Adaptive Context Network for Scene Parsing

Jun Fu [1,4] Jing Liu* [1] Yuhang Wang [1]
Yong Li [2] Yongjun Bao [2] Jinhui Tang [3] Hanqing Lu [1]

[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
[2]Business Growth BU, JD.com [3] Nanjing University of Science and Technology
[4] University of Chinese Academy of Sciences
{jun.fu,jliu,luhq}@nlpr.ia.ac.cn,wangyuhang.casia@gmail.com,
{liyong5,baoyongjun}@jd.com,jinhuitang@njust.edu.cn

Abstract

Recent works attempt to improve scene parsing performance by exploring different levels of contexts, and typically train a well-designed convolutional network to exploit useful contexts across all pixels equally. However, in this paper, we find that the context demands are varying from different pixels or regions in each image. Based on this observation, we propose an Adaptive Context Network (ACNet) to capture the pixel-aware contexts by a competitive fusion of global context and local context according to different per-pixel demands. Specifically, when given a pixel, the global context demand is measured by the similarity between the global feature and its local feature, whose reverse value can be used to measure the local context demand. We model the two demand measurements by the proposed global context module and local context module, respectively, to generate adaptive contextual features. Furthermore, we import multiple such modules to build several adaptive context blocks in different levels of network to obtain a coarse-to-fine result. Finally, comprehensive experimental evaluations demonstrate the effectiveness of the proposed ACNet, and new state-of-the-arts performances are achieved on all four public datasets, i.e. Cityscapes, ADE20K, PASCAL Context, and COCO Stuff.

1. Introduction

Scene parsing is a fundamental image understanding task which aims to perform per-pixel categorizations for a given scene image. Most recent approaches for scene parsing are based on Fully Convolutional Networks (FCNs) [24]. However, there are two limitations in FCN framworks. First, the consecutive subsampling operations like pooling and convolution striding lead to a significant decrease of the initial image resolution and make the loss of spatial details for scene parsing. Second, due to the limited receptive field [23, 25] or local context features, the per-pixel dense classification is often ambiguous. In the end, FCNs result in the problems of rough object boundaries, ignorance of small objects, and misclassification of big objects and stuff.

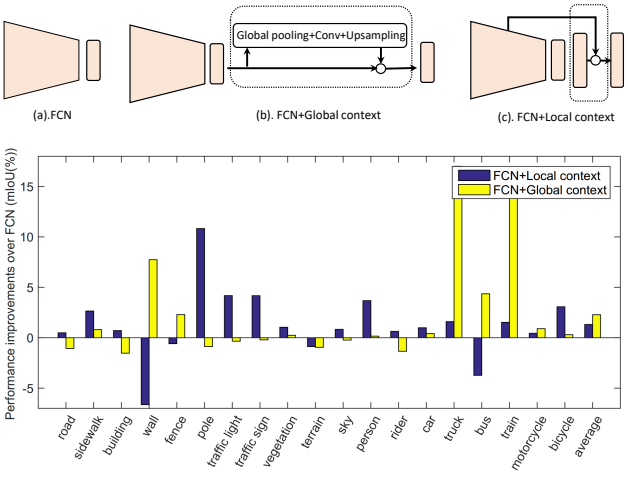

Figure 1. The performance improvements over the basic FCN (a. Dilated FCN) on Cityscapes val set with the help of global context (b. Dilated FCN+Global context) and local context (c. Dilated FCN+Local context). Specially, pixel-wise enhanced representation by the global average pooling feature are employed as the global context, and a concatenated representation with low-level features as the local context.

Throughout various FCN-based improvements to overcome the above limitations, effective strategies to utilize different levels of contexts (i.e., local context and global context) are the main directions. Specifically, some methods [22, 39, 34, 9] adopt "U-net" architectures, which ex-

*Corresponding Author

ploit multi-scale local contexts from middle-layers, to complement more visual details. Some methods [2, 35] employ dilated convolution layers to capture a wider context with a larger receptive field while maintaining the resolution. Besides, the image-level features obtained by global average pooling [23, 3, 40] are proposed as a global context to clarify local confusion. However, these FCN-based variants adopt per-pixel unified processing and overlook different per-pixel demands on different levels of contexts. That is, the local context from middle layers is essential for the class prediction of those pixels on edges or small objects, while the global context exploring the image-level representation is benefit to categorize large objects or stuff regions, especially for the case when the target region exceeds the receptive field of the network. We can also observe the necessity of pixel-sensitive context modeling from the results comparison shown in Figure 1, in which the local and global contexts achieve different improvements on different objects or stuff. Therefore, how to effectively capture such pixel- or region-aware contexts in an end-to-end training framework is an open but valuable research topic for comprehensively accurate scene parsing.

In this paper, we propose an Adaptive Context Network (ACNet) to capture the pixel-aware contexts for image scene parsing. Different from previous methods which fuse different-level contexts for each pixel equally, ACNet generates different per-pixel contexts, i.e., the context-based features are functions of the input data and also vary from different pixels. Such an adaptive context generation is achieved by a competitive fusion mechanism of the global context from image-level feature and local context from the middle-layer feature according to per-pixel different demands. In other words, with the more attention paid on global context for a certain pixel, the less attention is paid on local context, and vice versa.

