Semantic Segmentation of Parking Lot Scene Based on Surround View System

Jie Wu1,a, Ye He1, * and Xiaoan Chen1,b

1 College of Mechanical and Vehicle Engineering, Chongqing University, China
a email: 2495388321@qq.com, b email: xachen@cqu.edu.cn

*Corresponding author’s e-mail: 201932021002@cqu.edu.cn

Abstract. Semantic segmentation of parking lot scenes is the prerequisite of environment perception for automatic parking technology. It provides environmental semantic information for automatic parking of vehicles. However, due to the dim light of the parking lot environment, unclear ground signs, road reflections and other factors, semantic segmentation such as FCN The method is not yet able to segment ground signs such as background and lane lines to meet the perception needs of automatic parking. This paper proposes a parking lot scene semantic segmentation method based on a surround view system. The surround view system consists of 4 fisheye cameras. The images acquired by each camera are subjected to distortion correction, inverse perspective transformation and image stitching fusion to obtain a ring view. Based on the ring view, a semantic segmentation algorithm based on attention and feature fusion is proposed. Experiments are carried out on a self-made parking lot dataset with a size of 1280x960 pixels. The results show that the method proposed in this paper improves the mIoU of the FCN model by 12.3%.

1. Introduction
Auto-matic parking technology is developed from driving assistance technology. The early parking assistance system commonly used is the reversing image, but it has a large blind spot. With the development of autonomous driving technology, the automatic parking system has developed into a 360-degree surround view system to provide drivers with a full range of visual information[1]. The environment perception of automatic parking is based on the generated ring view, which can accurately segment the ground markings such as parking space lines, lane lines, and arrows to obtain semantic level information. Traditional image segmentation is mainly performed by extracting low-level features of the image, and the results obtained by segmentation do not have semantic level information. In recent years, with the enhancement of computer computing power, researchers have begun to pay attention to semantic-level image segmentation. At the same time, the development of deep learning has allowed the task of semantic segmentation to be solved better, far exceeding traditional methods in terms of accuracy and efficiency[2].

FCN[3] fist proposed a fully convolutional neural network in 2014, replacing the fully connected layer in the VGG-16 network with a convolutional layer, which can achieve end-to-end semantics for input images of any size Segmentation. DeepLab[4] combines the hole convolution and spatial pyramid pooling methods and proposes the ASPP module. The parallel sampling of the hole convolution in ASPP is equivalent to capturing the context of the image at multiple scales and can integrate multi-scale features. The literature[5] proposes a The Pyramid Pooling Module combines the features of 4 different pyramid scales and then obtains the unpoled size through bilinear interpolation, and finally the channels
are superimposed together, and the image pixel classification is done through the convolutional layer. SegNet\[6\] applies the maximum pooling index in the up-sampling process, the resolution of segmentation is improved, and the effect of semantic segmentation is improved. The literature\[7\] uses deconvolution instead of the bilinear interpolation algorithm of the up-sampling method in the decoder to restore some details of the down-sampling of the encoder.

These methods have proposed many new methods to improve the accuracy of semantic segmentation in terms of space and feature channels. However, due to the dim light of the parking lot, the serious wear of ground signs, and ground reflections, the use of hollow convolution will lose a lot of spatial information and local parts. Information, this article adopts the encoder-decoder structure, uses Resnet-101 as the backbone to extract context information, uses independent simple convolution to obtain spatial information, and then uses attention mechanism to obtain global context important features, and uses jump connections to achieve deep and shallow feature fusion, and finally output the semantic segmentation results of lane lines, parking spaces and other ground signs. The segmentation results show that the semantic segmentation algorithm in this paper has higher accuracy.

2. Method

2.1. Surround view system
Due to the lack of field of view in the reversing image and it does not meet the perception requirements of automatic parking, the current automatic parking systems of major auto manufacturers all use ring views to assist the driver in parking. The panoramic surround view system consists of 4 fisheye cameras, The front-view camera is installed under the car logo on the air intake grille; the rear-view camera is installed on the license plate; the left and right cameras are installed under the left and right rear-view mirrors respectively\[8\]. The generation process of the surround view image is shown in Figure 1, and it is mainly divided into the following three steps: calibration and correction of the fisheye camera, inverse perspective transformation, and image stitching and fusion.

