Multi-Receptive Atrous Convolutional Network for Semantic Segmentation

Mingyang Zhong and Brijesh Verma
Centre for Intelligent Systems
Central Queensland University
Brisbane, Australia
{m.zhong, b.verma}@cqu.edu.au

Joseph Affum
Transport Safety
Australian Road Research Board (ARRB)
Brisbane, Australia
joseph.affum@arrb.com.au

***Abstract*—Deep Convolutional Neural Networks (DCNNs) have enhanced the performance of semantic image segmentation but many challenges still remain. Specifically, some details may be lost due to the downsampling operations in DCNNs. Furthermore, objects may appear in an image at different scales, and extracting features using convolutional filters with large sizes is costly in computation. Moreover, in many cases, contextual information, such as global and background features, is potentially useful for semantic segmentation. In this paper, we address these challenges by proposing a Multi-Receptive Atrous Convolutional Network (MRACN) for semantic image segmentation. The proposed MRACN captures the multi-receptive features and the global features at different receptive scales of the input. MRACN can serve as a module easily being integrated into existing models. We adapt the ResNet-101 model as the backbone network and further propose a MRACN segmentation model (MRACN-Seg). The experimental results demonstrate the effectiveness of the proposed model on two datasets: a benchmark dataset (PASCAL VOC 2012) and our industry dataset.**

I. Introduction

Deep convolutional neural networks [1] have improved the performance of semantic image segmentation task that performs pixel-level classification in an image [2]–[5]. Many DCNN based models have been designed such as Full Convolutional Networks (FCNs) [2], and they usually employ a pre-trained classification network and output a probability map for categorizing every pixel of an input image. In DCNNs, consecutive downsampling operations are employed to reduce image resolution, which makes very deep architectures feasible [6]–[8]. However, these repeated combinations of the downsampling operations significantly reduce the spatial resolution of the feature maps, resulting the loss of the informative details that are potentially useful for semantic segmentation.

To solve the resolution reduction problem, many approaches have been proposed. One type of approaches is to recover the resolution by the upsampling operations or the deconvolutional layers that generate feature maps with high resolution [2], [3], [9]. This upsampling process restores the feature maps to the input resolution for classification at pixel level. However, the parameter size of the network cascaded with the deconvolutional and the unpooling layers is doubled, compared with the original convolutional network. For example, Long et al. [2] and Wang et al. [9] use the pre-trained VGG16 network [7] to generate the initial parameters. Furthermore, directly extending these networks to deeper architectures usually degrades their performance [8]. Another type of approach is to keep the input resolution unchanged by atrous convolutions, also known as dilated convolutions [4], [10], [11]. In atrous convolution, filters with large dilation rates are used to enlarge the receptive field while remaining the spatial resolution.

Another problem to semantic image segmentation is that objects exist in images at different scales. A natural idea is to enlarge the receptive field by stacking more convolutional layers or using convolutional filters with larger sizes for capturing contextual information at different receptive scales. As the term "multi-scale" is also used in methods that rescale the input images, such as [12], in this paper, we use "multi-receptive" to refer different sizes of receptive fields by changing filter sizes. Based on the convolutional filters with different sizes, the inception module [13] and the spatial pyramid pooling [14] are two widely adopted modules by many recent works for extracting multi-receptive features such as [8], [15], [16]. However, it incurs high computational cost when filters with large sizes are used. Chen et al. [4] also adopt this idea in their work using multiple atrous convolutions. In addition, recent works exploit the global/background information of the input images to enhance feature extraction processes [11], [17], [18].

In this work, we address the aforementioned problems together and propose a multi-receptive atrous convolutional network for semantic image segmentation. Inspired by the atrous/dilated convolution [4], [10], [11] and the inception module [13], in the proposed MRACN, we employ a combination of atrous convolutions for extracting the multi-receptive feature maps. These atrous convolutions with different dilation rates are able to capture the contextual information of the input map at different receptive scales without reducing the resolution. More importantly, we exploit the global features of an input map also in a "multi-receptive" fashion, rather than extracting only one global feature from the input feature map [11], and the underlay reasons are twofold. On one hand, intuitively, some objects would only exist in certain background. For example as shown in Fig. 1, 110 speed limit signs can only appear in highway scenes, while vehicles would have a large chance to be located on roads rather than on roadside grass. On the other hand, multiple global feature

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

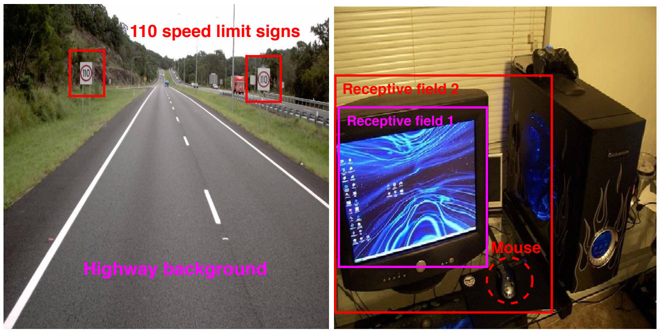


Fig. 1: Examples of object scale and its relation to the background.

