SPNet: Superpixel Pyramid Network for Scene Parsing

Bingbing Xu[1,2], Fei Yang[1,2], Jinfu Yang[1,2], Suishuo Wu[1,2],Yi Shan[1,2]
1. Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China
2. Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing 100124, China
e-mail: S201602088@emails.bjut.edu.cn

***Abstract*—Scene parsing is the important part of computer vision research. And the deep coding-decoding network is widely applied to scene parsing. However, there are still some problems, such as ambiguity of object edge segmentation and uncertainty when segmenting small-size-objects in scene analysis. In this paper, we propose Superpixel Pyramid Network for Scene Parsing. First, a deep coding-decoding network is used to learn image features. Then, multi-scale spatial pyramid pooling structure is employed to enhance the performance of small-size-objects. Next, the Superpixel Segmentation is also applied to cope with the problem of ambiguity of object edge. Finally, a two-layer neural network classifier is applied to identify the fused features pixel-by-pixel. Extensive experimental results over ADE20K, PASCAL VOC 2012, and Camvid, demonstrated that the proposed method can obtain better performance counterparts than other.**

Keywords—scene parsing, deep coding-decoding network, pyramid pooling structure, superpixel segmention

I. Introduction

Nowadays, the world has entered a new era of artificial intelligence. Computer vision is an important research field of artificial intelligence. Scene parsing is one of the hot spots in computer vision research. Scene parsing is applied in many areas including robot sensing, automatic driving, virtual reality, augmented reality, to name a few. Scene parsing can not only recognize the information of an overall scene category, but also has a comprehensive and detailed understanding of the local scene. The goal of scene parsing is to assign a category label to each pixel in the image. It needs to predict the position, category and shape of each pixel. Combine object detection, segmentation, and multiple tag identification issues in the same process. The final understanding of the entire image is completed.

In recent years, many scene parsing methods have emerged [1,2,3,4,6,27]. Most of the traditional scene parsing methods rely on pre-segmentation strategies for superpixel segmentation and segmentation of candidate regions. Then the features are extracted from the segmented regions or the combined regions of the segmented regions. Use conditional random fields or other types of graph models to train each segmented superpixel block to ensure global consistency of the label [2].

Recently, due to the rapid development and widespread use of deep convolutional neural networks (CNNs), more and more scene parsing methods have turned to algorithms based on deep CNN networks [3,4,6,27]. More representative examples are: full convolutional neural networks [3], segnet networks [4], and so on. The full convolutional neural network [3] performs scene analysis, an end-to-end semantic segmentation network, and the parsed results are blurred and smooth and insensitive to the details in the image. Deep coding-decoding network solves the problem of blockiness of pixel-by-pixel label results, but the edge of the object in the image segmented by this method is relatively blurred, and there is a great uncertainty in the segmentation of small-size-objects in the image. Superpixel segmentation can make the edges of objects in image segmentation more clear. Spatial pyramid pooling network takes advantage of the overall information and the global information of the subregion for an image. Therefore, we propose superpixel pyramid network for scene parsing.

In the paper, we firstly extract the image features using a deep coding-decoding network. Then, using the multi-scale spatial pyramid pooling structure pools features and integrates global information and local information to enhance the network's parsing of small-size-objects. Next, employing the method based on the graph segmentation segments the original image, so that the object edges are more clearly. Finally, classifiers are used for pixel-by-pixel classification. This article has three contributions as follows:

1) The method based on the coding-decoding network, which combined with pyramid structure introduced by us can extract more useful information for image scene parsing.

2) The proposed method can make the edge of the object in the image scene analysis clearer.

3) The proposed method can determine the position and outline of small-size-objects in terms of image scene analysis for an image.

This paper will be organized as follows. In section II, we review the related recent literature and introduce the methods used in each step of the scene parsing. We describe our proposed method in section III. Section IV shows the experiments about the method on ADE20K, PASCAL VOC 2012, and Camvid datasets. Finally, we conclude our paper in section V.

II. Related work

In the section, we will introduce the method steps of traditional scene parsing and the method of deep-based scene analysis.

