**CaseNet: Content-Adaptive Scale Interaction Networks for Scene Parsing**

Xin Jin\(^*\)  
USTC  
jinxustc@mail.ustc.edu.cn

Cuiling Lan  
MSRA  
culan@microsoft.com

Wenjun Zeng  
MSRA  
wezeng@microsoft.com

Zhizheng Zhang  
USTC  
zhiheng@mail.ustc.edu.cn

Zhibo Chen  
USTC  
chenzhibo@ustc.edu.cn

---

### ABSTRACT

Objects in an image exhibit diverse scales. Adaptive receptive fields are expected to catch suitable range of context for accurate pixel level semantic prediction for handling objects of diverse sizes. Recently, atrous convolution with different dilation rates has been used to generate features of multi-scales through several branches and these features are fused for prediction. However, there is a lack of explicit interaction among the branches to adaptively make full use of the contexts. In this paper, we propose a Content-Adaptive Scale Interaction Network (CaseNet) to exploit the multi-scale features for scene parsing. We build the CaseNet based on the classic Atrous Spatial Pyramid Pooling (ASPP) module, followed by the proposed contextual scale interaction (CSI) module, and the scale adaptation (SA) module. Specifically, first, for each spatial position, we enable context interaction among different scales through scale-aware non-local operations across the scales, i.e., CSI module, which facilitates the generation of flexible mixed receptive fields, instead of a traditional flat one. Second, the scale adaptation module (SA) explicitly and softly selects the suitable scale for each spatial position and each channel. Ablation studies demonstrate the effectiveness of the proposed modules. We achieve state-of-the-art performance on three scene parsing benchmarks Cityscapes, ADE20K and LIP.

### KEYWORDS

scene parsing, multi-scale information, contextual scale interaction, scale adaptation

---

**1 INTRODUCTION**

Scene parsing or semantic segmentation is a fundamental and challenging task. The purpose is to predict the semantic of each pixel including stuffs (e.g. sky, road) and objects (e.g. person, car). This study benefits various challenging tasks such as autonomous driving [34], robot sensing [28, 29], and image editing/captioning [19, 23].

With the development of the Fully Convolutional Network (FCN) [25], semantic image segmentation has achieved promising results with significantly improved feature representation. However, as shown in Figure 1, objects within a picture are typically diverse on scales. The CNNs with standard convolutions can not handle diverse scales due to fixed receptive fields. Objects that are larger than the receptive fields often have inconsistent parsing prediction while objects that are smaller than the receptive fields are often ignored/mislabeled [47].

---

\(^*\)This work is done when Xin Jin is an intern at MSRA.

---

Figure 1: Example of the scale variations of objects in a street scene. The same category of objects, such as cars, may vary largely in scale. Each pixel needs a suitable receptive field to catch the best context.

To deal with scale diversity, multi-scale context fusion schemes have been proposed [4, 43, 48]. The early representative work is the pyramid pooling module (PPM) in PSPNet [48], which pools the feature map at multiple rates and multiple effective field-of-views, capturing objects at multiple scales[5]. However, such a pooling operation reduces the feature map resolution and also leads to information loss. Atrous spatial pyramid pooling (ASPP) [4] and densely connected atrous spatial pyramid pooling (DenseASPP) [43] modules use atrous/dilated convolutions [3] to effectively enlarge the field of view of filters to have multi-scale features. Atrous/Dilated convolutions with different atrous rates lead to different receptive field sizes for generating features. All the above approaches concatenate these features followed by convolution operations to exploit multi-scale contexts. To some extent, these convolution filters promote the interactions and selections among features of different scales. However, after training, the filter weights are fixed and thus the interaction and selection are not highly content adaptive and flexible. Besides, for different spatial positions, the object/stuff categories and scales are different. The feature interaction among scales for exploiting contextual features of different scales should be content adaptive rather than position invariant.

To address the above problems, we propose a Content-Adaptive Scale Interaction Network (CaseNet) to adaptively (i.e., case by case for each position) exploit multi-scale features for scene parsing. Fig. 2 shows the overall flowchart of our framework. We build our framework on top of the classic Atrous Spatial Pyramid Pooling (ASPP) module, which provides \( K \) feature maps with each characterized by a different scale/receptive field.

First, to enable the feature interaction among scales to exploit the scale contextual information, we design a contextual scale interaction (CSI) module for feature refinement, where scale-aware non-local operations are performed across the \( K \) scales of features.
Figure 2: An overview of the Content-Adaptive Scale Interaction Network (CaseNet). We build it based on the classic Atrous Spatial Pyramid Pooling (ASPP) [5], followed by the proposed Contextual Scale Interaction (CSI) module, and the Scale Adaptation (SA) module. The fused feature is fed to a convolution layer followed by SoftMax for prediction. Note \( d \) denotes the dilation rate of the atrous convolution.