maps extracted from each multi-receptive feature maps are potentially helpful for semantic segmentation. For example, a computer screen would be sit on a desk with a mouse when increasing the receptive field, and the global features extracted from the multi-receptive features are able to capture the contextual information of the related objects regarding the computer screen in the global context. Furthermore, the proposed MRACN consumes significantly small parameter space than the inception module with the similar structure, and can serve as a module being easily integrated into existing models. Finally, we adapt the ResNet-101 model [8] as the backbone network and propose a MRACN segmentation model for semantic image segmentation. We evaluate the performance of the proposed models on PASCAL VOC 2012 dataset [19] and our industry dataset to demonstrate the effectiveness of the proposed models. The main contributions of this paper are listed as follows:

- We propose a MRACN that captures the multi-receptive features and the global features at different receptive scales of the input.
- We propose a MRACN-Seg for semantic image segmentation that integrates MRACN.
- We perform a thorough evaluation on a benchmark dataset and our industry dataset in comparison with the state-of-the-art methods, and the experimental results demonstrate the effectiveness of the proposed models.

In the following section, we review the most relevant state-of-the-art methods. Section III presents the proposed multi-receptive atrous convolutional network for semantic image segmentation. In Section IV, we detail the experiments and report the results. Finally, Section V concludes the paper.

II. Related Work

In the previous decade, semantic segmentation relied on hand-crafted features and flat classifiers, such as Support Vector Machines [20] and Random Forests [21]. With the advances of deep learning in recent years, DCNN based approaches achieved significant improvements on computer vision tasks. In this section, we review the recent DCNN based approaches.

DCNNs have demonstrated to be powerful to extract dense features from images for many computer vision tasks, such as image classification [7], [8], [15], [22], object detection [23]–[25] and semantic segmentation [26], [27]. For example, U-Net [26] exploits multi-level features by using the contracting path and the expanding path with skip connections. SegNet [27] uses a stack of the encoders, that are pipelines containing e.g. convolutions and poolings, and the decoders that upsample the encoded feature maps to obtain pixel-wise labelling. DCNN based approaches use repeated combinations of pooling operations that extract features with reduced resolutions. When the network goes deeper, the resolution can be reduced dramatically. Although the resolution can be restored by deconvolutions [2], [3], [28], the informative details of images are lost.

Atrous convolution [4], [10], [11], that inserts “holes” between filter weights for allowing the filter with a specified size to sample the field with a larger size than that with the specified size, has been proposed to address the resolution reduction problem. In DeepLab [4], Chen et al. replace the downsampling operations in last two stages of DCNNs by atrous convolutions. The advance of DeepLab makes it prevalent, and many following works [11], [29], [30] have been proposed based on atrous convolution. For example, Dai et al. [29] propose a deformable convolution network that generalizes the atrous convolution with additional offsets. Similar to [11], Wang et al. [30] propose a hybrid dilated convolution framework that enlarges the receptive field to aggregate contextual information.

Multi-scale, multi-receptive and global features are beneficial to semantic segmentation, as the rich contextual information can be obtained from them. Furthermore, integrating these features can make the deeper processing stages robust to scale changes [13]. Many models have been proposed to obtain the multi-scale features. For example, Chen et al. [12] propose an attention mechanism that learns the shared weights from the input images at multiple scales. Similarly, RefineNet [31] uses multiple paths to extract features from images with different resolutions and then generates high-resolution semantic features. Pinheiro et al. [32] and Amirul et al. [33] use multi-scale features to refine the performance of their segmentation models. For multi-receptive and global features, two widely-adopted models, the inception module [13] and the spatial pyramid pooling [14], use multiple convolutional filters with different sizes to extract the multi-receptive features and then fuse them by concatenation. However, when the networks become deeper or the sizes of the convolutional filters go larger, the complexity of the model parameter space usually becomes a problem. Chen et al. [4] propose DeepLab in which a atrous spatial pyramid pooling based on atrous convolutions is used to limit the parameter space.

III. Multi-Receptive Atrous Convolution

In this section, we first revisit the atrous convolution, and then present the proposed MRACN and the extraction of the multi-receptive features and the global features. Finally, we

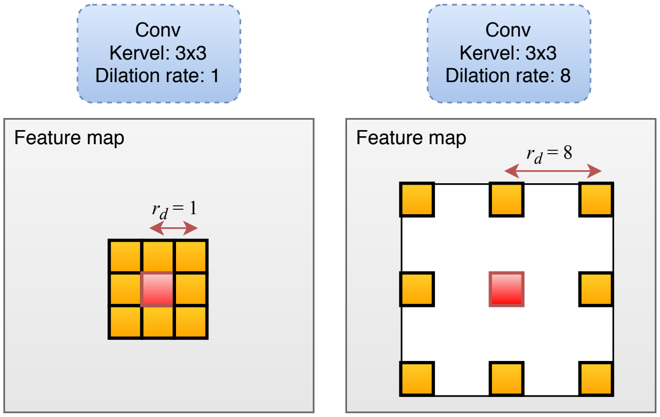


Fig. 2: Illustration of atrous convolution. Kernel size is 3×3 and dilation rates r_ds are 1 and 8. An atrous convolution with $r_d = 1$ is equivalent to a standard convolution.

describe our segmentation model MRACN-Seg in which the proposed MRACN has been integrated.