Usually, the global average pooled feature has a semantic guidance for large objects and stuff, but it lacks spatial information which makes it different from the features of details. Hence, we can match the global pooled feature with the feature of each pixel, obtaining the possibility of the pixel to be an element of large objects or spatial details. It can be further used as a pixel-aware context guidance to adaptively fuse the global features (global context) and low-level features (local context). Motivated by this intuition, we propose a *global context module* to adaptively capture global context. By measuring the similarity between the global feature and per-pixel feature, we can obtain the pixel-aware demanding extent, called as global gated coefficient. The larger gated coefficient indicates that more global context and the less local context could be fused to the pixel. Then we multiply the global feature with the pixel-aware global gated coefficient before adding it to the pixel feature, with which some mislabeling and inconsistent results can be further corrected.

We also propose a *local context module* to compensate spatial details according to the local context demands. Specifically, we find that the pixels with features dissimilar to global feature, trend to be detail parts of images and need more local context to obtain precise results. Hence, we regard the reverse value of the global gated coefficient as local gated coefficient and multiply it with the low-level feature to generate a local gated feature. It emphasizes pixel-aware local context to spatial details and avoids some noises to the pixels belonging to big objects. Furthermore, we reuse multiple local gated features, which is similar to a recurrent learning process and complements more detail information.

We jointly employ a global context module and a local context module as an adaptive context block, and import such blocks into different levels of network. The architecture of our proposed ACNet is shown in Figure 2. Finally, comprehensive experimental analyses on Cityscapes dataset [5], ADE20k[42], PASCAL Context [26], and COCO stuff[1] dataset demonstrate the effectiveness of ACNet.

The main contributions of this paper are as follows:

- We propose an Adaptive Context Network (ACNet) to improve contextual information fusion according to the context demands of different pixels.
- A novel mechanism is proposed to measure global context demand. Global pooled feature can be adaptively fused to the pixels which need large context, thus reducing misclassification for large objects or stuff.
- We improve local context fusion according to the local context demand and reuse local feature progressively, thus improving segmentation results on small objects and edges.
- ACNet achieves new state-of-the-art performance on various scene parsing datasets. In particular, our ACNet achieves a Mean IoU score of 82.3% on Cityscapes testing set without using coarse data, and 45.90% on ADE20K validation set, respectively.

2. Related Work

Global context embedding. Global context embedding have been proven its effectiveness to improve the categorization of some large semantic regions. ParseNet [23] employs the global average pooled feature to augment the features at each location. PSPNet [40] applies the global average pooling in their Spatial Pyramid Pooling module to collect global context. The work [15] captures global contexts by a gloabl context network based on scene similarities. BiSeNet [33] adds the global pooling on the top of the encoder structure to capture the global context. EncNet [37] employs an encoding layer to capture global context and selectively highlight the class-dependent featuremaps.

Local context embedding. U-net based methods often adopt local context from low- and middle-level visual features to generate sharp boundaries or small details for high-resolution prediction. RefineNet [22] utilizes an encoder-decoder framework and refines low-resolution segmentation with fine-gained low-level feature. ExFuse [39] assigns auxiliary supervisions directly to the early stages of the encoder network for improving low-level context. Deeplabv3+ [4] adds a simple decoder module to capture local context, refining the segmentation results.

Attention and gating mechanisms. Attention mechanisms have been widely used to improve the performance of segmentation task. PAN [17] uses a global pooling to generate global attention, which can select the channel maps effectively. LRR [10] generates a multiplicative gating to refine segment boundaries reconstructed from lower-resolution score maps. Ding et al.[7] proposes a scheme of RNN-based gating mechanism to selectively aggregate multi-scale score maps, which can achieve an optimal multi-scale aggregation. The works [16, 36, 14] adopt self-attention mechanism to model the relationship of features.

Different from these works,we introduce a data-driven gating mechanism to capture global context and local context according to pixel-aware context demand.

3. Adaptive Context Network

3.1. Overview

Contextual information is effective for scene parsing task, most of current methods fuse the different context to each pixel equally, ignoring the different demands of pixel-aware contexts. In this work, we propose a novel Adaptive Context Network (ACNet) to weigh the global and local context complemented to each pixel by a competitive fusion mechanism.

The overall architecture is shown in Figure 2, which adopts pretrained dilated ResNet [13], as the backbone network, and multiple adaptive context blocks to progressively generate high-resolution segmentation map. In the backbone network, we remove the downsampling operations and employ dilated convolutions in the last ResNet blocks, thus obtaining dense feature with output size 1/16 of the input image. It could achieve the balance between retaining spatial details and computation cost [4]. In upsampling process, three adaptive context blocks are employed with three different resolutions. Each adaptive context block consists of a global context module, an upsampling module and a local context module, where the global context module selectively captures global context from the high-level features and the local context module selectively captures local context from the low-level features.

