Dynamic Sampling Network for Semantic Segmentation

Bin Fu,[1,2] **Junjun He,**[1,2] **Zhengfu Zhang,**[1,3] **Yu Qiao***[1,2]
[1]ShenZhen Key Lab of Computer Vision and Pattern Recognition,
SIAT-SenseTime Joint Lab, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[2]SIAT Branch, Shenzhen Institute of Artificial Intelligence and Robotics for Society
[3]University of Chinese Academy of Science
{bin.fu, jj.he, zf.zhang, yu.qiao}@siat.ac.cn

Abstract

Sampling is a basic operation of modern convolutional neural networks (CNN) since down-sampling operators are employed to enlarge the receptive field while up-sampling operators are adopted to increase resolution. Most existing deep segmentation networks employ regular grid sampling operators, which can be suboptimal for semantic segmentation task due to large shape and scale variance. To address this problem, this paper proposes a Context Guided Dynamic Sampling (CGDS) module to obtain an effective representation with rich shape and scale information by adaptively sampling useful segmentation information in spatial space. Moreover, we utilize the multi-scale contextual representations to guide the sampling process. Therefore, our CGDS can adaptively capture shape and scale information according to not only the input feature map but also the multi-scale semantic context. CGDS provides a plug-and-play module which can be easily incorporated in deep segmentation networks. We incorporate our proposed CGDS module into Dynamic Sampling Network (DSNet) and perform extensive experiments on segmentation datasets. Experimental results show that our CGDS significantly improves semantic segmentation performance and achieves state-of-the-art performance on PASCAL VOC 2012 and ADE20K datasets. Our model achieves 85.2% mIOU on PASCAL VOC 2012 test set without MS COCO dataset pre-trained and 46.4% on ADE20K validation set. The codes will become publicly available after publication.

Introduction

Semantic segmentation is a fundamental and challenging task in computer vision, which aims at assigning one of predefined categories to each pixel in an image. It is a central task for various applications such as autonomous driving, image generation and robot sensing. In recent years, the performance has been significantly improved since various Deep Convolutional Neural Networks (DCNN) based methods have been developed such as Fully Convolutional Network (FCN) (Long, Shelhamer, and Darrell 2015), DeepLab (Chen et al. 2018a; 2017; 2018b) and PSPNet (Zhao et al. 2017). The standard philosophy of these models is to modify several successful classification networks (Krizhevsky, Sutskever, and Hinton 2012; He et al. 2016; Szegedy et al. 2017) pre-trained on ImageNet dataset (Russakovsky et al. 2015) to produce the segmentation map by replacing fully connected layers with convolutional layers and adapting several up-sampling layers to gradually recover original resolution of the input image. Although this straight-forward approach has obtained impressive performance on semantic segmentation task, the inherent difference between classification and segmentation tasks limits further improvement of segmentation performance. Since the holistic representation with large receptive field is essential for classification problem, the classification network employs convolutional and down-sampling layers to extract global information into a feature vector and then pass it to a classification sub-network to estimate class label. However, semantic segmentation task requires the model to assign category labels for every pixel in input image, thus both global information and local details are important for this task. Following this guidance, several approaches have been adopted to enlarge context field while preserve high resolution. For example, atron convolution (Chen et al. 2018a) introduces 'dilation' into convolution operator which can enlarge the receptive field without extra parameters. Although this approach can effectively enlarge receptive fields, the inherent limitation of CNN ar-

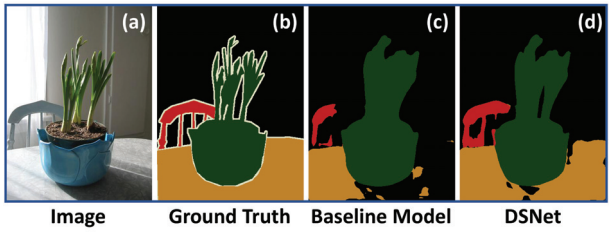

Figure 1: The shape and scale variance problem in semantic segmentation task. In above image, the shape and scale of plant, chair and table are different. The small convolution kernel can produce subtle segmentation boundaries for plant and a fragmentary map for table. While a large convolution kernel will give an unbroken map for table and a coarse map for plant.

*Corresponding author

chitecture hinders the further development of segmentation performance due to the large shape and scale variance of different objects.

