

# OCNet: Object Context Network for Scene Parsing

Yuhui Yuan Jingdong Wang  
 Microsoft Research  
 {yuyua, jingdw}@microsoft.com

## Abstract

*In this paper, we address the problem of scene parsing with deep learning and focus on the context aggregation strategy for robust segmentation. Motivated by that the label of a pixel is the category of the object that the pixel belongs to, we introduce an object context pooling (OCP) scheme, which represents each pixel by exploiting the set of pixels that belong to the same object category with such a pixel, and we call the set of pixels as object context.*

*Our implementation, inspired by the self-attention approach, consists of two steps: (i) compute the similarities between each pixel and all the pixels, forming a so-called object context map for each pixel served as a surrogate for the true object context, and (ii) represent the pixel by aggregating the features of all the pixels weighted by the similarities. The resulting representation is more robust compared to existing context aggregation schemes, e.g., pyramid pooling modules (PPM) in PSPNet and atrous spatial pyramid pooling (ASPP), which do not differentiate the context pixels belonging to the same object category or not, making the reliability of contextually aggregated representations limited. We empirically demonstrate our approach and two pyramid extensions with state-of-the-art performance on three semantic segmentation benchmarks: Cityscapes, ADE20K and LIP. Code has been made available at: <https://github.com/PkuRainBow/OCNet.pytorch>.*

## 1. Introduction

Scene parsing is a fundamental topic in computer vision and is critical for various challenging tasks such as autonomous driving and virtual reality. The goal is to predict the label of each pixel, i.e., the category label of the object that the pixel belongs to.

Various techniques based on deep convolutional neural networks have been developed for scene parsing since the pioneering fully convolutional network approach [18]. There are two main paths to tackle the segmentation problem. The first path is to raise the resolution of response

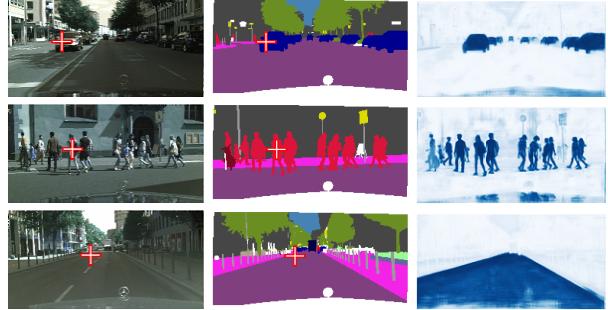


Figure 1: Illustrations of the object context maps. The first column illustrates example images sampled from the validation set of Cityscapes. Three pixels from object car, person and road are marked by  $\oplus$ . The second column illustrates ground truth segmentation maps. The third column illustrates object context maps of the three pixels. For each object context map, it can be seen that most of the weights are focused on the pixels belonging to the same category with the selected pixel.

maps for improving the spatial precision, e.g., through dilated convolutions [2, 31]. The second path is to exploit the context [2, 31, 34] for improving the labeling robustness, which our work belongs to.

Existing representative works mainly exploit the context formed from spatially nearby or sampled pixels. For instance, the pyramid pooling module in PSPNet [34] partitions the feature maps into multiple regions, and the pixels lying within each region are regarded as the context of the pixel belonging to the region. The atrous spatial pyramid pooling module (ASPP) in DeepLabv3 [3] regards spatially regularly sampled pixels at different atrous rates as the context of the center pixel. Such spatial context is a mixture of pixels that might belong to different object categories, thus the resulting representations obtained from context aggregation are limitedly reliable for label prediction.

Motivated by that the label of a pixel in an image is the category of the object that the pixel belongs to, we present a so-called object context for each pixel, which is the set of pixels that belong to the same object category with such a pixel. We propose a novel object context pooling (OCP) to aggregate the information according to the object context.

We compute a similarity map for each pixel  $p$ , where each similarity score indicates the degree that the corresponding pixel and the pixel  $p$  belongs to the same category. We call such similarity map as object context map, which serves as a surrogate of the true object context. Figure 1 shows several examples of object context map.

We exploit the object context to update the representation for each pixel. The implementation of object context pooling, inspired by the self-attention approach [14, 23], computes the weighted summation of the representations of all the pixels contained in the object context, with the weights from the object context map.

We further present two extensions: (i) pyramid object context, which performs object context pooling in each region in the spatial pyramid and follows the pyramid design introduced in PSPNet [34]. (ii) atrous spatial pyramid object context, which combines ASPP [3] and object context pooling. We demonstrate our proposed approaches by state-of-the-art performance on two challenging scene parsing datasets, Cityscapes and ADE20K, and the challenging human parsing dataset LIP.

## 2. Related Work

**Semantic Segmentation.** Semantic segmentation or scene parsing has achieved great progress with the recent works such as FCN [18], UNet [21], SegNet [1], ParseNet [16], PSPNet [34] and DeepLabv3 [3].

There exist two main challenges, (i) resolution: there exists a huge gap between the output feature map’s resolution and the input image’s resolution. (e.g., the output feature map of ResNet-101 is  $\frac{1}{8}$  or  $\frac{1}{32}$  of the input image’s size when we use dilated convolution [31] or not.) (ii) multi-scale: there exist objects of various scales, especially in the urban scene images such as Cityscapes [4]. Most of the recent works are focused on solving these two challenges.

