IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 27, 2020, accepted May 5, 2020, date of publication May 11, 2020, date of current version May 22, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2993570

Multi-Scale Convolutional Features Network for Semantic Segmentation in Indoor Scenes

YANRAN WANG[1], QINGLIANG CHEN[1,2], SHILANG CHEN[3], AND JUNJUN WU[3]

[1]Department of Computer Science, Jinan University, Guangzhou 510632, China
[2]Guangzhou Xuanyuan Research Institute Company, Ltd., Guangzhou 510006, China
[3]School of Mechatronics Engineering, Foshan University, Foshan 528000, China

Corresponding author: Junjun Wu (jjunwu@fosu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61603103 and Grant 61673125, in part by the National Key Research and Development Program of China under Grant 2018YFB1308000, in part by the Key Area Research Projects of Universities of Guangdong Province under Grant 2019KZDZX1026 and Grant 2018KZDXM074, and in part by the Natural Science Foundation of Guangdong Province under Grant 2016A030310293.

ABSTRACT Semantic segmentation is one of the most fundamental techniques for visual intelligence, which plays a vital role for indoor service robotic tasks such as scene understanding, autonomous navigation and dexterous manipulation. However, semantic segmentation of indoor environments poses great challenges for existing segmentation techniques due to the complex overlaps, heavy occlusions and cluttered scenes with objects of different shapes and scales, which may lead to the loss of edge information and insufficient segmentation accuracy. And most of the semantic segmentation networks are very complex and cannot be applied to mobile robot platforms. Thus, it is of significant importance for ensuring as few network parameters as possible while improving the detection of meaningful edges in indoor scenes. In this paper, we present a novel indoor scene semantic segmentation method that can refine the segmentation edges and achieve a balance between accuracy and model complexity for indoor service robots. Our approach systematically incorporates dilated convolution and rich convolutional features from the intermediate layers of Convolutional Neural Networks (CNN), which is based on two motivations: (1) The middle hidden layer of CNN contains a lot of potentially useful information for better edge detection which is, however, no longer present in latter layers in traditional structures. (2) The dilated convolution can change the size of receptive field and obtain multi-scale feature information without losing the resolution and introducing any additional parameters. Thus we propose a new end-to-end Multi-Scale Convolutional Features (MSCF) network to integrate the dilated convolution and rich convolutional features extracted from the intermediate layers of traditional CNN. Finally, the resulting approach is extensively evaluated on the prestigious indoor image datasets of SUN RGB-D and NYUDv2, and shows promising improvements over state-of-the-art baselines, both qualitatively and quantitatively.

INDEX TERMS Semantic segmentation, convolutional neural networks (CNN), hidden convolutional features, dilated convolution, indoor service robots.

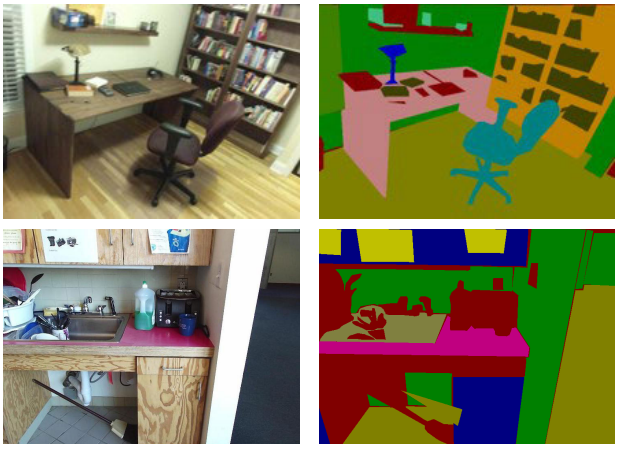
I. INTRODUCTION

Semantic segmentation is one of the most essential driving techniques to visual intelligence, which is defined as a segmentation that classifies each pixel according to the semantic content expressed by each pixel in an image. It is the most difficult and fundamental task in the current understanding of the scene. The significance of semantic segmentation technology to service robots lies in that the realization of semantic segmentation technology for indoor scenes enables the service robots to model the indoor environment, so as to achieve indoor map construction, autonomous navigation and path planning capabilities. In recent years, great progress has been made based on the powerful CNN structures [1]–[6]. Those CNN-based models have pushed semantic segmentation to a new high level against the traditional ones. However, unstructured targets, irregular shapes and object occlusion in images from complex real environments still pose great challenges to existing semantic segmentation approaches, greatly restricting their applicability.

With the availability of low-cost and compact 2.5/3D visual sensing devices, the research community is witnessing

The associate editor coordinating the review of this manuscript and approving it for publication was Jenny Mahoney.

a growing interest in visual scene understanding of indoor environments, which is a remarkably challenging benchmark for semantic segmentation, because there are a lot of complex overlaps between objects, heavy occlusions and cluttered scenes with targets of diversified outlines and scales. In these scenarios, the accurate edge detection for every object is much harder to obtain, and is, however, of vital importance for mobile robots since it is the key to a more accurate perception of their surroundings.



(a) Image (b) Ground Truth

FIGURE 1. The indoor scene in SUN RGB-D dataset. There are various shapes of objects that may be unstructured and occluded.

