Transactions on Computational and Scientific Methods | Vo. 4, No. 9, 2024
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Panoramic Image Segmentation Using Attention Mechanism with ResNet-50 Backbone and Multi-Task Learning

Lysander Eaton
Department of Computer Engineering, University of Tennessee, USA
lysander.e592@utk.edu

Abstract: Traditional image segmentation techniques, such as semantic and instance segmentation, often fall short in providing comprehensive scene understanding. Semantic segmentation fails to differentiate individual objects within the same category, while instance segmentation cannot identify distinct background regions. To address these limitations, panoramic segmentation was introduced, combining the strengths of both methods to assign semantic categories to each pixel while distinguishing objects of the same category. This paper proposes an improved panoramic segmentation approach based on attention mechanisms. Using ResNet-50 as the backbone, the method extracts features that are processed separately by semantic and instance segmentation branches. The semantic segmentation branch employs a fully convolutional network (FCN), while the instance segmentation branch uses Mask R-CNN, with information flow shared between branches. A cross-layer attention fusion module aggregates multi-scale features into a prototype mask module to enhance segmentation accuracy. The final results from both branches are fused heuristically to produce refined panoramic segmentation output, effectively addressing occlusion and improving scene comprehension.

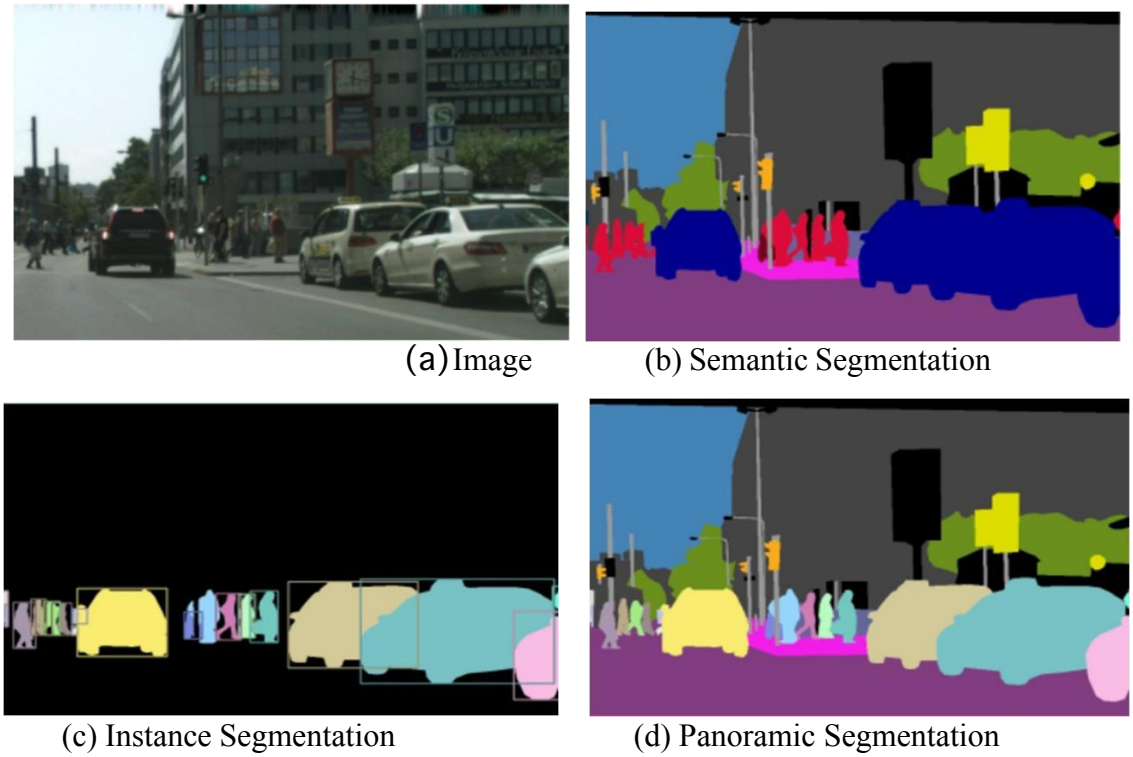
Keywords: Panoramic Segmentation; Spatial Sorting; Attention Mechanism.