To handle the problem of resolution, we adopt the dilated convolution within OCNet by following the same settings of PSPNet and DeepLabv3. Besides, it is important to capture information of multiple scales to alleviate the problem caused by multi-scale objects. PSPNet applies PPM (pyramid pooling module) while DeepLabv3 employs the image-level feature augmented ASPP (atrous spatial pyramid pooling). OCNet captures the multi-scale context information by employing object context pooling over regions of multiple scales.

**Context.** The context plays an important role in various computer vision tasks and it is of various forms such as global scene context, geometric context, relative location, 3D layout and so on. Context has been investigated for both object detection [5, 17] and part detection [8].

The importance of context for semantic segmentation is also verified in the recent works [16, 34, 3]. We define

the context as a set of pixels in the literature of semantic segmentation. Especially, we can partition the conventional context to two kinds: (i) nearby spatial context: the ParseNet [16] treats all the pixels over the whole image as the context, and the PSPNet [34] employs pyramid pooling over sub-regions of four pyramid scales and all the pixels within the same sub-region are treated as the context for the pixels belonging to the sub-region. (ii) sampled spatial context: the DeepLabv3 employs multiple atrous convolutions with different atrous rates to capture spatial pyramid context information and regards these spatially regularly sampled pixels as the context. Both these two kinds of context are defined over rigid rectangle regions and carry pixels belonging to various object categories.

Different from the conventional context, object context is defined as the set of pixels belonging to the same object category.

**Attention.** Attention is widely used for various tasks such as machine translation, visual question answering and video classification. The self-attention [14, 23] method calculates the context at one position as a weighted sum of all positions in a sentence. Wang *et al.* further proposed the non-local neural network [25] for vision tasks such as video classification, object detection and instance segmentation based on the self-attention method.

Our work is inspired by the self-attention approach and we mainly employ the self-attention method to learn the object context map recording the similarities between all the pixels and the associated pixel  $p$ . The concurrent DANet [6] also exploits the self-attention method for segmentation, and OCNet outperforms the DANet on the test set of Cityscapes and DANet is not evaluated on the ADE20K and LIP benchmarks.

Besides, the concurrent work PSANet [35] is also different from our method. The PSANet constructs the pixel-wise attention map based on each pixel independently while OCNet constructs the object context map by considering the pair-wise similarities among all the pixels.

## 3. Approach

Given an image  $I$ , the goal of scene parsing is to assign a label to each pixel, where the label is the category of the object the pixel belongs to, outputting a segmentation (or label) map  $L$ .

**Pipeline.** Our approach feeds the input image  $I$  to a fully convolution network (e.g., a part of a ResNet), outputting a feature map  $X$  of size  $W \times H$ , then lets the feature map  $X$  go through an object context module, yielding an updated feature map  $\bar{X}$ , next predicts the label for each pixel according to the updated feature map, and up-samples the label map for  $8 \times$  times at last. The whole structure is called OCNet,