A. Atrous convolution

Rather than the repeated combinations of convolutions and pooling operations at consecutive layers that significantly reduce the resolution of the feature maps [2], [6], [7] or the convolutional filters with large sizes that increase the parameter space [13], [14], atrous convolution [4], [10], [11] uses the resolution of the feature maps and allows convolve a receptive field with a larger size and without increasing the parameter space. Fig. 2 shows the idea of atrous convolution. Using the same kernel size, atrous convolution enlarges the receptive field by increasing the dilation rate r_d.

Formally, when convolving a two-dimensional input feature map $\mathbf{x}$, for each location l on the output feature map $\mathbf{f}$ and a filter $\mathbf{w}$, atrous convolution is applied over the input feature map $\mathbf{x}$, and the output feature map $\mathbf{f}$ is computed:

$$\mathbf{f}[l] = \sum_{k} \mathbf{x}[l + r_d \cdot k]\mathbf{w}[k] \qquad (1)$$

where k is the size of the filter $\mathbf{w}$. Therefore, by changing the dilation rate r_d, we are able to sample the receptive field at different scales. For example, an atrous convolution with a dilation rate $r_d = 8$ is equivalent to convolving an input with enlarged filters in which 7 zeros $(r_d - 1)$ have been inserted between two consecutive filter values along each spatial dimension. An atrous convolution with $r_d = 1$ is equivalent to a standard convolution.

B. Multi-receptive atrous convolutional network

The architecture of the proposed MRACN is shown in Fig. 4 (a). We employ a 1×1 convolution and multiple atrous convolutions with different dilation rates in a parallel fashion to extract the multi-receptive feature maps. After that, we obtain the global feature maps from each multi-receptive feature map, and then fuse the output feature maps from all branches by concatenation.

The multi-receptive feature maps $\mathbf{f}^{mr}$s are computed by using Eq. 1. The global feature maps are extracted similarly to [11], [17] but rather than extracting one global feature map from the input feature map, we obtain multiple global feature maps at multiple receptive scales. Specifically, we first apply global average pooling on the multi-receptive feature map. Using the i-th multi-receptive feature map $\mathbf{f}_i^{mr}$ with $h \times w \times d$ dimension as an example, the global average $\mathbf{g}_i[j]$ for each $h \times w \times 1$ dimensional map $\mathbf{f}_i^{mr}[j]$ of the i-th multi-receptive feature map $\mathbf{f}_i^{mr}$ can be computed:

$$\mathbf{g}_i[j] = \frac{\sum_{h',w'}^{h,w} \mathbf{f}_i^{mr}[j][h',w']}{h \times w} \qquad (2)$$

where $\mathbf{f}_i^{mr}[j][h',w']$ is the value at the location (h',w') on $\mathbf{f}_i^{mr}[j]$.

Then, bilinear upsampling is applied on the $1 \times 1 \times d$ dimensional output $\mathbf{g}_i$ to obtain the global feature map $\mathbf{f}_i^{g}$ of the i-th multi-receptive feature map $\mathbf{f}_i^{mr}$. Note that, as the $h \times w \times d'$ dimensional input feature map $\mathbf{x}$ of MRACN could be from other pre-trained networks such as ResNet [8], d' can be significantly larger than d. If we directly use the global features $\mathbf{x}^{g}$ with the dimension of $h \times w \times d'$ extracted from $\mathbf{x}$, $\mathbf{x}^{g}$ would dominate the performance of MRACN. Thus, we apply a 1×1 convolution (with d filters and batch normalization [34]) on $\mathbf{x}^{g}$ to generate a dense global feature map $\mathbf{x}^{g'}$ with the desired spatial resolution $h \times w \times d$.

After that, we fuse the output feature maps from all branches by concatenation:

$$\mathbf{f}^{fuse} = \oplus(\mathbf{x}', \mathbf{f}_1^{mr}, ..., \mathbf{f}_n^{mr}, \mathbf{x}^{g'}, \mathbf{f}_1^{g}, ..., \mathbf{f}_n^{g}) \qquad (3)$$

where $\oplus$ is the concatenation operation, and n is number of the atrous convolutions ($n = 3$ in Fig. 4 (a)). Finally, the fused feature map is fed to another 1×1 convolution (with d filters and batch normalization) to generate the final output $\mathbf{f}^{out}$ of the proposed MRACN that captures the multi-receptive features and the global features at multiple receptive scales of the input.

C. MRACN based segmentation

With the proposed MRACN, we further proposed the MRACN segmentation model for semantic image segmentation, as shown in Fig. 3. Specifically, we adapt the ResNet-101 model [8] as our backbone network. Similarly to [4], we find the last "block" of the ResNet-101 model and remove all subsequent layers. Then, we integrate the proposed MRACN with the adapted ResNet in cascade. Finally, the output of the MRACN is passed through the final 1×1 convolution that