1. Introduction

Traditional image segmentation tasks can be divided into semantic segmentation and instance segmentation, which are interrelated and independent of each other. However, all instance objects contained in the same semantic category cannot be distinguished by semantic segmentation, while instance segmentation cannot distinguish different background regions. So, neither of them can fully describe the image. Meanwhile, with the continuous growth of large-scale data and the complexity of image scene, semantic segmentation or instance Segmentation cannot achieve a refined scene understanding of the image.

With the continuous development of deep learning, Alexander Kirillov et al. proposed the first panoramic Segmentation framework based on deep learning and a new segmentation task -- Panoramic Segmentation by combining semantic features with instance features.

In order to achieve rich and complete scene segmentation of image, panoramic segmentation not only needs to assign a semantic category to each pixel, but also needs to distinguish different instance objects of the same semantic category. The schematics of different image segmentation tasks are shown in Figure 1



(a) Image (b) Semantic Segmentation

(c) Instance Segmentation (d) Panoramic Segmentation

Figure 1. Diagram of different image segmentation tasks

The panoramic segmentation technique is divided into two sub-tasks, namely semantic segmentation and instance segmentation. In recent years, deep learning has been widely used in image segmentation. Semantic segmentation techniques based on deep learning can be divided into the following two categories: one is semantic segmentation method based on region classification; The other is semantic segmentation method based on pixel classification. Case segmentation techniques based on deep learning can be divided into the following two categories: one is single-stage case segmentation; The other is two-stage instance segmentation. Two - stage case segmentation can be divided into top-down and bottom-up methods. Top-down method means first detection and then segmentation, while bottom-up method means first marking pixels and then clustering. at present, subtask fusion methods of panoramic segmentation networks are divided into heuristic fusion and Panoptic Head fusion.

1.1 Semantic Segmentation

Semantic segmentation gives a category label to each pixel in the image, but it cannot split different objects of the same category. Semantic segmentation techniques based on deep learning can be divided into two types: region classification and pixel classification.

Since the region classification method has some problems such as low segmentation accuracy and slow speed, the semantic segmentation method of pixel classification is proposed in the following part. FCN is one of the classical networks based on pixel classification.

FCN treats each pixel as a training sample, not only to predict its category, but also to calculate classification losses. The network structure of FCN is shown in Figure 2. In order to accept input images of any size, FCN replaces the full connection layer with the convolutional layer on the basis of CNN, and then uses the deconvolution layer for reverse learning to up-sample the feature graph replaced by the convolutional layer, so that the output feature graph can be restored to the same size as the original one. In this way, the position and spatial information of the original input image can

be retained. Then category prediction is made for each pixel on the feature map obtained from the up-sampling. Finally, the SoftMax function is used to calculate the loss per pixel.

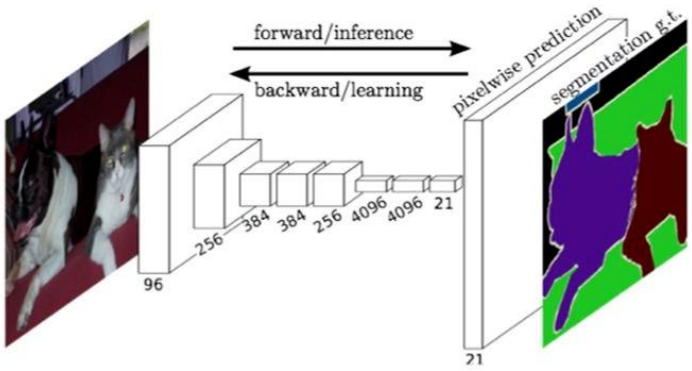


Figure 2. FCN network structure

1.2 Instance Segmentation

Instance segmentation is to detect specific objects in the image and accurately segment the detected objects. It combines the results of two tasks, object detection and semantic segmentation. Case segmentation techniques based on deep learning can be divided into single-stage segmentation and two-stage segmentation. The two-stage segmentation is divided into the top-down detection and segmentation method and the bottom-up labeling and clustering method.

The top-down detection before segmentation method is to detect the object with a boundary box first, and then segment the object. Mask RCNN is one of the most successful methods. The structure of Mask RCNN is shown in Figure 3. First, convolutional neural network is used to extract features from input images. Here, ResNet residual network is used as the basic network. Then, the FPN is used to generate N suggestion boxes for each image, and the suggestion boxes are mapped to the last convolutional feature map of CNN, and the feature maps of the same size are generated for the N suggestion boxes by the RoI Align layer. Finally, category classification, border prediction and mask regression are carried out by full connection.