This paper aims to address this challenge, proposing a new Multi-Scale Convolutional Features neural network for semantic segmentation of indoor service robots, who may encounter a typical scene shown in Figure 1 where there are a variety of unstructured and occluded objects with different contours and sizes. To tackle the complexity in indoor scenes, the proposed MSCF model effectively utilizes the features captured in the middle layers in CNN to perfect the identification of edges for objects, and incorporates the dilated convolution [7] to further facilitate the segmentation. Our framework is based on the pre-trained model of VGG-16 [8] or ResNet-101 [9], and we introduce a new module for extracting useful features in middle layers of CNN, inspired by the Richer Convolutional Features (RCF) [10] model for edge detection.

To be more specific, the general pipeline of our approach is as follows. We firstly introduce a new module to extract the features of each convolutional layer of CNN, incorporate these features, and then get the fusion loss through the loss layer which can be trained by back-propagation. It is evident this approach can clearly enhance separations of objects because it carries the function for more accurate edge detection. However, this module only uses general convolutions with sizes 3×3 and 1×1, and so it can only identify a few features that are not enough to recognize objects well. Therefore, we go one step further to insert the part of dilated convolution to enrich the expressive power of the network, enabling it to learn multi-scale information to cope with the challenges of varied sizes and irregular shapes of objects in indoor scenes.

The main contributions of the paper can be summarized as follows:

1) We introduce a new Hidden Convolutional Feature (HCF) module for extracting valuable information from all hidden layers in CNN to capture the coarse high-level semantic features as well as fine-grained low-level ones, which can better the separations of objects because it can provide the information to identify their boundaries more accurately.
2) We encapsulate dilated convolution as Pyramid Convolution Module (PCM) to get multi-scale information for different objects in indoor scenes, and systematically incorporate HCF and PCM in a unified network, which is conducive to the production of high-resolution output.
3) A Multi-Scale Convolutional Features structure for indoor service robot scene understanding has been proposed to crystallize the ideas above. It strikes a good balance between accuracy and model complexity, and the prototype has been implemented and evaluated on the prestigious SUN RGB-D [11] and NYUDv2 [12] indoor scene datasets, showing promising results with respect to state-of-the-art baselines, both qualitatively and quantitatively.

The structure of the paper is as follows. The related work is first discussed in the next section, followed by the introduction of the proposed approach, showing how to combine the HCF and dilated convolutional features together to produce the resulting MSCF model. Then the experimental results are presented to show the efficacy of the model, compared to different state-of-the-art baselines. Finally, we conclude the paper with future research.

II. RELATED WORK

Traditional semantic segmentation methods, such as [13], [14], use manual annotation of semantic information to promote image comprehension tasks after segmentation of objects with shallow visual features. With the rise of deep learning models, researchers began to introduce CNN into semantic segmentation, which has been empirically proved to be highly effective. Therefore, the current mainstream frameworks are based on CNN that acts as the underlying building block to classify and recognize objects. And a lot of variants along with the optimization techniques are put forward for different scenarios.

The Fully Convolutional Network (FCN) [15] proposed by Long *et al.* is the first CNN-based network to implement pixel-level semantic segmentation. This landmark model is based on classic CNN such as VGG [8], ResNet [9] and GoogleNet [16], providing the capability of precise semantic segmentation in complex environments. However, its continuous pooling operation results in low output image resolution. In order to solve this problem, Chen *et al.* [17] took

Conditional Random Fields (CRF) as the post-processing module of the FCN network, and refined the segmentation results with CRF, which has achieved satisfactory results. But FCN and CRF modules are not integrated well, violating the end-to-end training policy. Furthermore, Zheng *et al.* proposed CRF as the Recurrent Neural Network (RNN) model [18] that can successfully combine CRF and RNN into a complete end-to-end framework, significantly improving the segmentation accuracy of FCN.

In addition to FCN, there are other influential encoder-decoder architectures. Badrinarayanan *et al.* [19] presented a typical encoding-decoding neural network for road and vehicle segmentation. The advantage of this network is that the pooling layer is capable of preserving the spatial location of pixel points. And these pixels can be mapped back to the corresponding position when restoring image resolution. However, this method does not segment object boundaries well. The DeconvNet [20] proposed by Noh *et al.* mirrored the convolutional layer to form the encoder-decoder structure, and improved FCN with this new structure. Paszke *et al.* [21] introduced another approach that also applied the encoder-decoder framework, in which they added BN layers and ReLU between convolution, and obtained enhanced results.

Recently, the utilization of multi-scale information to improve the accuracy of semantic segmentation has been paid much attention to [7], [22]–[27]. Chen *et al.* proposed the DeepLab series of models [17], [22], [23], [28], using the Atrous Spatial Pyramid Pooling (ASPP) structure to make full use of the multi-scale information. Lin *et al.* introduced RefineNet [29], which adopted a chain residual connection that can effectively fuse the missing information in the down-sampling to produce predicted images with high-resolution. but the network parameters scale is large. Zhao *et al.* [30] designed a novel model that combined dilated convolution and a pyramid pooling module to capture global information. Yu *et al.* [31] presented the Convolutional Block Attention Module (CBAM) [32] in order to select more discriminative features, effectively solving two basic problems of semantic segmentation: intra-class inconsistency and inter-class indistinction. And Zhang *et al.* [33] put forward a contextual semantic coding module and a class prediction module, alleviating the problem of unbalanced samples between classes in semantic segmentation.

