Blind sidewalk segmentation based on the lightweight semantic segmentation network

Xingjian Liu, Xinge Zhao* and Sizhan Wang
The Northeastern University, College of Information Science and Engineering, Shenyang, China

*Corresponding author. Email: 20172303@stu.neu.edu.cn

Abstract. A lightweight semantic segmentation network is proposed to solve the problem of real-time segmentation of blind sidewalk. On the basis of U-Net's encode-decoding structure, the inverse residual block composed of deep separable convolution is used to replace the ordinary convolution for feature extraction, and the number of convolution layer channels is compressed. Experiments show that the blind segmentation method used can effectively overcome the disadvantages of traditional methods that are affected by the environment easily. Compared with U-Net, the parameter amount is reduced by 25.2 times, and the reasoning speed is increased by 3.6 times, while the loss of precision is just less than 2%, which basically meet the requirement of real-time inference

Keywords. lightweight, U-Net, deepth separable convolution, blind sidewalk.

1. Introduction
According to statistics from the WHO, there are 140 million visually impaired people in the world. Visual defects can seriously hinder people's daily activities. However, there is no convenient and low-cost blind guide device. Due to the ability to obtain sufficient information, vision-based blind guidance technology has become a research focus. [1] Among them, how to accurately segmenting the blind sidewalk area from the complex environment is an indispensable technology in visual guidance.

Traditional blind sidewalk segmentation methods mainly construct a classifier by extracting the color or texture features of the blind sidewalk, thereby segmenting the blind sidewalk area. [2] But these methods are either susceptible to environmental interference or require a longer calculation time to run [3], which cannot be well applied to actual blind sidewalk segmentation.

In recent years, deep learning has made great breakthroughs in image semantic segmentation and other fields. The fully convolutional network (FCN) [4] replaced the fully connected layer at the end of the traditional convolutional neural network with a convolutional layer, formally introduce convolutional neural networks into the field of semantic segmentation. Ronneberger et al. proposed U-Net for the first time and applied it to medical image segmentation. [5] Yang et al. constructed a deep learning network in the form of encoding-decoding to segment stairs, sidewalks, etc. from RGB-D images. [6] Yu et al. proposed a lightweight neural network BiSeNet, which has a spatial path (SP) and a context path (CP) that uses a lightweight model to quickly obtain a larger receptive field. [7]

In order to solve the problem that traditional blind sidewalk segmentation methods are greatly affected by the environment and lack of universality, we propose a blind sidewalk condition detection...
model based on a lightweight semantic segmentation network. To be precise, we use U-Net's encoding-decoding structure to build the network and reduce model parameters by replace the convolutional layer in U-Net with the inverse residual structure to complete the feature extraction operation. The experimental results show that the segmentation effect of the improved network is significantly better than the traditional segmentation method. Compared with U-Net, the model parameters are reduced by 25.2 times, and the running speed is increased by 3.6 times, while the accuracy of the model is only slightly reduced.

2. Proposed methods

U-Net is a segmentation network with an encoding-decoding structure which has outstanding performance in small sample size segmentation tasks. In this paper, we use U-Net-based semantic segmentation network for blind segmentation.

![Figure 1. The structure of proposed model.](image1)

The semantic segmentation network we proposed is shown in figure 1. We replace the ordinary convolution of the feature extraction part of U-Net with reverse residual structure composed of deep separable convolution. The depth separable convolution decomposes the convolution operation into two steps: deep convolution with channel separation and 1x1 point-wise convolution, which can effectively reduce the parameter amount. Since each channel performs a convolution operation separately, the depth separable convolution will cause information loss, so the residual structure is often combined with the inverse residual structure.

As shown in figure 2, on the basis of the residual structure, the reverse residual structure first uses 1×1 to expand the number of channels to retain richer information and then performs convolution with a convolution kernel size of K×K in the depth direction. Finally, 1×1 convolution is used to compress the number of channels to the original size. The inverse residual structure can combine the advantages of the deep-separated convolution and residual structure to make up for a certain information loss when reducing the amount of convolutional layer parameters.

![Figure 2. Reverse residual structure with deep separable convolution.](image2)

In addition, U-Net performs 4 down-sampling operations in the feature extraction stage and uses a larger number of channels in convolution layers. It ensures a certain accuracy, but leads to a large amount of model parameters, which sacrificed the speed of reasoning to a certain extent. In order to further reduce the weight of the model, we reduce the downsampling operation to 3 times. Compared with U-
Net, the number of channels of the convolutional layer is compressed by 1/4 time, which effectively reduces the amount of parameters without sacrificing too much accuracy.

3. Experiment and analysis

3.1. Datasets
Currently, there is no dataset for blind sidewalk segmentation. Therefore, We first collect 400 blind sidewalk images in different scenes, lighting and shadows. Then, we used Labelme annotation tool to manually annotate the dataset, and each pixel in the image is divided into blind sidewalk area or background area. In order to avoid the poor classification performance caused by over-fitting due to the small number of training samples, data enhancement should be carried out on the training set. [9] We randomly rotate each image between -15° and 15°, followed by random translation of 10-40 pixels in four directions and randomly cropped the image. In the end, the dataset is enlarged to 2000 images with the size of 512x512.