For the feature of a scale at a spatial position, the interaction is achieved by computing the response as a weighted sum of the features at all the scales of the same spatial position. Note that this module is functionally different from the non-local block [37] from two aspects. First, we use the non-local operations to exploit the scales while non-local block [37] use them to exploit the long-range spatial/temporal contexts. Second, we explicitly leverage prior scale information, i.e., scale/branch index, together with the embedding features to learn the weights (edges) which are used for computing the sum. The learned weights are different for different spatial positions and thus content adaptive.

Second, we design a spatial and channel adaptive scale adaptation (SA) module to adaptively select the appropriate receptive fields/scales for objects/stuffs of different sizes. For each spatial position, we simultaneously learn scale and channel attention. Ours is different from the attention in [6, 16]. First, we relax the sum 1 constraint on the attention weights across the scales, which provides more flexible optimization. On the other hand, since the sum of the attention weights in a spatial position is not restricted to 1, the energy on different spatial position can vary and play a role of spatial attention. Besides, to be more flexible, our attention is also channel adaptive.

Our main contributions are summarized as follows:

- We propose a simple yet effective content-adaptive scale interaction network (CaseNet) to efficiently exploit the multi-scale features for scene parsing.
- Two content adaptive modules, contextual scale interaction (CSI) module and scale adaptation (SA) module, are proposed to boost the interaction of multi-scale features. The CSI module enables the scale context interaction for feature refinement while the SA module facilitates the appropriate scale selection adapted to spatial positions and channels.
- We validate the effectiveness of the proposed modules through adequate ablation studies. We achieve state-of-the-arts results on three scene parsing benchmarks including Cityscapes [8], ADE20K [51], and LIP datasets [12].

2 RELATED WORK

Multi-Scale Feature Exploration. Scene parsing has achieved great progress with the development of Fully Convolutional Networks (FCNs) [25]. To alleviate the local receptive field (RF) issue of convolution operations in FCNs, several network variants are proposed to generate and aggregate multi-scale features. Motivated by the spatial pyramid matching [18], PSPNet [48], Deeplabv2 [4], and Deeplabv3 [5] are proposed to concatenate features of multiple receptive field sizes together for the semantics prediction. PSPNet [48] employs four spatial pyramid pooling (down-sampling) layers in parallel, named pyramid pooling module (PPM), to aggregate information from multiple receptive field sizes. Deeplabv2 [4] and Deeplabv3 [5] both adopt atrous spatial pyramid pooling (ASPP) to concatenate features from multiple atrous convolution layers with different dilation rates arranged in parallel.

Besides the above popular multi-branch structures, there are some other works for exploiting spatial contexts to have different receptive field sizes. Peng et al. [30] enlarge the kernel size with a decomposed structure for global convolution. Zhang et al. [47] propose a scale-adaptive convolution to acquire flexible-size receptive fields by adding a scale regression layer. Lin, et al. [10, 22] utilize the encoder-decoder structure to fuse mid-level and high-level features. DAG-RNN [33] employs a recurrent neural network to capture the contextual dependencies over local features. EncNet [46] introduces a global encoding layer to learn whether some category of object/stuff exists and enforce this as channel attention over the score maps.

In considering the simple structure yet superior performance of ASPP [4, 5], our design follows the multi-branch structure as ASPP. However, for the set of works [4, 5, 48], there is a lack of content adaptive interaction among the different scaled features. In this paper, we enable this by introducing a contextual scale interaction (CSI) module.

Non-Local Mean. Non-local mean has proven effective for many tasks, such as image denoising [2, 9], and texture synthesis [11]. It calculates the feature at one position as a weighted sum of all other positions to exploit spatial context. Wang et al. extend the
We propose a Content-Adaptive Scale Interaction Network (CaseNet) we extend the attention to be both channel and spatial adaptive to weight is shared across all channels. Weights are normalized by Attention. Attention that aims to enhance important features weights that control the amount of information passing.

At the same time, the feature of boat may be more important than the features from different scales can provide useful contextual information to the current scale. For example, a receptive field being a little larger than the boat size may involve the surrounded water. Thus, a mixed receptive field, as illustrated in Fig. 3 which allocates large importance levels near the center and smaller ones over places far from the center, is more ideal. Fortunately, we can achieve this by simply enabling the interaction among multi-scale features and weighted combine the feature responses come from different receptive field. We propose a Contextual Scale Interaction (CSI) module which performs the interaction among the multi-scale features and weighted combine the feature responses come from different receptive field. We propose a Contextual Scale Interaction (CSI) module which performs the interaction among scales by scale-aware non-local operations. For scene parsing, an attention mechanism is proposed that enable the interaction among different feature scales to exploit the scale contexts. Specially, we leverage the scale prior, i.e., scale/branch index, as part of the features for determining the weights that control the amount of information passing.

