Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 5, No. 1, 2023

A Survey of Image Semantic Segmentation Algorithm Based on Deep Learning

Jian Chen[1], Fen Luo[1, *]

[1] School of Software, Henan Polytechnic University, China

[*] **Corresponding author**: Fen Luo (Email: luofenjsj@hpu.edu.cn)

Abstract: Image semantic segmentation technology is one of the core research contents in the field of computer vision, and has a wide range of applications in production and life. With the improvement of computer performance and the continuous development of deep learning technology, researchers have increasingly high research enthusiasm for the performance of image semantic segmentation. This paper summarizes the research status of image semantic segmentation based on deep learning and introduces the common datasets used in the field of semantic segmentation. Finally, we point out the existing problems and future development trend of image semantic segmentation algorithms.

Keywords: Image semantic segmentation, Deep learning.

1. Introduction

Image semantic segmentation is an important research direction in the field of computer vision that aims to assign each pixel in an image to the target category to which it belongs. Image semantic segmentation is important for solving many practical problems, such as autonomous driving, medical image analysis, remote sensing image analysis, robot navigation, etc. Traditional segmentation methods, such as super-pixel, watershed algorithm and graph algorithms, are based on the physical properties of the images.

In recent years, with the development of deep learning, the image semantic segmentation technology has been rapidly improved. These algorithms use large amounts of training data to learn the features of the target object and perform semantic segmentation through Convolutional Neural Network (CNN). A series of CNN-based semantic segmentation methods have made various breakthroughs in segmentation accuracy and improved the segmentation performance. This paper summarizes the image semantic segmentation methods based on deep learning.

2. Methods

2.1. General Semantic Segmentation

Fully Convolutional Network (FCN) [1] is the first work of using deep learning technology to realize the end-to-end semantic segmentation model by removing fully connected layers. Based on FCN [1], some semantic segmentation methods employ an encoder-decoder architecture network to encode more spatial information. Most of these methods are designed into a U-shape network structure. U-Net [2] introduces the useful dense skip connections to exploit the spatial details, resulting in different variants of the encoder-decoder structures. Global Convolution Network [3] introduce global convolution operation to extract more global information in the case of larger receptive field, so as to improve the segmentation accuracy. SegNet [4] with pooled index strategy, RefineNet [5] with multi-path refinement, LRR [6] with Laplacian pyramid reconstruction, DFN [7] with channel attention block and HRNet [8] with multi-branches utilize the encoder-decoder backbone network to further recover the detailed information.

To enlarge the receptive field, PSPNet [9] , DeepLab v2 [10], and Deeplab v3 [11] adopt dilated convolution to maintain the spatial size of the feature map. However, both types of dilation backbone and encoder-decoder structure care less about the inference speed and computational cost, which are mainly designed for general semantic segmentation tasks.

2.2. Real-time Semantic Segmentation

Most deep learning methods have high requirements for computational resources and storage space, so their applications in mobile terminals, and embedded systems are limited. The emergence of real-time semantic segmentation algorithm brings higher efficiency and scalability to the semantic segmentation task. SegNet [4] utilizes the skip-connected method and transposed convolutions for upsampling to preserve fine-grained spatial information. It has fewer parameters, faster training speed, and does not need to use a large amount of preprocessing data.

ICNet [12] utilizes a multi-resolution cascade architecture and fused features extracted from image pyramids to achieve better accuracy than previous. ERFNet [13] uses factorized convolutional layers to reduce the number of parameters while maintaining expressive power. It also incorporates residual connections and dilated convolutions to improve information flow. Based on depth-wise separable convolutions [13] and inverted residual blocks [14], BiSeNetV1 [15] uses a spatial path with a small receptive field to capture local information, and a context path with a large receptive field to capture global information. BiSeNetV2 [16] further improves upon this architecture by incorporating a new feature fusion module, which enhances the fusion of features from the spatial and context paths.

3. Datasets

In this part, Cityscapes [17], CamVid [18] and COCO-Stuff [19] datasets are introduced, which as shown in Figure 1.

Cityscapes: As a famous dataset focusing on urban street scene parsing, Cityscapes [17] is taken from a car perspective. The dataset collected from 50 different cities in Germany contains 30 categories, including roads, vehicles, buildings, trees, etc. The Cityscapes dataset includes 5000 finely

annotated images, of which 2975 for training, 500 for validation, and 1525 for testing. We usually use the fine annotated images whose resolution is 2048×1024 for a fair comparison.

CamVid: The CamVid street scene dataset [18] is a video-based semantic segmentation dataset, which only has 11 semantic categories. From the perspective of a driving automobile, it includes 701 images and is separated into training, validation, and test datasets, with 367, 101, and 233 images, respectively. The resolution of these images is 720×960.

