Indoor Scene Semantic Segmentation Based on RGB-D Image and Convolution Neural Network

Guitang Wang¹, Ziyu Wang¹, Yongbin Chen², Guozhen Wang² and Jianqiang Chen³

1. School of Electromechanical Engineering, Guangdong University of Technology, No. 100, West Ring Road, Guangzhou University Higher Mega Center, Xiaoguwei street, Panyu District, Guangzhou, China.
2. Foshan Cangke Intelligent Technology Co., Ltd. Foshan, China.
3. Guangzhou Cangheng Automatic Control Technology Co., Ltd. Guangzhou, China.
Email:13380039269@163.com

Abstract. In recent years, convolutional neural network has been widely used in image semantic segmentation and achieved great success. In this paper, a semantic segmentation network of indoor scene based on rgb-d image is proposed: SRNET (Strong supervision Residual Net). In this network model, the original data is processed by separate training and gradual fusion, and the mandatory supervision module is added in the decoding stage, which effectively improves the accuracy of semantic segmentation. At the same time, anti residual decoding method and jump structure are introduced to reduce information loss. Experimental results show that the segmentation accuracy of this model is better than most of the current segmentation algorithms.

1. Introduction

As the key technology for robot to realize scene, semantic segmentation algorithm is the key for robot to realize artificial intelligence and interact with the outside world. In essence, the semantic segmentation algorithm is actually a classification algorithm. The algorithm judges the category of each pixel in the input image, then calibrates it, and finally obtains the category of the region where the object is located in the image.

At present, there are two mainstream semantic segmentation algorithms: traditional segmentation algorithm and deep learning algorithm. The traditional segmentation algorithm uses interval information, image features, classification algorithm and other methods to segment the input data semantically. For example, Zhang et al. [1] proposed the hidden Markov random field (MRF) model, which uses the indirect estimation of the random process generated by MFR to segment the image, and Hebert [2] And Kumar et al. [3] proposed the conditional random field (CRF) model in the field of image semantic segmentation. In addition, felzenszwalb et al [4] proposed a method of applying the hog feature operator to semantic segmentation. Karsnas et al. [5] use region based segmentation method to segment the image into small blocks with the same properties. Yang et al [6] proposed an image segmentation algorithm based on hierarchical model using support vector machine algorithm. Cohen et al. [7] used several color spaces, such as RGB, YCbCr, HSL, lab and YIQ, to segment images.

The semantic segmentation of base and deep learning is the mainstream in the research of semantic segmentation algorithm. Long et al [8] proposed FCN structure in CVPR computer vision conference, replacing the full connection layer with convolution layer in traditional neural network, and introduced two structures of up sampling and jump structure. Noh et al [9] proposed that the network structure of...
Deconvnet adopts the structure of encoding decoding referring to the model of FCN. Badrinarayanan et al [10] proposed SegNet network and adopted a network structure similar to Deconvnet to realize smoother de pooling operation at decod side. He [11] et al. Proposed a deep convolution neural network model, ResNet, introduced a residual module to deal with the problems in depth learning such as gradient explosion.

However, FCN has been difficult to get high precision in the semantic segmentation of indoor environment. This is because the color selection of RGB image blurs the boundary between objects, so that a lot of space information in the scene is lost. McCormac et al. [12] directly combine depth information and RGB information into 4-channel information, It improves the effect of semantic segmentation and the accuracy of segmentation. Hazerbas et al. [13] proposed a processing method of gradual fusion of training and image processing through subsequent experiments, and achieved better experimental results than depth information stack fusion.

In this paper, an optimized network structure based on ResNet is proposed. In the lower sampling stage, the image information is extracted by the method of training and gradual fusion respectively. In the up sampling process, the strengthened supervision module is added to optimize the result of semantic segmentation. Finally, the result of semantic segmentation in indoor scene is very good.

![SRNet structure](image)

**Figure 1. SRNet structure**
2. SRNet Structure
Proposed semantic segmentation network is SRNet (Strong supervision Residual Net). As shown in Figure 1, the network is a deep neural network based on ResNet-34 [11], which has two branches: RGB image training branch (i.e. main branch) and deep image training branch (i.e. secondary branch). This is the structure of gradual integration of separate training mentioned above.