**Attention.** Attention that aims to enhance important features and suppress irrelevant features are widely used in many tasks [1, 7, 14, 35, 41, 42]. Hu et al. [14] introduce a channel-wise attention mechanism through a squeeze-and excitation (SE) block to modulate the channel-wise feature responses for image classification. For scene parsing, an attention mechanism is proposed that learns to weight the multi-scale features at each pixel location on score maps [6, 16]. For a spatial position at a scale, each attention weight is shared across all channels. Weights are normalized by SoftMax/Gumbel Softmax across scales.

Our scale adaptation (SA) can be viewed as an attention, but ours is different from them in two aspects. First, we relax the sum 1 constraint over the attention weights (across the scales). Second, we extend the attention to be both channel and spatial adaptive to enable the network to flexibly select the appreciate scales.

### 3 PROPOSED CASENET

We propose a Content-Adaptive Scale Interaction Network (CaseNet) to efficiently exploit the multi-scale features for scene parsing. Fig. 2 illustrates the overall framework. For a given input image, the ASPP module generates multi-scale features (e.g. $K = 5$ scales) from the feature map extracted by a backbone FCN. For the multi-scale features, the proposed contextual scale interaction (CSI) module (see subsection 3.1) facilitates the mutual interaction among scales for refining features. Over the refined multi-scale features, the proposed scale adaptation (SA) module (see subsection 3.2) adaptively determines suitable scales for channels and spatial positions to fuse multi-scale features. The fused feature is led to a convolution layer followed by SoftMax for prediction. Note that the entire network is end-to-end trained.

#### 3.1 Contextual Scale Interaction (CSI) Module

The features from different scales can provide useful contextual information to the current scale. For example, a receptive field being a little larger than the boat size may involve the surrounded water. Such context could help reduce the ambiguity of being car or boat. At the same time, the feature of boat may be more important than that with mixed water feature. Thus, a mixed receptive field, as
where $\theta_k(x^{(i,j)}_k, k)$, $\phi_m(x^{(i,j)}_m, m)$ represent two individual embedding functions implemented by two $1 \times 1$ convolutional layers followed by batch normalization (BN) and ReLU activation. For each spatial position $(i, j)$, from $x^{(i,j)}_k$, $x^{(i,j)}_m$ with $k, m \in \{1, \cdots, K\}$, we can obtain an affinity matrix $G^{(i,j)}$ of $K \times K$ for guiding the interaction. It is worth noting that we embed the prior information of scale index $k$ to better learn the affinity among features of various scales.

Then, we obtain five refined feature maps (tensors), $\{\hat{X}_1, \cdots, \hat{X}_5\}$ as the refined feature map with $\text{size} = 5 \times C$ attention values (best viewed in color).

3.2 Scale Adaptation (SA) Module

We propose a spatial and channel adaptive scale adaptation (SA) module. This enables the adaptive selection of the appropriate receptive fields/scales for different sized objects/stuffs. As illustrated in Fig. 5, the refined feature maps (of $\{\hat{X}_1, \hat{X}_2, \hat{X}_3, \hat{X}_4, \hat{X}_5\}$ with $\hat{X}_k \in \mathbb{R}^{H \times W \times C}$ of width $W$, height $H$, and $C$ channels) from the CSI module is the input of SA module. For each spatial position $(i, j)$, we concatenate the $K$ scale features to derive the attention over each channel of each scale, i.e. $5 \times C$ attention values.

For each spatial position $(i, j)$, we represent the feature vectors from the five scale feature maps as $s^{(i,j)}_k = X_k(i,j) \in \mathbb{R}^C$, where $k = 1, \cdots, 5$. We concatenate the feature vectors to have $s^{(i,j)} = [s_1^{(i,j)}, s_2^{(i,j)}, \cdots, s_5^{(i,j)}]$. Specifically, the channel adaptive scale attention vector $a^{(i,j)}_k$ over the $k^{th}$ refined feature map $\hat{X}_k$ at position $(i, j)$ is obtained by:

$$a^{(i,j)}_k = \text{Sigmoid}(W_k \text{ReLU}(W_s s^{(i,j)})), \quad (3)$$

where $W$ and $W_k$ are implemented by 1x1 convolution followed by batch normalization. $W$ shrinks the channel dimension by a rate $r$ (we experimentally set it to $r = 4$), and $W_k$ transforms the channel dimension to $C$. Then, the attention element-wisely modulates the channel dimensions for each scale.