The performance of DCNN heavily depends on the network architectures and their corresponding parameters. In most previous works, the network architectures are predefined and the parameters are optimized through back propagation (BP) algorithm with training data. In other words, the connections between different neurons are fixed given particular network architecture and only the corresponding parameters are updated by fitting the network to specific task. However, this can be a suboptimal solution for segmentation task. As shown in Figure 1, the dense connections in a small region are beneficial for segmenting plant and chair with subtle boundary, while large-scale sparse connections are helpful for segmenting table with larger context. Therefore, the dynamic connections are preferable to adaptively incorporate useful information into segmentation feature maps. Inspired by this observation, this paper proposes Dynamic Sampling Convolution (DSC) by re-interpreting convolution operator from a dynamic sampling perspective. Then we propose a Context Guided Dynamic Sampling (CGDS) module to obtain an effective representation with rich shape and scale information by adaptively sampling useful segmentation information in spatial space. Since multi-scale contextual information is significant for segmentation task, we employ the high level feature as a prior to guide our sampling process. Finally, we incorporate our CGDS module into Dynamic Sampling Network (DSNet) and perform extensive experiments on PASCAL VOC 2012 (Everingham et al. 2010), ADE20K (Zhou et al. 2017) and PASCAL Context (Mottaghi et al. 2014) datasets. Experimental results demonstrate the effectiveness of our proposed methods. The main contributions of this paper are as follows:

- We introduce Context Guided Dynamic Sampling module to obtain an effective representation with rich shape and scale information by adaptively sampling segmentation information in spatial space. Moreover, we employ high level semantic information as a prior to guide the dynamic sampling process.
- CGDS provides a plug-and-play module which can be easily incorporated into any segmentation networks and be trained in an end-to-end way.
- Our DSNet achieves state-of-the-art performance on PASCAL VOC 2012 and ADE20K datasets. For PASCAL VOC 2012 test set, our DSNet achieves new record 85.2% without MS COCO dataset (Lin et al. 2014) pre-trained. Moreover, our model achieves 46.4% on ADE20K validation set.

Related Work

Multi-scale Contextual Information

Convolutional Neural Network (CNN) based methods have made great progress on semantic segmentation task. Fully Convolutional Network (FCN) (Long, Shelhamer, and Darrell 2015) first converting classification network to generate segmentation map by replacing fully connected layers with convolutional layers. People witness the context information is important for segmentation task and thus various models are developed to enhance multi-scale contextual information for segmentation task. RefineNet (Lin et al. 2017) collect context information from all earlier stages by employing a multi-path refinement network. Several methods employ a probability model, conditional random field (CRF) (Krähenbühl and Koltun 2011), to output a structure predication for each pixels by using dense connected structure to capture long range dependencies such as (Chen et al. 2018a; Zheng et al. 2015). The DeepLab (Chen et al. 2018a; 2017; 2018b) adopt atrous spatial pyramid pooling (ASPP) module to enhance multi-scale contextual information by employing different dilated convolutions in a parallel fashion. Since the global average pooling operation will bring a significant enhancement for context information, PSPNet (Zhao et al. 2017) develops a pyramid pooling module to collect useful global information. EncNet (Zhang et al. 2018) employs a encoding layer to learn an inherent dictionary of the semantic context information. CFNet (Zhang et al. 2019) introduces a co-occurrent feature model to capture co-occurrent contextual information in a given image. APCNet (He et al. 2019) adaptively constructs multi-scale contextual representations under the guidance of the global information. OCNet (Yuan and Wang 2018) and DANet (Fu et al. 2018) adaptively collect local features under the guidance of long range dependencies between different positions. Moreover, DMNet (He, Deng, and Qiao 2019) extract multi-scale context information by utilizing dynamic filter (Jia et al. 2016).

Spatial Adaptive for Shape and Scale Variations

The basic operations such as convolution, pooling and bilinear interpolation operations employ the regular grid sampling approach on input feature map. This brings a serious problem in classification and segmentation tasks since there are large shape and scale variations of different objects. To address this issue, several works have been developed along two different directions. The first direction employs spatial transformation to warp the feature map into the same scale. This approach is first put forward by Spatial Transformer Networks (Jaderberg et al. 2015) which uses a sub-network to learning a set of global parameters for spatial transformation. Although this scaling method achieves appealing improvement in various tasks, it is inherent inefficient to collect complex boundary information of different objects. The second direction solves shape and scale problem by designing novel convolution operations. The active convolution (Jeon and Kim 2017) and deformable convolution (Dai et al. 2017) generate a set of offsets by a convolutional layer and use them to generate sampling points for convolution operation. Moreover, several convolution methods such as Scale-Adaptive Convolutions (Zhang et al. 2017) and Dynamic Gaussian Filter (Shelhamer, Wang, and Darrell 2019) have been developed with different sampling philosophies.

Different from above methods, we develop a new sampling method for semantic segmentation task which can efficiently collect local information under the guidance of multi-scale contextual information.

Method

In this section, we will present the proposed method in details. We firstly formulate dynamic sampling convolution and then introduce Context Guided Dynamic Sampling (CGDS) module for semantic segmentation task. Finally, we incorporate our CGDS module into Dynamic Sampling Network (DSNet).