The approaches mentioned above have played a significant role in promoting semantic segmentation of images. However, they mainly focus on the feature information in the last layer of CNN, and ignore somewhat useful and hierarchical features along with the hidden layers. Based on this observation, we propose a new approach to address this deficit by taking advantage of CNN's hierarchical features from the hidden layers. Moreover, multi-scale information from dilated convolution is also systematically integrated to improve the accuracy, leading to a new end-to-end network. Finally, we choose the challenging benchmark of indoor scenes to justify its efficacy.

III. THE PROPOSED APPROACH

For intelligent service robots, being able to recognize various types of objects indoors, such as tables, cups and chairs, is an indispensable ability. However, most of the existing semantic segmentation methods cannot segment these objects finely, and the network is too large to be applied to mobile robot platforms. In order to refine the boundaries of indoor object segmentation, while light-weighting the semantic segmentation method, we propose a MSCF model that has exhibited good empirical performance on standard benchmarks. The strength of MSCF comes from two new proposed modules within that we introduce: Hidden Convolutional Feature Module and Pyramid Convolution Module. HCF can preserve the discriminative information and prevent them from losing gradually in the process of downward transmission in CNN, while PCM can deal with the situation where the shapes and sizes of objects vary substantially and hence different sizes of receptive field are badly needed. Therefore, the integration of the two modules can empower the proposed MSCF to effectively segment an indoor image on a high-resolution feature map.

A. HIDDEN CONVOLUTIONAL FEATURE MODULE (HCF)

As is known, a CNN is always composed of many convolutional layers and some pooling layers, which attenuates the input image resolution layer by layer to produce high-level features and global information. As the example shown in Figure 2, with the increasing of the convolutional layer, the overall structure information extracted by the network is gradually clear, but the details are becoming less and less.

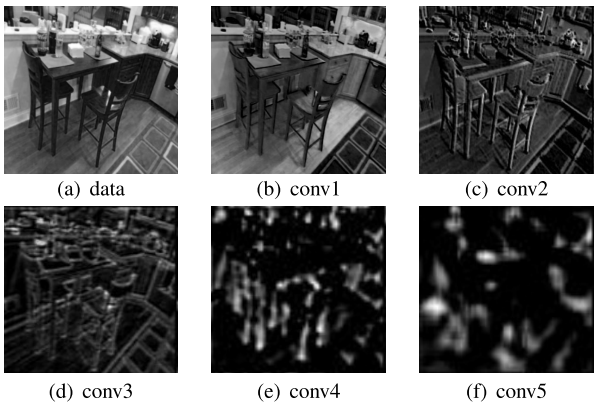


FIGURE 2. Features extracted by every convolutional layer on SUN RGB-D data. The convolutional feature becomes coarser as the convolutional layer increases.

Although the features extracted by the network decrease with the increase in the number of CNN layers, it is known that useful features are captured in actually each layer of CNN. For a high-precision segmentation task, the final rough frame features provided by just the last layer of CNN are definitely not sufficient. So a better feature map containing more detailed features from the intermediate hidden layers of

CNN should be taken into account as well, generating rich feature reserves for the downstream segmentation task.

The idea described above has actually been proved to be effective by [10], in which RCF was proposed using hidden layer features of CNN for edge detection. The results justified that features from hidden convolutional layers can indeed empower the network to identify more details of target boundaries.

Based on the above motivation, we propose a novel HCF module to extract the convolutional features from the hidden layers in CNN. The proposed HCF is a more advanced version of RCF, since the task of RCF is just for edge detection, and there is no need to identify every object, which is equivalent to a binary classification task, that is, the edge is one class and the non-edge is another class. Therefore, a relatively simple 1×1 convolution is used by RCF to preserve the features from hidden convolutional layers. However, for the more complex task of semantic segmentation, it is necessary to keep the details of every target of different shapes and sizes in the pixel level, and thus the convolution of 1×1 is definitely not sufficient. So the proposed HCF will go one step forward to combine the convolution of 1×1 and 3×3 to enlarge the receptive field. The 1×1 convolution is applied first and then the output is fed into a 3×3 hole convolution [7] in a parallel structure, with 3×3 hole convolutions in different rates. This kind of operation can extract features for targets in different scales, and hence can sharpen our model to recognize targets' boundaries with varied sizes.

B. PYRAMID CONVOLUTION MODULE (PCM)

After generating good convolutional features, the next key step in segmentation is to appropriately increase the receptive field and capture the context information.

When performing semantic segmentation, the pooling operation is often carried out to reduce the image size while increasing the receptive field. A common understanding is that the lost information caused by pooling in down-sampling is irreversible. However, the emergence of dilated convolution [7] makes it possible to allow for greater receptive field without losing any information. Thus, dilated convolution becomes a common convolution technique for the pixel-wise output model, which can preserve the internal data structure of the image while avoiding the loss of spatial stratification information.

Inspired by the ASPP structure [17] in which dilated convolutions are incorporated very well, we apply pyramidal convolution layers in two modules in our MSCF model. At the HCF module, pure hole convolution [7] is used to extract features for targets in different scales with receptive fields of different sizes. In the PCM module, a combination of hole convolutions and ordinary convolutions is executed. We first apply hole convolutions to extract the context information in different scales, and then use ordinary convolutions for further feature extraction and recognition. The PCM enables the MSCF model to easily obtain multi-scale features of an image without introducing any extra parameters. Finally, all the features of different scales from the two proposed modules are integrated to form a high-resolution feature map.