Similar idea but with shared channel attention has been proposed in [6, 16] to spatially adaptively determine the suitable feature scales. However, they all constrain the sum of the attention weights across scales to be 1, e.g., using SoftMax activation function. We relax the sum 1 constraint on the attention weights by using Sigmoid activation function. This provides more flexible optimization space. Intuitively, since the sum of the attention weights in a spatial position is not restricted to be 1, the energy on different spatial position can vary and this is equivalent to play a role of spatial attention. Further, to be more flexible, our scale attention is also channel adaptive.

4 EXPERIMENTS

To evaluate the proposed CaseNet, we carry out comprehensive experiments on three widely-accepted scene parsing datasets including Cityscapes [8], ADE20K [51] and LIP [12]. In the following sections, we will first introduce three datasets and the implementation details, respectively. Then we perform extensive ablation studies on the Cityscapes dataset to demonstrate the effectiveness of our designs. Finally, we further evaluate CaseNet’s performance on ADE20K [51] and LIP [12].

4.1 Datasets

Cityscapes. The dataset is comprised of a large, diverse set of high-resolution (1024x2048) images recorded in streets from 50 different cities, where 5000 of these images have high quality pixel-level labels of 19 classes and are finely annotated. Following the standard setting of Cityscapes, the finely annotated 5000 images are divided into 2975, 500, and 1525 images for training, validation, and testing. We do not use coarse data (20000 coarsely annotated images) in our experiments.

ADE20K. The dataset has been used in ImageNet scene parsing challenge 2016, including 150 classes and diverse scenes with 1038 image-level labels, which is divided into 20K/2K/3K images for training, validation, and testing, respectively.

LIP. The dataset has been used in the Look into Person (LIP) challenge 2016 for single human parsing, including 50,462 images with 19 semantic human part classes and 1 background class.

4.2 Implementation Details

We employ a pre-trained FCN (ResNet) [35] with the dilated strategy [4] as the backbone to extract the feature map. To capture objects of different scales, we further organize five dilated convolutional branches in a parallel fashion, which provides five feature maps with each characterized by a different scale/receptive field. Referring to ASPP [4, 5], we set the dilation rates for the five branches as $d = 1, 6, 12, 24, 36$, respectively. We implement our method based on Pytorch, and all experiments are performed on 4x Tesla P40 GPUs.

4.2.1 Cityscapes. Referring to the public settings in previous work, we set the initial learning rate as 0.01 and weight decay as 0.0005. The original image size is 1024x2048 and we crop it to 769x769 following PSPNet [48]. Training batch size is set to 8 and we train our model with InPlaceABNSync [32] to synchronize the mean and standard-deviation of BN across multiple GPUs in all the experiments. We employ 40K/80k training iterations when training...
Table 1: Ablation study on Cityscapes validation set. CSI represents the contextual scale interaction (CSI) module, SA represents the scale adaptation (SA) module, and CaseNet represents the complete version with both CSI and SA.

| Methods                        | Train. mIoU (%) | Val. mIoU (%) |
|--------------------------------|-----------------|---------------|
| ResNet-101 Baseline            | 83.75           | 74.82         |
| ResNet-101 + ASPP [5]          | 85.91           | 78.51         |
| ResNet-101 + ASPP + CSI        | 87.74           | 80.17         |
| ResNet-101 + ASPP + SA         | 87.66           | 80.08         |
| CaseNet                        | **88.33**       | **81.04**     |

without/with validation set. Similar to the previous works ASPP [4] and DenseASPP [43], we also employ the poly learning rate policy where the learning rate is multiplied by $(1 - \frac{iter}{total\_iter})^{0.9}$.

For the data augmentation, we apply random flipping horizontally and random scaling in the range of [0.5, 2]. For loss functions, following [4], we employ class-balanced cross entropy loss on both the final output and the intermediate feature map output, where the weight of the final loss is 1 and the auxiliary loss is 0.4. Following previous works [38], we also adopt the online hard example mining (OHEM) strategy.

4.2.2 ADE20K. We set the initial learning rate as 0.02 and weight decay as 0.0001. The input image is resized to 300, 375, 450, 525, 600 randomly since the images are of various sizes on ADE20K. Training batch size is set to 8 and we also train our model with InPlaceABNSync [32]. We train the model for 100k iterations. By following the previous works [5, 48], we employ the same poly learning rate policy and data augmentation and employ a supervision in the intermediate feature map.

4.2.3 LIP. Following the CE2P [24], we set the initial learning rate as 0.007 and weight decay as 0.0005. The original images are of various sizes and we resize all of them to 473x473. Training batch size is set to 40 and we also train our model with InPlaceABNSync [32]. We train the model for 110k iterations. The poly learning rate policy, data augmentation methods and deep supervision in the intermediate feature map are consistent with the experiments on Cityscapes and ADE20K.