Dynamic Sampling Convolution

The convolution operation (LeCun et al. 1989) with a $k \times k$ kernel on feature map I can be expressed as:

$$S(i,j) = \sum_m \sum_n I(i+m, j+n)\, K(m,n), \qquad (1)$$

with $m, n \in \left[-\frac{(k-1)}{2}, \frac{(k-1)}{2}\right]$. To introduce the sampling process, we modify convolution operation into the form

$$S(i,j) = \sum_m \sum_n I(i+p_m, j+p_n)\, K(m,n), \qquad (2)$$

and the corresponding sampling process can be defined as

$$Z = \mathrm{samp}\,(X)\,, \qquad (3)$$

where the tied weights $K(p_m, p_n) = K(m,n)$ have been used to reduce parameters of kernel function K; $p_m, p_n \in Z$ are sampling points which can be generated by different sampling methods and the notation 'samp' stands for the sampling method. The regular grid sampling method is used in tradition convolution operation.

From Eq. (1), we can find that the connections of different neurons between two adjacent layers are sparse connected within a predefined $k \times k$ region. Compared with Eq. (1), the network connections in Eq. (2) are governed by the sampling process since the neurons in current layer are connected by the sampled neurons in former layer. Therefore, the network behaviours are determined by sampling process. The network will show 'static' behaviours under regular grid sampling approach while show 'dynamic' behaviours under dynamic sampling process. We regard the convolution operation in Eq. (2) as dynamic sampling convolution since different neurons are dynamically connected with each other.

In the following section, we will focus on the selection of sampling methods for dynamic sampling convolution.

Context Guided Dynamic Sampling Module

In this section, we give a detailed analysis of various sampling strategies and propose Context Guided Dynamic Sampling (CGDS) module for semantic segmentation task. The formulation in Eq. (3) indicates that two key factors have influence on sampling process.

The first factor is the input feature X to be sampled. Inspired by attention mechanism (Vaswani et al. 2017; Wang et al. 2018), we sort different sampling methods into two categories according to the input feature map, self-sampling method and general-sampling method. The input feature for sampling and for dynamic sampling convolution are the same in self-sampling method while are different in general-sampling method. By this definition, the deformable convolution (Dai et al. 2017; Zhu et al. 2019) and active convolution (Jeon and Kim 2017) both belong to the self-sampling method. However, the self-sampling method has several drawbacks for classification and segmentation tasks, since these methods only exploit cues on the current feature map. It does not take full advantage of useful knowledge provided in other layers. Moreover, this method introduces high order correlations into neural networks and these high order correlations may make the optimization procedure unstable since the gradient vanishing or explosion will become a serious problem in training procedure. These disadvantages motivate us to explore more powerful and helpful features for sampling process. Unlike classification task, the low level information and high level information are both useful for segmentation task. The low level features contain rich shape, color, texture and boundary information while high level features contain semantic category and multi-scale contextual information. Since the semantic category and large-scale contextual information can provide useful and robust cues to guide dynamic sampling convolution of low level feature maps, our CGDS module employs high level context feature to dynamically generate sampling positions and sample features on low level feature map.

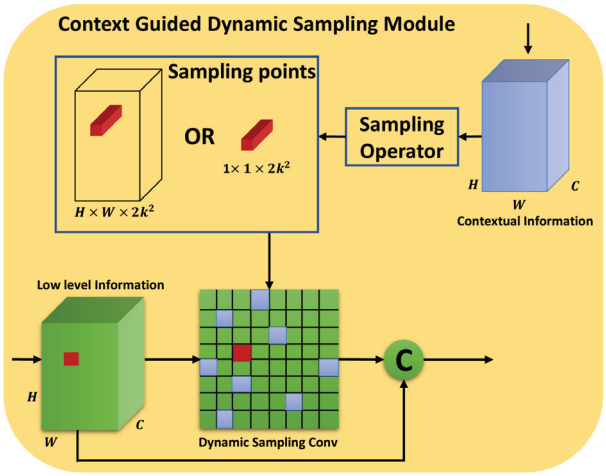

Figure 2: Architecture of Context Guided Dynamic Sampling module. The notation 'C' represents concatenate operator and k is kernel size of dynamic sampling convolution. This module employs a pair of feature maps as the input feature. It employs multi-scale contextual information to generate sampling points for low level feature map. The dynamic sampling convolution is adopt to collect context information on low level feature map.

The second factor is how to estimate sampling positions for input feature maps. Two different approaches can be employed as the sampling operation for our CGDS module. The first approach employs a convolutional layer to estimate sampling positions while the second approach uses a subnetwork to extract it. In our sampling model, we choose the first approach to generate sampling points since the complex structure will make the framework difficult to be optimized.