The first step in scene parsing is to perform image segmentation. Most methods adopt superpixel segmentation algorithms [5]. The superpixel is an image block composed of adjacent pixels which have similar texture, color, brightness, etc. There are two types of superpixel algorithms One is a graph theory-based approach and another is a gradient-based approach. Common graph-based methods [7] are superpixel lattice algorithms [8] and entropy-based methods [9]. The common methods based on the gradient-decreasing superpixel algorithm have Turbopixels algorithm [10] and SLIC algorithm [11]. In recent years, these superpixel algorithms are widely applied on image

978-1-7281-1312-8/18/$31.00 ©2018 IEEE

segmentation. and have already achieved good segmentation results.

After extracting features from segmented image regions, most methods rely on Markov Random Fields [28], Conditional Random Fields [29], or other types of graph models to label each region [12,13,19,20]. However, the methods of scene parsing described above have the characteristics of artificially defining the extraction of the segmented regions. The artificially defined features cannot fully express the implicit information and often discard some of the discriminative information for an image. And an entire image parsing process is more complicated, increasing the difficulty of scene analysis.

In recent years, CNN based methods [1,2,3,4,6,27] have already obtained outstanding achievement on scene parsing and semantic segmentation tasks. Farabet et al [2] raised the employ of convolutional neural networks to learn the hierarchical features of scene images for scene analysis. This method uses the CNN network which extracts image features. These image features are more comprehensive, and multi-scale feature fusion is used to increase the spatial context information of images, which greatly improve performance of scene analysis. Girshick et al [1] proposed object detection and semantic segmentation network which is based on R-CNN structure. This method come up with a new idea for the field of object detection. Compared with other methods, the object detection performance has a greatly improvement. Simultaneously, the semantic segmentation performance has also been enhanced. Long et al [3] raised a full convolutional neural network dedicated to pixel-by-pixel labeling tasks. They replaced convolutional neural networks with convolutional layers and directly output class predictions for images. The network which is an end-to-end semantic segmentation network is enable to input images of any size and output a segmented image of the same size as the input image. However, analytical results obtained are blurred, smooth, and insensitive to the details in the image. Since the network only considers the classification of each pixel. It does not consider the relationship between pixels. Thus, segmented images lack spatial consistency. Badrinarayanan et al [4] proposed a deep convolutional coding-decoding structure for scene-by-pixel labeling. By recording the parameters in the convolution pooling process, this method avoids directly copying all the pixel values in a block when upsampling, and solves the problem of block-by-pixel labeling. At the same time, the image context information has been increased and good results have been achieved. However, the edge of the object in the image segmented by the method is blurred, and there is a great uncertainty in the segmentation of the small-size-objects in the image.

To sum up, we propose superpixel pyramid network for scene parsing. This method not only makes the edge of the object in the image more clearly, but also has more certainty for the segmentation of small-size-objects in the image. This enhances understanding of complex scenes.

III. Method

In the section, we will describe the raised method framework. Different from traditional segmentation approaches, our method applies a deep coding-decoding network, combined with the spatial pyramid module, and introduces a superpixel segmentation algorithm to enhance the edge information of the object. Our proposed method explained in Fig.1 is described to improve property for sharpness of target edges and detection of small-size-objects. First, we use the deep coding-decoding network to obtain the original image feature map. Then, the acquired feature map is pooled by using the 4-level spatial pyramid, and a new feature map in which the global prior and the original feature map are connected in series is obtained, and is trained as an input of a two-layer neural network classifier. At the same time, we make the edge information of the small-size-objects in the image more clear by superpixel segmentation for the original image. Next, the pooled features are superimposed with the superpixel-divided features, and together as an input, the trained two-layer neural network classifier is used for pixel-by-pixel classification.

A. Pyramid Pooling Module

In deep neural network, the scale of receptive field can approximately indicate that we utilize the amount of context information. In especial, for high-level layers, the empirical acceptive field of CNN is much smaller than the theoretical one. Most of the methods adopted take the form of directly copying the pixel values in one block of the feature map for upsampling, but this is easy to cause image pixel block problem. In order to solve problems, we have adopted the deep coding-decoding network [4], which enhances the understanding of the scene by recording the pooling parameters. At the same time, the deep coding-decoding network also produces two problems. One is ambiguity when dealing with object edges. The other is uncertainty when segmenting small-size objects in scene analysis. And the network does not take advantage of global and sub-regional information of the scene. We address this issue by introducing a spatial pyramid pooling module.