4.3 Ablation Study

We perform all ablation studies on the Cityscapes datasets. We split this section into three sub-sections for better illustration: (1) CaseNet versus Baseline. (2) Study on the Contextual Scale Interaction (CSI) module. (3) Study on the Scale Adaptation (SA) module.

4.3.1 CaseNet versus Baseline. In order to verify the effectiveness of each module in CaseNet, we perform comprehensive ablation studies on the Cityscapes validation datasets. We use the ResNet-101 Baseline [13] to represent the traditional FCN with the dilated strategy [4], which can be viewed as the “FCN” in our pipeline (Fig. 2). To ensure the fairness, the ResNet-101 + ASPP scheme follows DeepLabv3 [5] with some modifications in our designs: we remove the original image-level pooling branch, and employ five 3x3 dilated convolution branches with dilation rates of 1, 6, 12, 24 and 36, respectively (as shown in Fig. 2).

Because our key contributions lie in the contextual scale interaction (CSI) module and scale adaptation (SA) module, we verify their effectiveness separately: ResNet-101 + ASPP + CSI and ResNet-101 + ASPP + SA. The complete version, i.e. ResNet-101 + ASPP + CSI + SA, is abbreviated as CaseNet. For quantitative evaluation, mean of class-wise Intersection over Union (mIoU) is used.

Experiment results are reported in Table 1, where all the results are based on single scale testing. The performance of ResNet-101 + ASPP is comparable to the numbers in the original DeepLabv3 paper [5]. We make the following observations.

Our contextual scale interaction (CSI) module and scale adaptation (SA) module both significantly improve performance over other two powerful baselines. ResNet-101 + ASPP + CSI outperforms the ResNet-101 Baseline and ResNet-101 + ASPP by 5.35% and 1.66% respectively in mIoU, which verifies the effectiveness of the feature interaction among scales to exploit scale contextual information. In addition, ResNet-101 + ASPP + SA outperforms the ResNet-101 Baseline and the ResNet-101 + ASPP by 5.26% and 1.57% respectively in mIoU, which verifies the effectiveness of the spatial and channel adaptive scale attention.

We find that we can further improve performance by combining the CSI and SA modules together. For example, the complete version CaseNet achieves the best **81.04%** on the validation set based on single scale testing and improves by **0.87%** over ResNet-101 + ASPP + CSI and **0.96%** over ResNet-101 + ASPP + SA.

4.3.2 Study on Contextual Scale Interaction (CSI) Module. The influences of Contextual Scale Interaction (CSI) module can be observed in Fig. 6. We observe that some inconspicuous objects are well-captured and object boundaries are clearer with the CSI module included, such as the ‘fence’ in the first row and the ‘pole’ in the second row, which demonstrates that the contextual interaction over multi-scale features enhances the discrimination of details.

Meanwhile, some salient objects exhibit their intra-class consistency because some misclassified regions are now correctly classified with the CSI module, such as the ‘truck’ in the third row and the ‘bus’ in the fourth row, which further verifies that the feature interaction of CSI is powerful and can promote the content adaptation.
Figure 7: Visualization of scale attention (SA) map on Cityscapes validation set.

Table 2: Influence of the scale prior (scale index) in the CSI module. For fairness of comparison, we implement different designs on top of our baseline ResNet-101 + ASPP.

| Methods                  | Train. mIoU (%) | Val. mIoU (%) |
|--------------------------|-----------------|---------------|
| CSI without scale index  | 86.68           | 79.51         |
| CSI with scale index (Ours) | 87.74         | 80.17         |

Influence of Scale Index. The scale index (branch index with each corresponding to different dilation rate) in our CaseNet acts as an important scale prior to explicitly represent its corresponding receptive-field size. We verify its effectiveness in Table 2.

Sharing Embedding or Not? We have compared the individual embedding and the shared embedding function among scales which is used for calculating the affinities within the CSI module. We find that the former achieves about 0.45% gain over the latter on the validation set. We note that features of different scales should be embedded into different high dimensional feature spaces.

4.3.3 Study on Scale Adaptation (SA) Module. In this subsection, extensive experiments are designed to study how different designs of the Scale Adaptation (SA) module influence performance quantitatively. We perform the studies on the Cityscapes’ validation set. For fairness of comparison, we implement different designs on top of our baseline ResNet-101 + ASPP.

Proposed Scale Adaptation versus Other Attentions. Table 3 compares the performance of our scale adaptation (SA) modules with the other attention designs. (1) Channel Attention Alone. Following the attention design in Squeeze-and-Excitation module (SE) [14], we use spatially global average-pooled features to compute channel-wise attention for the five branches, by using two fully connected (FC) layers with the non-linearity. In comparison with SE, our SA achieves 1.33% gain in mIoU, which demonstrates the importance of the spatially-adaptive attention for scene parsing. (2) Spatial Attention Alone. For using only spatial attention (Spatial-A) designs, we change our attention output from the original C dimensions to 1 dimension. In comparison with Spatial-A, our SA achieves 0.58% gain in mIoU. By using both spatial and channel attention across scales, our final scheme achieves the best performance.