Different from the traditional encoding-decoding neural network structure, the residual module structure is introduced in the process of down sampling and up sampling. At the same time, the enhanced supervision module is introduced in the process of up sampling

2.1. Residual and Unresidual Modules
The standard residual module is used in this paper, as shown in Figure 2. Standard residual module, as shown in Figure 2 (a), where conv refers to convolution operation, [(3,3), 1, 1] refers to convolution kernel is 3 × 3, convolution step is 1, and channel coefficient is 1. As shown in Figure 2 (b), there is one more subsampling branch for the residual module with subsampling compared with the standard residual module, which is equivalent to a subsampling of the input size. In order to keep the output of the main branch and the secondary branch can be added directly, the size of its space size and the number of feature channels must be converted.

![Figure 2. Residual module, (a) Standard residual module(b) Residual module with subsampling](image)

Similar to the residual module, the unresidual module has two forms, as shown in Figure 3. As shown in Figure 3 (a), the standard unresidual module has the same network structure as the common residual module. As shown in Figure 3 (b), the structure of the unresidual module with up sampling is similar to that of the residual module with down sampling. The main difference lies in the selection of convolution steps and channel coefficients in the convolution process. One is the encoding process and the other is the decoding process.

![Figure 3. Unresidual module, (a) standard unresidual module(b) unresidual module with up sampling](image)

2.2. Complete Algorithm Structure
This model adopts the structure of encoding decoding. In the first half of the network, the input image is sampled down to reduce the spatial size of the feature layer. In the second half, the input image is sampled up to enlarge the spatial size of the feature layer. The final output segmentation result is the same size as the
input image. In addition, the network also adopts the jump structure proposed by long and so on, which effectively improves the accuracy of semantic segmentation by transferring the fine features of the underlying information to the high-level abstract information.

In the coding phase, SRNet uses the method of training and gradual fusion to fuse the information of RGB image and depth image. In the main branch of the model, the structure from the first convolution layer conv1 to the residual layer layer4 and the structure of the secondary branch adopt the resnet-34 structure proposed by he [11]. The specific operation of training the gradual fusion of information is as follows: the output of the main branch pool1 layer and the output of the secondary branch pool1-d layer are fused in the form of adding elements, and the fusion result is taken as the input of the main branch layer1 layer. By analogy, the input data of layer2, layer3, layer4 and TRANS1 in the main branch are respectively from the fusion of the operation results of layer1, layer2, layer3, layer4 in the main branch and layer1-d, layer2-d, layer3-d and layer4-d in the secondary branch. After the fusion of layer4 and layer4-d, all the data information is collected into the main branch, the secondary branch is finished, the model coding is finished, and there is only one main branch in the subsequent up sampling decoding stage.

In the decoding phase, SRNet adopts three jump structures to retain fine features to improve the segmentation accuracy. But the structure after layer 4 is replaced by the deconvolution layer with the unresidual module, which is not the traditional ResNet structure. As shown in Figure 1. The input of trans2 layer is generated by the combination of the demerit recording of layer 3 and layer 3-D information fusion and the output information of Trans1 layer. In turn, for example, this pattern runs through the decoding stage of the up sampling and finally optimizes the segmentation details of semantic inference. The reason why only three jump structures are designed and the results of trans4 and pool1 and pool1-d are not fused as the input of trans5 is that the results of pool1 and pool1-d are only convoluted once, and the detailed information is very original, which is similar to that of trans4 There is a big difference between the output of the layer and that of the final output. Direct fusion will reduce the segmentation accuracy, so we only do three jumps fusion. Among them, layer1 and trans5 use ordinary residual module, layer2-4 use residual module with down sampling, trans1-4 use residual module with up sampling, and the final trans layer is unconvolution layer.

In this paper, the enhanced supervision module is added in the decoding stage, as shown in Figure 1. Its function is to strengthen the supervision of the activation layers of TRANS1 to 4 in the sampling stage of the model, force each convolution layer to extract useful features for segmentation, and optimize the final semantic segmentation results layer by layer. The supervision module is also a convolution layer. The convolution kernel of convfs1 to convfs4 is 1 × 1, the stride is 1, and the output channel is 37. That is to say, it does not change the spatial size of the input data, but classifies the objects at this position, and the input is TRANS1 to trans4, four unresidual layers.