The pyramid pooling module [6] is behind the final layer feature map of deep coding-decoding network, which is explained in part blue dotted box of Fig.1.It contains feature maps which are produced by four different levels image scales. According to image segmentation results, the pooling type we use is average pooling. The top level expressed in red is whole pooling to produce an ordinary export. The feature map is divided into unlike subregions and represent the different location by this module. In the module, different sizes feature maps are acquired by the output of different levels. Then, we utilize 1×1 convolution layer behind each pyramid level to lower the dimension of context expression so that we obtain the weight of global feature. Next, we upsample the low dimensional feature maps and obtain the same size feature as the original feature map by bilinear interpolation. Finally, the ultima pyramid pooling global feature is composed of the different gradations of features. These features are classified pixel by pixel as input to a two-layer BP neural network. The different subregions are obtained by using different size pooling cores in some steps. Hence, a reasonable gap should exist between multiple cores. The module is a four-level one that is composed of the sizes of 1×1, 2×2, 3×3 and 6×6.

B. Superpixel segmentation

As mentioned earlier, before we use the spatial pyramid pooling of features, we use the deep coding-decoding network to extract features from an original picture. The network is prone to blurring the edges of objects and difficult to detect small-size-objects when segmenting images. The

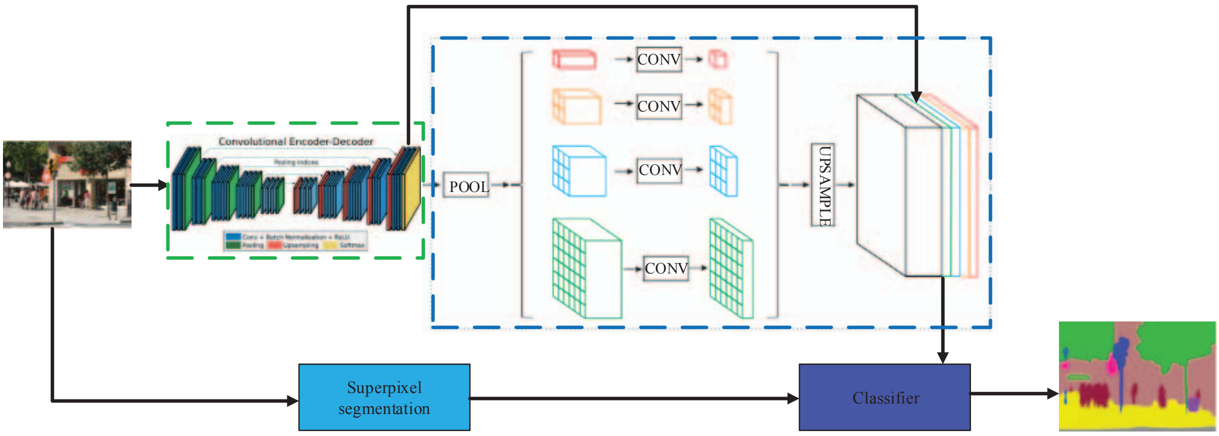

Figure 1. An overview of our proposed superpixel pyramid network.

previous pyramid pooling module can solve the detection problem of small-size-objects by using the information of the global and sub-regions. In order to address the problem of object edge ambiguity when segmenting images, we use the method of superpixel segmentation algorithm and deep coding-decoding network fusion to analyze the image. We adopt graph-based over-segmentation to improve the problem that blurs the edge of the object in the image when the coding-decoding network segments image.

