A new deep distortion convolutional neural network for semantic segmentation of panoramic images

Xing Hu¹, Yi An¹*, Cheng Shao¹ and Pan Qin¹

¹ Control Science and Engineering, Dalian University of Technology, Dalian, Liaoning, 116024, China

*Corresponding author’s e-mail: anyi@dlut.edu.cn

Abstract. Semantic segmentation of panoramic images plays a crucial role in many applications, such as scene understanding, autonomous navigation, and community security. However, the traditional deep learning algorithms achieve lower accuracy for panoramic images due to the serious image distortion. This paper proposes a new deep convolutional neural network to segment panoramic images of outdoor scenes semantically. The proposed network includes two branches: the semantic segmentation sub-branch (SS-branch) and the feature enhancement sub-branch (FE-branch). The SS-branch contains two parts: Encoder and decoder. In the encoder, a new distortion convolution layer (DCL) is defined to reduce the distortion of panoramic images. In the FE-branch, a canny is used to detect the edge and a lightweight network is designed to enhance the features of image boundaries. Finally, a weighted loss function is used. The experimental results show that the proposed semantic segmentation method has better performance for different outdoor scenes.

1. Introduction

The rapid developments of semantic segmentation methods have allowed us to accurately recognize various objects in an image [1]. The panoramic images acquired by a camera contain a broader view than common images. Nowadays, panoramic images are widely used in many applications, especially in the image recognition, object location and detection [2], visual navigation [3]. However, traditional semantic segmentation algorithms are not adequate for panoramic images because of the significant distortion effect. Therefore, the segmentation of panoramic images is very important to the scene understanding and environment perception.

Although the semantic segmentation methods of normal images have been developed rapidly in recent years, those networks are awkward to process panoramic images. The most networks processing normal images underlies that the images have the planer grid architectures. However, the panoramic image defines image on different structures form the normal image. Therefore, the convolutional neural networks (CNN) can not be directly applied to panoramic images.

The commonly used methods for the semantic segmentation of the panoramic images can be divided into two categories: the direct methods [4]-[6] and indirect methods [7]-[9]. In the direct methods, the special deep learning networks are designed for the semantic segmentation of panoramic images end-to-end, which doesn't need the image distortion correction in advance. In the indirect methods, the panoramic images are corrected into the normal images at first, and then the traditional deep learning networks are used for the semantic segmentation of the normal images.

In the direct methods, Varga et al. [4] proposed a method to segment the panoramic images by using the VGG neural network. Yang et al. [5] proposed a PASS architecture for the semantic
segmentation of panoramic images. The PASS adopts the ERF-PSPNet network [6] to predict the semantical feature maps and uses multisource data to train the network.

In the indirect methods, Budvytis et al. [7] developed a semantic localization method, which first trains semantic segmentation network using the collected dataset and then recognizes surrounding objects. Fernandez et al. [8] proposed a line estimation method for the semantic segmentation of panoramic images. This method segments panoramic images using the geometry features, and generates layout hypotheses of the room assuming Manhattan world. Suzuki et al. [9] presented a new method for the semantic change detection of panoramic images. This method detects panoramic images using normal images of different scales.

In recent studies, many researchers have devoted to the enhancement of feature maps. Zhao et al. [10] proposed an improved FCN network for the segmentation of images, which combines features in different stages. Zhou et al. [11] developed a high-resolution encoder-decoder network for the segmentation of medical images. Hu et al. [12] designed a classification network for high-resolution remote sensing images by using a fully convolutional network. This network adopts the global and local attention modules to enhance feature maps. Xu et al. [13] proposed a multichannel network for the gland instance segmentation. This method reduces the loss of boundary information during downsampling by the introduction of sub-network.

Therefore, a new semantic segmentation network for panoramic images is proposed, which is called the deep distortion convolutional network (DDCN). DDCN contains SS-branch and FE-branch. The SS-branch contains two parts: Encoder and decoder. To reduce the distortion effect of images, the DCL is designed in the front end of encoder. To enhance the features of image boundaries, the FE-branch is developed using a canny edge detection method [14]. Then, a weighted loss function is used to improve the training process.

2. Semantic segmentation network of panoramic images

In this section, a new semantic segmentation network for panoramic images is proposed. The architecture of our semantic segmentation network is shown in figure 1. To reduce the effect of image distortion and change the distortion sampling coordinates according to the degree of image distortion, the DCL is designed in the front end of encoder. The FE-branch is developed to enhance the features of image boundaries by designing a lightweight network. Then, a weighted loss function is used to improve the training process.