The fisheye camera calibration is mainly to determine the conversion relationship between the world coordinate system and the camera image coordinate system. Generally, a checkerboard is used to calibrate the cameras. Firstly, each camera is calibrated separately, and the internal parameters and distortion coefficients of each camera are calculated. Then, a calibration board is laid around the entire vehicle to realize the joint calibration of 4 cameras. The vehicle coordinate system XYZ of the ring view is established, and then the coordinate mapping relationship of the other four images under this coordinate system to the ring view is established, so that the four images can be processed and displayed in the same plane. Use direct linear transformation\[9\] to determine the projection matrix of the inverse perspective transformation, use internal and external parameters to perform distortion correction and inverse perspective transformation on the fisheye image, transform the points of the world coordinate system to the image coordinate system to generate a bird's-eye view, mathematical expressions as follows:
\[
M^{-1}N^{-1} = \begin{bmatrix}
u \\
v \\
1
\end{bmatrix} = \begin{bmatrix}1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{bmatrix}\begin{bmatrix}X_w \\
Y_w \\
Z_w
\end{bmatrix}
\]

among them:

\[
N = \begin{bmatrix}a_x & 0 & u_0 \\
0 & a_y & v_0 \\
0 & 0 & 1
\end{bmatrix}
\]

\(N\) is the internal parameter matrix, \(M\) is the rotation and translation matrix:

\[
M = \begin{bmatrix}r_1 & r_2 & r_3 & t
\end{bmatrix} = \begin{bmatrix}r_1 & r_2 \\
t
\end{bmatrix}
\]

Because the projection surface is a plane, \(Z = 0\), the generality of the representative plane equation is not lost, so the rotation component \(r_3\) can be omitted from \(M\).

The stitching of the ring view uses the coordinate transformation relationship of the 4 fields of view to build the mapping relationship between the 4 fields of view under the fisheye camera to the ring view, and then generates a lookup table corresponding to each other’s coordinates\(^{10}\), and completes the stitching through the lookup table filling. The original field of view and the ring view have the following mapping relationship: the vector from the origin to the point under the ring view is proportional to the vector from the origin to the point on the perspective surface, and the formula is as follows:

\[
\begin{bmatrix}u \\
v \\
f(\rho)
\end{bmatrix} = sM\begin{bmatrix}X_w \\
Y_w \\
Z_w
\end{bmatrix}
\]

\(M\) and \(f(\rho)\) can be obtained by calibration:

\[
M = \begin{bmatrix}b_{11} & b_{12} & b_{13} \\
b_{21} & b_{22} & b_{23} \\
b_{31} & b_{32} & b_{33}
\end{bmatrix}
\]

\[
\rho = u^2 + v^2
\]

\[
f(\rho) = \frac{\rho Z_c}{\sqrt{X_c^2 + Y_c^2}}
\]

Among them, \(X_c, Y_c, Z_c\) is the three-dimensional coordinates of the points on the perspective surface. Therefore, the corresponding relationship can be established from the surrounding image point to the next point in the world coordinate system, and the surrounding view stitching can be completed according to the mapping relationship between the ring view and the corresponding point of the original field of view. It can be seen from Figure 1 that the generated panoramic ring view can reproduce the environment around the car in the form of a bird’s-eye view, and has a wide visual range, even in a place far away from the camera, it can be spliced together well.
2.2. Network structure

The overall structure of the algorithm in this paper is shown in Figure 2, where 1/4 and 1/8 represent the size ratio of the feature map after subsampled, and ×8 represents the upsampling ratio of the feature map. Each of these colored squares represents different network layer. The network mainly includes Convolutional layer, activation layer, batch normalization layer, pooling layer and residual layer. The global pooling layer will change the input spatial dimension to 1. The fusion of spatial information prevents overfitting and improves generalization ability. U-Net\(^{11}\), the portability of using skip connections to restore the full spatial resolution has been proven, but its skip connections are only to force aggregation on the same-scale feature maps of the encoder and decoder, which is an unnecessary constraint. What we propose the network also uses this popular encoder-decoder architecture, uses ResNet-101\(^{12}\) to extract contextual information, uses attention mechanism and jump connections to achieve feature fusion decoders to restore semantic information, and finally outputs the parking lot semantic segmentation results.