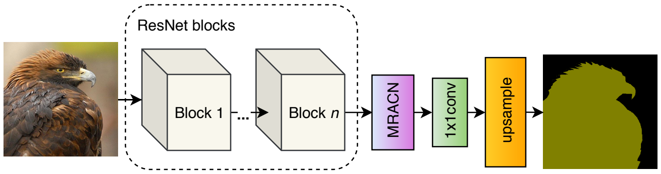
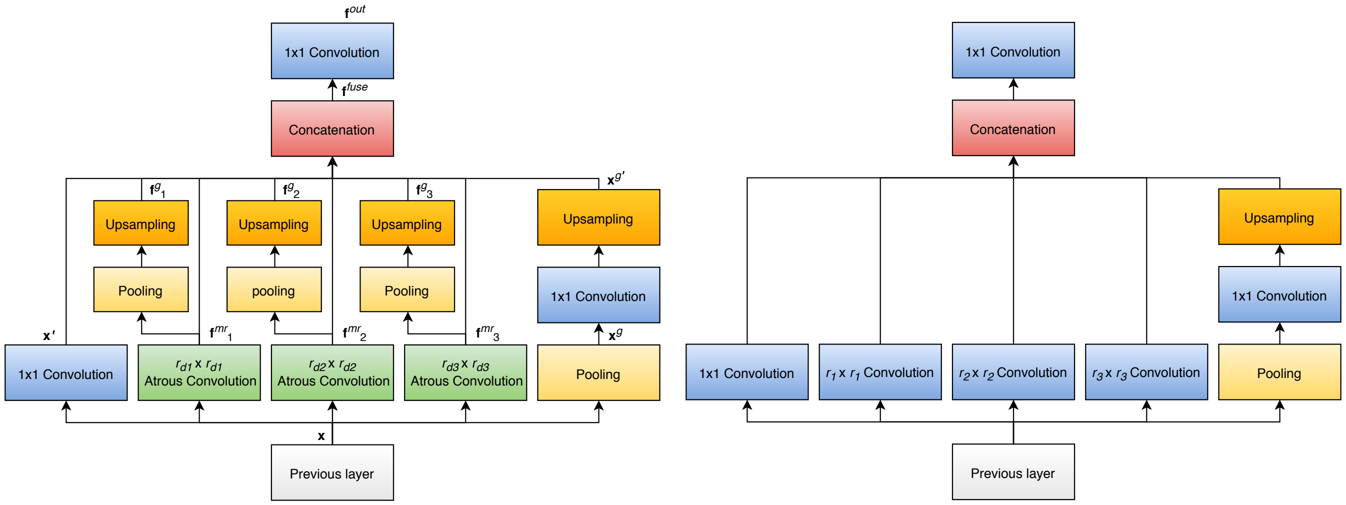


Fig. 3: Architectures of the proposed MRACN-Seg.



(a) MRACN

(b) Inception module [13]

Fig. 4: Architectures of the proposed MRACN and the inception module.

generates the final logits. Step-wise process of the proposed MRACN-Seg is as follows.

- *Step 1*: Feed the train set (original images and the pixel-wise labels) to MRACN-Seg.
- *Step 2*: Pass the train data to ResNet blocks and output the feature maps **x**.
- *Step 3*: Pass the feature maps **x** to MRACN and output the feature maps $\mathbf{f}^{out}$ that capture the multi-receptive features and the global features at different receptive scales of the input.
- *Step 4*: Pass the feature maps $\mathbf{f}^{out}$ to the 1×1 convolution and finally upsample to the original resolution.
- *Step 5*: Back propagate to optimize the model parameters.
- *Step 6*: Save the trained weights of MRACN-Seg.
- *Step 7*: Load the trained model and test with the test set.

Note that, the reasons that we do not adapt other models as the backbone network are twofold. Firstly, we want to keep the simplicity of the network structure to validate the effectiveness of the proposed MRACN. Secondly, our segmentation model MRACN-Seg is similar to replacing the ASPP module in DeepLabv2 [4] by MRACN. Furthermore, the proposed MRACN can be easily integrated into other state-of-the-art models, such as [11] and [5], in cascade or in parallel.

IV. Experiments

In this section, we first analyze the complexity of the proposed MRACN in terms of the parameter space. Then, we evaluate the proposed MRACN-Seg on two datasets: PASCAL VOC 2012 dataset [19] and our DTMR-DVR dataset. Intersection-over-Union (IoU) and averaged IoU across all classes (mIoU) are used to measure the performance of the proposed MRACN-Seg.

A. Complexity analysis of MRACN

In order to analyze the efficiency of the proposed MRACN, we investigate the complexity of the proposed model in terms of the parameter space. To conduct fair comparison, we compare MRACN demonstrated in Fig. 4 (a) with the inception module with the similar structure shown in Fig. 4 (b). We use the *same* convolutions with 256 filters for all convolutions, and set the receptive fields of the convolutions in the two models at the same scales ($r_{d1} = r_1 = 3$, $r_{d2} = r_2 = 6$ and $r_{d3} = r_3 = 12$). The atrous convolutions in MRACN use 3×3 kernels. For example, assuming the input dimension is $16 \times 16 \times 4096$, for both models, the parameter space of the 1×1 convolution on the input **x** is $16 \times 16 \times 256 \times 4096 = 268,435,456 \approx 268M$. We calculate the parameter spaces of all convolutions, reported in Table I. We can see that the parameter space of MRACN is significantly smaller than that of the inception module, 7,651M and 51,089M respectively, even at small receptive scales (3, 6 and 12) in this example. When increasing the receptive scales, the computational cost of the inception module increases dramatically while the parameter space of the proposed MRACN remains the same.