![Figure 1. Deep distortion convolutional network proposed in this paper.](image-url)
2.1. Distortion convolutional layer (DCL)

A CNN is usually composed of convolution layers, pooling layers and a full connection layer. In a common structure, the convolution layers and pooling layers are unique to a convolutional neural network. The basic operation of CNN is convolution, which filters each small regions of images, and obtains features of these small regions. The process of convolution is

$$S(i, j) = \sum_{c=1}^{C} \sum_{m=0}^{M} \sum_{n=0}^{N} I_c(i+m, j+n)K_c(m,n)$$  \hspace{1cm} (1)

where $K_c$ is the convolutional kernel, $I_c$ is a color channel of the input image $I$, and $S$ is the output feature map.

The regular sampling of pixels in the traditional convolution that is $I_c(i+m, j+n)$, which is for normal images. It is not suitable for panoramic images. Therefore, we propose a new distortion convolution.

The distortion convolutional layer is

$$S(i, j) = \sum_{c=1}^{C} \sum_{m=0}^{M} \sum_{n=0}^{N} I_c(i+m, j+n+\Delta n)K_c(m,n)$$  \hspace{1cm} (2)

The new distortion sampling of pixels $I_c(i+m, j+n+\Delta n)$ is as shown in figure 2. $(0, \Delta n)$ is the distortion sampling coordinate for panoramic images which is used to offset the effects of objects distortion in panoramic images. In our network, the size of distortion sampling region is set as $5 \times 7$. 

Figure 2. Distortion sampling: (a)-(c) is for the top half part of a panoramic image, (d)-(f) is for the half bottom of the panoramic image.

Figure 3. Edge detecting: (a) image of an outdoor scene, and (b) edge detecting of the image.
2.2. Distortion sampling coordinate

To determine $\Delta n$ in (2), the edges are extracted from panoramic images using the canny edge detection algorithm [14], as shown in figure 3. Then, the edge in the panoramic image that is the horizontal line in the real world is selected, such as the curb of the road or the straight marking line. In order to calculate the mean curvature of an edge $E$, we use the method which is dedicatedly illustrated in [15]. To obtain accurate mean curvature, $K$ edges are selected from different panoramic images.

In [15], the edge $E$ is composed of a series of discrete points $\{p_i = (x_i, y_i) | 1 \leq i \leq n_i\}$. Then, the discrete curvature at each point $(x_i, y_i)$ has been calculated. The mean curvature $\bar{\kappa}$ of the $k$-edge $E$ is

$$\bar{\kappa} = \frac{\sum_{i=1}^{n_k} \kappa_i}{n_k}$$

(3)

where $\kappa_i$ is the discrete curvature of the point $(x_i, y_i)$.

The adjustment factor $\Delta f$ is designed to change with respect to the distortion degree of different panoramic images, as the following

$$\Delta f = p \times \left( \frac{\sum_{i=1}^{n_k} \bar{\kappa}_i}{K} \right)$$

(4)

where $\bar{\kappa}_i$ is the mean curvature of the $k$-edge and $p$ is a constant. Then, the distortion sampling coordinate $\Delta n$ is calculated as with

$$\Delta n = \begin{cases} - \left\lfloor \Delta f \ast \Delta c \right\rfloor & \text{if } j < w/2 \\ \left\lfloor \Delta f \ast \Delta c \right\rfloor & \text{if } j \geq w/2 \end{cases}$$

(5)

where $\Delta c$ is the set distortion variable, $w$ is the height of the panoramic image.

2.3. FE-branch

In the traditional segmentation networks for boundaries, the sub-branches which can be trained independently from the main-branch are commonly used, such as the DCAN [13] and the deep multichannel side supervision (DMCS) [16]. However, the method in those paper need the additional annotation of boundaries, which is less in the datasets of outdoor scenes.

Therefore, a new FE-branch is proposed which is as shown in figure 1. In the FE-branch, a canny method is used to extract edges of images. Then, the labels for edges of images are extracted, which are used as the labels of the FE-branch.

In the semantic segmentation of images, an overlarge time of upsampling will dedicatedly loss the texture information, which can decrease the segmentation accuracy. In the FE-branch, the high-level features from the bottom of the encoder are upsampled by 4-time and convoluted by a $1 \times 1$ convolution successively. Then, the output feature maps are sequentially upsampled by 4-time and convoluted by a $1 \times 1$ convolution successively. Finally, the segmentation result of edges are obtained.