The superpixel segmentation algorithm treats the image as a weighted map. Think of each pixel in the image as a vertex in the graph. The edges between the vertices represent the dissimilarity of the two pixels. The color dissimilarity of the pixel points is calculated and merged into regions. Then, the regions are fused on the basis of an adaptive threshold. First, for each pixel in the graph, calculate the dissimilarity of its 4 neighborhoods. Then, the obtained dissimilarity values are arranged in order from small to large and the smallest dissimilarity is selected. Finally, combine the edges of the smallest dissimilarity. If the two pixels in the vertex connecting the side satisfy that they do not belong to the same area, and the dissimilarity between the two areas is smaller than the dissimilarity inside the two areas respectively, the update threshold and the class labels are performed. Otherwise, reorder and select the next edge to merge and judge until all edges have been traversed.

C. *Appling Network*

The convolution pooling feature in CNN makes the output feature map size much smaller than the input image. In theory, the acceptive domain of ResNet [26] is yet larger than an input image. However, in fact, in especial on high-level layers, the acceptive field of CNN is smaller than the theoretical one

Most of the methods adopt the method of directly copying the pixel values in one block of the feature map for upsampling, which causes the image segmentation to be blockified. We use the deep coding-decoding network to solve this problem. The method of extracting the original image and training the learning model is the same as the reference [4]. It contains a coder network, a relevant decoder network, the connecting pixel-level classification layer, and a decoding network is the same as the 13 convolutional layer of VGG16. The output of the ultima decoder is provided with a soft-max classifier which generates independently class probabilities of each pixel.

Our classifier adopts a two-layer BP neural network. First, features of the original image are extracted using a deep coding-decoding network. Then, spatial pyramid pooling module is performed on the obtained features. Next, train a two-layer BP neural network classifier, randomize weights, and use the gradient descent method for training. Finally, in the verification stage, we perform Gaussian filtering on the original image to de-noising and then perform superpixel segmentation, and linearly interpolate the segmented feature map and the spatial pyramid pooled features. As an input to the trained classifier, this gives the category labels for all pixels. The process of classifying a classifier by a trained classifier is as follows:

1) The output activation value y_i is obtained using the trained classifier, and the pixel-by-pixel distribution $\hat{d}_k$ in the super pixel k is predicted.

$$y_i = W_2 \tanh(W_1 F_i + b) \tag{1}$$

Where W_1, W_2 is the weight of the neural network, b is the offset, and F_i is the input pixel value.

2) When the calculation category is a ,the predicted probability distribution $\hat{d}_{i,a}$ is normalized.

$$\hat{d}_{i,a} = \frac{e^{y_{i,a}}}{\sum_{b \in classes} e^{y_{i,b}}} \tag{2}$$

3) Calculate the predicted probability distribution $\hat{d}_{k,a}$ in the superpixel k .

$$\hat{d}_{k,a} = \frac{1}{s(k)} \sum_{i \in k} \hat{d}_{i,a} \tag{3}$$

Where $s(k)$ is the surface area of the superpixel k .

4) Calculate the label of each superpixel.

$$l_k = \underset{a \in classes}{\arg\max}\, \hat{d}_{k,a} \tag{4}$$

Where l_k is the probability value that pixel k belongs to category a .

IV. Experimental Results

In this section, we evaluate our proposed method on three different datasets, including ADE20K dataset [15], PASCAL VOC2012 semantic segmentation dataset [23] and Camvid