TABLE I: Comparison on the parameter spaces of the proposed MRACN and the inception module [13].

MRACN		**Inception module**	
Convolution	**Parameters**	**Convolution**	**Parameters**
1×1 on **x**	268M	1×1	268M
$r_{d1} \times r_{d1}$ on **x**	2,416M	$r_1 \times r_1$	2,416M
$r_{d2} \times r_{d2}$ on **x**	2,416M	$r_2 \times r_2$	9,664M
$r_{d3} \times r_{d3}$ on **x**	2,416M	$r_3 \times r_3$	38,656M
1×1 on $\mathbf{x}^g$	1M	1×1	1M
1×1 on $\mathbf{f}^{fuse}$	134M	1×1	84M
Total	7,651M	**Total**	51,089M

B. Datasets

PASCAL VOC 2012 dataset [19]: A widely used benchmark that contains 20 object classes, such as aeroplane, person, cat and car, and one background class. The original dataset contains a train set, a validation set and a test set with 1,464, 1,449 and 1,456 images, respectively. Hariharan et al. [35] have provided extra annotations for creating the augmented train set with 10,582 images.

DTMR-DVR dataset: This dataset is provided by the Department of Transport and Main Roads (DTMR), Queensland, Australia. Vehicle mounted cameras are used to collect the Digital Video Recording (DVR) data. We have manually created our DTMR-DVR dataset for semantic segmentation. Specifically, we extract image frames from the provided videos, and then use the Adobe Photoshop to annotate the extracted images for generating the pixel-wise class labels. Finally, the DTMR-DVR dataset contains a train set, a validation set and a test set with 400, 100 and 100 images, respectively. The dataset contains 13 roadside object classes, such as electric pole, speed limits and road, and one background class, listed in Table V.

C. Training protocol

We adopt the similar training protocol to [4], [11].

Image size: As MRACN is able to work with large receptive scales, images with large size is required. Otherwise, filters with large dilation rates are mostly applied to the zero padding regions. Therefore, we resize the images to the size of 513×513 for both training and testing. The three dilation rates, r_{d1}, r_{d2} and r_{d3}, in MRACN are set to 6, 12 and 18, respectively.

Learning rate: Similarly to [4], [11], [17], we employ the "poly" learning rate policy in which the initial learning rate is multiplied by $(1 - \frac{iter}{max\_iter})^p$ where $p = 0.9$.

Batch normalization: In MRACN-Seg, all added modules on top of ResNet have included batch normalization [34], except the final 1×1 convolution for generating the logits. The batch size is set to 10, and the batch normalization decay is set to 0.9997. For PASCAL VOC 2012 dataset, similarly to [11], after 30K iteration training on the augmented train set with the initial learning rate set to 0.007, we freeze the batch normalization parameters, and then double the dilation rates ($r_{d1} = 12$, $r_{d2} = 24$ and $r_{d3} = 36$) and train on the original train set for another 30K iterations. Note that, MRACN enables us to control the receptive scales at different training stages without changing the parameter space of the model, demonstrated in Section IV-A. For DTMR-DVR dataset, we train on the train set with 30K iterations.

Data augmentation: To alleviate overfitting, similar data augmentation strategy [11] has been applied. During training, images have been randomly re-scaled with a scale factor $\in$ [0.5, 2], and been randomly left-right flipped.

We use the High Performance Computing (HPC) facilities of our university as our experiment platform, with the allocated resources (CPU: Intel Xeon Skylake 6126, and GPU: Nvidia Tesla P100).

D. Experiments on PASCAL VOC 2012 dataset

1) Ablation study: We investigate the effects of the multi-receptive features and the global features of the proposed models. We first deactivate all the multi-receptive features and the global features by removing the proposed MRACN from MRACN-Seg. Then, we only activate the multi-receptive features and fix the receptive scales of the three parallel 3×3 atrous convolution branches by setting r_{d1}, r_{d2} and r_{d3} to 6, 12 and 18, respectively. After that, we add another parallel branch with $r_{d4} = 24$ for larger receptive context. Finally, we activate the global features. Table II shows the performance of the variants of our segmentation model on PASCAL VOC 2012 validation set. We can see that either of the multi-receptive features and the global features (activating either of them) can improve the performance. Interestingly, adding another parallel branch with $r_{d4} = 24$ decreases the performance slightly by 0.11%. Activating both features improves the performance by 2.28% (from 74.52% to 76.80%).

TABLE II: Effects of the multi-receptive features and the global features of the proposed MRACN. ✓: activated, and ✗: deactivated.

r_d**s**: (6, 12, 18)	r_d**s**: (6, 12, 18, 24)	**Global features**	**mIoU**
✗	✗	✗	74.52
✓	✗	✗	76.07
✗	✓	✗	75.96
✓	✗	✓	76.80

We also investigate the effects of the data augmentation strategies. As shown in Table III, the data augmentation strategies further improve the performance by 1.33% (from 76.80% to 78.13%).