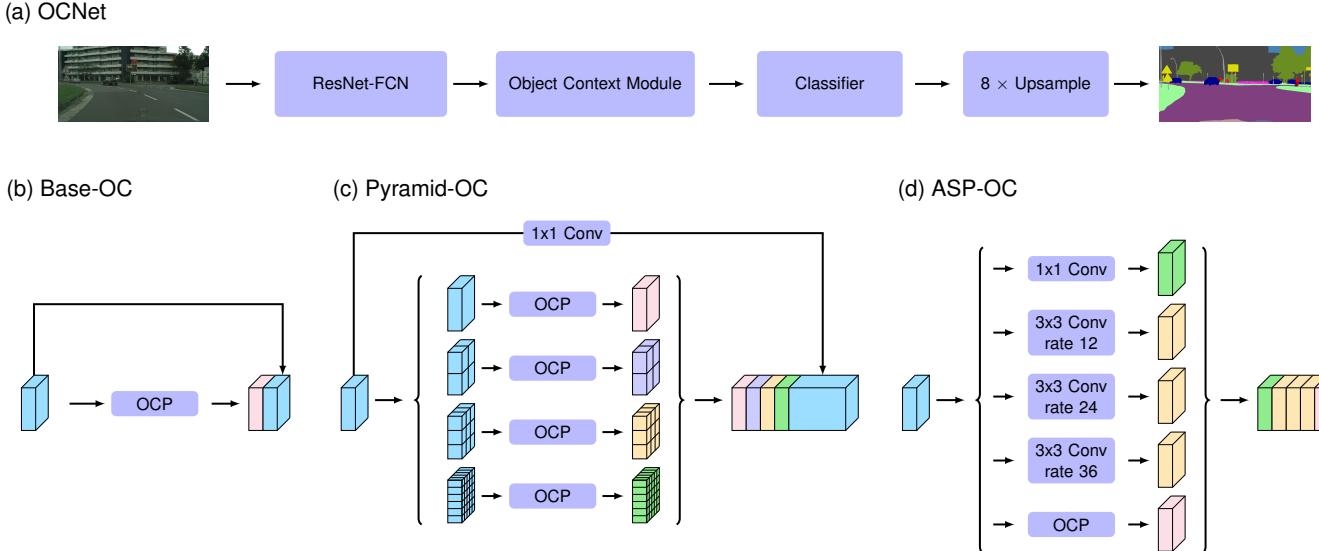


Figure 2: (a) The overall network structure of OCNet: Given an input image, we employ a fully convolution network (FCN) to extract a feature map, then employ an object context module on the feature map and output an updated feature map. Based on the updated feature map, we employ a classifier to predict the pixel-wise label map and employ bilinear method to up-sample the label map for  $8\times$  times as the final prediction. (b) Base-OC: Given an input feature map, we employ an object context pooling (OCP) on it, then we concatenate the output feature map of OCP and the input feature map as the output feature map. (c) Pyramid-OC: Given an input feature map, we employ four parallel OCPs independently. Each branch partitions the input to different pyramid scales, and the object context pooling is shared within each branch, then we concatenate the four output feature maps with a new feature map that is generated by increasing the channels of the input feature map. (d) ASP-OC: Given an input feature map, we employ an OCP and four dilated convolutions (these four branches are the same with the original ASPP), then we concatenate the five output feature maps as the output.

and our key contribution to scene parsing lies in the object context module. The pipeline is given in Figure 2 (a).

### 3.1. Object Context

The intuition of the object context is to represent a pixel by exploiting the representations of other pixels lying in the object that belongs to the same category.

The key component of object context module is the object context pooling (OCP), and the design of OCP is inspired by the self-attention approach [14, 23]. The object context pooling includes two main steps: object context estimation and object context aggregation.

**Object context pooling.** (i) *Object context estimation.* The *object context* for each pixel  $p$  is defined as a set of pixels that belong to the same object category as the pixel  $p$ . We compute an object context map, denoted in a vector form by  $\mathbf{w}_p^1$  for the pixel  $p$ , indicating the degrees that each other pixel and the pixel  $p$  belong to the same object category. The object context map is a surrogate for the true object context. The computation of object context map is given as

<sup>1</sup>We use the vector form to represent the 2D map for description convenience.

follows,

$$w_{pi} = \frac{1}{Z_p} \exp(f_q(\mathbf{x}_p)^\top f_k(\mathbf{x}_i)), \quad (1)$$

where  $\mathbf{x}_p$  and  $\mathbf{x}_i$  are the representation vectors of the pixels  $p$  and  $i$ . The normalization number  $Z_p$  is a summation of all the similarities:  $Z_p = \sum_{i=1}^N \exp(f_q(\mathbf{x}_p)^\top f_k(\mathbf{x}_i))$ , where  $N = W \times H$ .  $f_q(\cdot)$  and  $f_k(\cdot)$  are the query transform function and the key transform function.

(ii) *Object context aggregation.* We construct the object context representation of the pixel  $p$  by aggregating the representations of the pixels according to the object context map as below,

$$\mathbf{c}_p = \sum_{i=1}^N w_{pi} \phi(\mathbf{x}_i), \quad (2)$$

where  $\phi(\cdot)$  is the value transform function following the self-attention.

**Base object context.** We employ an object context pooling to aggregate the object context information according to the object context map of each pixel, and concatenate the output feature map by OCP with the input feature map as the output. We call the resulting method as *Base-OC*. More details are illustrated in Figure 2 (b).

**Pyramid object context.** We partition the image into regions using four pyramid scales:  $1 \times 1$  region,  $2 \times 2$  regions,  $3 \times 3$  regions, and  $6 \times 6$  regions, which is similar to PSPNet [34], and we update the feature maps for each scale by feeding the feature map of each region into the object context pooling separately, then we combine the four updated feature maps together. The pyramid object context module has the capability of purifying the object context map by removing spatially far but appearance similar pixels that belong to different object categories. Finally, we concatenate the multiple pyramid object context representations with the input feature map. We call the resulting method as *Pyramid-OC*. More details are illustrated in Figure 2 (c).

**Combination with ASPP.** The atrous spatial pyramid pooling (ASPP) consists of five branches: an image-level pooling branch, a  $1 \times 1$  convolution branch and three  $3 \times 3$  dilated convolution branches with dilation rates being 12, 24 and 36, respectively over the feature map with output stride of 8. We connect four among the five branches except the image-level pooling branch and our object context pooling in parallel, resulting in a method which we name as *ASP-OC*. More details are illustrated in Figure 2 (d).

### 3.2. Network Architecture

**Backbone.** We use the ResNet-101 pretrained over the ImageNet dataset as the backbone, and make some modifications by following PSPNet [34]: replace the convolutions within the last two blocks by dilated convolutions with dilation rates being 2 and 4, respectively, so that the output stride becomes 8.

**Object context module.** We construct the Base-OC module, Pyramid-OC module and ASP-OC module by employing an extra  $1 \times 1$  convolution on the output feature map of Base-OC, Pyramid-OC and ASP-OC.