2.4. Weighted Loss Function (WLF)

To train the SS-branch and the FE-branch simultaneously, the loss function in the training of the network is formulated as

$$L_{total} = \lambda L_{FE} + (1 - \lambda)L_{SS}, \lambda \in (0, 1)$$

(7)

where $L_{SS}$ and $L_{FE}$ are the loss of the SS-branch and the FE-branch respectively, and $\lambda$ is the weighted factor.
3. Experiment results

![Figure 4. Semantic segmentation results of Panoramic Image I-II. In the above figures, red, green, yellow, blue, purple, and black represent tree, ground, building, car, person, and others respectively. BS and FE represent the baseline and FE-branch respectively.]

Table 1. Semantic segmentation results for panoramic images.

| Network        | Acc (%) | mIoU (%) | fwIoU (%) | tree (%) | ground (%) | building (%) | car (%) | person (%) | others (%) |
|----------------|---------|----------|-----------|----------|------------|--------------|---------|------------|------------|
| Baseline       | 88.87   | 70.18    | 80.93     | 90.39    | 96.39      | 86.08        | 88.77   | 73.18      | 70.92      |
| Baseline+DCL   | 89.26   | 70.92    | 81.51     | 91.75    | 97.33      | 90.79        | 89.21   | 73.52      | 72.29      |
| DDCN           | 89.60   | 71.66    | 81.77     | 90.71    | 97.95      | 92.95        | 91.46   | 73.58      | 73.46      |
| ERF-PSNet [6]  | 88.87   | 70.26    | 80.95     | 90.71    | 96.34      | 86.02        | 88.83   | 73.01      | 70.23      |
| PASS [5]       | 89.01   | 70.38    | 80.99     | 90.95    | 96.61      | 86.00        | 88.41   | 73.34      | 71.29      |

3.1. Setting

To further illustrate the effectiveness of our method, we present the results from both segmentation indicators and segmentation scenes. The experiment environment is Centos 7.0 system, RAM 32G, and GPU Tesla K40M. To train our network, the IMPART dataset is used, which is synthesized using the Cityscape dataset [17]. The IMPART dataset contains 5000 annotated panoramic images.

The loss function of SS-branch and the FE-branch are both used the CrossEntropy. Momentum and weight decay are set to 0.9 and 0.0005 respectively. The learning rate is 0.007 and $\lambda$ is 0.3. The annotation contains six common classes of ground, building, tree, car, person, and others.

3.2. Ablation study

To evaluate the performance of each module in DDCN, the experiments are conducted in the setting: without the DCL, with the DCL, with the DCL+FE-branch+WLF. Our baseline is the ResNet101+ASPP. Table 1 and figure 4-figure 5 show the semantic segmentation results of the ablation study. From Table 1, the baseline+DCL exceeds the baseline 0.39/0.74/0.38 in terms of Acc, mIoU and fwIoU. As marked by red boxes in figure 4 and figure 5, the baseline+DCL has better segmentation effects than the baseline. This is because the specially designed module DCL can reduce the distortion effects and improve the feature extraction of panoramic images.
We can find the performance of our method can be further improved by the addition of the FE-branch and the WLF. Compared with the baseline, the baseline+DCL+FE-branch+WLF, i.e. DDCN, outperform it by 0.73/1.48/0.84 in terms of Acc, mIoU and fwIoU. As marked by green and yellow boxes in figure 4 and figure 5, the DDCN has better segmentation effects for images boundaries than the baseline+DCL. This is because the developed FE-branch and WLF can enhance the feature extraction of image boundaries. These results demonstrate that the modules DCL, FE-branch, and WLF enhance the performance of the semantic segmentation network, including distortion reduction, feature extraction, and feature enhancement.

3.3. Comparison study
To compare our DDCN with the semantic segmentation methods of panoramic images, two classical methods, i.e. ERF-PSPNet [6] and PASS [5], are chosen for comparison. Table 1 and figure 4-5 show the segmentation results of panoramic images with two methods and DDCN. As can be observed, our DDCN achieves Acc 89.60 and mIoU 71.66. Our method is about 1 higher than the other two methods respectively. As marked by the yellow boxes in figure 4-5, the semantic segmentation results of our DDCN are much closer to the ground truth than those of the other methods.