Table 3: Performance comparisons of our Scale Adaptation (SA) with other attention designs on the validation set of Cityscapes. We implement different attention designs on top of our baseline ResNet-101 + ASPP for fair comparison.

| Methods                  | Train. mIoU (%) | Val. mIoU (%) |
|--------------------------|-----------------|---------------|
| ResNet-101 + ASPP        | 85.91           | 78.51         |
| ResNet-101 + ASPP + SE   | 86.13           | 78.75         |
| ResNet-101 + ASPP + Spatial-A | 86.71       | 79.5          |
| ResNet-101 + ASPP + SA   | 87.66           | 80.08         |

Table 4: Influence of relaxing the constraint on attention values in the SA module. We implement the different designs on top of our baseline ResNet-101 + ASPP for fair comparison.

| Methods                  | Train. mIoU (%) | Val. mIoU (%) |
|--------------------------|-----------------|---------------|
| SA with Softmax          | 86.82           | 79.56         |
| SA with Sigmoid (Ours)   | 87.66           | 80.08         |

Visualization. In order to further understand our Scale Adaptation (SA) module intuitively, we randomly choose some examples from the validation set of Cityscapes and visualize the Scale Attention (SA) map in Fig. 7. For each branch in CaseNet, the overall scale attention map has a size of \(H \times W \times C\) because it is spatial and channel adaptive, so we show the average scale attention map along the channel dimension to see whether they could capture different scaled objects at different spatial position.

As illustrated in Fig. 7, we can find that each branch has different degrees of response to the objects of different scales. For example, for the first and second branch with dilation rates being 1 and 6, their scale attention maps “Attention map #1” and “Attention map #2” mainly focus on small inconspicuous objects and object boundaries, such as the ‘traffic sign’ in the first row and second row. With the increase of the dilation rates, the third and fourth branch with dilation rates being 12 and 24 both have larger receptive-fields, so their corresponding scale attention maps “Attention map #3” and “Attention map #4” mainly focus on salient objects, such as the ‘train, sidewalk’ in the first row and the ‘terrain’ in the second row.
Table 5: Category-wise comparisons with state-of-the-art approaches on the Cityscapes test set. CaseNet outperforms existing methods and achieves 81.9% in Mean IoU.

| Methods          | Mean IoU (%) |
|------------------|--------------|
| road             | 97.4         |
| sidewalk         | 78.4         |
| building         | 89.2         |
| wall             | 34.9         |
| fence            | 44.2         |
| pole             | 47.4         |
| traffic light    | 60.1         |
| traffic sign     | 65           |
| vegetation       | 91.4         |
| terrain          | 69.3         |
| sky              | 93.9         |
| person           | 77.1         |
| rider            | 51.4         |
| car              | 92.6         |
| truck            | 35.3         |
| bus              | 48.6         |
| train            | 46.5         |
| motorcycle       | 51.6         |
| bicycle          | 66.8         |

Table 6: Comparisons with state-of-the-art approaches on the validation set of the ADE20K dataset.

| Methods          | Publication | Backbone | Val. mIoU (%) |
|------------------|-------------|----------|---------------|
| RefineNet [22]   | CVPR2017    | ResNet-101 | 40.20         |
| RefineNet [22]   | CVPR2017    | ResNet-152 | 40.70         |
| PSPNet [48]      | CVPR2017    | ResNet-101 | 43.29         |
| PSPNet [48]      | CVPR2017    | ResNet-152 | 43.51         |
| PSPNet [48]      | CVPR2017    | ResNet-269 | 44.94         |
| SAC [47]         | ICCV2017    | ResNet-101 | 44.30         |
| PSANet [49]      | ECCV2018    | ResNet-101 | 43.77         |
| UperNet [40]     | ECCV2018    | ResNet-101 | 42.66         |
| DSSPN [21]       | CVPR2018    | ResNet-101 | 43.68         |
| EncNet [46]      | AAAI2019    | ResNet-101 | 44.65         |
| CaseNet          | –           | ResNet-101 | 45.28         |

Table 7: Comparisons with state-of-the-art approaches on the validation set of the LIP dataset.

| Methods          | Publication | Backbone | Val. mIoU (%) |
|------------------|-------------|----------|---------------|
| Attention+SSL [12] | CVPR2017    | ResNet-101 | 44.73         |
| JPPNet [20]      | TPAMI2018   | ResNet-101 | 51.37         |
| SS-NAN [50]      | CVPR2017    | ResNet-101 | 47.92         |
| MMAN [26]        | ECCV2018    | ResNet-101 | 46.81         |
| MuLA [27]        | ECCV2018    | ResNet-101 | 49.30         |
| CE2P [24]        | AAAI2019    | ResNet-101 | 53.10         |
| CaseNet          | –           | ResNet-101 | 54.38         |