In summary, as shown in Figure 2, our CGDS module

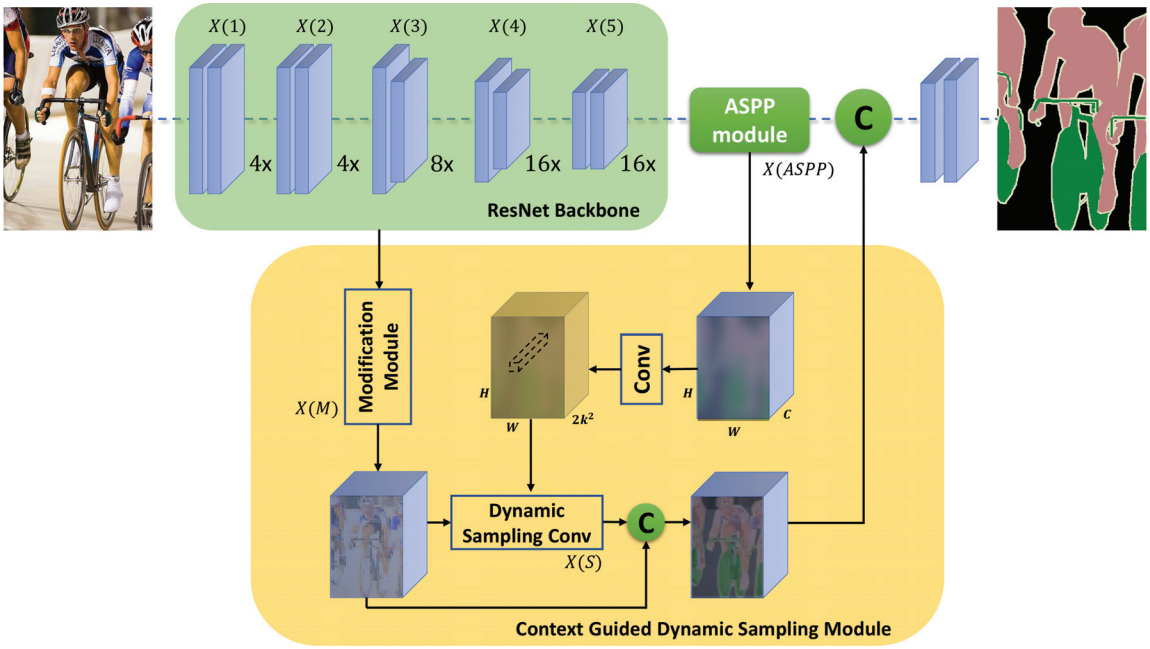

Figure 3: Architecture of our Dynamic Sampling Network. The notation 'C' represents concatenate operator and k is kernel size of dynamic sampling convolution. We employs ImageNet pre-trained ResNet as our backbone. The low level feature maps $X(i)$ are extracted from the ith stage of ResNet. The modification module is employed to adaptively adjust low level feature to reduce the difference with high level contextual feature. We employ ASPP module to further extract multi-scale contextual information and pass the resulting feature map to our Context Guided Dynamic Sampling module.

employs a low level/high level feature map pair as the input. The sampling operation determines sampling positions according to multi-scale contextual high level feature map and then generates a set of sampling points. These sampling points are employed to extract the corresponding information on low level feature map.

Dynamic Sampling Network

In this section, we incorporate the proposed CGDS module into semantic segmentation framework.

The architecture of our semantic segmentation framework is shown in Figure 3. Let $I \in R^{H\times W\times 3}$ be the input image for semantic segmentation. We employ ResNet-50 or ResNet-101 (He et al. 2016) as our backbone, and change the output stride from stride 32 to stride 16 following (Long, Shelhamer, and Darrell 2015; Chen et al. 2018a). A series of feature maps $X(i)$ are extracted from this backbone where the notation $X(i)$ denotes the feature map generated at the ith stage of ResNet. In order to obtain high level feature with rich semantic information, the Atrous Spatial Pyramid Pooling (ASPP) module (Chen et al. 2018b) has been employed to incorporate multi-scale and global contextual information into the resulting feature map $X(ASPP)$. We employ this feature map as the high level guidance of our CGDS module. Meanwhile, the feature maps $X(i)$ at early stages of ResNet contain rich shape, boundary and texture information. Adaptively sampling on these feature maps under the semantic prior will collect useful shape and texture information into the final segmentation feature map. However, directly sampling on low level feature maps $X(i)$ is problematic since there are huge differences between high level information and low level information. To settle this issue, a modified low level feature map $X(M)$ is generated by employing a convolutional layer as the feature modification module to adaptively adjust low level feature map. Therefore, our CGDS module uses high level information $X(ASPP)$ as the semantic prior to generate sampling points Z and employs dynamic sampling convolution in Eq.(2) to collect shape and texture cues according to the generated sampling points. Specifically, we first up-sample $X(ASPP)$ to the same spatial size with $X(M)$ (if necessary) and then employ a convolution layer to generate sampling points from $X(ASPP)$ for each position in $X(M)$. For a specific position, the sampling points are expressed by a $2 \times k \times k$ vector which includes $k \times k$ sampling points with x and y coordinate. Then we perform Dynamic Sampling Convolution on feature map $X(M)$. Since coordinates of sampling points maybe not integer, we use bilinear interpolation to obtain feature vector for each sampling points. Finally, we obtain segmentation feature maps $X(F)$ by concatenating the modified feature $X(M)$, the sampled feature $X(S)$ with $X(ASPP)$, and then pass them to pixel-wise classification sub-network.