In the following subsections, we will elaborate the designing details of the global context module, the local context module and their aggregation within an adaptive context block.

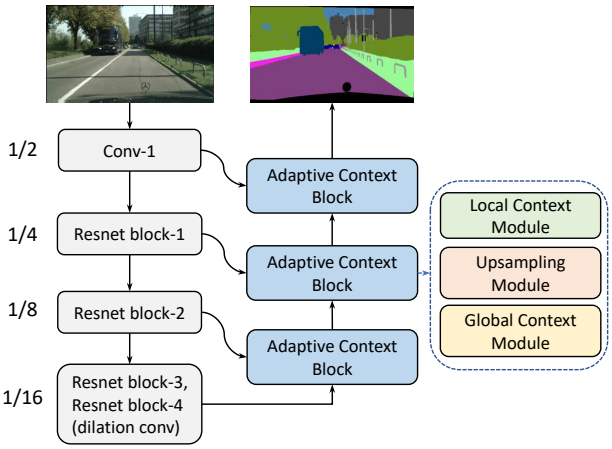

Figure 2. Overview of Adaptive Context Network. (Best viewed in color)

3.2. Global Context Module

Global context can provide global semantic guidance for overall scene images, thus rectifying misclassification and inconsistent parsing results. However, the benefit from global context is different for large objects and spatial details. it is necessary to treat each pixel differently when exploring the global context, that is to say, some pixels need more global context for categorization, while others may do not. Based on the intuition that the global pooled feature prefers to the large objects and stuff and lacks spatial information, we can match the global feature with the feature in each pixel, obtaining its possibility to be as an element of large objects or spatial details. Then we can exploit it to adaptively fuse the global context. To this end, we propose a Global Context Module (GCM) as follows.

Given an input feature map $\mathbf{A} \in \mathbb{R}^{C\times H\times W}$, we use a global average pooling following a convolution layer to generate a global feature $p \in \mathbb{R}^{C\times 1\times 1}$. In order to obtain the pixel-aware demand for global context (global gated coefficient), we first measure the feature similarity by calculating the euclidean distance $\mathbf{D} \in \mathbb{R}^{H\times W}$ between global feature p and the features $a_i \in \mathbf{A}$ for each pixel i, $d_i \in \mathbf{D}$ is denoted as:

$$d_i = \|a_i - p\|_2 \tag{1}$$

where $a_i \in \mathbf{A}$, $i \in [1, 2, ..., H \times W]$ is i^{th} location in $\mathbf{A}$. Noted that the smaller d_i indicates that the feature at i^{th} locations is closer to the global feature. Then we generate a global gated coefficient $\mathbf{W^g} \in \mathbb{R}^{H\times W}$ which is smoothed by an exponential function, $w_i^g \in \mathbf{W^g}$ is denoted as:

$$w_i^g = exp(-\frac{d_i - k}{\delta}) \tag{2}$$

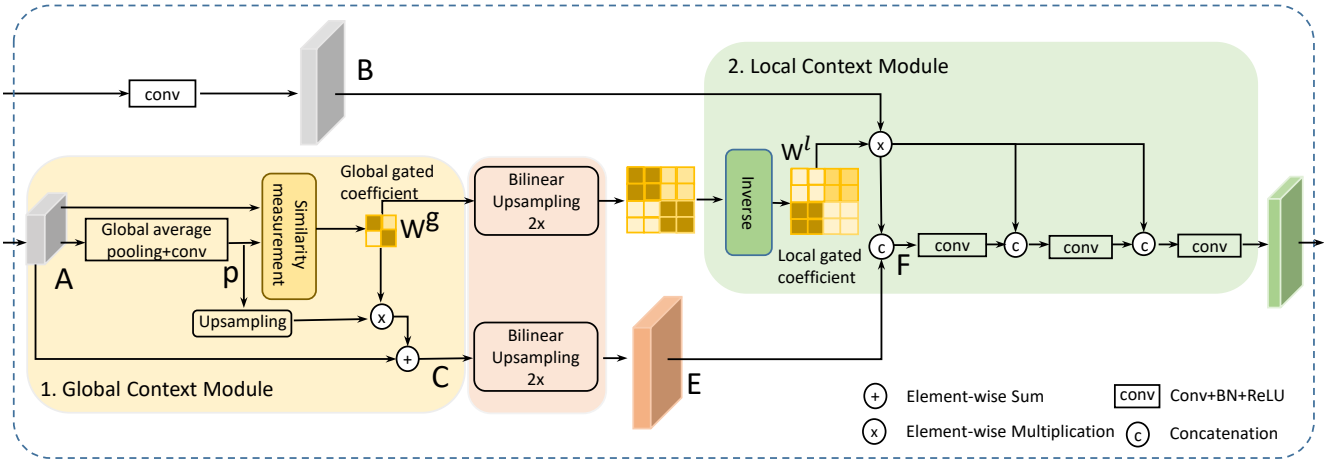

Figure 3. The details of Adaptive Context Block including (1) Global Context Module and (2) Local Context Module. (Best viewed in color)

where k is set to $\min_{i=1}^{H\times W}(d_i)$ for limiting the range of $w_i^g \in (0, 1]$. And δ is a hyperparameter, which controls the amplitude of the difference between high response and low response.