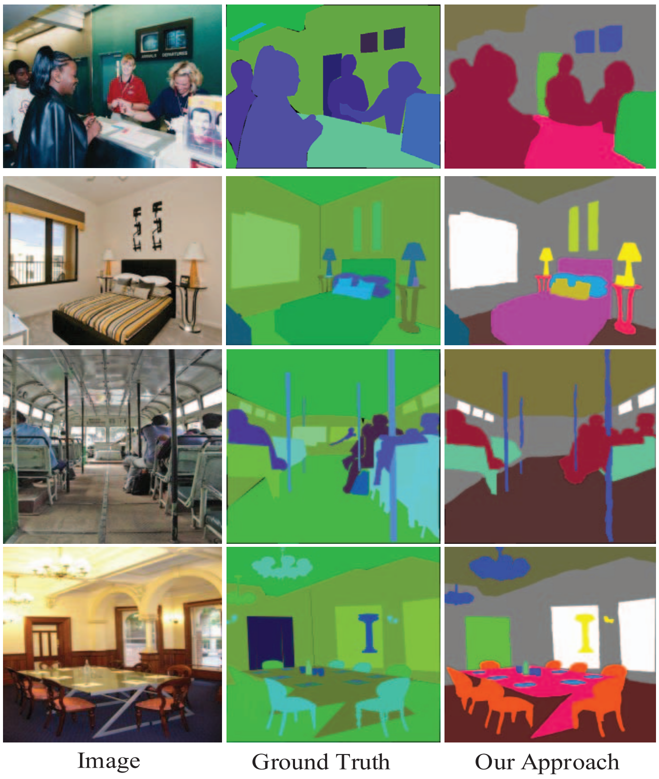

Fig.2. Visual improvements on ADE20K, our method generates precise and detailed results.

dataset [14]. For evaluation, we use pixel-wise accuracy (Pixel Acc.) and mean of class-wise intersection over union (Mean IoU).

Note that all the experiments are implemented with NVIDIA TITAN X GPU with 12-GB memory under CUDNN V5 and CUDA 7.5. Encouraged by [24], we use the "poly" learning rate policy where current learning rate equals to the base one multiplying $(1-\frac{iter}{max\_iter})^{power}$ with base learning rate 0.01 and 0.9. Weight decay and momentum are set to 0.0001 and 0.9 respectively. In order to reduce overfitting, we adopt data augmentation techniques, such as random mirror and random resizing between 0.5 and 2.

A. ADE20K

The ADE20K dataset [15] is a newly dataset for scene parsing. It includes 150 object category labels including objects (e.g., car, chair, etc.) and stuff (e.g., wall, road, etc.), and 1038 image-level scene descriptors. The data has 20K/2K/3K images for training, verifying and testing. More, it also be used to analyze both objects and content in the image. There, parsing the data is more difficult than other datasets.

To evaluate our method, we adopt pixel Acc and mean IoU as our evaluation indicators. We reveal some detailed analyses on the validation set of ADE20K in Table Ⅰ. The proposed method yields the best results 40.2%/79.6% on mean IoU and pixel Acc, which outperforms other prior methods. It exceeds the SegNet by 18.6% and 8.6%. Our framework produces more detailed and accurate results.

Fig.2 reveals another serveal parsing results on validation set of ADE20K. For "person" in the first and third rows, our method is very good at separating the edges of the human body. For "pillow" and "bed" in the second row, we are completely separated. For "bedside lamp" and "bottle" in the second and fourth rows, the method we proposed can detect them completely. Our approach executes well on small-size-object classes in images.

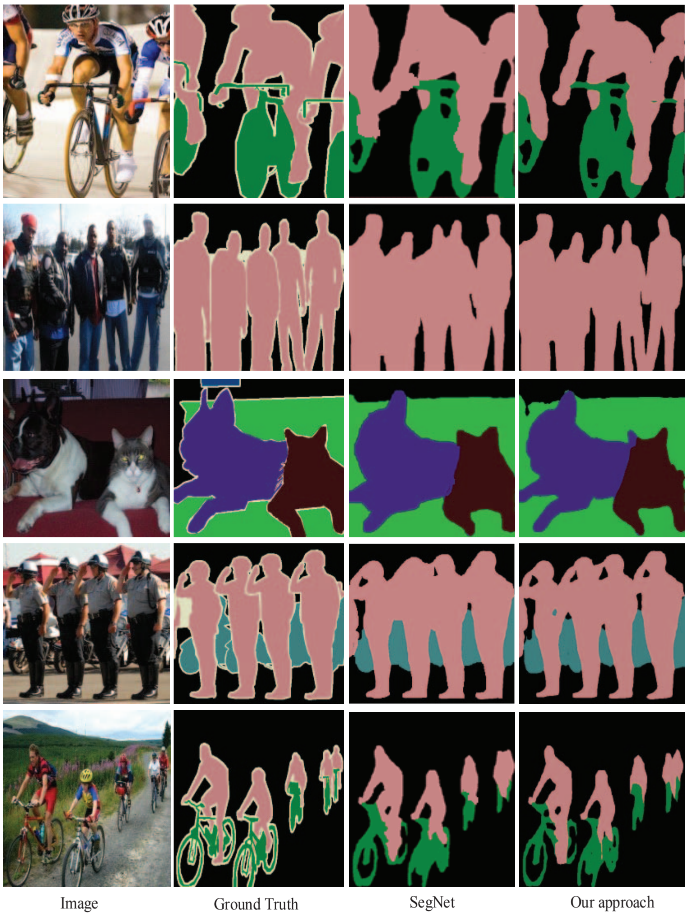

Fig.3. Visual improvements on PASCAL VOC 2012 dataset.