3. Construction and Training of Experimental Network

In this paper, SUN RGB-D image data set is used [14]. This is the most widely used RGB-D semantic segmentation dataset. The dataset has 10335 RGB images and depth images synchronized data. The data includes 20 scenarios and 37 categories. Because of the use of four rgb-d image acquisition devices (intelrealsense, ASUS xton, Kinect V1, and Kinect V2) in data acquisition, the data has four different sizes. The size of the input data in this paper is set as 640 × 480. In the dataset, 5285 pictures are used as training sets, 1000 of which are verification sets, and the rest 5050 are test sets.

3.1. Parameter Initialization

SRNet is an optimization algorithm based on ResNet-34, so when initializing the parameters, we introduce the idea of migration learning [15] to assign the network parameters. Resnet-34 is used to train the object classification task on the ImageNet data set. After the training, the parameters of previous conv1 to Layer4 are initially the corresponding values in the training. Because Conv1 input RGB three channel image, and in the secondary branch, Conv1-d input is single channel depth information, so we can not simply assign Conv1 parameters to Conv1-d. Different convolution checks are more sensitive to different colors in the acquisition of RGB three-channel information, so the convolution core parameters that are only sensitive to gray-scale image can be obtained by averaging the parameters of RGB three components.
3.2. Network Training

When training the model, we need to use loss function to measure the performance of the model. The lower the loss function, the better the performance of the training model. The cross entropy loss function is used to evaluate the model. The formula of cross entropy function is as follow:

\[ P(x = k) = \frac{\exp(a_i)}{\sum_{i=1}^{L} \exp(a_i)}, i = 1, 2, L K \]  

\[ L = -\log \left( P(x = k) \right) = -\log \left( \frac{\exp(a_i)}{\sum_{i=1}^{L} \exp(a_i)} \right), i = 1, 2, \ldots K \]  

In the above formula (1), \( P(x = k) \) is the probability that a pixel belongs to its correct category \( K \), \( K \) represents the number of categories in the classification algorithm, and \( A_i \) represents the eigenvalue of the \( i \)th category. Equation (2) is the formula form of cross entropy when Softmax function is used in the last layer of the network. The loss function of this model is the sum of five cross entropy functions, which are constructed from convfs1 to convfs4 and the final output.

When training the model, the parameters are updated by the random gradient descent method of the coefficient of driving quantity. The initial momentum coefficient is 0.9, and the initial learning rate is 0.002. After 100 times of training with the whole data, the learning rate will be multiplied by the coefficient of 0.9 for attenuation.

The experiment is carried out on the laboratory platform, the data of which are as follows: Intel i7-7700k central processor, NVIDIA Geforce 1080ti graphics card, Ubuntu 16.04 operating system.

4. Analysis Of Experimental Results

The experimental results of semantic segmentation are shown in Figure 4.

It can be seen from Figure 4 that the network clearly defines the contour of common indoor objects, such as the ground, wall, chair, bed, box, etc., and clearly demarcates the segmentation area.

There are various evaluation criteria for semantic segmentation. The three most commonly used criteria are:
1) PA (Pixel Accuracy), Percentage of accuracy correctly classified in all pixels:

\[ P = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k} t_i} \]  

(3)

2) MA (Mean Accuracy), Percentage of accuracy correctly classified in all pixels average of pixel accuracy of class objects:

\[ M = \frac{1}{k} \sum_{i=0}^{k} \frac{p_{ii}}{t_i} \]  

(4)

3) Area intersection precision (IoU), the percentage of the intersection of the correct segmentation area of various objects and the algorithm output segmentation area in the total area of the two:

\[ IoU = \frac{1}{k} \sum_{i=0}^{k} \frac{p_{ii}}{t_i + \sum_{j=0}^{k} p_{ij} - p_{ii}} \]  

(5)

Among them, \( p_{ii} \) is the number of pixels classified correctly by the algorithm in class I, \( p_{ij} \) is the number of pixels classified into class j (including the case of \( i = j \)), \( t_i \) is the total number of pixels in the correct class i region, and \( k \) is the number of all classes.

### Table 1. Comparison of accuracy between SRNet and other networks

|                  | Pixel | Mean | IoU |
|------------------|-------|------|-----|
| FCN-32s[8]       | 68.4  | 41.2 | 28.9|
| FCN-16s[8]       | 67.6  | 38.7 | 27.2|
| SegNet[10]       | 71.2  | 46.0 | 30.5|
| Context-CRF[17]  | 78.5  | 53.2 | 42.2|
| FuseNet[13]      | 76.3  | 48.3 | 37.3|
| RefineNet[18]    | 80.6  | 57.8 | 45.8|
| SRNet0           | 80.4  | 55.6 | 45.1|
| SRNet            | 80.9  | 58.4 | 46.9|

Table 1 shows the comparison between SRNet and common network segmentation accuracy. Under the same experimental conditions, we first use the classical semantic segmentation network to complete the segmentation task and get the segmentation accuracy, then train the SRNet to get the segmentation accuracy. It should be noted that SRNet0 does not adopt the enhanced supervision module, and SRNet adopts the enhanced supervision module.