Finally, we multiply the global feature p by w_i^g and a scale parameter α, and then perform an element-wise sum operation with the features $\mathbf{A}$ to obtain the final output $\mathbf{C} \in \mathbb{R}^{C\times H\times W}$, $c_i \in \mathbf{C}$ is denoted as:

$$c_i = \alpha w_i^g p + a_i \tag{3}$$

where α is a learned factor and initialized as 1. Here, we adopt sum operation instead of concatenation for saving memory. The details of global context module is shown in Figure.3 (1).

It can be inferred from the above formulation that the feature $\mathbf{C}$ at different position obtain different global context according to global gated coefficient $\mathbf{W^g}$. With this design, GCM could selectively enhance semantic consistency and reduce the misclassification and inconsistent predictions for large objects or stuff.

3.3. Local Context Module

Local context contributes to refine object boundaries and details. However, many methods fuse local context to all pixels without considering the different demand for local context. To solve this problem, we propose a Local Context Module (LCM) to selectively fuse local context for better refined segmentation.

As mentioned in Section 3.2, the global gated coefficient with high response indicates the pixels belong to large objects and stuff while low response indicates the pixels belong to spatial details. Based on this observation, we could obtain the local gated coeffcient by reversing value of global gated coefficient, where the global gated coefficient have been upsampled, formulated as:

$$\mathbf{W^l} = \mathbf{1} - up(\mathbf{W^g}) \tag{4}$$

where $up(\cdot)$ denotes a bilinear interpolation operation. In this way, the local gated coefficient indicates the more possibility the pixels belong to spatial details, the more local contexts are required, and vice versa. Then we obtain pixel-aware local context (gated local features) by multiply the local feature $\mathbf{B} \in \mathbb{R}^{C\times H\times W}$ from middle-layer features with the local gated coefficient and a scale parameter β. Finally, we concat the feature with the upsampled feature $\mathbf{E} \in \mathbb{R}^{C\times H\times W}$ to generate a refined feature $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$, $f_i \in \mathbf{F}$ is denoted as:

$$f_i = cat(\beta w_i^l b_i, e_i) \tag{5}$$

where $cat(\cdot)$ denotes a concatenation operation, and β is a learned factor and initialized as 1. We adopt concatenation operation to combine the gated local feature and high-level feature, and a convolution layer is employed to fuse them. The details of local context module is shown in Figure.3(2). With this design, we can selectively aggregate the local context according to the context demand of each pixel.

In addition, we find that it is useful to introduce gated local context directly multiple times. Specifically, we reuse gated local features by a concatenation operation followed by a convolution layer for three times. Such a recurrent learning process complements more spatial details for each position, and achieve a coarse-to-fine performance improvement. Noted that, it haven't been discussed in previous works [22, 39, 35, 4]. And we also verify the effectiveness of this process in experiments.

3.4. Adaptive Context Block

Based on GCM and LCM, we further design an Adaptive Context Block to selectively capture global and local contextual information simultaneously.

Adaptive context block is built upon a cascaded architecture, the high-layer features are first fed into a global context module to selectively fuse global context to each pixel. Then passed sequentially through a bilinear upsampling layer and a local context module for learning a restoration of refined features. In order to obtain resolution corresponding to low-level feature, we also enlarge spatial resolution of the global gated coefficient by a bilinear upsampling operation before feeding into the local context module. Following [4], we apply a convolution layer on the low-level features to reduce the number of channels, thus refining the low-level features.

In the adaptive context block, we introduce a competitive fusion mechanism to capture global and local context according to their correlation of gated coefficient, thus suitable context can be adaptively fused to each pixels for better feature representation.

4. Experiments

The proposed method are evaluated on Cityscapes [5], ADE20K [42], PASCAL Context [26], COCO Stuff [1]. Experimental results demonstrate that ACNet achieves new state-of-the-art performance on these datasets. In the next subsections, we first introduce the datasets and implementation details, then we make detail comparisons to evaluate our approaches on Cityscapes dataset. Finally, we present our results compared with state-of-the-art methods on ADE20K, PASCAL Context, COCO Stuff dataset.

4.1. Datasets

Cityscapes The dataset is a well-known road scene dateset collected for scene parsing, which has 2,979 images for training, 500 images for validation and 1,525 images for testing. Each image has a high resolution of 2048×1024 pixels with 19 semantic classes. Noted that no coarse data is employed in our experiments.

ADE20K The ADE20K dataset is a vary challenge scene understanding dataset, which contains 150 classes (35 stuff classes and 115 discrete object classes). The dataset is divided into 20, 210/2, 000/3, 352 images for training, validation and testing.

PASCAL Context The dataset is widely used for scene parsing, which contains 4,998 images for training and 5,105 images for testing. Following previous works [22, 37], we evaluate the method on 60 categories (59 classes and one background category).