Table 1 ResNet-101 network structure of each layer

| layer name | output size | channels | 101-layer | t |
|------------|-------------|----------|-----------|---|
| input      | 1280x960    | 3        | -         | - |
| conv1      | 640x480     | 64       | 7x7, stride 2 | 1 |
| res2       | 320x240     | 256      | 3x3 max pool, stride 2 | 1 |
| res3       | 160x120     | 512      | [1x1, 3x3, 1x1] | 3 |
| res4       | 80x60       | 1024     | [1x1, 3x3, 1x1] | 4 |
| Res5       | 40x30       | 2048     | [1x1, 3x3, 1x1] | 3 |

In this paper, the backbone uses ResNet-101 as shown in Table 1, where t is the number of bottleneck designs, 101-layer represents the convolution operation of each layer. The initial block conv1 consists of 7x7 Convolutional layer; then use the maximum pooling layer and 4 residual layers to gradually reduce the resolution, while increasing the number of feature map channels, the last residual layer res5 reduces the resolution of the feature map to 1/32 of the input image.

In the decoder, the last three layers res2, res3, and res4 of ResNet-101 are first used to obtain the characteristic information of important channels through the attention mechanism module to suppress the output of error information or redundant information\(^{13}\). This article uses the principle of attention mechanism which is similar to human capturing information, and often only focuses on the important
part to obtain important information, so as to enhance the role of effective features and improve efficiency. The attention mechanism module of this article first uses the global feature map Average pooling is used to capture the global context information, and then the 1x1 convolution operation and the sigmoid activation layer are used to obtain the attention vector, and finally the initial feature layer is multiplied to obtain a new feature map. The attention mechanism module provides pixel-level attention to the feature map, making it focus on more informative feature points, that is, expressing important information as much as possible with limited parameters to improve the accuracy of semantic segmentation. BiseNet\cite{14} proposes a Spatial Path to maintain the spatial size of the original input image and encode the spatial information of the image, thereby using a smaller calculation to retain the spatial information, and then fusion with the context information obtained by the Resnet network to achieve high precision Semantic segmentation. This paper also uses a three-layer simple network independent of Resnet-101 to obtain spatial features. Each layer is made up of conv layer, BN and ReLU with a step of 2, and the extracted feature map is 1/8 of the original image.

Feature fusion is a combination of a global feature and a relatively local feature map. Due to the inaccurate semantic information in the shallow features, directly adding the deep feature map and the shallow feature map will cause the error information in the shallow feature or redundant information is added to deep features, which affects the algorithm's segmentation accuracy. This paper adds feature fusion methods in res3, res4 and res5, using feature maps of different sizes to obtain more feature information and further improve the segmentation effect. In order to ensure the same spatial dimensions and feature channels when performing feature fusion, the deep features are processed for attention and up-sampling to the shallow feature map size using linear interpolation, and the deep features and shallow features are concatenated and then used in two consecutive layers. Because a three-layer network independent of Resnet-101 is added to obtain spatial information, there is no need to perform feature fusion on the conv1 and res2 layers, just the final fusion feature of the res3 layer. Just do fusion, which can greatly reduce the amount of calculation. Through this fusion method, the feature details caused by subsampled can be fully restored, and the segmentation effect of small objects and occluded objects is also greatly improved.

This paper uses auxiliary loss functions to supervise the training of our proposed method. We use the main loss function to monitor the output of the entire network. In addition, we added three specific auxiliary loss functions to supervise the output of feature extraction and fusion. All the loss functions are Softmax loss, as shown in equation (8).

\[
\text{loss} = \frac{1}{N} \sum_{i} L_{i} = \frac{1}{N} \sum_{i} - \log \left( \frac{e^{p_{i}}}{\sum_{j} e^{p_{j}}} \right)
\]  

(8)

In addition, we use the parameter \( \lambda \) to balance the weight of the main loss and the auxiliary loss, as shown in equation (9).

\[
L(X;W) = l_{p}(X;W) + \lambda \sum_{i=2}^{5} l_{i}(X_{i};W)
\]  

(9)

Where \( l_{p} \) is the loss predicted by the network output, \( l_{i} \) is the auxiliary loss of the \( i \) stage, and \( K \) is equal to 4.