The detailed architecture of Base-OC module is given as follows. Before feeding the feature map into the OCP, we employ a dimension reduction module (a  $3 \times 3$  convolution) to reduce the channels of the feature maps output from the backbone from 2048 to 512. Then we feed the updated feature map into the OCP and concatenate the output feature map of the OCP with the input feature map to the OCP. We further employ a  $1 \times 1$  convolution to decrease the channels of the concatenated feature map from 1024 to 512.

For the Pyramid-OC module, we also employ a  $3 \times 3$  convolution to reduce the channels from 2048 to 512 in advance, then we feed the dimension reduced feature map to the Pyramid-OC and employ four different pyramid partitions ( $1 \times 1$  region,  $2 \times 2$  regions,  $3 \times 3$  regions, and  $6 \times 6$  regions) on the input feature map, and we concatenate the four different output object context feature maps output by the four parallel OCPs. Each one of the four object context feature maps has 512 channels. We employ a  $1 \times 1$  convolution

to increase the channel of the input feature map from 512 to 2048 and concatenate it with all the four object context feature maps. Lastly, we employ a  $1 \times 1$  convolution on the concatenated feature map with 4096 channels and produce the final feature map with 512 channels.

For the ASP-OC module, we only employ the dimension reduction within the object context pooling branch, where we employ a  $3 \times 3$  convolution to reduce the channel from 2048 to 512. The output feature map from object context pooling module has 512 channels. For the other four branches, we exactly follow the original ASPP module and employ a  $1 \times 1$  convolution within the second above branch and  $3 \times 3$  dilated convolution with different dilation rates (12, 24, 36) in the remained three parallel branches except that we change the output channel from 256 to 512 in all of these four branches. To ensure the fairness of our experiments, we also increase the channel dimension from 256 to 512 within the original ASPP in all of our experiments. Lastly, we concatenate these five parallel output feature maps and employ a  $1 \times 1$  convolution to decrease the channel of the concatenated feature map from 2560 to 512.

## 4. Experiments

### 4.1. Cityscapes

**Dataset.** The Cityscapes dataset [4] is tasked for urban scene understanding, which contains 30 classes and only 19 classes of them are used for scene parsing evaluation. The dataset contains 5,000 high quality pixel-level finely annotated images and 20,000 coarsely annotated images. The finely annotated 5,000 images are divided into 2,975/500/1,525 images for training, validation and testing.

**Training settings.** We set the initial learning rate as 0.01 and weight decay as 0.0005 by default, the original image size is  $1024 \times 2048$  and we choose crop size as  $769 \times 769$  following PSPNet [34], all the baseline experiments only use the 2975 train-fine images as the training set without specification, the batch size is 8 and we choose the InPlaceABN-Sync [22] to synchronize the mean and standard-deviation of BN across multiple GPUs in all the experiments. We employ 40K training iterations, which take about  $\sim 20$  hours with  $4 \times$ P100 GPUs.

Similar to the previous works [3], we employ the "poly" learning rate policy, where the learning rate is multiplied by  $(1 - \frac{iter}{iter_{max}})^{0.9}$ . For the data augmentation methods, we only apply random flipping horizontally and random scaling in the range of [0.5, 2].

**Loss function.** We employ class-balanced cross entropy loss on both the final output of OCNet and the intermediate feature map output from res4b22, where the weight over the final loss is 1 and the auxiliary loss is 0.4 following the original settings proposed in PSPNet [34].

Table 1: Comparison to global pooling (GP), pyramid pooling module (PPM) in PSPNet [34], and atrous spatial pyramid pooling (ASPP) in DeepLabv3 [3] on the validation set of Cityscapes.

Method	Train. mIoU (%)	Val. mIoU (%)
ResNet-101 Baseline	84.26 $\pm$ 0.23	75.69 $\pm$ 0.20
ResNet-101 + GP [16]	85.02 $\pm$ 0.14	77.60 $\pm$ 0.22
ResNet-101 + PPM [34]	85.26 $\pm$ 0.12	77.84 $\pm$ 0.44
ResNet-101 + ASPP [3]	85.64 $\pm$ 0.15	78.65 $\pm$ 0.17
ResNet-101 + Base-OC	85.16 $\pm$ 0.12	78.80 $\pm$ 0.26
ResNet-101 + Pyramid-OC	85.10 $\pm$ 0.11	78.78 $\pm$ 0.30
ResNet-101 + ASP-OC	<b>85.72 <math>\pm</math> 0.12</b>	<b>79.58 <math>\pm</math> 0.24</b>

Table 2: The effect of the OHEM, Ms+Flip, Training w/ the validation set and Fine-tuning, we report the results on the test set of Cityscapes.

OHEM	Ms + Flip	w/ Val	Fine-tuning	Test. mIoU (%)
$\times$	$\times$	$\times$	$\times$	78.22
$\checkmark$	$\times$	$\times$	$\times$	78.90 ( $\Delta$ 0.68)
$\checkmark$	$\checkmark$	$\times$	$\times$	80.06 ( $\Delta$ 1.16)
$\checkmark$	$\checkmark$	$\checkmark$	$\times$	81.54 ( $\Delta$ 1.48)
$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	81.67 ( $\Delta$ 0.13)

Table 3: Comparison to state-of-the-art on the test set of Cityscapes.