COCO Stuff The dataset has 171 categories including 80 objects and 91 stuff annotated to each pixel. Following previous works [7, 27, 22], we adopt 9,000 images for training and 1,000 images for testing.

4.2. Implementation Details

We employ a dilated pretrained ResNet architecture as our backbone network, where the dilated rates in the last ResNet block is set to (2,2,2). Following [37, 40], we apply a 3×3 convolution layer with BN, ReLU on the outputs of the last ResNet block to reduce the number of channels to 512 before feeding into the first adaptive context block. In addition, we adopt the outputs of ResNet block-1 and ResNet block-2 as the low-level features, which provide local context for the first two adaptive context blocks. And we only adopt a global context module in last adaptive context block. In the first two adaptive context block,we employ a 3×3 convolution layer on the low-level featues before feeding it into local context module. The other convolution layers in the first two adaptive context block are composed of a 3×3 convolution operation with 448 and 256 kernels respectively followed by BN and ReLU. Pytorch is used to implement our method.

During training phase, we employ a poly learning rate policy where the initial learning rate is multiplied by $(1 - \frac{iter}{total\_iter})^{0.9}$ after each iteration, and enable synchronized batch normalization [37]. The base learning rate is set to 0.005 for Cityscapes and ADE20K, 0.001 for PASCAL Context and COCO stuff. Momentum and weight decay coefficients are set to 0.9 and 0.0001 respectively. Following [40], auxiliary loss is adopted when we adopt the bockbone ResNet101. In addition, we apply random cropping and random left-right flipping during training phase, and the randomly scaling for data augmentation is not employed if not mentioned on Cityscapes dataset.

4.3. Results on Cityscape dataset

Global Context Module: Firts of all, we design a global context module to adaptively aggregate global context according to pixel-aware demands. Specifically, we follow [2] and build two dilated networks (ResNet-50) which yield the final feature maps with the 1/8 and 1/16 size of the original image. Next, the global context are added on the top of the networks with two different settings, which are GC and GCM respectively (GC denotes that we directly sum the global features to each pixel equally, GCM represents the global context module).

Experimental results are shown in Table. 1, we can see that GCM (global context module) achieves better performance than GC in both two settings, especially for the output 1/8 size of the original image. It shows the effectiveness of GCM and also indicates that the improvement will be more obvious if the higher-resolution global gated coefficient is produced in GCM based on the dilated FCN. In addition, we also provide a discussion about δ, which controls the amplitude of the difference between high response and low response of the global gated coefficient (mentioned in Sec.3.1). When we set δ to 5, the gloabl context module

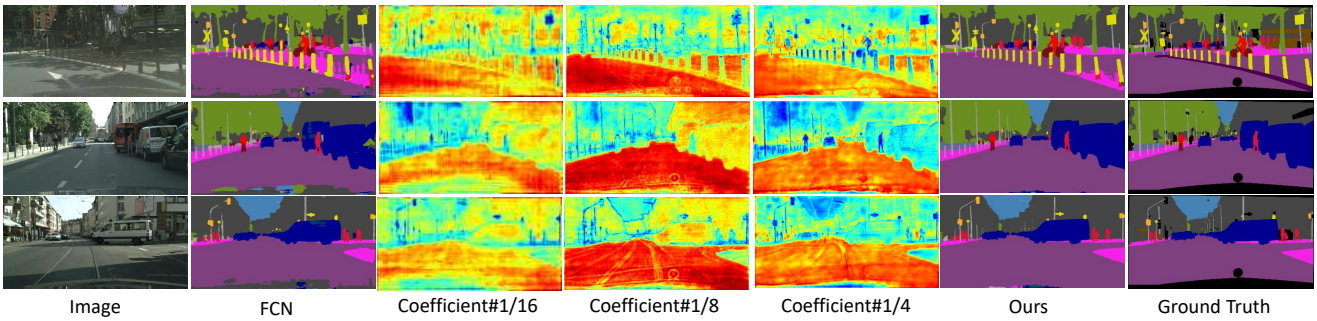

Figure 4. Visualization results of global gated coefficient from global context module with 1/16, 1/8 and 1/4 resolution respectively, we can find that the pixels with large global gated coefficient perfer to dominant stuff and large objects. Compared with FCN, our method enhance semantic guidance with global context in the regions with large coefficient and provide more local context in other regions, thus obtaining accurate segmentation results. (Best viewed in color)

Method	mIOU(%)	
Outputsize	1/16	1/8
Res-50	69.15	70.83
Res-50+GC	71.24	72.77
Res-50+GCM ($\delta = 2$)	72.36	74.21
Res-50+GCM ($\delta = 5$)	**72.45**	**74.50**
Res-50+GCM ($\delta = 10$)	71.87	74.30

Table 1. Ablation experiments of Global Context Module on Cityscapes validation set, δ denotes the amplitude of difference distribution of the global gated coefficient.