From the data analysis in Table 1, it can be seen that the segmentation accuracy of SRNet is higher than that of other common semantic segmentation networks, with outstanding performance. From the data comparison between SRNet0 and SRNet, it can be concluded that the enhanced supervision module can improve the network accuracy to some extent.

## 5. Summary

In summary, this paper proposes an optimized network model based on ResNet.

1. In the process of down sampling, the information of the main and secondary branches is fused by means of training and gradual fusion respectively, which effectively retains the original information in the data and improves the accuracy of the semantic segmentation network.

2. In the process of up sampling, we add the module of strengthening supervision, and supervise the decoding layer of each layer.

The result shows that this module improves the segmentation accuracy of the network.

3. In the process of encoding and decoding, the residual module is added, which effectively solves the problem of gradient dispersion when parameters are updated.
6. Acknowledgement

This study was supported by: Guangzhou Science and technology plan Pearl River science and technology new star project (201806010128), Foshan GDUT research institute innovation and Entrepreneurship Talent Team plan project.

7. References

[1] Zhang Y, Brady M, Smith S. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm[J]. IEEE Transactions on Medical Imaging, 2002, 20(1):45-57.

[2] Hebert K S. Discriminative random fields: A discriminative framework for contextual interaction in classification[C]. IEEE International Conference on Computer Vision, Nice: IEEE Computer Society, 003:1150—1157.

[3] Kumar S, Hebert M. Discriminative fields for modeling spatial dependencies in natural images[C]. Advances in Neural Information Processing Systems. Vancouver: Neural Information Processing Systems Foundation, 2004:1531—1538.

[4] Felzenszwalb P F, Girshick R B, Mcallester D, et al. Object Detection with Discriminatively Trained Part-Based Models[J]. IEEE Transactions on Software Engineering, 2010, 32(9):1627-1645.

[5] Andreas Kärnsä, Dahl A L, Larsen R. Learning histopathological patterns[J]. Journal of Pathology Informatics, 2011.

[6] Yang Y, Hallman S, Ramanan D, et al. Layered Object Models for Image Segmentation[J]. IEEE Transactions on Software Engineering, 2012, 34(9):1731-1743.

[7] Memory based active contour algorithm using pixel-level classified images for colon crypt segmentation[J]. Computerized Medical Imaging and Graphics, 2015, 43:150-164.

[8] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[A]. Computer Vision and Pattern Recognition. Boston: IEEE Computer Society, 2015: 3431—3440.

[9] Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation[C]. IEEE International Conference on Computer Vision, Santiago: IEEE Computer Society, 2016: 1520—1528.

[10] Badrinarayanan V, Kendall A, Cipolla R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Scene Segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017:1-1.

[11] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[J]. 2015.

[12] McCormac J, Handa A, Davison A, et al. SemanticFusion: Dense 3D Semantic Mapping with Convolutional Neural Networks[J]. 2016.

[13] Hazirbas C, Ma L, Domokos C, et al. FuseNet: Incorporating Depth into Semantic Segmentation via Fusion-Based CNN Architecture[J]. 2016.

[14] Song S, Lichtenberg S P, Xiao J. [IEEE 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) - Boston, MA, USA (2015.6.7-2015.6.12)] 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) - SUN RGB-D: A RGB-D scene understanding benchmark suite[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition.

[15] Thrun S, Pratt L. Learning to learn[M],[S.l.]: Springer Science & Business Media, 2012. Recognition (CVPR). IEEE, 2015:567-576.

[16] 项前,唐继婷,吴建国. 多级上采样融合的强监督 RGBD 显著性目标检测[J]. 计算机工程与应用, 2019,20:1-9.

[17] Lin G, Shen C, Hengel A v d, et al. Exploring Context with Deep Structured models for Semantic Segmentation[J]. arXiv preprint arXiv:1603.03183, 2016.

[18] Lin G, Milan A, Shen C, et al. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation[A]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C], 2017