Relation to Other Approaches

In this subsection, we offer a comparison between our Dynamic Sampling Network and other similar approaches. Dynamic Filter Network (Jia et al. 2016) employs a sub-

network to dynamically generate parameters for convolution kernel while our approach generates sampling points. Deformable Convolutional Network (Dai et al. 2017; Zhu et al. 2019) employs self-sampling method and incorporates deformable convolution into segmentation framework by replacing traditional convolution operation in last stage of ResNet backbone. Although this approach achieves great improvement in classification task, it cannot obtain appealing state-of-the-art performance in semantic segmentation task. Our DSNet considers the characteristic of segmentation task and designs CGDS module to solve this problem. Moreover, our model achieves state-of-the-art performance on segmentation task.

Experiments

In this section, we provide extensive experiments to demonstrate the effectiveness of our framework on three challenging semantic segmentation datasets, including PASCAL VOC 2012 (Everingham et al. 2010), ADE20K (Zhou et al. 2017) and PASCAL Context (Mottaghi et al. 2014). Experimental results demonstrate that our proposed model achieves state-of-the-art performance on PASCAL VOC 2012 and ADE20K datasets. In the following, we will introduce implementation details of our model and then perform several ablation experiments on PASCAL VOC 2012 dataset. Finally, the experimental results on three datasets will be given.

Implementation Details

We use ImageNet (Russakovsky et al. 2015) pre-trained ResNet (He et al. 2016) as our backbone. Following (Long, Shelhamer, and Darrell 2015; Chen et al. 2018a), we modify ResNet structure to a FCN-liked structure with atrous convolutions. Specifically, the stride of last stages is removed and the atrous convolution is employed to enlarge receptive field by setting dilation as 2, thus the output size of ResNet has been enlarged from $1/32$ to $1/16$ compared with the origin image. The ASPP (Chen et al. 2018b) module has been employed to extract multi-scale contextual information on top of backbone. We employ this context feature as high level feature while the feature of stage three as low level feature for our proposed CGDS module. All experiments are performed on the Pytorch (Paszke et al. 2017) platform. During the training process, we employ the poly learning rate policy $\mathrm{lr} = \mathrm{initial\_lr} \times \left(1 - \frac{\mathrm{iter}}{\mathrm{total\_iter}}\right)^{0.9}$ (Chen et al. 2017; Zhang et al. 2018) for the ResNet backbone of our DSNet where the initial learning rate is 0.01 for PASCAL VOC 2012 and ADE20K datasets, 0.005 for PASCAL Context dataset. While the learning rate for the remaining parts of our DSNet have been setted as $0.1 \times \mathrm{lr}$. The network is optimized by stochastic gradient descent (SGD) (Bottou 2010) for 80 epochs on PASCAL VOC 2012 and PASCAL Context datasets, for 150 epochs on ADE20K dataset with momentum 0.9 and weight decay 0.0001. We set the batch size as 32 for PASCAL VOC 2012 and PASCAL Context datasets, 24 for ADE20K dataset. The Cross-GPU Batch Normalization developed by Zhang (Zhang et al. 2018) has been implemented in our framework. At training stage, random horizontally flipping and image scaling in range 0.5 to 2.0 are used as data augmentation, and then a fixed rectangle region is random cropped form it. The crop size of input images is chosen as 512×512 for PASCAL VOC 2012 and PASCAL Context datasets, 576×576 for ADE20K dataset. When evaluating our model, horizontally flipping and multi-scale resizing are used to further boost segmentation performance.

LLF	LLF OS	HLF	HLF OS	mIoU%
$X(3)$	8	$X(5)$	16	77.4
$X(3)$	8	$X(ASPP)$	16	**79.0**
$X(4)$	16	$X(ASPP)$	16	76.7
$X(5)$	16	$X(ASPP)$	16	78.1

Table 1: Comparison between different selections of low level/high level pairs for our Context Guided Dynamic Sampling Module on PASCAL VOC 2012 validation set with ImageNet pretrained **ResNet-50** backbone. We denote the feature map at ith stage of ResNet backbone as $X(i)$. **LLF** represents low level feature map while **HLF** represents high level feature map for our CGDS Module. **OS** represents output stride of the corresponding feature map compared with input image.

PASCAL VOC 2012

The PASCAL VOC 2012 is a widely used benchmark (Everingham et al. 2010) for semantic segmentation, including 20 foreground object classes and one background class. The original dataset contains $1,464$ images for training, $1,449$ for validation, and $1,456$ for test. The training set has been augmented to $10,582$ images by extra annotations from (Hariharan et al. 2015) and thus we use this augmented training set in our experiments. In the following, we first conduct ablation experiments on this dataset and then show state-of-the-art performance of our DSNet.