Method	Conference	Backbone	mIoU (%)
PSPNet [34] <sup>†</sup>	CVPR2017	ResNet-101	78.4
PSANet [35] <sup>†</sup>	ECCV2018	ResNet-101	78.6
AAF [9] <sup>†</sup>	ECCV2018	ResNet-101	<b>79.1</b>
OCNet <sup>†</sup>	-	ResNet-101	<b>80.1</b>
RefineNet [13] <sup>‡</sup>	CVPR2017	ResNet-101	73.6
SAC [33] <sup>‡</sup>	ICCV2017	ResNet-101	78.1
DUC-HDC [24] <sup>‡</sup>	WACV2018	ResNet-101	77.6
BiSeNet [29] <sup>‡</sup>	ECCV2018	ResNet-101	78.9
PSANet [35] <sup>‡</sup>	ECCV2018	ResNet-101	80.1
DFN [30] <sup>‡</sup>	CVPR2018	ResNet-101	79.3
DSSPN [12] <sup>‡</sup>	CVPR2018	ResNet-101	77.8
DepthSeg [10] <sup>‡</sup>	CVPR2018	ResNet-101	78.2
DenseASPP [28] <sup>‡</sup>	CVPR2018	DenseNet-161	<b>80.6</b>
OCNet <sup>‡</sup>	-	ResNet-101	<b>81.7</b>

<sup>†</sup> Training with only the train-fine datasets.

<sup>‡</sup> Training with both the train-fine and val-fine datasets.

**Object context vs. PPM and ASPP.** To evaluate the effectiveness of OCNet, we conduct a set of baseline experiments on Cityscapes. Especially, we run all of the experiments for three times and report the mean and the variance to ensure that our results are reliable. We use the ResNet-101 + GP to represent employing the global average pooling based context following ParseNet [16], ResNet-101 + PPM represents the PSPNet that applies pyramid pooling module on feature maps of multiple scales and ResNet-101 + ASPP follows the DeepLabv3 that incorporates the image-level global context into the ASPP module except that we increase the output channel of ASPP from 256 to 512 in all of our experiments to ensure the fairness.

We compare these three methods with the object context

module based methods such as Base-OC, Pyramid-OC and ASP-OC. The related experimental results are reported in Table 1, where all the results are based on single scale testing. The performance of both PSPNet and DeepLabv3 are comparable with the numbers in the original paper.

According to the performance on the validation set, we find that our basic method ResNet-101 + Base-OC can outperform the previous state-of-the-art methods such as PSPNet and DeepLabv3. We can further improve the performance with the ASP-OC module. For example, the ResNet-101 + ASP-OC achieves about 79.58 on the validation set based on single scale testing and improves about 1.0  $\uparrow$  point over DeepLabv3 and 2.0  $\uparrow$  points over PSPNet.

**Ablation study.** Based on the ResNet-101 + ASP-OC method (mIoU=79.58/78.22 on the Val./Test. set), we adopt the online hard example mining (OHEM), multi-scale (Ms), left-right flipping (Flip) and training with validation set (w/ Val) to further improve the performance on the test set. All the related results are reported in Table 2.

- **OHEM:** Following the previous works [26], the hard pixels are defined as the pixels associated with probabilities smaller than  $\theta$  over the correct classes. Besides, we need to keep at least  $\mathcal{K}$  pixels within each mini-batch when few pixels are hard pixels. e.g., we set  $\theta = 0.7$  and  $\mathcal{K} = 100000$  on the Cityscapes and improves +0.83 mIoU on validation set and +0.68 mIoU on test set.
- **Ms + Flip:** We further apply the left-right flipping and multiple scales including  $[0.75 \times, 1 \times, 1.25 \times]$  to improve the performance from 78.90 to 80.06 on the test set.
- **Training w/ validation set:** We can further improve the performance on test set by employing the validation set for training. We train the OCNet for 80K iterations on the mixture of training set and validation set and improve the performance from 80.06 to 81.54 on the test set.
- **Fine-tuning:** We adopt the finetuning strategy proposed by DeepLabv3 [3] and DenseASPP [28] to fine-tune the model with the fine-labeled dataset for extra epochs and further boost the performance to 81.67 on the test set with only fine-labeled training set.

**Results.** We compare the OCNet with the current state-of-the-art methods on the Cityscapes. The results are illustrated in Table 3 and we can see that our method achieves better performance over all the previous methods based on ResNet-101. OCNet without using the validation set achieves even better performance than most methods that employ the validation set. Through employing the validation set and fine-tuning strategies, OCNet achieves new

Table 4: Comparison to global pooling (GP), pyramid pooling module (PPM) in PSPNet [34], and atrous spatial pyramid pooling (ASPP) in DeepLabv3 [3] on the validation set of ADE20K.

Method	mIoU (%)	Pixel Acc (%)
ResNet-50 Baseline	$34.35 \pm 0.10$	$76.41 \pm 0.10$
ResNet-50 + GP [16]	$41.17 \pm 0.38$	$79.87 \pm 0.20$
ResNet-50 + PPM [34]	$41.34 \pm 0.10$	$79.96 \pm 0.10$
ResNet-50 + ASPP [3]	$42.53 \pm 0.17$	$80.44 \pm 0.10$
ResNet-50 + Base-OC	$40.66 \pm 0.26$	$79.77 \pm 0.17$
ResNet-50 + Pyramid-OC	$42.28 \pm 0.28$	$80.21 \pm 0.17$
ResNet-50 + ASP-OC	<b><math>43.06 \pm 0.15</math></b>	<b><math>80.70 \pm 0.10</math></b>

Table 5: Comparison to state-of-the-art on the validation set of ADE20K.

