

[18] Chunrui Han, Shiguang Shan, Meina Kan, Shuzhe Wu, and Xilin Chen. Face recognition with contrastive convolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 118–134, 2018. 3

[19] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 447–456, 2015. 5

[20] Junjun He, Zhongying Deng, Lei Zhou, Yali Wang, and Yu Qiao. Adaptive pyramid context network for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7519–7528, 2019. 2, 5, 8

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[22] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *Computer Science*, 2015. 5

[23] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 667–675. Curran Associates, Inc., 2016. 3

[24] Xiaodan Liang, Hongfei Zhou, and Eric Xing. Dynamic-structured semantic propagation network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 752–761, 2018. 8

[25] Di Lin, Yuanfeng Ji, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Multi-scale context intertwining for semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 603–619, 2018. 8

[26] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017. 7, 8

[27] Guosheng Lin, Chunhua Shen, Anton Van Den Hengel, and Ian Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3194–3203, 2016. 2, 7, 8

[28] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid

networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017. 2

[29] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 2

[30] Wei Liu, Andrew Rabinovich, and Alexander C Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015. 5, 8

[31] Ziwei Liu, Xiaoxiao Li, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Semantic image segmentation via deep parsing network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1377–1385, 2015. 7

[32] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1, 7, 8

[33] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2014. 4, 5, 7

[34] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015. 7

[35] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 5

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1, 2

[37] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 4

[38] Gabriel Schwartz and Ko Nishino. Material recognition from local appearance in global context. *arXiv preprint arXiv:1611.09394*, 2016. 5

[39] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 1, 3, 4

[40] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell. Understanding convolution for semantic segmentation. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1451–1460. IEEE, 2018. 1

[41] Jialin Wu, Dai Li, Yu Yang, Chandrajit Bajaj, and Xiangyang Ji. Dynamic filtering with large sampling field for convnets. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 185–200, 2018. 3

[42] Zifeng Wu, Chunhua Shen, and Anton van den Hengel. Bridging category-level and instance-level semantic image segmentation. *arXiv preprint arXiv:1605.06885*, 2016. 8

[43] Zifeng Wu, Chunhua Shen, and Anton van den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *arXiv preprint arXiv:1611.10080*, 2016. 7

[44] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. *arXiv preprint arXiv:1807.10221*, 2018. 8

[45] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3684–3692, 2018. 1

[46] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Learning a discriminative feature network for semantic segmentation. *arXiv preprint arXiv:1804.09337*, 2018. 5

[47] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 8

[48] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 472–480, 2017. 4

[49] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7151–7160, 2018. 4, 5, 7, 8

[50] Rui Zhang, Sheng Tang, Yongdong Zhang, Jintao Li, and Shuicheng Yan. Scale-adaptive convolutions for scene parsing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2031–2039, 2017. 2, 3, 8

[51] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 1, 2, 3, 4, 5, 6, 7, 8

[52] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 267–283, 2018. 1, 8

[53] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1529–1537, 2015. 7, 8

[54] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 4, 5, 8

[55] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 9308–9316, 2019. 3