Method	mIOU(%)
Res-50	69.15
Res-50+GCM	72.45
Res-50+GCM+LC	73.48
Res-50+GCM+LCM(1)	74.03
Res-50+GCM+LCM(2)	74.56
Res-50+GCM+LCM(3)	**74.67**

Table 2. Ablation experiments of Local Context Module on Cityscapes validation set, (n) denotes the n times of fusion of local gated features in the LCM.

yields the best performance. We fix this value and employ the lowest resolution output 1/16 size of the original image in following experiments.

Local Context Module: We also propose a local context module to refine spatial details. Since we need to generate the local gated coefficient for each pixel by inversing global gated coefficien, the local context module is built on the global context module. Specifically, experiments are conducted on a dilated ResNet-50 with a GCM, then we cascade local features from the outputs of ResNet block-2 with (LCM) and without (LC) local gated coefficient .

Results are shown in Table 2, we can see that the local context improves the performance from 72.45% to 73.48%. When we adopt the local gated coefficient to selectively fuse local context for each pixel once, the performance is further improved to 74.03%. Reusing local gated feature brings continuous improvements of the performance from 74.03% to 74.67%.

Method	mIoU(%)
Res-50+ACB#1	74.67
Res-50+ACB#2	75.98
Res-50+ACB#3	**76.53**
Res-101+ACB#3	77.42
Res-101+ACB#3+MG	78.50
Res-101+ACB#3+MG+DA	80.09
Res-101+ACB#3+MG+DA+OHEM	80.89
Res-101+ACB#3+MG+DA+OHEM+MS	**82.00**

Table 3. Ablation experiments of Adaptive Context Block on Cityscapes validation set, #n denotes the number of Adaptive Context Block, MG denotes multi-grid dilated convolution, DA denotes data augmentation with multi-scale input during training phase, MS denotes multi-scale testing.

Adaptive Context Block: We further build an adaptive context block and cascade it for three times to obtain high resolution predictions. Results are listed in Table 3. When we employ three adaptive context blocks (ACB#3), the performance is improved to 76.53%, which verifies the effectiveness of our method.

In addition, we visualize the global gated coefficients in three adaptive context blocks with different resolutions as shown in Figure 4. The images are from the validation set of Cityscapes. We can find that the pixels with large global gated coefficient perfer to dominant stuff and large objects, such as the "road" in the first row and "car" in the last two rows. These stuff and objects are improved in our method. In addition, the pixels with small global gated coefficient perfer to small objects and edges, such as the "traffic sign"

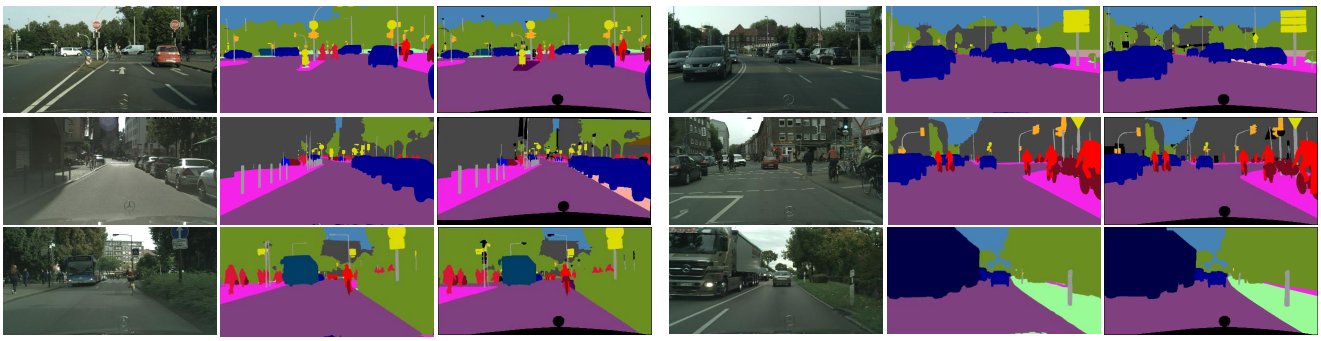

Figure 5. Example results of ACNet on Cityscapes validation set. (Best viewed in color)

Methods	mIoU	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle
RefineNet [22]	73.6	98.2	83.3	91.3	47.8	50.4	56.1	66.9	71.3	92.3	70.3	94.8	80.9	63.3	94.5	64.6	76.1	64.3	62.2	70
DUC [29]	77.6	98.5	85.5	92.8	58.6	55.5	65	73.5	77.9	93.3	72	95.2	84.8	68.5	95.4	70.9	78.8	68.7	65.9	73.8
ResNet-38 [30]	78.4	98.5	85.7	93.1	55.5	59.1	67.1	74.8	78.7	93.7	72.6	95.5	86.6	69.2	95.7	64.5	78.8	74.1	69	76.7
PSPNet [40]	78.4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
BiSeNet [33]	78.9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
PSANet [41]	80.1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
DenseASPP [32]	80.6	98.7	87.1	93.4	60.7	62.7	65.6	74.6	78.5	93.6	72.5	95.4	86.2	71.9	96.0	78.0	90.3	80.7	69.7	76.8
CCNet [14]	81.4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
DANet [16]	81.5	98.6	86.1	93.5	56.1	63.3	69.7	77.3	81.3	93.9	72.9	95.7	87.3	72.9	96.2	76.8	89.4	86.5	72.2	78.2
ACNet	**82.3**	98.7	87.1	93.9	61.6	61.8	71.4	78.7	81.7	94.0	73.3	96.0	88.5	74.9	96.5	77.1	89.0	89.2	71.4	79.0

Table 4. Category-wise comparison with state-of-the-art methods on Cityscapes testing set.

and "pole","person", etc. These spatial details are also be refined in our results. A similar trend is also spotted in other images.