TABLE III: Effects of data augmentation. MS: multi-scale input, and Flip: left-right flipped input. ✓: activated, and ✗: deactivated.

r_d**s**: (6, 12, 18)	**Global features**	**MS**	**Flip**	**mIoU**
✓	✓	✗	✗	76.80
✓	✓	✓	✗	77.78
✓	✓	✗	✓	77.10
✓	✓	✓	✓	78.13

2) Comparison with the state-of-the-arts: We compare MRACN-Seg with the state-of-the-art methods on PASCAL VOC 2012 test set. We fine tune our model on PASCAL VOC 2012 train and validation sets, and report the results on the test set.

As shown in Table IV, MRACN-Seg is able to outperform the compared baselines. Note that, further improvement could be achieved by pre-training the model on MS-COCO dataset [43], demonstrated by [4], [40], [41]. We qualitatively visualize the segmentation results of MRACN-Seg in Fig. 5.

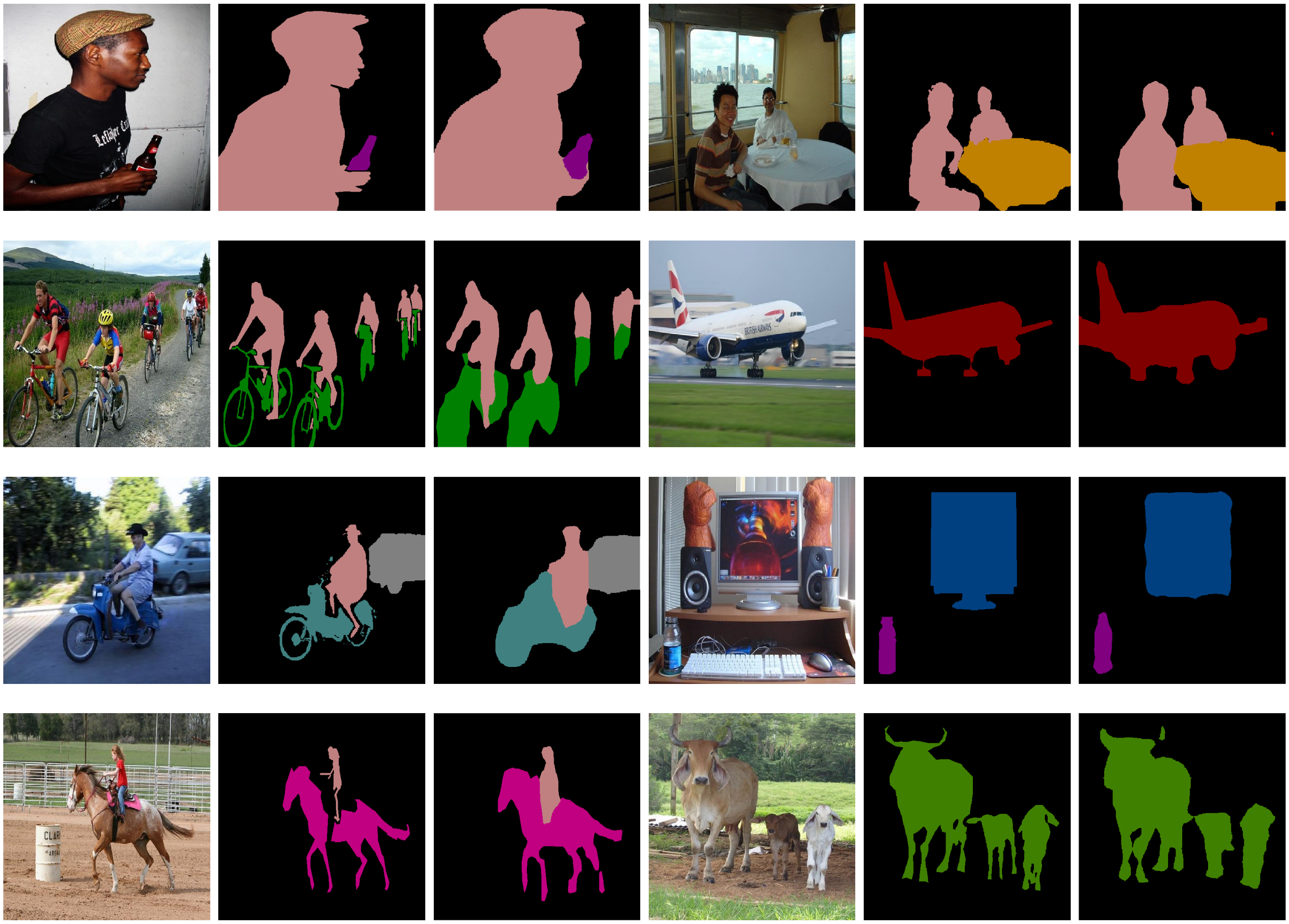


Image Ground truth Segmentation result Image Ground truth Segmentation result

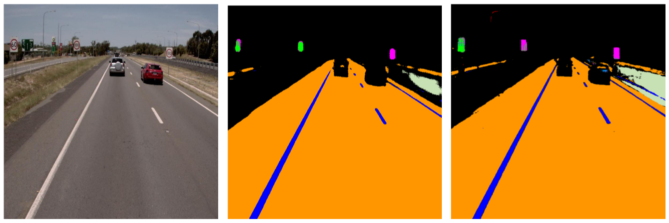
Fig. 5: Segmentation results on PASCAL VOC 2012 validation set.