4.4 Comparison with State-of-the-Art

Results on Cityscapes Dataset. We compare our method with existing methods on the Cityscapes test set. Specifically, we train our CaseNet with only finely annotated data (including validation set for training) for 80k training iterations and submit our test results (with multi-scale [0.75x, 1.0x, 1.25x] testing strategy) to the official evaluation server. Results are shown in Table 5. We observe that CaseNet significantly outperforms existing approaches. In particular, our model outperforms DenseASPP by 1.3% in mIoU, even though DenseASPP uses a more powerful and more complex pretrained DenseNet model [15] as the backbone network.

Results on ADE20K Dataset. We further carry out experiments on the ADE20K dataset to evaluate the effectiveness of our method. Comparisons with previous state-of-the-art methods on the ADE20K validation set are shown in Table 6. Results show that our CaseNet achieves the best performance of 45.28% in Mean IoU.
Our CaseNet outperforms the ResNet-101 based state-of-the-art method EncNet by 0.63% in mIoU, and improves over the ResNet-269 based PSPNet by 0.34% in mIoU.

Results on LIP Dataset. We also conduct experiments on the LIP dataset. Comparisons with previous state-of-the-art methods are reported in Table 7. We observe that the proposed CaseNet achieves 54.38% in mIoU, which outperforms previous methods by a large margin, i.e. 1.28%. This indicates that the proposed CaseNet does well on the human parsing task.

5 CONCLUSION
In this paper, we propose a simple yet effective Content-Adaptive Scale Interaction Network (CaseNet) to adaptively exploit multi-scale features through contextual interaction and adaptation. Specifically, we build the framework based on the classic Atrous Spatial Pyramid Pooling (ASPP), followed by the proposed contextual scale interaction (CSI) module which enables feature interaction among scales, and a spatial and channel adaptive scale adaptation (SA) module to facilitate the appropriate scale selection. Our ablation studies demonstrate the effectiveness of the proposed CSI and SA modules, leading to more precise parsing results. In addition, CaseNet achieves state-of-the-art performance on Cityscapes, ADE20K, and LIP datasets.