C. NETWORK ARCHITECTURE

We now present the MSCF network to fuse the HCF module and PCM module into an end-to-end model, whose architecture is shown in Figure 3.

We use a pre-trained VGG-16 or ResNet-101 as the underlying building block, with HCF and PCM modules to produce the rich feature map. Here, we remove all fully connected

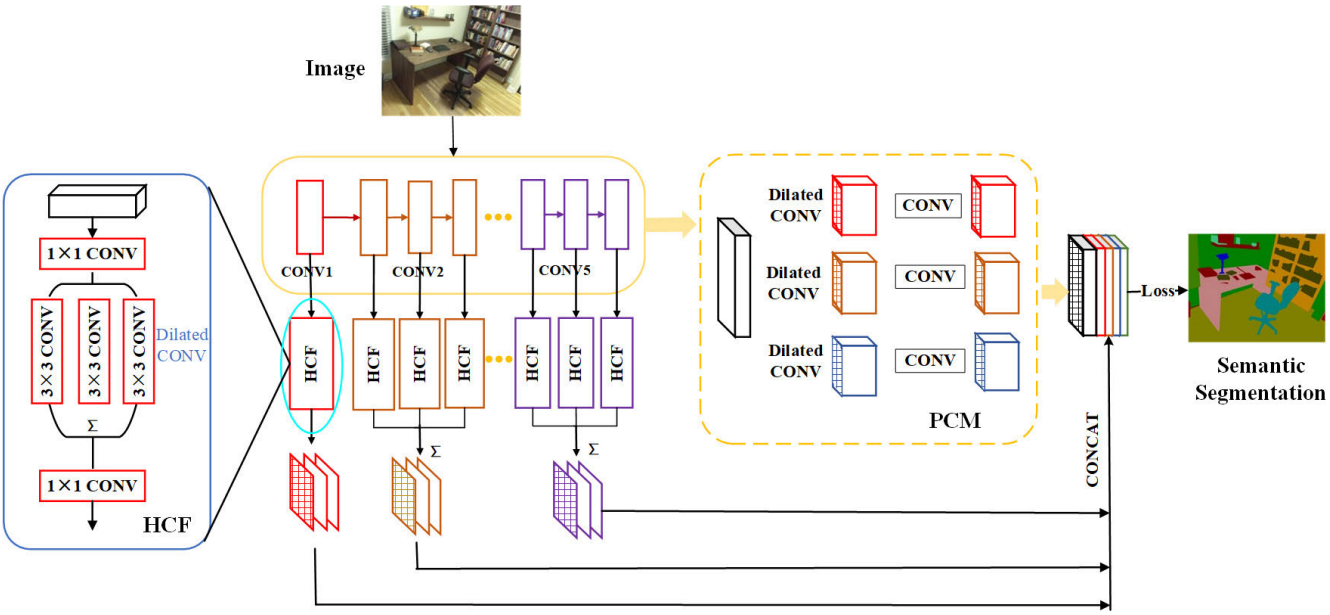


FIGURE 3. The proposed MSCF network. This network consists of a basic network, an HCF module and a pyramid structure of PCM, which can make full use of convolutional features to construct high-resolution feature maps.

layers and the last pooling layer of VGG-16 and ResNet-101. The details of our model can be summarized as below:

- Our basic network uses five stages of VGG16 or ResNet101. The input of the conv1 stage is the image size, and the output is one-half of the original image. The output of the conv2 stage is a quarter of the original image. the output of the conv3 stage is one-eighth of the original image. The output of both conv4 and conv5 stage is one-sixteenth of the original image. In each conv phase, the size of the feature map does not change.
- After each convolutional layer of basic network, an HCF module composed of convolutions with kernel sizes of 1×1 and 3×3 is connected to obtain the additional detailed feature information.
- The convolutional features are fused by *add* in each stage, which is to compute element-wise operations, such as product and sum, along multiple input blobs.
- Each stage contains a *cross-entropy loss/sigmoid* layer for monitoring network training process.
- 3×3 hole convolutions with different rates are applied, and they are connected to the last layer of the basic network in a pyramid structure.
- The features learned by the two modules are merged into a rich feature map and trained through back propagation.

Finally, we use the *multiclass cross-entropy loss function* to get the fusion loss and train the neural network by back-propagation, which is defined as the negative log-likelihood function over an entire data set of independent samples:

$$E_{\text{cross-entropy}} = -\sum_{c=1}^{M} y_o^c \ln p_o^c \tag{1}$$

where M is the number of classes, y is the correct class label for pixel o, and p is the predicted probability of pixel o being of class c.

IV. EXPERIMENTS

In this section, we present experimental results that we have conducted on SUN RGB-D [11] and NYUDv2 [12] datasets to evaluate the proposed model. We use the standard evaluation method of semantic segmentation: mean Pixel Accuracy (mPA) [15], and mean Intersection over Union (mIoU) [15] to carry out quantitative studies on the network performance. Moreover, we illustrate the output segmentation graph of the network on the test set to show the qualitative comparisons with respect to different state-of-the-art counterparts.