TABLE I. SEGMENTATION RESULTS ON THE ADE20K DATASET.

Experimental Approach	**Experimental Results**	
	Mean IoU(%)	Pixel Acc(%)
FCN [3]	29.4	71.3
SegNet [4]	21.6	71.0
DilatedNet [22]	32.3	73.5
CascadeNet [15]	34.9	74.5
Our approach	**40.2**	**79.6**

B. PASCAL VOC 2012

The proposed approach performs startling on segmantic segmentation. We conduct tests on PASCAL VOC 2012 segmentation dataset. It includes 20 object categories and one background class. Inspired by literature [11], we use 10582 images, 1449 images and 1456 images for training, validation and testing respectively.

Per-class results are demonstrate in TableⅡ, we contrast our method with best-performing approaches on the testing set based on without pre-trained on MS-COCO dataset [25]. When our proposed method is trained with VOC 2012 dataset, the accuracy reaches 88.2%. It is 5.5% higher than the best method. We get the highest accuracy on 14 out of the 20 classes. In particular, the segmentation of small-size objects in the image, such as chair, sofa, boat and bottle, etc., our method greatly improves the degree of overlap of the segments compared to other methods.

TABLE II. THE RESULTS ON PASCAL VOC 2012 TESTING SET.

Method	**Experimental Results**					
	FCN [3]	SegNet [4]	CRF-RNN [16]	GCRF [17]	PSPNet [6]	Ours
aero	76.8	83.3	87.5	85.2	**91.8**	91.1
bike	34.2	42.1	39.0	43.9	**71.9**	70.8
bird	68.9	79.5	79.7	83.3	**94.7**	91.9
boat	49.4	56.9	64.2	65.2	71.2	**87.6**
bottle	60.3	67.4	68.3	68.3	75.8	**88.3**
bus	75.3	86.1	87.6	89.0	**95.2**	93.2
car	74.7	79.8	80.8	82.7	89.9	**91.4**
cat	77.6	81.3	84.4	85.3	**95.9**	93.2
chair	21.4	35.1	30.4	31.1	39.3	**75.6**
cow	62.5	75.2	78.2	79.5	90.7	**92.9**
table	46.8	60.1	60.4	63.3	71.7	**87.1**
dog	71.8	78.6	80.5	80.5	90.5	**92.7**
horse	63.9	73.5	77.8	79.3	**94.5**	91.9
mbike	76.5	82.5	83.1	85.5	88.8	**89.7**
person	73.9	80.4	80.6	81.0	89.6	**90.8**
plant	45.2	57.3	59.5	60.5	72.8	**82.6**
sheep	72.4	80.7	82.8	85.5	89.6	**93.0**
sofa	37.4	47.1	47.8	52.0	64.0	**86.2**
train	70.9	75.6	78.3	77.3	85.1	**92.2**
tv	55.1	61.9	67.1	65.1	76.3	**87.6**
mIoU	62.2	69.2	72.0	73.2	82.6	**88.2**

To exhibit our performance intuitively, we show several examples in Fig.3. For “cyclists” in row one, the SegNet produces disjoint object segments, while our method demonstrates precise boundaries. For “cat” in the third row, our approach executes well on small-size-object classes in images relative with other method. For “person”, “saluting police” and “cyclists in the distance” in the second, fourth and fifth rows, our approach can divide the overall boundaries of people. This implies that the superpixel segmentation is beneficial to semantic segmentation. These features from the baseline highly localized boundary information.