3. Experiment

3.1. Dataset and evaluation

At present, there is no public semantic segmentation data set of parking lot scenes in China. In order to verify the experimental effect of semantic segmentation in real parking lot scenes, this paper uses a four-channel fisheye camera to record the original video in the parking lot and compare the collected data. After the video data is framed and stitched around, 960 top views with a size of 1280x960 are obtained,
and the corresponding semantic segmentation label map (ground truth, GT) is obtained by manual labeling using the Labelme\textsuperscript{[15]}. We follow the ratio of 0.8:0.1:0.1 Randomly allocate the data set into training set, validation set and test set. The image data set (720 images in the training set, 90 images in the validation set, 90 images in the test set and the images in the data sets are not repeated) includes a total of 6 categories of classification objects. The labeled categories include background, parking space, lane lines, drivable area, For the arrows and the center line, the corresponding GT pixel values are shown in Table 2:

![Table 2](image)

Figure 3 shows an example of a multi-label image segmentation data set in a parking lot scene. Each image in the data set includes lane lines, parking spaces and other objects. The image lighting conditions are complex, the ground is reflective and there are obstacles.

Figure 3 A few sample images and corresponding pixel-annotations

Backbone in this article is Resnet-101 pre-trained on the ImageNet\textsuperscript{[16]} dataset. The software environment is Windows10 64bit operating system, the deep learning software framework is Pytorch 1.8.1, CUDA 10.2, the hardware environment is Intel(R) Core(TM) i5-7300HQ, CPU is 2.5GHz processor, 8GB memory, the GPU is an NVIDIA GTX 1060 6GB. The initial learning rate of the model is set to 1e-5, momentum is set to 0.99, weight decay weight decay is set to 0.0005, the batch size is set to 8 and the number of iterations is 35,000. In order to highlight the improvement in accuracy of the method in this article, network structures such as FCN, SegNet and DeepLabv3 are used for comparison.

The evaluation of the experimental results uses the average intersection ratio for evaluation\textsuperscript{[17]}, where IoU represents the intersection area of the prediction result and the labeled information compared to the union area, and mIoU is the average value of all types of IoU. The specific calculation method is as follows:
\[ mIoU = \frac{1}{N} \sum_{i=1}^{N} \frac{X_i}{T_i} + \sum_{j=1}^{N} \left( \frac{X_{ij} - X_i}{T_i} \right) \]

Among them, \( N \) represents the number of categories of image pixels; \( T_i \) represents the total number of pixels in the \( i \) category; \( X_i \) represents the total number of pixels with the actual type \( i \) and the prediction type \( i \); \( X_{ij} \) represents the total number of pixels with the actual type \( i \) and the prediction type \( j \).

### 3.2. Experimental results and analysis

In order to verify the effectiveness of the algorithm in this paper, a quantitative comparison experiment was performed on segmentation accuracy and speed with the commonly used semantic segmentation networks FCN, SegNet, DeepLabv3, etc. As shown in Table 3, DeepLabv3 uses ResNet-101 as its basis Network, mIoU reached 70.2, but it takes 0.74s to process a 1280x960 picture. SegNet takes 0.28s because of the small network model, but the accuracy is much lower than Deeplav3, mIoU is only 65.6. Original The FCN network does not achieve good results in terms of accuracy and speed. The method proposed in this paper has a greater improvement in accuracy compared with other methods. The mIoU reaches 75.8, which is 5.6% higher than DeeplabV3, and the speed reaches 0.12s, compared with other methods, there is also a significant improvement.

| Method    | BaseModel | mIoU/\% | t/s |
|-----------|-----------|---------|-----|
| FCN       | VGG16     | 63.5    | 0.40|
| SegNet    | VGG16     | 65.6    | 0.28|
| DeepLabv3 | ResNet-101| 71.9    | 70.2| 0.74|
| Ours      | ResNet-101| 76.4    | 75.8| 0.12|

Figure 4 shows the comparison of the results of different semantic segmentation models on the parking lot surround view image on the test set. From left to right are the original image, the label, the FCN test result, the SegNet test result, the DeepLabv3 test result and the semantic segmentation algorithm of this article. Test results. It can be seen from the qualitative comparison that FCN and SegNet have poor segmentation effects, and the segmentation contours of the arrows are not obvious enough. Compared with these methods, the segmentation quality of the algorithm in this paper is better, and the occluded arrows and parking spaces are higher. The segmentation accuracy is better and more complete for the segmentation of lane lines.
The results of semantic segmentation and several superimposed effects of the original ring view are shown in Figure 5. It can be seen that the algorithm in this paper can clearly segment the parking space line, lane line and other ground signs. It can also reasonably divide the driving area, which provides an essential sensory condition for the automatic parking technology in the parking scene.