TABLE IV: Performance comparison on PASCAL VOC 2012 test set.

Method	**mIoU**
FCN [2]	62.2
Zoom-out [36]	69.6
DeepSN-CRF [37]	70.1
DeepLab [10]	71.6
CRF-RNN [38]	72.0
DeconvNet [3]	72.5
GCRF [39]	73.2
DPN [40]	74.1
Piecewise [41]	75.3
VeryDeep FCRN [42]	79.1
PSPNet+Hierarical attention [5]	79.5
DeepLabv2 [4]	79.7
Our MRACN-Seg	80.2

E. Experiments on DTMR-DVR dataset

To further validate the proposed models, we evaluate our MRACN-Seg on our DTMR-DVR dataset. We employ the similar training protocol on PASCAL VOC 2012 dataset, described in Section IV-C. We compare our MRACN-Seg with the FCN baseline [2], and the results are shown in Table V. We can see that our MRACN-Seg achieves substantial improvement over the baseline by 14.4% (mIoU).



(a) Image (b) FCN [2] (c) MRACN-Seg

Fig. 6: Example of the FCN baseline and our MRACN-Seg.

TABLE V: Performance of MRACN-Seg on DTMR-DVR test set, compared with the FCN baseline. IoU per semantic classes and mean IoU (mIoU) are used to measure the performance. S_60: Speed limit sign 60, S_100: Speed limit sign 100, S_110: Speed limit sign 110, M_bar: Metal barrier, C_bar: Concrete barrier, B_path: Bicycle path, M_concrete: Median concrete, M_grass: Median grass, and T_light: Traffic light.

Method	mIoU	Road	Line	S_60	S_100	S_110	Pole	Tree	M_bar	C_bar	B_path	M_concrete	M_grass	T_light
FCN [2]	46.0	91.8	63.7	35.1	30.5	49.3	43.8	23.0	48.3	28.6	88.2	34.2	18.8	42.9
MRACN-Seg	60.4	97.3	79.7	56.4	49.0	62.7	57.8	39.5	63.9	47.8	94.0	48.1	29.4	59.6

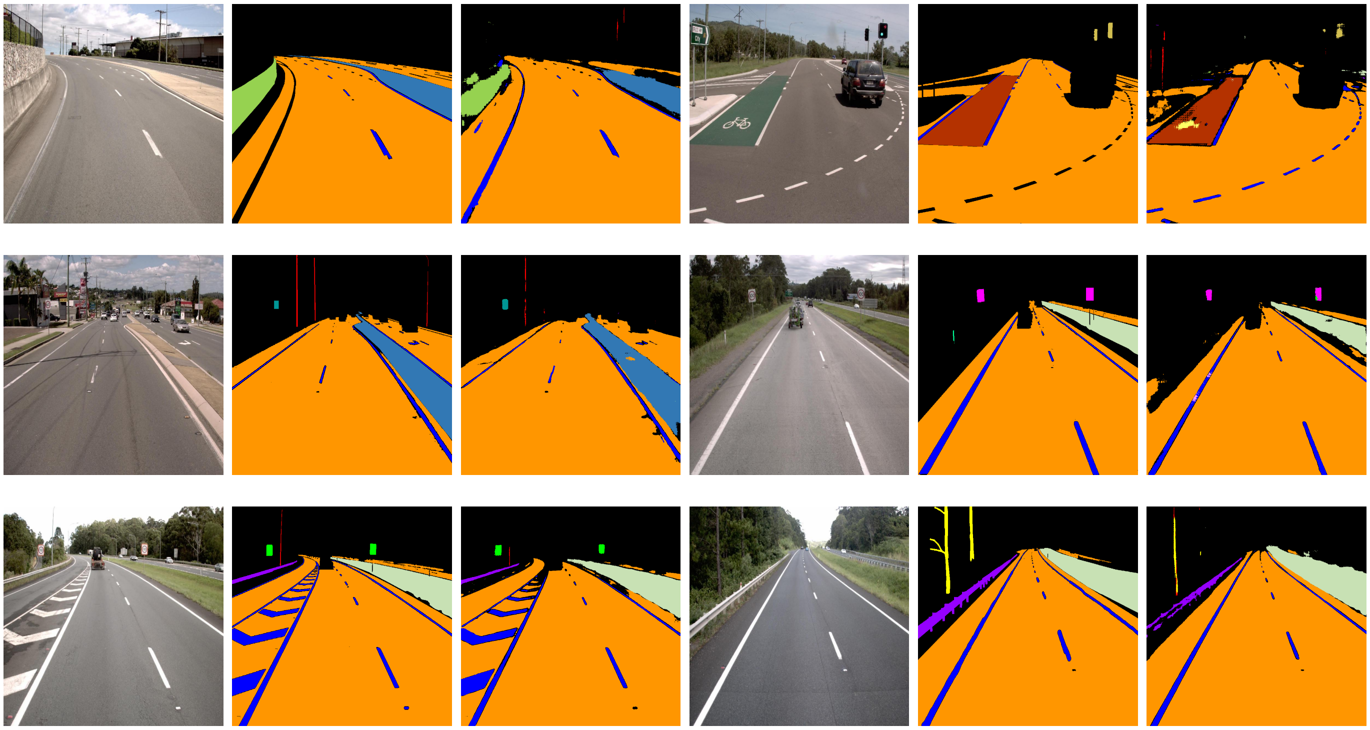


Image Ground truth Segmentation result Image Ground truth Segmentation result

Fig. 7: Segmentation results on DTMR-DVR validation set.