C. CamVid

The CamVid [14] is a dataset collected for semantic segmentation on road scenes. It contains 11 foreground classes and 1 background class, such as: Road, building, vehicle, pedestrian, etc. We carry out experiments on the CamVid dataset, which includes 377 training images and 233 testing image with different conditions. Following Ref.4, we also report our results with mIoU metric.

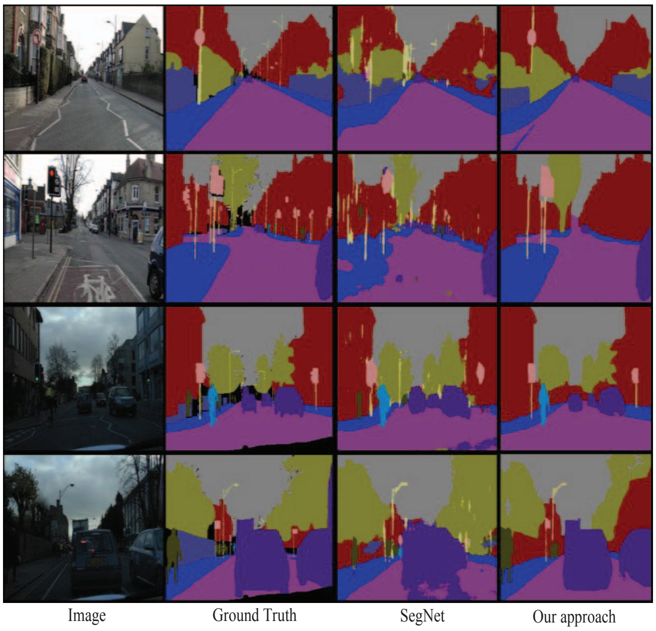

Fig.4. Examples of our method on CamVid dataset.

We compare our approach with traditional methods, containing SegNet [4], Dense Depth Maps [18], Super parsing [20], etc., over the CamVid dataset. As shown in Table Ⅲ, our method executes the best on 8 of the 11 classes and shows nearly 7.8% improvement in mean accuracy. For small-size objects, such as: sign-symbol, bicycle, etc., we get the best accuracy than other methods. Our approach shows excellent performance on these small-size objects.

In order to exhibit our unique contribution, a few visual examples are shown in Fig.4. In the second row, the sidewalk is connected to the edge of the road, while our method is able to separate them well. For the two cars that are connected together and have blurred edges in the fourth row, our method is very good to separate the edges of the two cars to make their edges clearer. These are enough to show that our approach strengthens the edges of the object and accurately segments the object. At the same time, we also saw the problem. For the "street lights" in the first and third rows, there is room for improvement in our approach.

V. Conclusion

We propose a superpixel pyramid network for scene parsing. First, the original image is segmented by the superpixel segmentation method, the edge information of the object in the image is clarified, and the edge blurring in the traditional deep coding-decoding network structure is supplemented. Then, we use a spatial pyramid pooling structure that fuses global and local information for complex objects with similar shapes and sizes, as well as the difficulty of segmentation for some small-size-objects in the scene. This increases the additional spatial contextual information, enhances the use of global information to the whole scene and improves the ability of the cognitive understanding of the scene. Extensive experiments focusing on scene parsing and semantic segmentation show that our approach can obtain better performances than counterparts in terms of accurate parsing and precise segmentations.

TABLE III. THE RESULTS OF OUR METHOD AND OTHER APPROACHES ON THE CAMVID 11 ROAD DATASET.

Method	Experimental Results				
	Dense Depth Maps[18]	Super parsing[20]	Boosting+higher order[21]	SegNet [4]	Ours
building	85.3	**87**	84.5	75	81
tree	57.3	67.1	72.6	84.6	**86.7**
sky	**95.4**	96.9	97.5	91.2	93.9
car	69.2	62.7	72.7	82.7	**85.1**
sign-symbol	46.5	30.1	34.1	36.9	**68.3**
road	**98.5**	95.9	95.3	93.3	95.5
Pedestrian	23.8	14.7	34.2	55	**62.1**
fence	44.3	17.9	45.7	37.5	**51.3**
column-pole	22	1.7	8.1	44.8	**60.8**
sidewalk	38.1	70	77.6	74.1	**80.1**
bicycle	28.7	19.4	28.5	16	**40.7**
mIoU	82.1	83.3	83.8	84.3	**92.1**

REFERENCES

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation[C]// CVPR, 2014, 580-587.