Mean Pixel Accuracy (mPA) [15], and mean Intersection over Union (mIoU) [15] are defined as follows:

$$\text{mPA} = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}} \tag{2}$$

$$\text{mIoU} = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \tag{3}$$

Here $k+1$ is the number of classes, including k labeled classes and one unlabeled class. p_{ij} represents the number of pixels that belong to class i but are predicted to be class j. So p_{ii}, p_{ij}, p_{ji} are defined as true positives, false positives and false negatives, respectively.

A. TRAINING DETAILS

Our model is based on the pre-trained network of VGG-16 or ResNet-101. All weights of convolutional layers that do not belong to the underlying VGG or ResNet network are initialized to a Gaussian distribution with zero mean and 0.01 variance, while the biases are initialized with 0. We use Caffe to implement our model, which is trained and fine-tuned on the NVIDIA GTX1080 Ti GPU. We adopt the poly learning policy in Caffe and set the initial learning rate to 0.001 and the power to 0.9, so that the learning rate will decrease as the number of iterations increases. Finally, the momentum and weight_decay were set to 0.9 and 0.0005, respectively. In the experiment, we cropped the original images to the size of 321. For data augmentation, we randomly mirror and scale (0.5, 0.75, 1, 1.25, 1.5) the input images.

B. QUANTITATIVE EVALUATION

SUN RGB-D dataset [11] was derived from four RGB-D sensors, containing 10335 RGB-D images and 37 categories of dense annotations. This dataset contains images from NYU Depth V2 [12], Berkeley B3DO [34] and the SUN3D [35] dataset, which is very suitable for scene understanding tasks. We use only RGB images to train our network without using depth information. Then, we apply the standard split of 5285 and 5050 images for training and testing.

We first tested the effect of dilated convolution kernel rate in the HCF module on segmentation accuracy. We set three rates, which are rate $=(4, 8, 12)$, rate $=(6, 12, 18)$ and rate $=(8, 16, 24)$, respectively, and the comparative results are shown in Table 1. The results are obtained on the testing set using a single-scale input.

As presented in Table 1, we show the ablation study with HCF and PCM modules, respectively that use different field-of-view sizes. Here we use the VGG-16 as the basic network for training. For the HCF module, we first compare the performance of 1×1 kernel size with the 3×3 one. The results indicate that the convolution of size 1×1 is better than that of 3×3, which may be due to the fact that 1×1 convolution kernel can integrate the feature graph without changing the feature resolution. Compared to 3×3 convolution, 1×1 convolution results in less feature information loss. Subsequently, we add dilated convolution with

TABLE 1. Semantic segmentation accuracy on SUN RGB-D with different rates of dilated convolution.

Method	1×1	3×3	rate=(4,8,12)	rate=(6,12,18)	rate=(8,16,24)	Mean Acc	Mean IoU
VGG-16						24.32	17.61
+HCF	✓					23.07	17.18
		✓				22.45	16.65
		✓	✓			23.66	17.70
	✓		✓			23.80	17.83
	✓			✓		**29.06**	**22.00**
	✓				✓	22.78	17.61
+PCM	✓		✓			35.08	26.70
	✓			✓		35.50	26.92
	✓				✓	**36.46**	**26.94**

TABLE 2. Pixel accuracy per-class on NYUDv2.

	Wall	Floor	Cabinet	Bed	Chair	Sofa	Table	Door	Window	Book shelf
Silberman et al. [12]	60.7	77.8	33.0	40.3	32.4	25.3	21.0	5.9	29.7	22.7
Ren et al. [36]	60.0	74.4	37.1	42.3	32.5	28.2	16.6	12.9	27.7	17.3
Saurabh et al. [37]	**67.9**	**81.5**	45.0	60.1	41.3	47.6	29.5	12.9	34.8	18.1
HrNetV2 [38]	64.6	71.7	46.0	49.1	36.4	40.8	28.8	23.3	33.0	27.3
MSCF (*ResNet-101*)	64.5	74.9	**49.7**	**61.0**	**50.6**	**54.0**	**37.9**	**26.5**	**35.1**	**31.7**
	Picture	Counter	Blinds	Desk	Shelves	Curtain	Dresser	Pillow	Mirror	Floor mat
Silberman et al. [12]	35.7	33.1	40.6	4.7	3.3	27.4	13.3	18.9	4.4	7.1
Ren et al. [36]	32.4	38.6	26.5	10.1	6.1	27.6	7.0	19.7	17.9	20.1
Saurabh et al. [37]	40.7	**51.7**	41.2	6.7	5.2	26.9	25.0	32.8	21.2	25.8
HrNetV2 [38]	43.7	41.2	48.5	7.5	5.5	26.2	15.9	23.6	16.0	18.4
MSCF (*ResNet-101*)	**45.2**	47.6	**52.5**	**14.3**	**12.7**	**35.9**	**30.0**	**37.6**	**29.3**	**28.8**
	Clothes	Ceiling	Books	Fridge	Tele-vision	Paper	Towel	Shower curtain	Box	White board
Silberman et al. [12]	6.5	73.2	5.5	1.4	5.7	12.7	0.1	3.6	0.1	0.0
Ren et al. [36]	9.5	53.9	14.8	1.9	18.6	11.7	12.6	5.4	3.3	0.2
Saurabh et al. [37]	7.7	**61.2**	7.5	11.8	15.8	14.7	20.0	4.2	1.1	10.9
HrNetV2 [38]	2.1	45.1	21.3	28.3	38.9	17.2	18.7	18.3	2.3	**57.5**
MSCF (*ResNet-101*)	**13.7**	27.6	**27.5**	**50.1**	**55.2**	**20.8**	**31.4**	**21.6**	**5.6**	31.0
	Person	Night stand	Toilet	Sink	Lamp	Bathtub	Bag	Other str	Other prop	Other prop
Silberman et al. [12]	6.6	6.3	26.7	25.1	15.9	0.0	0.0	6.4	3.8	22.4
Ren et al. [36]	13.6	9.2	35.2	28.9	14.2	7.8	1.2	5.7	5.5	9.7
Saurabh et al. [37]	1.4	17.9	48.1	45.1	31.1	19.1	0.0	7.6	3.8	22.6
HrNetV2 [38]	12.3	20.5	49.1	40.6	27.3	**26.9**	2.3	13.5	9.8	21.8
MSCF (*ResNet-101*)	**61.3**	**28.6**	**72.1**	**49.7**	**31.3**	26.2	**9.8**	**25.8**	**15.3**	**30.7**