3.3. Improvements

In the semantic segmentation of parking lot scenes, the algorithm in this paper has achieved good results in terms of accuracy, but experiments have found that the parking lot scenes are complex, such as obstacles, severe ground reflections and dark light, as shown in Figure 6, (a) Obstacles, (b) Reflective, (c) The light is relatively dark. To solve these problems still need to do more research, high-definition cameras as part of unmanned perception is an important sensor, it has on light conditions demand is
higher, however, cameras and other sensors data fusion is an inevitable trend in the field of unmanned, such as fusion of millimeter wave radar, ultrasonic, laser radar, etc. To some extent, this method overcomes the limitation of having only a single sensor, and provides more high-dimensional spatial information, which can improve the robustness of the algorithm and the system.

(a)Obstacles                        (b)Ground reflection                        (c)Low light

Figure 6 Three types of poor experimental results

4. Conclusions
The semantic segmentation method of parking lot scene based on surround view system proposed in this paper provides environmental semantic information for automatic parking of vehicles. A surround view system composed of 4 fisheye cameras. First, perform distortion correction on the images acquired by each camera, then inverse perspective transformation, and finally stitch and merge to obtain a ring view. The ring view based on the parking lot scene uses the attention mechanism and features. The fusion method improves the mIoU accuracy of segmentation. This paper collects and manually annotates 960 parking lot data with a size of 1280 x 960 pixels and conducts experiments on this dataset. The results show that the method proposed in this paper is 12.3% higher than the mIoU of the FCN and it can better segment the ground signs such as lane lines and parking spaces. However, in the process of research, we found that the categories of this article are not detailed enough. For example, obstacles such as vehicles are simply classified as background, and no additional treatment is taken for the reflective and dark parts of the ring view, so we focus on the following improvement. We will improve the label category, no longer limited to the segmentation of ground signs, but also separate the obstacles such as vehicles and pedestrians from the background, so as to build a more reasonable semantic segmentation of parking scenes.

References
[1] Csurka G, Perronnin F. An efficient approach to semantic segmentation[J]. International Journal of Computer Vision, 2011,95(2): 198-212.
[2] He Y H, Wang H, Zhang B. Color-based road detection in urban traffic scenes[J]. IEEE Transactions on Intelligent Transportation Systems, 2004, 5(4): 309-318.
[3] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3431–3440, 2015.
[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence, 40(4):834–848, 2018.
[5] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In IEEE Conf. on Computer Vision and Pattern Recognition(CVPR), pages 2881–2890, 2017.

[6] V. Badrinarayanan, A. Kendall and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 12, pp. 2481-2495, 1 Dec. 2017, doi: 10.1109/TPAMI.2016.2644615.

[7] L. Mou, P. Ghamisi and X. X. Zhu, "Unsupervised Spectral–Spatial Feature Learning via Deep Residual Conv–Deconv Network for Hyperspectral Image Classification," in IEEE Transactions on Geoscience and Remote Sensing, vol. 56, no. 1, pp. 391-406, Jan. 2018, doi: 10.1109/TGRS.2017.2748160.

[8] Ying X., Hu Z., Zha H. (2006) Fisheye Lenses Calibration Using Straight-Line Spherical Perspective Projection Constraint. In: Narayanan P.J., Nayar S.K., Shum HY. (eds) Computer Vision – ACCV 2006. ACCV 2006. Lecture Notes in Computer Science, vol 3852. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11612704_7.

[9] Liang Chen, Charles W. Armstrong, Demetrios D. Raftopoulos, An investigation on the accuracy of three-dimensional space reconstruction using the direct linear transformation technique. Journal of Biomechanics, Volume 27, Issue 4, 1994, Pages 493-500, ISSN 0021-9290.

[10] T. Ho and M. Budagavi, "Dual-fisheye lens stitching for 360-degree imaging," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 2172-2176, doi: 10.1109/ICASSP.2017.7952541.

[11] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation[C] //Proceedings of Medical Image Computing and Computer Assisted Intervention. Heidelberg: Springer, 2015: 234-241.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CVPR 2016, 2016.

[13] Lin T, Dollar P, Girshick R, et al. Feature pyramid networks for object detection[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 936-944.

[14] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao,Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. arXiv preprint arXiv:1808.00897, 2018.

[15] Russell, B.C., Torralba, A., Murphy, K.P. et al. LabelMe: A Database and Web-Based Tool for Image Annotation. Int J Comput Vis 77, 157–173 (2008). https://doi.org/10.1007/s11263-007-0090-8.

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In CVPR, 2009.

[17] Rezatofighi, S. H. (2019) Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In: CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach. pp.658-666.