REFERENCES
[1] Dmitriy Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014).
[2] Antoni Buades, Bartomeu Coll, and J-M Morel. 2005. A non-local algorithm for image denoising. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2. 60–65.
[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2018. Semantic image segmentation with deep convolutional nets and fully connected crfs. International Conference on Learning Representations.
[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2018. Deeplab: Semantic image segmentation with deep convolutional nets and atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence 40, 4 (2018), 834–848.
[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. 2017. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017).
[6] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. 2016. Attention to scale: Scale-aware semantic image segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition. 3213–3223.
[7] Jan K Chorowski, Dmitriy Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. In NeurIPS. 577–585.
[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE conference on computer vision and pattern recognition. 3213–3223.
[9] Kostadim Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. 2007. Image Denoising by Sparse 3-D Transform-Domain Collaborative Filtering. Transactions on image processing 16, 8 (2007), 2080–2095.
[10] Hengshu Ding, Xiaodong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. 2018. Context contrasted feature and gated multi-scale aggregation for scene segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2393–2402.
[11] Alexei A Efros and Thomas K Leung. 1999. Texture synthesis by non-parametric sampling. In IEEE International Conference on Computer Vision. Vol. 2. 1083–1088.
[12] Ke Gong, Xiaodan Liang, Dongyin Zhang, Xiaohui Shen, and Liang Lin. 2017. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 932–940.
[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition. 770–778.
[14] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition. 7312–7319.
[15] Guo Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Deeply connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition. 4700–4708.
[16] Shu Kong and Charless Fowlkes. 2019. Pixel-wise Attentional Gating for Scene Parsing. In IEEE Winter Conference on Applications of Computer Vision. 1024–1033.
[17] Shu Kong and Charless C Fowlkes. 2018. Recurrent scene parsing with perspective understanding in the loop. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 956–965.
[18] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2. 2169–2178.
[19] Xiangyang Li, Xinhong Song, Luis Herranz, Yaohui Zhu, and Shuqiang Jiang. 2016. Image captioning with both object and scene information. In Proceedings of the 24th ACM international conference on Multimedia. ACM, 1107–1110.
[20] Xiaodan Liang, Ke Gong, Xiaobin Shen, and Liang Lin. 2018. Look into person: Joint body parsing & pose estimation network and a new benchmark. IEEE transactions on pattern analysis and machine intelligence (2018).
[21] Xiaodan Liang, Hongjie Zhou, and Eric Xing. 2018. Dynamic-structured semantic propagation network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 752–761.
[22] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. 2017. Refinedet: Multi-path refinement networks for high-resolution semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition. 1925–1934.
[23] Daqing Liu, Zheng-Jun Zha, Hanwang Zhang, Yongdong Zhang, and Feng Wu. 2018. Context-Aware Visual Policy Network for Sequence-Level Image Captioning. In 2018 ACM Multimedia Conference on Multimedia Conference. ACM, 1436–1442.
[24] Ting Liu, Tao Ruan, Zilong Huang, Yunchao Wei, Shikui Wei, Yao Zhao, and Thomas Huang. 2018. Devii in the details: Towards accurate single and multiple human parsing. arXiv preprint arXiv:1809.05996 (2018).
[25] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition. 3431–3440.
[26] Yawei Luo, Zherdong Zheng, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. 2018. Macro-micro adversarial network for human parsing. In Proceedings of the European Conference on Computer Vision. 418–434.
[27] Xuexing Nie, Jiushi Feng, and Shuicheng Yan. 2018. Mutual Learning to Adapt for Joint Human Parsing and Pose Estimation. In Proceedings of the European Conference on Computer Vision. 502–517.
[28] Yoon Jung Park, Yoonai Kyu, Hyocheol Ro, JungHyun Byun, Seougho Chea, and Tack Don Han. 2018. Meet AR-bot: Meeting Anywhere, Anytime with Movable Spatial AR Robot. In 2018 ACM Multimedia Conference on Multimedia Conference. ACM, 1295–1299.
[29] Yoon Jung Park, Yoonai Kyu, Hyocheol Ro, Jinwoun Cha, Kyunmi Kim, and Tack Don Han. 2018. ChildAR-bot: Educational Playing Projection-based AR Robot for Children. In 2018 ACM Multimedia Conference on Multimedia Conference. ACM, 1278–1282.
[30] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. 2017. Large Kernel Matters–Improve Semantic Segmentation by Global Convolutional Network. In Proceedings of the IEEE conference on computer vision and pattern recognition. 4353–4361.
[31] Tobias Pohlen, Alexander Hermans, Markus Matias, and Bastian Leibe. 2017. Full-resolution residual networks for semantic segmentation in street scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 4151–4160.
[32] Samuel Rota Bulò, Lorenzo Porzi, and Peter Kontschieder. 2018. In-place activated batchnorm for memory-optimized training of dnns. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 5639–5647.
[33] Bing Shuai, Zhen Zuo, Bing Wang, and Gang Wang. 2018. Scene segmentation with dag-recurrent neural networks. IEEE transactions on pattern analysis and machine intelligence 40, 6 (2018), 1480–1493.
[34] Marvin Teichmann, Michael Weber, Marius Zoellner, Roberto Cipolla, and Raquel Urtasun. 2018. Multinet: Real-time joint semantic reasoning for autonomous driving. In 2018 IEEE Intelligent Vehicles Symposium (IV). IEEE, 1013–1020.
[35] Fei Wang, Mengjiao Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. 2017. Rethinking atrous convolution for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition. 3156–3164.
[36] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Xiaoou Tang. 2017. Residual attention network for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 5639–5647.
[38] Zifeng Wu, Chunhua Shen, and Anton van den Hengel. 2016. High-performance semantic segmentation using very deep fully convolutional networks. arXiv preprint arXiv:1604.04339 (2016).

[39] Zifeng Wu, Chunhua Shen, and Anton Van Den Hengel. 2019. Wider or deeper: Revisiting the resnet model for visual recognition. Pattern Recognition 90 (2019), 119–133.

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

[41] Huijuan Xu and Kate Saenko. 2016. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In European Conference on Computer Vision. Springer, 451–466.

[42] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In ICML. 2048–2057.

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

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

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

[46] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyou Zhang, Xiaogang Wang, Ambesh Tyagi, and Amit Agrawal. 2018. Context encoding for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 7151–7160.

[47] Ruix Zhang, Sheng Tang, Yongdong Zhang, Jintao Li, and Shuicheng Yan. 2017. Scale-adaptive convolutions for scene parsing. In Proceedings of the IEEE International Conference on Computer Vision. 2031–2039.

[48] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. 2017. Pyramid scene parsing network. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2881–2890.

[49] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. 2018. Pspnet: Point-wise spatial attention network for scene parsing. In Proceedings of the European Conference on Computer Vision. 267–283.

[50] Jian Zhao, Jiashu Li, Xuecheng Nie, Fang Zhao, Yunpeng Chen, Zhecan Wang, Jiashi Feng, and Shuicheng Yan. 2017. Self-supervised neural aggregation networks for human parsing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 7–15.

[51] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene parsing through ade20k dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 633–641.