Ablation Study for Context Guided Dynamic Sampling Module We implement our CGDS module on top of the dilated backbone as shown in Figure 3. As we have discussed in last section, the input feature for sampling process and the sampling operation for generating sampling points are both important for our CGDS module. In this subsection, we conduct several ablation experiments with different settings to demonstrate the effectiveness of our proposed model.

We first perform ablation experiments on the selection of input feature map pairs for our proposed module. Specifically, we employ a convolutional layer as the sampling operation and explore the different low level/high level feature map pairs for our CGDS module. Experimental results have been shown in Table 1. From this table, we can obtain the following conclusions: (1). The multi-scale contextual information offers a useful guidance on sampling points generating process which is important for our CGDS module, since employing feature map $X(5)$ as high level feature gives poor performance on validation set. (2). Employing feature map in early stage of ResNet backbone ($X(3)$)

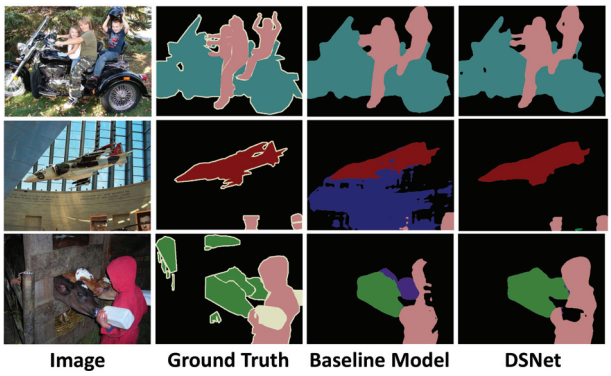

Figure 4: Visualization of segmentation results of DSNet.

Method	DS	Flip	MS	FT	Mean IoU%
Baseline	✓				80.06
DSNet	✓				80.98
DSNet	✓	✓			81.63
DSNet	✓	✓	✓		82.53
DSNet	✓	✓	✓	✓	83.74

Table 2: Comparison between different training and evaluation strategies on PASCAL VOC 2012 validation set with ImageNet pre-trained **ResNet-101** backbone. **DS** represents deep supervision (Zhao et al. 2017), **Flip** represents horizontally flipping in evaluation stage, **MS** represents multi-scale inputs during inference and **FT** represents fine tune the trained model on PASCAL VOC 2012 original training set.

as the input for CGDS module significantly improves segmentation performance. Our proposed model can enhance segmentation feature by dynamically collect shape and texture information under multi-scale contextual guidance. (3). The detailed shape and texture information is lost in high level features, therefore employing low level feature to recover object details is important for semantic segmentation task. Experimental results in Table 1. demonstrate the effectiveness of our CGDS module. In the following experiment, we take feature maps $X(3)$ and $X(ASPP)$ as our low level/high level pairs for CGDS module.

We conduct experiments on the selection of sampling operations in our CGDS Module. Apart from employing a single convolutional layer to calculate sampling points, we employ non-local structure (Wang et al. 2018) as a sub-network to generate sampling points. Experiment shows that the proposed non-local sub-network only achieves 76.6% mIOU with ResNet-50 backbone. We think the reason for this poor performance is that the sub-network structure is more difficult to optimize than a single convolutional layer. Therefore, we employ a convolutional layer as our sampling operation in the following experiment.

Comparing with Related Works

Since our work is similar with deformable convolution (Dai et al. 2017; Zhu et al. 2019) and dynamic filter (Jia et al. 2016), we conduct extensive experiments to compare our DSNet with Deformable Convolution v2 (DCv2) and Dynamic Filter (DF). We implement DCv2 and DF on top of ASPP module with ResNet50 backbone. For DCv2, we employ a convolution layer to generate offsets and masks. For DF, we employ Conv (followed with BN and ReLU), Adaptive Average Pooling and Conv layers to generate kernel weight. The performance for DCv2 and DF are 75.4% and 74.6%, respectively, which demonstrate the effectiveness of our DSNet.

Ablation Study for Training and Evaluation Strategies The performance will be further improved by employing some training and evaluation strategies. Following (Chen et al. 2018b; Zhang et al. 2018; He et al. 2019), we adopt several widely-used strategies: (1). Flip: horizontally flip input image in evaluation stage; (2). MS: average the score map from multi-scale images $\{0.75, 1.0, 1.25, 1.5, 1.75, 2.0\}$ in evaluation stage; (3). FT: fine tune the trained model on PASCAL VOC 2012 original training set.