Method	Conference	Backbone	mIoU (%)
RefineNet [13]	CVPR2017	ResNet-101	40.20
RefineNet [13]	CVPR2017	ResNet-152	40.70
PSPNet [34]	CVPR2017	ResNet-101	43.29
PSPNet [34]	CVPR2017	ResNet-152	43.51
PSPNet [34]	CVPR2017	ResNet-269	<b>44.94</b>
SAC [33]	ICCV2017	ResNet-101	44.30
PSANet [35]	ECCV2018	ResNet-101	43.77
UperNet [27]	ECCV2018	ResNet-101	42.66
DSSPN [12]	CVPR2018	ResNet-101	43.68
EncNet [32]	CVPR2018	ResNet-101	44.65
OCNet	-	ResNet-101	<b>45.45</b>

state-of-the-art performance of 81.7 on the test set and outperforms the DenseASPP based on DenseNet-161 by over  $1.0 \uparrow$  point.

## 4.2. ADE20K

**Dataset.** The ADE20K dataset [37] is used in ImageNet scene parsing challenge 2016, which contains 150 classes and diverse scenes with 1,038 image-level labels. The dataset is divided into 20K/2K/3K images for training, validation and testing.

**Training setting.** We set the initial learning rate as 0.02 and weight decay as 0.0001 by default, the input image is resized to the length randomly chosen from the set  $\{300, 375, 450, 525, 600\}$  due to that the images are of various sizes on ADE20K. The batch size is 8 and we also synchronize the mean and standard-deviation of BN cross multiple GPUs. We employ 100K training iterations, which take about  $\sim 30$  hours with ResNet-50 and  $\sim 60$  hours with ResNet-101 based on  $4 \times$ P100 GPUs.

The experiments on ADE20K are based on the open-source implementation [37]. By following the previous works [34, 3], we employ the same "poly" learning rate policy and data augmentation methods and employ the deep supervision in the intermediate feature map output from res4b22.

**Object context vs. PPM and ASPP.** We follow the same settings as the previous comparison experiments on Cityscapes. We also re-run all of the experiments for three

Table 6: Comparison to state-of-the-art on the validation dataset of LIP.

Method	Conference	Backbone	mIoU (%)
Attention+SSL [7]	CVPR2017	ResNet-101	44.73
JPPNet [11]	PAMI2018	ResNet-101	51.37
SS-NAN [36]	CVPR2017	ResNet-101	47.92
MMAN [19]	ECCV2018	ResNet-101	46.81
MuLA [20]	ECCV2018	ResNet-101	49.30
CE2P [15]	AAAI2019	ResNet-101	53.10
OCNet	-	ResNet-101	<b>54.72</b>

times and report the mean and the variance. We compare the ResNet-50 + GP, ResNet-50 + PPM and ResNet-50 + ASPP with ResNet-50 + Base-OC, ResNet-50 + Pyramid-OC and ResNet-50 + ASP-OC. The related experimental results on ADE20K are reported in Table 4, where all the results are based on single scale testing.

The performance of both PSPNet and DeepLabv3 is comparable with the numbers reported in the original paper. We can see that both ResNet-50 + Pyramid-OC and ResNet-50 + ASP-OC achieve better performance compared with the ResNet-50 + Base-OC, which verifies the effectiveness of considering the multi-scale context information. Especially, ResNet-50 + Pyramid-OC improves the ResNet-50 + PPM by about  $1.0 \uparrow$  point while ResNet-50 + ASP-OC improves the ResNet-50 + ASPP by about  $0.5 \uparrow$  points.

**Results.** To compare with the state-of-the-art, we replace the ResNet-50 with ResNet-101 and further employ the multi-scale, left-right flipping strategies to improve the performance. According to the reported results in Table 5, OCNet improves the previous ResNet-101 based state-of-the-art method EncNet by about  $0.8 \uparrow$  points, and OCNet also improves the PSPNet based on ResNet-269 by about  $0.5 \uparrow$  points.

## 4.3. LIP

**Dataset.** The LIP (Look into Person) dataset [7] is employed in the LIP challenge 2016 for single human parsing task, which contains 50,462 images with 20 classes (19 semantic human part classes and 1 background class).

**Training setting.** We set the initial learning rate as 0.007 and weight decay as 0.0005 following the CE2P [15]. The original images are of various sizes and we resize all the images to  $473 \times 473$ . The batch size is 40 and we also employ the InPlaceABNSync. We employ 110K training iterations, which take about  $\sim 45$  hours with  $4 \times$ P100 GPUs. We also employ the same (i) "poly" learning rate policy, (ii) data augmentation methods and (iii) deep supervision in the intermediate feature map output from res4b22 following the experiments on Cityscapes and ADE20K.