We notice that our MRACN has made significant improvements on speed limit signs. We visualize a mis-classified example by showing the output from the FCN baseline and MRACN-Seg in Fig. 6. We can see that FCN misclassifies the 100 speed limit sign in the middle of the image as a 60 speed limit sign, while our MRACN-Seg is able to produce the correct segmentation. We also notice that both models have limited performance on the tree class. The reason is that, in our DTMR-DVR dataset, we focus on the roadside objects relating to road safety. For the tree class, we are only interested in the trees with thick stems that would case serious traffic accidents, and only the thick stems of the trees have been annotated in our dataset. Therefore, the tree class may be similar to the pole class. Finally, the qualitative visualization of our MRACN-Seg is shown in Fig. 7.

V. Conclusion

In this paper, we proposed a multi-receptive atrous convolutional network for semantic image segmentation. MRACN uses the contextual information of the input, and captures the multi-receptive features and the global features at different receptive scales. MRACN requires a smaller parameter space than the inception module with the similar structure. Finally, the proposed MRACN segmentation model (MRACN-Seg) has been evaluated on PASCAL VOC 2012 dataset and DTMR-DVR dataset, and the experimental results demonstrate the effectiveness of the proposed model, compared with other state-of-the-art models. In the future, more roadside objects will be annotated and added to our DTMR-DVR dataset.

Acknowledgment

This research was supported under Australian Research Council's Linkage Projects funding scheme (project number: LP170101255).

References

[1] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[2] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[3] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1520–1528.

[4] L. 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," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[5] H. Lu, Z. Deng, and X. Liu, "Semantic image segmentation based on attentions to intra scales and inner channels," in *Proceedings of the IEEE International Joint Conference on Neural Networks*, 2018, pp. 1–8.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[9] Y. Wang, J. Liu, Y. Li, J. Yan, and H. Lu, "Objectness-aware semantic segmentation," in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2016, pp. 307–311.

[10] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014.

[11] L. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[12] L. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 3640–3649.

[13] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[15] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.

[16] Z. He, Y. Cao, L. Du, B. Xu, J. Yang, Y. Cao, S. Tang, and Y. Zhuang, "Mrfn: Multi-receptive-field network for fast and accurate single image super-resolution," *IEEE Transactions on Multimedia*, 2019.

[17] W. Liu, A. Rabinovich, and A. C. Berg, "Parsenet: Looking wider to see better," *arXiv preprint arXiv:1506.04579*, 2015.

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

[19] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *Proceedings of the International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, 2015.

[20] B. Fulkerson, A. Vedaldi, and S. Soatto, "Class segmentation and object localization with superpixel neighborhoods," in *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2009, pp. 670–677.

[21] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[22] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[23] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[24] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 91–99.

[25] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 379–387.

[26] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[27] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[28] J. Fu, J. Liu, Y. Wang, J. Zhou, C. Wang, and H. Lu, "Stacked deconvolutional network for semantic segmentation," *IEEE Transactions on Image Processing*, 2019.

[29] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 764–773.

[30] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell, "Understanding convolution for semantic segmentation," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*. IEEE, 2018, pp. 1451–1460.

[31] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings of the IEEE conference on Computer Cision and Pattern Recognition*, 2017, pp. 1925–1934.

[32] P. O. Pinheiro, T. Lin, R. Collobert, and P. Dollár, "Learning to refine object segments," in *Proceedings of the European Conference on Computer Vision*. Springer, 2016, pp. 75–91.

[33] M. Amirul Islam, M. Rochan, N. D. Bruce, and Y. Wang, "Gated feedback refinement network for dense image labeling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3751–3759.

[34] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[35] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2011, pp. 991–998.

[36] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich, "Feedforward semantic segmentation with zoom-out features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3376–3385.

[37] D. Lai, Y. Deng, and L. Chen, "Deepsqueezenet-crf: A lightweight deep model for semantic image segmentation," in *Proceedings of the IEEE International Joint Conference on Neural Networks*, 2019, pp. 1–8.

[38] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr, "Conditional random fields as recurrent neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1529–1537.

[39] R. Vemulapalli, O. Tuzel, M.-Y. Liu, and R. Chellapa, "Gaussian conditional random field network for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3224–3233.

[40] Z. Liu, X. Li, P. Luo, C.-C. Loy, and X. Tang, "Semantic image segmentation via deep parsing network," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1377–1385.

[41] G. Lin, C. Shen, A. Van Den Hengel, and I. Reid, "Efficient piecewise training of deep structured models for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3194–3203.

[42] Z. Wu, C. Shen, and A. v. d. Hengel, "Bridging category-level and instance-level semantic image segmentation," *arXiv preprint arXiv:1605.06885*, 2016.

[43] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proceedings of the European Conference on Computer Vision*. Springer, 2014, pp. 740–755.