Some improvement strategies: we follow the common procedure of [3, 16, 12, 8, 6, 11] to further improve the performance of ACNet: (1) A deeper and powerful network ResNet-101. (2) MG: Different dilated rates (4,8,16) in the last ResNet block. (3) DA: We transform the input images with random scales (from 0.5 to 2.2) during training phase. (4) OHEM: The online hard example mining is also adopted. (5) MS: we apply the multi-scale inputs with scales {0.5 0.75, 1, 1.25, 1.5, 1.75, 2, 2.25} as well as their mirrors for inference.

Experimental results are shown in Table 3, when employing a deeper backbone (ResNet101), ACNet obtains 77.42% in terms of mean IoU. Then multi-grid dilated convolutional improves the performance by 1.08%. Data augmentation with multi-scale input (DA) brings another 1.59% improvement. OHEM increases the performance to 80.89%. Finally, using multi-scale testing, we attains the best performance of 82.00% on the validation set.

Compared with state-of-art methods: We also compare our method with state-of-the-art methods on Cityscapes test set. Specifically, we fine tune our best model of ACNet with only fine annotated trainval data, and submit our test results to the official evaluation server. For each method, we report the accuracy for each class and the average class accuracy, which are reported in the original paper. Results are shown in Table. 4. We can see that our ACNet achieve a new state-of-the-art performance of 82.3% on the test set. With the same backbone ResNet-101, our model outperforms DANet[16]. Moreover, ACNet also surpasses DenseASPP [32] , which uses more powerful pretrained models, and is heigher than Deeplabv3+ [4] (82.1%), which uses extra the coarse annotations in training phase.

4.4. Results on ADE20K dataset

In this subsection, we conduct experiments on the ADE20K dataset to validate the effectiveness of our method. Following previous works [14, 18, 37, 40, 41], data augmentation with multi-scale input and multi-scale testing are used. We evalute ACNet by pixel-wise accuracy (PixelAcc) and mean of class-wise intersection over union (mIoU). Quantitative results are shown in Table.5. With ResNet50, the dilated FCN obtains 37.32%/77.78% in terms of mIoU and PixelAcc. When adopting our method, the performance is improved by 5.69%/3.23%. When employing a deeper backbone ResNet101, ACNet achieves a new state-of-the-art performance of 45.90%/81.96%, which outperforms the previous state-of-the-art methods. In ad-

Backbone	Method	mIoU (%)	PixAcc%
Res-50	Dilated FCN	37.32	77.78
	EncNet[37]	41.11	79.73
	GCU[18]	42.60	79.51
	PSPNet[40]	42.78	80.76
	PASNet[41]	42.98	80.92
	ACNet	**43.01**	**81.01**
Res-101	UperNet[31]	42.66	81.01
	PSPNet[40]	43.29	81.39
	DSSPN[20]	43.68	81.13
	PASNet[41]	43.77	81.51
	SGR [19]	44.32	81.43
	EncNet[37]	44.65	81.19
	GCU[18]	44.81	81.19
	ACNet	**45.90**	**81.96**

Table 5. Results of semantic segmentation on ADE20K validation set.

Method	Final score(%)
PSPNet269 (1st in place 2016)	55.38
PSANet-101[41]	55.46
CASIA_IVA_JD (1st in place 2017)	55.47
EncNet-101 [37]	55.67
ACNet-101	**55.84**

Table 6. Results of semantic segmentation on ADE20K testing set.

Backbone	Method	mIoU (%)
Res-101	Ding et al.[7]	51.6
	EncNet [37]	51.7
	SGR [19]	52.5
	DANet [16]	52.6
	ACNet	**54.1**
Res-152	RefineNet [22]	47.3
	MSCI[21]	50.3
Xception-71	Tian et al.[28]	52.5

Table 7. Segmentation results on PASCAL Context testing set.

dition, we also fine tune our best model of ACNet-101 with trainval data, and submit our test results on the test set. The with single model of ACNet-101 gets final score as 55.84%. Among the approaches, most of methods [40, 37, 18, 38, 41, 14] attemp to explore the global information by aggregation variant and relationship of the feature on the the top of the backbones. While our method focuses on capturing the pixel-aware contexts from high and low-level features and achieves better performance.