TABLE 3. Results on different baselines.

Network	mPA	mIoU
MSCF (*VGG-16*)	36.46	26.94
MSCF (*ResNet-101*)	**44.19**	**33.56**

TABLE 4. Comparisons for different models on SUN RGB-D.

Network	mPA	mIoU	Parameters
Liu et al. [39]	10.0	-	-
Ren et al. [36]	36.3	-	-
DeconvNet [20]	33.3	22.6	-
FCN-8s [15]	38.4	27.4	-
SegNet [19]	43.9	31.8	**30M**
DeepLab v2 [28]	42.2	32.1	44M
DeepLab v3+ [23]	42.5	31.7	59M
HrNetV2 [38]	31.0	22.0	66M
MSCF (*ResNet-101*)	**44.2**	**33.6**	45M

different rates after 1×1 convolution. Since the feature extracts in HCF are now further equipped with information of different scales, the average pixel precision and average IoU are greatly improved. They grow by about 5% and 4%, respectively. As for the PCM module, we also compare the effects of different rates on network accuracy, where the model performs best when rate = (8, 16, 24).

Then we present the comparative results of our model on the SUN RGB-D dataset in Table 3 and Table 4. Since our network structure is flexible and not bounded by a specific pre-trained network, we compare two different basic networks: VGG-16 and ResNet-101 in Table 3. The results show that the deeper the network is, the better the results are. Our VGG-16 based MSCF obtained 26.9% of mIOU, while MSCF on ResNet-101 exhibits an improvement of approximately 7% over the VGG-16 based network model.

In Table 4, we verify the efficacy of our network with comparative experiments. We compare the segmentation accuracy and parameters of our MSCF with other state-of-the-art approaches such as DeconvNet [20], FCN [15], SegNet [19], DeepLab v2 [28], DeepLab v3+ [23] and HrNetV2 [38] on the SUN RGB-D dataset. Here we use the deeper ResNet-101 network as the underlying pre-trained network for the experiments. The results indicate that our model strikes a good balance between accuracy and model complexity.

The performance of the latest representative models DeepLab v3+ and HrNetV2 on the SUN RGB-D dataset is not satisfactory, indicating that blindly using encoder-decoder methods to improve the resolution of feature maps, or using high-resolution feature maps throughout, is not very effective for semantic segmentation of complex indoor environment with irregular shapes and mutual occlusion. Our method that makes full use of all hidden convolutional layer's details information and enriches the rough high-resolution feature maps of the last layer, has shown advantages over the latest prestigious landmark model of DeepLab v3+ and HrNetV2 in both accuracy and number of parameters.

NYUDv2 [12] is another commonly used 2.5-dimensional dataset with depth information, including 40 categories of indoor objects, 1449 RGB-D images captured by Microsoft Kinect device. This dataset focuses on depicting indoor scenes and can be used for training tasks of home robots. In this benchmark, we use the standard split of 795 and 654 images for training and testing, and again, no depth

(a) Image (b) Ground Truth (c) RCF (d) DeepLab v2 (e) Our MSCF

FIGURE 4. Qualitative results of our MSCF method compared to RCF and DeepLab v2 on SUN RGB-D. It can be seen that RCF pays too much attention to the details of the edges in objects, and thus is good at drawing the outline, such as tables, beds, and bicycles, but the black in the figure indicates that RCF fails to recognize the separated objects. On the contrary, DeepLab v2 can correctly recognize most of the segmented objects. But the segmentation effect is not as fine as RCF, since the edge of the segmented object tends to be blurred compared to RCF, as shown in (d). Due to the elaborate fusion of the strength from both RCF and DeepLab v2, our model outperforms DeepLab v2 in edge segmentation, and RCF in object recognition accuracy.

information is used. The details of the evaluation of our network on the NYUDv2 dataset are illustrated in Table 5.

Statistics in Table 5 show that our MSCF network outperforms other counterparts. It achieved 47.3% and 36.0% in mPA and mIoU, respectively. This result justifies that our proposed network also performs well on the NYUDv2 dataset.