COCO-Stuff: The COCO-Stuff [19] provides a large number of labeled image data for semantic segmentation tasks. By augmenting the COCO 2017 dataset with dense stuff annotations, the dataset includes 91 stuff categories for evaluation and 1 category 'unlabeled'. Among all 164K images, 118K images are used for training, 5K images are used for validation, 20K images are used for test development, and 20K images are used for test-challenge.

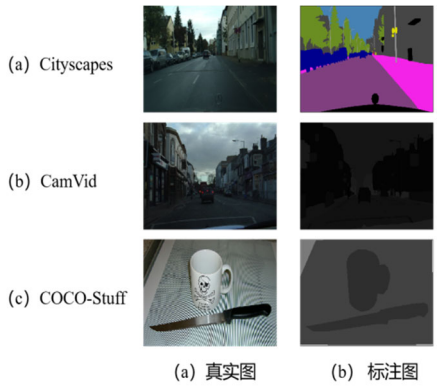

Figure 1. Small part of pictures in three datasets

4. Conclusion

Although the image semantic segmentation algorithm based on deep learning has made significant progress, there are still some challenges to overcome. One of the important challenges is how to deal with the factors that affect the result of image semantic segmentation, such as occlusion, illumination change and image resolution. In the future, image semantic segmentation based on deep learning will continue to make progress in the direction of real time and be widely used in more fields. With the continuous improvement of technology, it is expected to achieve more accurate and efficient image semantic segmentation.

References

[1] Long, J., Shelhamer, E., Darrell, T.: 'Fully convolutional networks for semantic segmentation', Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.

[2] Ronneberger, O., Fischer, P., Brox, T.: ; Springer. 'U-net: Convolutional networks for biomedical image segmentation', International Conference on Medical image computing and computer-assisted intervention, 2015, pp. 234–241.

[3] Peng, C., Zhang, X., Yu, G., Luo, G., Sun, J.: 'Large kernel matters–improve semantic segmentation by global convolutional network', Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4353–4361.

[4] Badrinarayanan, V., Kendall, A., Cipolla, R.: 'Segnet: A deep convolutional encoder-decoder architecture for image segmentation', IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39, pp. 2481–2495.

[5] Lin, G., Milan, A., Shen, C., Reid, I.: 'Refinenet: Multi-path refinement networks for high-resolution semantic segmentation', Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1925–1934.

[6] Ghiasi, G., Fowlkes, C.C.: ; Springer. 'Laplacian pyramid reconstruction and refinement for semantic segmentation', European conference on computer vision,2016, pp. 519–534.

[7] Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: 'Learning a discriminative feature network for semantic segmentation', Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 1857–1866.

[8] Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., et al.: 'Deep highresolution representation learning for visual recognition', IEEE transactions on pattern analysis and machine intelligence, 2020, 43, (10), pp. 3349–3364.

[9] 16 Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: 'Pyramid scene parsing network', Proceedings of the IEEE conference on computer vision and pattern recognition, 2017,pp. 2881–2890.

[10] Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: 'Deeplab:Semantic image segmentation with deep convolutional nets, atrous convolution,and fully connected crfs', IEEE transactions on pattern analysis and machine intelligence, 2017, 40, (4), pp. 834–848.

[11] Florian, L.C., Adam, S.H.: 'Rethinking atrous convolution for semantic image segmentation', Conference on Computer Vision and Pattern Recognition (CVPR) IEEE/CVF, 2017.

[12] Zhao, H., Qi, X., Shen, X., Shi, J., Jia, J.: 'Icnet for real-time semantic segmentation on high-resolution images', Proceedings of the European conference on computer vision (ECCV), 2018, pp. 405–420.

[13] Romera, E., Alvarez, J.M., Bergasa, L.M., Arroyo, R.: 'Erfnet: Efficient residual factorized convnet for real-time semantic segmentation', IEEE Transactions on Intelligent Transportation Systems, 2017, 19, (1), pp. 263–272.

[14] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: 'Mobilenetv2: Inverted residuals and linear bottlenecks', Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4510–4520.

[15] Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: 'Bisenet: Bilateral segmentation network for real-time semantic segmentation', Proceedings of the European c.onference on computer vision (ECCV), 2018, pp. 325–341

[16] Yu, C., Gao, C., Wang, J., Yu, G., Shen, C., Sang, N.: 'Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation', International Journal of Computer Vision, 2021, 129, (11), pp. 3051–3068.

[17] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R.,et al.: 'The cityscapes dataset for semantic urban scene understanding', 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016,pp. 3213–3223.

[18] Brostow, G.J., Shotton, J., Fauqueur, J., Cipolla, R.: 'R.: Segmentation and recognition using structure from motion point clouds', In: ECCV, 2008.

[19] Caesar, H., Uijlings, J., Ferrari, V.: 'Coco-stuff: Thing and stuff classes in context',Proceedings of the IEEE conference on computer vision and pattern recognition,2018, pp. 1209–1218.