4. Conclusion
This paper proposes a new deep convolutional neural network for segmenting panoramic images of outdoor scenes semantically. The main technical contributions included the definition of the DCL, the design of the FE-branch, and the development of the WLF. The DCL reduced the distortion of panoramic images. The FE-branch enhanced the feature maps. The WLF improved the training process. These technical features made our semantic segmentation method accurate and reliable. The experimental results show that the proposed semantic segmentation method has good performance for different outdoor scenes.
Acknowledgments
This work was supported in part by the National Natural Science Foundation of China under Grant 61673083 and in part by the Science and Technology Major Project of Shanxi Province under Grant 20191101014.

References
[1] Liang C.C., George P. (2017) Rethinking Atrous Convolution for Semantic Image Segmentation. arXiv: 1706.05587.
[2] An Y., Li B., Hu H., Zhou X. (2019) Building an Omnidirectional 3-D Color Laser Ranging System through a Novel Calibration Method. IEEE Transactions on Industrial Electronics, 66: 8821-8831.
[3] An Y., Li B., Wang L., Zhang C. (2019) Calibration of a 3D laser Rangefinder and a Camera based on Optimization Solution. Journal of Industrial & Management Optimization, 13: 1-19.
[4] Varga R., Costea A., Florea H., Giosan I. and Nedevschi S. (2017) Super-sensor for 360-degree Environment Perception: Point Cloud Segmentation using Image Features. 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), Yokohama, pp. 1-8, doi: 10.1109/ITSC.2017.8317846.
[5] Yang K., Hu X., Bergasa L. M., Romera E. and Wang K. (2020) PASS: Panoramic Annular Semantic Segmentation. in IEEE Trans. on Intelligent Transportation Systems, 21: 4171-4185, doi: 10.1109/TITS.2019.2938965.
[6] Sáez, Bergasa Á., López-Guillén L., Romera E., Tradacete E., Gómez-Huélamo M., del Egido C., Javier (2019) Real-Time Semantic Segmentation for Fisheye Urban Driving Images Based on ERFNet. Sensors, 19: 503-522; https://doi.org/10.3390/s19030503.
[7] Budvytis I., Sauer P., Cipolla R. (2018) Semantic Localization via Globally Unique Instance Segmentation, British Machine Vision Conference 2018, https://doi.org/10.17863/CAM.43095.
[8] Fernandez-Labrador C., Perez-Yus A., Lopez-Nicolas G. and Guerrero J. J. (2018) Layouts from Panoramic Images with Geometry and Deep Learning. IEEE Robotics and Automation Letters, 3: 3153-3160.
[9] Suzuki T., Minoguchi M., Suzuki R. et al. (2018) Semantic Change Detection. 2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV), pp.1785-1790.
[10] Zhao Y., Zhang Y., Yi K., Lei Y. (2019) Understanding and Improving the Intermediate Features of FCN in Semantic Segmentation. In: 2019 International Conference on Communications, Information System and Computer Engineering (CISCE), Haikou, China, pp. 415-419.
[11] Zhou S., Nie D., Adeli E., Yin J., Lian J., Shen D. (2020) High-Resolution Encoder–Decoder Networks for Low-Contrast Medical Image Segmentation. IEEE Transactions on Image Processing, 29: 461-475.
[12] Hu H., Li Z., Li L., Yang H., Zhu H. (2020) Classification of Very High-Resolution Remote Sensing Imagery Using a Fully Convolutional Network With Global and Local Context Information Enhancements. IEEE Access, 8: 14606-14619.
[13] Xu X. et al. (2018) Breast Region Segmentation being Convolutional Neural Network in Dynamic Contrast Enhanced MRI. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, pp. 750-753.
[14] Akinlar C., Chome E. (2015) CannySR: Using smart routing of edge drawing to convert Canny binary edge maps to edge segments. In: 2015 International Symposium on Innovations in Intelligent SysTems and Applications (INISTA), Madrid, pp. 1-6.
[15] An Y., Shao C., Wang X., Li Z. (2017) Geometric properties estimation from discrete curves using discrete derivatives. Computers & Graphics, 35: 916-930.

[16] Xu Y. et al. (2017) Gland Instance Segmentation Using Deep Multichannel Neural Networks. IEEE Transactions on Biomedical Engineering, 64: 2901-2912.

[17] Cordts M. et al. (2016) The Cityscapes Dataset for Semantic Urban Scene Understanding. in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 3213–3223.