[2] Farabet, C Couprie, L Najman, et al. Learning Hierarchical Features for Scene Labeling[J].IEEE Transactions on Pattern Analysis & Machine Intelligence, 2013, 35(8):1915-1929.

[3] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation[C]// CVPR, 2015: 3431-3440.

[4] Badrinarayanan V, Kendall A, Cipolla R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Scene Segmentation [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, PP (99):1-1.

[5] X Ren, J Malik. Learning a Classification Model for Segmentation[C]// ICCV, 2003:10-17.

[6] H. Zhao et al., “Pyramid scene parsing network,” in IEEE Conf. On Computer Vision and Pattern Recognition (CVPR), pp. 6230–6239(2017).

[7] P F Felzenszwalb, D P Huttenlocher. Efficient Graph-Based Image Segmentation [J]. International Journal of Computer Vision, 2004, 59(2):167-181.

[8] A P Moore, S J D Prince, J Warrell, et al. Superpixel lattices [J]. Dooral H L Nvry Ollg London, 2008:1-8.

[9] M Y Liu, O Tuzel, S Ramalingam, et al. Entropy rate superpixel segmentation[C]//CVPR,2011:2097-2104.

[10] A Levinshtein, A Stere, K N Kutulakos, D J Fleet, S J Dickinson, K Siddiqi. TurboPixels:Fast Superpixels Using Geometric Flows[J]. IEEE Trans. Pattern Anal. Mach. Intell, 2009, 31(12): 2290-2297.

[11] R Achanta, A Shaji, K Smith, Lucchi A, Fua P, and Süsstrunk S. SLIC superpixels compared to state-of-the-art superpixel methods [J]. IEEE Trans. Pattern Anal. Mach. Intell, 2012, 34(11): 2274-2282.

[12] M P Kumar, D Koller. Efficiently selecting regions for scene understanding[C]//CVPR, 2010:3217-3224.

[13] D Munoz, J A Bagnell, M Hebert. Stacked Hierarchical Labeling [J]. Lecture Notes in Computer Science, 2010, 6316:57-70.

[14] G J Brostow, J Fauqueur, R Cipolla. Semantic object classes in video: A high-definition ground truth database [J]. Pattern Recognition Letters, 2009, 30(2):88-97.

[15] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Semantic understanding of scenes through the ADE20K dataset. CoRR, abs/1608.05442, 2016.

[16] 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[C]// ICCV, 2015:1529–1537.

[17] R Vemulapalli, O Tuzel, M Y Liu, et al. Gaussian Conditional Random Field Network for Semantic Segmentation[C]// CVPR, 2016:3224-3233.

[18] C Zhang, L Wang, R Yang. Semantic Segmentation of Urban Scenes Using Dense Depth Maps[C]// ECCV, 2010:708-721.

[19] V Lempitsky, A Vedaldi, A Zisserman. A Pylon Model for Semantic Segmentation [J]. Advances in Neural Information Processing Systems, 2012:1485-1493.

[20] J Tighe, S Lazebnik. SuperParsing: Scalable Nonparametric Image Parsing with Superpixels [C] // ECCV, 2013:352-365.

[21] P Sturgess, K Alahari, L Ladicky, et al. Combining Appearance and Structure from Motion Features for Road Scene Understanding[C]// BMVC, 2009.

[22] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. CoRR, abs/1511.07122, 2015.

[23] M. Everingham, L. J. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman. The pascal visual object classes VOC challenge. IJCV, 2010.

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

[25] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick. Microsoft coco: Common objects in context. In ECCV, 2014.

[26] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2016:770-778.

[27] Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2018, 40(4):834-848.

[28] Liu Z, Li X, Luo P, et al. Semantic Image Segmentation via Deep Parsing Network[C]// IEEE International Conference on Computer Vision. IEEE, 2016:1377-1385.

[29] Liu M, Salzmann M, He X. Discrete-Continuous Depth Estimation from a Single Image[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2014:716-723.