**Results.** We evaluate the OCNet (ResNet-101 + ASP-OC) on the LIP benchmark and report the related results in Table 6. We can observe that the OCNet improves  $1.6 \uparrow$  points

over the previous state-of-the-art methods on the validation set of LIP. Especially, the human parsing task is different from the previous two scene parsing task as it is about labeling each pixel with the part category that it belongs to. The state-of-the-art results verify that OCNet generalizes well to the part-level semantic segmentation tasks.

#### 4.4. Visualization of object context maps

We randomly choose some examples from the validation set of Cityscapes and visualize the object context map learned within OCNet in the first five rows of Figure 3, where each object context map corresponds to the pixel marked with red in both the original images and ground-truth segmentation maps.

As illustrated in Figure 3, we can find that the estimated object context maps for most classes capture the object context that mainly consists of pixels of the same categories. Take the 1<sup>st</sup> image on the 2<sup>nd</sup> row as an example, it can be seen that the object context map corresponding to the pixel on the object bus distributes most of the weights on the pixels lying on the object bus and thus the object bus’s context information can help the pixel-wise classification.

Besides, we also illustrate some examples from the ADE20K and LIP in the middle three rows and the last three rows of Figure 3. It can be seen that most of the weights within each object context map are focused on the objects or parts that the selected pixel belongs to.

### 5. Conclusions

In this work, we present the concept of object context and propose the object context pooling (OCP) scheme to construct more robust context information for semantic segmentation tasks. We verify that the predicted object context maps within OCP distribute most of the weights on the true object context by visualizing multiple examples. We further demonstrate the advantages of OCNet with state-of-the-art performance on three challenging benchmarks including Cityscapes, ADE20K and LIP.

### References

- [1] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *PAMI*, 2017.
- [2] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *PAMI*, 40(4):834–848, 2018.
- [3] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. arXiv:1706.05587, 2017.
- [4] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016.
- [5] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert. An empirical study of context in object detection. In *CVPR*, pages 1271–1278. IEEE, 2009.
- [6] J. Fu, J. Liu, H. Tian, Z. Fang, and H. Lu. Dual attention network for scene segmentation. arXiv:1809.02983, 2018.
- [7] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *CVPR*, July 2017.
- [8] A. Gonzalez-Garcia, D. Modolo, and V. Ferrari. Objects as context for detecting their semantic parts. In *CVPR*, 2018.
- [9] T.-W. Ke, J.-J. Hwang, Z. Liu, and S. X. Yu. Adaptive affinity fields for semantic segmentation. In *ECCV*, September 2018.
- [10] S. Kong and C. C. Fowlkes. Recurrent scene parsing with perspective understanding in the loop. In *CVPR*, 2018.
- [11] X. Liang, K. Gong, X. Shen, and L. Lin. Look into person: Joint body parsing & pose estimation network and a new benchmark. *PAMI*, 2018.
- [12] X. Liang, H. Zhou, and E. Xing. Dynamic-structured semantic propagation network. In *CVPR*, 2018.
- [13] G. Lin, A. Milan, C. Shen, and I. D. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, volume 1, page 5, 2017.
- [14] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio. A structured self-attentive sentence embedding. 2017.
- [15] T. Liu, T. Ruan, Z. Huang, Y. Wei, S. Wei, Y. Zhao, and T. Huang. Devil in the details: Towards accurate single and multiple human parsing. arXiv:1809.05996, 2018.
- [16] W. Liu, A. Rabinovich, and A. C. Berg. Parsenet: Looking wider to see better. arXiv:1506.04579, 2015.
- [17] Y. Liu, R. Wang, S. Shan, and X. Chen. Structure inference net: Object detection using scene-level context and instance-level relationships. In *CVPR*, 2018.
- [18] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.
- [19] Y. Luo, Z. Zheng, L. Zheng, G. Tao, Y. Junqing, and Y. Yang. Macro-micro adversarial network for human parsing. In *ECCV*, pages 424–440, 2018.
- [20] X. Nie, J. Feng, and S. Yan. Mutual learning to adapt for joint human parsing and pose estimation. In *ECCV*, pages 502–517, 2018.
- [21] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proc. Medical Image Computing and Computer-Assisted Intervention, MICCAI*, pages 234–241. Springer, 2015.
- [22] S. Rota Bul, L. Porzi, and P. Kotschieder. In-place activated batchnorm for memory-optimized training of dnns. In *CVPR*, 2018.
- [23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.
- [24] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell. Understanding convolution for semantic segmentation. In *WACV*, 2018.