Finally, the precision results for each category compared to the other four state-of-the-art baselines are shown in Table 2. Our results indicate that most categories are significantly improved by the effective use of the generated rich convolutional features in our model, especially in categories with distinct geometric distinctions, such as *Curtain*, *Fridge*, *Person*, and *Toilet*. The recognition rate of *Box* category is relatively low, because there are fewer pictures containing *Box* category in the dataset and it is difficult to distinguish *Box* from other categories, such as *Books*, *Bag* and other objects shaped like *Box*, even with rich convolutional features.

C. QUALITATIVE ANALYSIS

We now present qualitative comparisons of our MSCF model with RCF [10] and DeepLab v2 [28] in Figure 4. The outputs are all results run on the same machine using the same dataset. Both RCF and DeepLab's code and related parameter settings are derived from the open source code on Github described in the published articles.

TABLE 5. Comparisons for different models on NYUDv2.

Network	mPA	mIoU
Silberman et al. [12]	17.5	-
Ren et al. [36]	20.2	21.4
Saurabh et al. [37]	29.6	30.7
Gupta et al. [40]	35.1	-
Eigen et al. [41]	45.1	34.1
FCN [15]	46.1	34.0
Wang et al. [42]	47.3	-
HrNetV2 [38]	39.8	28.3
MSCF (*ResNet-101*)	**47.3**	**36.0**

RCF uses 1×1 convolution to obtain the hidden layer information, and the receptive field is too small to correctly identify large objects such as *beds* and *tables*, as shown in Figure 4. However, it is better than DeepLab v2 at recognizing edges and small objects such as *chair legs*, *lamps* and *wheels*. DeepLab v2 works well on objects such as *tables*, *chairs*, *roofs* and *picture frames*, but it is not good enough in boundary segmentation. Our approach systematically combines the advantages of the two models to achieve better results than both of them. However, the segmentation accuracy of small targets is still poor, which needs to be further enhanced in future work.

V. CONCLUSION

This paper proposes a new semantic segmentation model of MSCF for indoor intelligent service robots, which can make full use of all convolutional features from hidden layers in traditional CNN and achieve a balance between accuracy and model complexity. This model contains a proposed HCF module and a PCM module. The HCF module applied to each convolutional layer can extract coarse high-level semantic features and fine-grained low-level ones while the PCM module can acquire multi-scale information for images in the form of a pyramid. Then, the high-resolution feature map can be produced by the systematic fusion of the two modules and hence enable better segmentation. Finally, comparative evaluation studies on two prestigious indoor image benchmarks show that the proposed MSCF model can achieve better results than state-of-the-art baselines.

In future research, we will consider the introduction of the binarization network [43] to further improve the efficiency of our model, develop real-time semantic segmentation techniques, and also explore more powerful encoder-decoder structures for the rich contextual information in the images.

REFERENCES

[1] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell, "Understanding convolution for semantic segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1451–1460.

[2] J. Fu, J. Liu, Y. Wang, J. Zhou, C. Wang, and H. Lu, "Stacked deconvolutional network for semantic segmentation," *IEEE Trans. Image Process.*, early access, Jan. 25, 2019, doi: 10.1109/TIP.2019.2895460.

[3] Á. Sáez, L. Bergasa, E. López-Guillén, E. Romera, M. Tradacete, C. Gómez-Huélamo, and J. del Egido, "Real-time semantic segmentation for fisheye urban driving images based on ERFNet," *Sensors*, vol. 19, no. 3, p. 503, 2019.

[4] X. Zhang, Z. Chen, Q. M. J. Wu, L. Cai, D. Lu, and X. Li, "Fast semantic segmentation for scene perception," *IEEE Trans. Ind. Informat.*, vol. 15, no. 2, pp. 1183–1192, Feb. 2019.

[5] I. Alonso and A. C. Murillo, "Semantic segmentation from sparse labeling using multi-level superpixels," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 5785–5792.

[6] S. Lee, S.-J. Park, and K.-S. Hong, "RDFNet: RGB-D multi-level residual feature fusion for indoor semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4980–4989.

[7] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. 4th Int. Conf. Learn. Represent. (ICLR)*, May 2016, pp. 1–13.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, May 2015, pp. 1–14.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[10] Y. Liu, M.-M. Cheng, X. Hu, J.-W. Bian, L. Zhang, X. Bai, and J. Tang, "Richer convolutional features for edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1939–1946, Aug. 2019.

[11] S. Song, S. P. Lichtenberg, and J. Xiao, "SUN RGB-D: A RGB-D scene understanding benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 567–576.

[12] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 7576. Berlin, Germany: Springer, 2012, pp. 746–760.

[13] C. Zhang, Z. Xue, X. Zhu, H. Wang, Q. Huang, and Q. Tian, "Boosted random contextual semantic space based representation for visual recognition," *Inf. Sci.*, vol. 369, pp. 160–170, Nov. 2016.

[14] J. Pont-Tuset, P. Arbelaez, J. T. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping for image segmentation and object proposal generation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 128–140, Jan. 2017.

[15] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[16] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[17] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, May 2015, pp. 1–14.

[18] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr, "Conditional random fields as recurrent neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1529–1537.