3.2. Experimental environment
The model training in this paper uses Titan RTX GPU on the server with 12G memory. The server system is Ubuntu18.04 with CUDA version 10.0. We used the deep learning framework Tensorflow1.15, the programming language is Python 3.6, in addition to Numpy, OpenCV and other common libraries.

3.3. Performance indices
In order to evaluate the segmentation effect, we used pixel accuracy (PA), Recall rate and average crossover ratio (MIoU) as evaluation metrics to measure the model performance. [10] The three metrics can be expressed as in (1), (2) and (3)

\[
PA = \frac{TP + TN}{TP + TN + FP + FN}
\]

Recall = \frac{TP}{TP + FN}

\[
MIoU = \frac{1}{k + 1} \sum_{i=0}^{k} \frac{TP}{TP + FP + FN}
\]

In the above formulas, TP is the number of correctly classified blind pixels, TN is the number of background pixels correctly classified, FP is the number of blind pixels that are incorrectly classified as background, and FN is the number of background pixels that are wrongly divided as blind.

3.4. Results and analysis
We use the same dataset and training parameters to train U-Net our model. The segmentation results obtained are shown in figure 3.
Figure 3. Blind sidewalk segmentation results. (a) represents the original image, (b) is manually labeled ground truth, (c) is the segmentation result obtained by using threshold-based method, (d) is the U-Net segmentation result, and (e) is the segmentation result obtained by proposed methods.

Comparing the segmentation results of the first three lines in figure 3, it can be found that the traditional threshold-based segmentation method is greatly affected by the illumination. When the blind sidewalk color is similar to the surroundings, the traditional method will also lose its segmentation ability. 3.(c), 3.(d) and 3.(e) show that using U-Net and the network proposed in this article can effectively overcome the shortcomings of traditional segmentation methods. The segmentation results are less affected by the surrounding environment, which means it can effectively segment the blind sidewalk in a more complex environment. What’s more, comparing 3.(d) and 3.(e), the segmentation effect of our model is similar to U-Net although the model parameters have been greatly reduced.

In order to further prove the effectiveness of the proposed method, this paper tests the performance indicators of the model and compares them with the U-Net network. The results are shown in Table 1.

|       | PA     | Recall | MIoU  | Time(cpu) | Weight size |
|-------|--------|--------|-------|-----------|-------------|
| U-Net | 0.9828 | 0.9324 | 0.9416| 1.56s     | 12.6M       |
| Proposed | 0.9674 | 0.9211 | 0.9273| 0.43s     | 0.5M        |

In the feature extraction stage, by replacing ordinary convolution with deep separable convolution and compressing the number of convolutional layer channels in the feature extraction stage, the model parameters can be reduced by 25.2 times. The inference speed of a single 512x512 input image is accelerated to 0.43s, which is 3.6 times faster on cpu, while the accuracy is only reduced by about 1.5%.

4. Conclusion

In order to segment the blind sidewalk area from the image in real time, a lightweight semantic segmentation network based on the U-Net’s encoding-decoding structure is designed. The inverse residual composed of depth separable convolution is used to replace ordinary convolution operation, and the number of channels of the convolutional layer is compressed to reduce the parament amount. Experimental results show that the proposed method can effectively reduce the parameters and increase the speed of inference with little accuracy loss. It can be further deployed and applied to the mobile terminal as an auxiliary means for the visually impaired to walk safely on the blind sidewalk.
References

[1] A. Budrionis, Plikynas, et al. Smartphone-based computer vision travelling aids for blind and visually impaired individuals: A systematic review. *Assistive Technology*. 6, 60 (2020).

[2] Y. Peng, J. Xue, and Y. Guo, Blind sidewalk recognition algorithm based on color and texture information. *Journal of Computer Applications*. 12, 37 (2014)

[3] T. Wei, L. Yuan, Highly real-time blind sidewalk recognition algorithm based on boundary tracking. *Opto-Electronic Engineering*. 7, 676-684 (2017)

[4] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition. 3431-3440 (2015)

[5] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention. 234-241 (2015)

[6] K. Yang, X. Hu, L. Bergasa, et al. Pass: Panoramic annular semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*. 10, 4171-4185 (2019)

[7] C. Yu, J. Wang, C. Peng, C. Gao, et al. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In Proceedings of the European conference on computer vision (ECCV). 325-341 (2018)

[8] M. Sandler, A. Howard, M. Zhu, et al. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 4510-4520 (2018)

[9] C. Shorten, T. Khoshgoftaar, A survey on image data augmentation for deep learning. *Journal of Big Data*. 6, 60 (2019)

[10] B. Bischke, P. Helber, J. Folz, D. Borth, A. Dengel, Multi-task learning for segmentation of building footprints with deep neural networks. In IEEE International Conference on Image Processing (ICIP). 1480-1484 (2019)