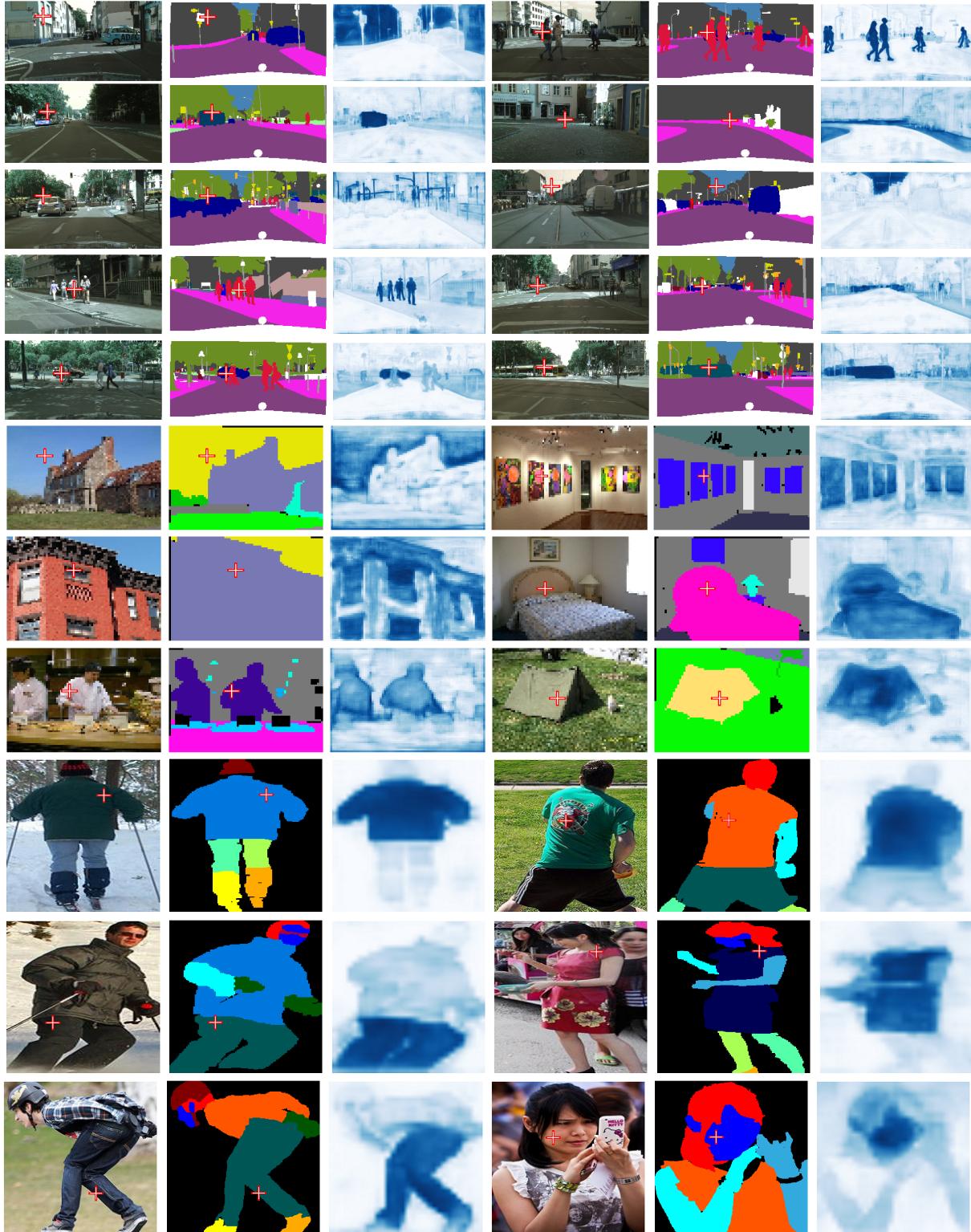


Figure 3: Visualization of object context map predicted by OCNet. The first five rows illustrate 10 examples from the validation set of Cityscapes, the next three rows illustrate 6 examples from the validation set of ADE20K, and the last three rows illustrate 6 examples from the validation set of LIP. (Best viewed in color)

- [25] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *CVPR*, 2018. [2](#)
- [26] Z. Wu, C. Shen, and A. v. d. Hengel. High-performance semantic segmentation using very deep fully convolutional networks. arXiv:1604.04339, 2016. [5](#)
- [27] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun. Unified perceptual parsing for scene understanding. 2018. [6](#)
- [28] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang. Denseaspp for semantic segmentation in street scenes. In *CVPR*, 2018. [5](#)
- [29] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. *ECCV*, 2018. [5](#)
- [30] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang. Learning a discriminative feature network for semantic segmentation. In *CVPR*, 2018. [5](#)
- [31] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *ICLR*, 2016. [1](#), [2](#)
- [32] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal. Context encoding for semantic segmentation. In *CVPR*, 2018. [6](#)
- [33] R. Zhang, S. Tang, Y. Zhang, J. Li, and S. Yan. Scale-adaptive convolutions for scene parsing. In *ICCV*, Oct 2017. [5](#), [6](#)
- [34] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *CVPR*, pages 2881–2890, 2017. [1](#), [2](#), [4](#), [5](#), [6](#)
- [35] H. Zhao, Z. Yi, L. Shu, S. Jianping, C. C. Loy, L. Dahua, and J. Jia. Psanet: Point-wise spatial attention network for scene parsing. *ECCV*, 2018. [2](#), [5](#), [6](#)
- [36] J. Zhao, J. Li, X. Nie, F. Zhao, Y. Chen, Z. Wang, J. Feng, and S. Yan. Self-supervised neural aggregation networks for human parsing. In *CVPRW*, pages 7–15, 2017. [6](#)
- [37] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017. [6](#)