[19] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[20] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1520–1528.

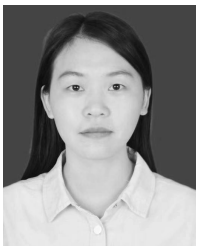
[21] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "ENet: A deep neural network architecture for real-time semantic segmentation," 2016, *arXiv:1606.02147*. [Online]. Available: http://arxiv.org/abs/1606.02147

[22] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*. [Online]. Available: http://arxiv.org/abs/1706.05587

[23] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with Atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 11211. Cham, Switzerland: Springer, 2018, pp. 833–851.

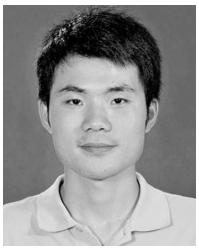
[24] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 11217. Cham, Switzerland: Springer, 2018, pp. 334–349.

[25] A. Kirillov, R. Girshick, K. He, and P. Dollár, "Panoptic feature pyramid networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6399–6408.
[26] Y. Wang, G. Wang, C. Chen, and Z. Pan, "Multi-scale dilated convolution of convolutional neural network for image denoising," *Multimedia Tools Appl.*, vol. 78, no. 14, pp. 19945–19960, Jul. 2019.
[27] Y. Wang, S. Hu, G. Wang, C. Chen, and Z. Pan, "Multi-scale dilated convolution of convolutional neural network for crowd counting," *Multimedia Tools Appl.*, vol. 79, nos. 1–2, pp. 1057–1073, Jan. 2020.
[28] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, Atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
[29] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5168–5177.
[30] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.
[31] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Learning a discriminative feature network for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1857–1866.
[32] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 11211. Cham, Switzerland: Springer, 2018, pp. 3–19.
[33] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, "Context encoding for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7151–7160.
[34] A. Janoch, S. Karayev, Y. Jia, J. T. Barron, M. Fritz, K. Saenko, and T. Darrell, "A category-level 3D object dataset: Putting the Kinect to work," in *Consumer Depth Cameras for Computer Vision*. London, U.K.: Springer, 2013, ch. 8, pp. 141–165.
[35] J. Xiao, A. Owens, and A. Torralba, "SUN3D: A database of big spaces reconstructed using SfM and object labels," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1625–1632.
[36] X. Ren, L. Bo, and D. Fox, "RGB-(D) scene labeling: Features and algorithms," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2759–2766.
[37] S. Gupta, P. Arbeláez, R. Girshick, and J. Malik, "Indoor scene understanding with RGB-D images: Bottom-up segmentation, object detection and semantic segmentation," *Int. J. Comput. Vis.*, vol. 112, no. 2, pp. 133–149, Apr. 2015.
[38] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, and J. Wang, "High-resolution representations for labeling pixels and regions," 2019, *arXiv:1904.04514*. [Online]. Available: http://arxiv.org/abs/1904.04514
[39] C. Liu, J. Yuen, and A. Torralba, "SIFT Flow: Dense correspondence across scenes and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 978–994, May 2011.
[40] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from RGB-D images for object detection and segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 8695. Cham, Switzerland: Springer, 2014, pp. 345–360.
[41] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2650–2658.
[42] J. Wang, Z. Wang, D. Tao, S. See, and G. Wang, "Learning common and specific features for RGB-D semantic segmentation with deconvolutional networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 9909. Cham, Switzerland: Springer, 2016, pp. 664–679.
[43] X. Lin, C. Zhao, and W. Pan, "Towards accurate binary convolutional neural network," in *Proc. Annu. Conf. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 345–353.

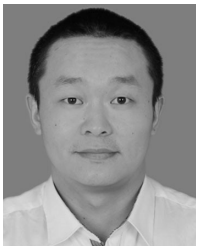


YANRAN WANG received the bachelor's degree in computer science from the Hubei University of Chinese Medicine. She is currently pursuing the degree with the Department of Computer Science, Jinan University. She has been a Core Member of several research grants funded by the National Natural Science Foundation of China. Her research interests are machine learning, computer vision, and robotics.

QINGLIANG CHEN received the Ph.D. degree in computer science from Sun Yat-sen University. He was a Postdoctoral Fellow with Peking University from 2010 to 2012. He is currently a Professor with the Department of Computer Science, Jinan University. He has published more than 30 articles in peer-reviewed international journals and conferences such as AAAI, IJCAI, and *The Computer Journal*. He has served as the Principal Investigator for three research grants from the National Natural Science Foundation of China. His research interests are machine learning, deep neural networks, and pattern recognition.



SHILANG CHEN received the bachelor's degree in medical information engineering from the Hubei University of Chinese Medicine. He is currently pursuing the degree with the School of Mechatronics Engineering, Foshan University. His research interests are visual SLAM, robotics, and computer vision.



JUNJUN WU received the Ph.D. degree from the School of Mechanical and Automotive Engineering, South China University of Technology. He is currently an Associate Professor with the School of Mechatronics Engineering, Foshan University. He has published more than 30 articles in intelligent robot journals and conferences such as ICRA, ROBIO, and ACTA AUTOMATICA SINICA. He has been the Principal Investigator of two research grants from the National Natural Science Foundation of China. His research interests are robot vision and visual SLAM.