4.5. Results on PASCAL Context Dataset

We also carry out experiments on the PASCAL Context dataset to further demonstrate the effectiveness of ACNet. We employ the ACNet-101 network with the same training strategy on ADE20K and compare our model with previous state-of-the-art methods. The results are reported in Table 7. ACNet obtains a Mean IoU of 54.1%, which surpasses previous published methods. Among the approaches, the recent methods[21, 28] use more powerful network(e.g. ResNet-152 and Xception-71) as encoder network and fuse high-and low-level feature in decoder network, our method outperforms them by a relatively large margin.

Backbone	Method	mIoU(%)
Res-101	RefineNet [22]	33.6
	Ding et al.[7]	35.7
	DSSPN[20]	38.9
	SGR [19]	39.1
	DANet[16]	39.7
	ACNet	**40.1**

Table 8. Segmentation results on COCO Stuff testing set.

4.6. Results on COCO stuff Dataset

Finally, we demonstrate the effectiveness of ACNet on the COCO stuff dataset. The ACNet-101 network is also employed. The COCO stuff results are reported in Table 8. ACNet achieves performance of 40.1% Mean IoU, which also outperforms other state-of-the-art methods.

5. Conclusion

In this paper, we present a novel network of ACNet to capture pixel-aware adaptive contexts for scene parsing, in which a global context module and a local context module are carefully designed and jointly employed as an adaptive context block to obtain a competitive fusion of the both contexts for each position. Our work is motivated by the observation that the global context from high-level features helps the categorization of some large semantic confused regions, while the local context from lower-level visual features helps to generate sharp boundaries or clear details. Extensive experiments demonstrate the outstanding performance of ACNet compared with other state-of-the-art methods. We believe such an adaptive context block can also be extended to other vision applications including object detection, pose estimation, and fine-grained recognition.

Acknowledgement: This work was supported by National Natural Science Foundation of China (61872366 and 61872364) and Beijing Natural Science Foundation (4192059)

References

[1] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1209–1218, 2018.

[2] 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.

[3] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017.

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 801–818, 2018.

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016.

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

[7] 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.

[8] 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.

[9] Jun Fu, Jing Liu, Yuhang Wang, and Hanqing Lu. Stacked deconvolutional network for semantic segmentation. *arXiv preprint arXiv:1708.04943*, 2017.

[10] Golnaz Ghiasi and Charless C. Fowlkes. Laplacian pyramid reconstruction and refinement for semantic segmentation. In *the European Conference on Computer Vision*, pages 519–534, 2016.

[11] Junjun He, Zhongying Deng, and Yu Qiao. Dynamic multi-scale filters for semantic segmentation. In *Proceedings of the International Conference on Computer Vision*, 2019.

[12] 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.

[13] 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.

[14] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. *arXiv preprint arXiv:1811.11721*, 2018.

[15] Wei-Chih Hung, Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Scene parsing with global context embedding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2631–2639, 2017.

[16] Fu Jun, Liu Jing, Tian Haijie, Li Yong, Bao Yongjun, Fang Zhiwei, and Lu Hanqing. Dual attention network for scene segmentation. In *Proceedings of thr IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[17] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang. Pyramid attention network for semantic segmentation. *arXiv preprint arXiv:1805.10180*, 2018.

[18] Yin Li and Abhinav Gupta. Beyond grids: Learning graph representations for visual recognition. In *Advances in Neural Information Processing Systems*, pages 9245–9255, 2018.

[19] Xiaodan Liang, Zhiting Hu, Hao Zhang, Liang Lin, and Eric P Xing. Symbolic graph reasoning meets convolutions. In *Advances in Neural Information Processing Systems*, pages 1858–1868, 2018.

[20] 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.

[21] Di Lin, Yuanfeng Ji, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Multi-scale context intertwining for semantic segmentation. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[22] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian D. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5168–5177, 2017.

[23] Wei Liu, Andrew Rabinovich, and Alexander C. Berg. Parsenet: Looking wider to see better. In *the International Conference on Learning Representations*, 2016.

[24] 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.

[25] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *Advances in neural information processing systems*, pages 4898–4906, 2016.

[26] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan L. Yuille. The role of context for object detection and semantic segmentation in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2014.

[27] Bing Shuai, Zhen Zuo, Bing Wang, and Gang Wang. Scene segmentation with dag-recurrent neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 1480–1493, 2018.

[28] Zhi Tian, Tong He, Chunhua Shen, and Youliang Yan. Decoders matter for semantic segmentation: Data-dependent decoding enables flexible feature aggregation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'19)*, 2019.

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

[30] 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.

[31] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434, 2018.

[32] 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.

[33] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 325–341, 2018.

[34] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Learning a discriminative feature network for semantic segmentation. In *CVPR*, 2018.

[35] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *the International Conference on Learning Representations*, 2016.

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

[37] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[38] 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.

[39] Zhenli Zhang, Xiangyu Zhang, Chao Peng, Xiangyang Xue, and Jian Sun. Exfuse: Enhancing feature fusion for semantic segmentation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 273–288, 2018.

[40] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6230–6239, 2017.

[41] 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.

[42] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5122–5130, 2017.