Experimental results have been presented in Table 2 and several conclusions can be drawn from this table. The horizontally flipping is useful and it will bring 0.65% improvement in our experiment. The multi-scale input will improve performance about 0.90%, since the correct segmentation information will be enhanced by different scales. Finally, fine tune the model on the original training set boosts the result to 83.74% mIOU on validation set due to the different distributions between the original dataset and augmented training set. Moreover, it is worthy to mention that, compare to Deeplab v3+ (Chen et al. 2018b), our CGDS module brings 1.14% improvement on mIOU which is a significant achievement in semantic segmentation task.

Visualization of Segmentation Results We further analyse our DSNet by visualizing the segmentation map in Figure 4. Compared with baseline model, for large-scale objects, our DSNet can obtain a intact and unbroken segmentation map; for small object, subtle boundaries can be recovered by our DSNet. Therefore, the multi-scale contextual guidance offers a strong prior for collecting shape and texture information of different objects which can significantly remove mistake segmentation information.

Comparing with State-of-the-art In this section, we demonstrate the effectiveness of our proposed model on PASCAL VOC 2012 testing set. We adopt deep supervise (Zhao et al. 2017) as additional supervision signal and train our model on augmented training set. We combine the training set and validation set of original PASCAL VOC 2012 dataset as the trainval set, and fine tune our model on this trainval set for 80 epoches. The best model on validation set is selected as the final model. At testing stage, we horizontally flip and crop the test image into $\{0.75, 1.0, 1.25, 1.5, 1.75, 2.0\}$ scales. Average fusion on prediction map are used to obtain final result. We submit our test result to the official evaluation server. Evaluation results are shown in Table 3. Our model achieves **85.2%** mIOU without MS COCO (Lin et al. 2014) pre-trained and outperforms existing approaches with a significant improve-

Methods	mIoU	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
FCN (Long et al. 2015)	62.2	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
DeepLabv2(Chen et al. 2018a)	71.6	84.4	54.5	81.5	63.6	65.9	85.1	79.1	83.4	30.7	74.1	59.8	79.0	76.1	83.2	80.8	59.7	82.2	50.4	73.1	63.7
CRF-RNN(Zheng et al. 2015)	72.0	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
PSPNet(Zhao et al. 2017)	82.6	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
EncNet(Zhang et al. 2018)	82.9	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
APCNet(He et al. 2019)	84.2	95.8	75.8	84.5	76.0	80.6	96.9	90.0	96.0	42.0	93.7	75.4	91.6	95.0	90.5	89.3	75.8	92.8	61.9	88.9	79.6
CFNet(Zhang et al. 2019)	84.2	95.7	71.9	95.0	76.3	82.8	94.8	90.0	95.9	37.1	**92.6**	73.0	**93.4**	94.6	89.6	88.4	74.9	**95.2**	63.2	**89.7**	78.2
DMNet (He et al. 2019)	84.4	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
Ours(DSNet)	**85.2**	**96.4**	**79.5**	86.7	74.6	81.0	96.5	92.1	**96.5**	**47.6**	**92.6**	75.1	92.0	94.6	**92.4**	**90.4**	75.2	92.5	**66.8**	87.4	**82.3**

Table 3: Experimental results on PASCAL VOC 2012 testing set. DSNet outperforms all existing approaches and achieves 85.2% in Mean IoU without MS COCO (Lin et al. 2014) pre-trained.

Method	Backbone	mIoU (%)
CascadeNet(Zhou et al. 2017)		34.90
RefineNet(Lin et al. 2017)	ResNet152	40.70
PSPNet(Zhao et al. 2017)	ResNet296	44.94
EncNet(Zhang et al. 2018)	ResNet101	44.65
OCNet(Yuan and Wang 2018)	ResNet101	45.08
APCNet(He et al. 2019)	ResNet101	45.38
CFNet(Zhang et al. 2019)	ResNet101	44.89
Ours	ResNet101	**46.37**

Table 4: Experimental results on ADE20K validation set. Our DSNet achieves state-of-the-art performance on this dataset.

Method	Backbone	mIoU (%)
CRF-RNN(Zheng et al. 2015)		39.3
DeepLabv2(Chen et al. 2018a)	ResNet101	45.7
RefineNet(Lin et al. 2017)	ResNet152	47.3
EncNet(Zhang et al. 2018)	ResNet101	51.7
DANet(Fu et al. 2018)	ResNet101	52.6
APCNet(He et al. 2019)	ResNet101	**54.7**
Ours	ResNet101	54.2

Table 5: Experimental results on PASCAL Context dataset.

ment.

ADE20K

The ADE20K dataset (Zhou et al. 2017) is a challenging large-scale dataset released by ImageNet Large Scale Visual Recognition Challenge 2016 (ILSVRC2016). It contains 150 semantic classes for scene parsing, with $20,210$ images for training, $2,000$ images for validation and $3,351$ images for testing. This dataset is more challenging due to two factors. The first factor is the high diverse range of semantic categories which the segmentation boundaries in feature space are difficult to separate well. The second factor is the large image size of this dataset, since it is more difficult to achieve the balance between global information and spatial details.

We conduct experiment on ADE20K dataset, and the performance is presented on Table 4. Our DSNet achieves new state-of-the-art segmentation performance **46.4%** mIOU on this dataset which is the first record above 46% mIOU.

PASCAL Context

The PASCAL Context (Mottaghi et al. 2014) contains 60 semantic classes (59 most frequent categories and setting other categories as background) with $4,998$ images in the training set and $5,105$ images in the validation set. We implement our DSNet on this dataset and the performance is presented in Table 5. Our model achieves 54.2% mIOU which is a comparable result to state-of-the-art performance.

Conclusion

In this paper, we discuss the weakness of convolutional neural network for semantic segmentation task and propose Dynamic Sampling Network to introduce dynamic connections in segmentation framework. The texture and shape information are adaptively selected to strengthen segmentation feature maps under the guidance of multi-scale contextual information. Experimental results demonstrate that our DSNet can recover more details of segmentation boundaries and obtain precision segmentation maps of the objects with different scales. Furthermore, the effectiveness of our proposed framework is demonstrated by state-of-the-art performance on PASCAL VOC 2012 and ADE20K datasets.

Acknowledgments

This work is partially supported by the National Key Research and Development Program of China (No. 2016YFC1400704), and National Natural Science Foundation of China (61876176, U1713208), Shenzhen Basic Research Program (JCYJ20170818164704758, CXB201104220032A), the Joint Lab of CAS-HK, Shenzhen Institute of Artificial Intelligence and Robotics for Society.

References

Bottou, L. 2010. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*. Springer. 177–186.

Chen, L.-C.; Papandreou, G.; Schroff, F.; and Adam, H. 2017. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2018a. 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.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018b. Encoder-decoder with atrous separable convolution for semantic image segmentation. *arXiv preprint arXiv:1802.02611*.

Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 764–773.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision* 88(2):303–338.

Fu, J.; Liu, J.; Tian, H.; Fang, Z.; and Lu, H. 2018. Dual attention network for scene segmentation. *arXiv preprint arXiv:1809.02983*.

Hariharan, B.; Arbeláez, P.; Girshick, R.; and Malik, J. 2015. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 447–456.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, J.; Deng, Z.; Zhou, L.; Wang, Y.; and Qiao, Y. 2019. Adaptive pyramid context network for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7519–7528.

He, J.; Deng, Z.; and Qiao, Y. 2019. Dynamic multi-scale filters for semantic segmentation. In *The IEEE International Conference on Computer Vision (ICCV)*.

Jaderberg, M.; Simonyan, K.; Zisserman, A.; et al. 2015. Spatial transformer networks. In *Advances in neural information processing systems*, 2017–2025.

Jeon, Y., and Kim, J. 2017. Active convolution: Learning the shape of convolution for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4201–4209.

Jia, X.; De Brabandere, B.; Tuytelaars, T.; and Gool, L. V. 2016. Dynamic filter networks. In Lee, D. D.; Sugiyama, M.; Luxburg, U. V.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 29*. Curran Associates, Inc. 667–675.

Krähenbühl, P., and Koltun, V. 2011. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, 109–117.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

LeCun, Y.; Boser, B.; Denker, J. S.; Henderson, D.; Howard, R. E.; Hubbard, W.; and Jackel, L. D. 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation* 1(4):541–551.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Lin, G.; Milan, A.; Shen, C.; and Reid, I. 2017. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1925–1934.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Mottaghi, R.; Chen, X.; Liu, X.; Cho, N.-G.; Lee, S.-W.; Fidler, S.; Urtasun, R.; and Yuille, A. 2014. 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*, 891–898.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115(3):211–252.

Shelhamer, E.; Wang, D.; and Darrell, T. 2019. Blurring the line between structure and learning to optimize and adapt receptive fields. *arXiv preprint arXiv:1904.11487*.

Szegedy, C.; Ioffe, S.; Vanhoucke, V.; and Alemi, A. A. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7794–7803.

Yuan, Y., and Wang, J. 2018. Ocnet: Object context network for scene parsing. *arXiv preprint arXiv:1809.00916*.

Zhang, R.; Tang, S.; Zhang, Y.; Li, J.; and Yan, S. 2017. Scale-adaptive convolutions for scene parsing. In *Proceedings of the IEEE International Conference on Computer Vision*, 2031–2039.

Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; and Agrawal, A. 2018. Context encoding for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7151–7160.

Zhang, H.; Zhang, H.; Wang, C.; and Xie, J. 2019. Co-occurrent features in semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 548–557.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890.

Zheng, S.; Jayasumana, S.; Romera-Paredes, B.; Vineet, V.; Su, Z.; Du, D.; Huang, C.; and Torr, P. H. 2015. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE international conference on computer vision*, 1529–1537.

Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; and Torralba, A. 2017. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 633–641.

Zhu, X.; Hu, H.; Lin, S.; and Dai, J. 2019. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9308–9316.