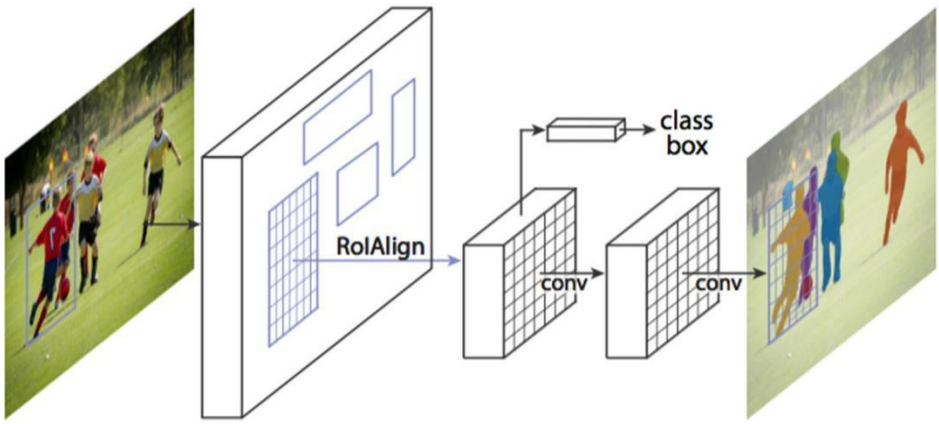


Figure 3. Mask R-CNN Network structure

1.3 Panoramic Segmentation

Different from semantic Segmentation and instance segmentation, Panoptic Segmentation requires that each pixel in an image must be assigned a semantic label and an instance id. Semantic tags refer to the class of objects, and instance ids correspond to different numbers of the same class of objects. Panoramic segmentation tasks can be divided into object instance segmentation sub-tasks and stuff segmentation sub-tasks. The panoramic segmentation method usually consists of three independent

parts: the object instance segmentation part, the stuff segmentation part, and the result fusion part of two sub-branches.

Usually, the object instance segmentation network and the stuff segmentation network are independent from each other, and no parameters or image features are shared between the networks. This method not only leads to high computation overhead, but also leads to high computation overhead. it also forces the algorithm to use a separate post-processing program to merge the two predictions, and makes panoramic segmentation unusable in industry.

2. Model

Panoramic segmentation network uses convolutional neural network to extract features from input images,In order to solve the impact of multi-objective overlap in complex scenes, this paper refers to the spatial sorting module, and introduces the attention mechanism module into the feature extraction network module to improve the feature extraction of the model. A feature sharing network is added between two sub-branch modules to realize sharing system parameters between different tasks.

2.1 Global Network Framework

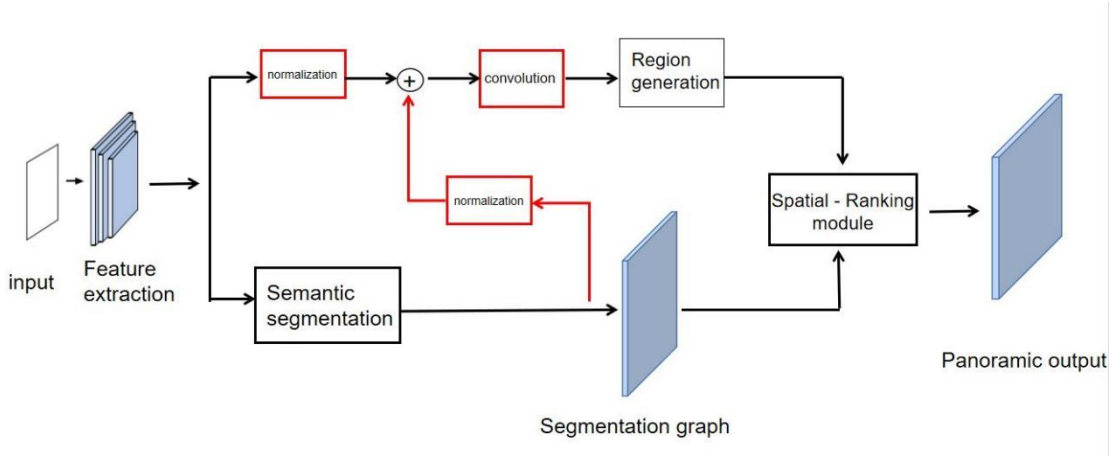


Figure 4. Global network framework

The overall framework of panoramic segmentation network mainly includes three parts, and the specific network structure is shown in Figure 4. Firstly, ResNet-50 is used as the basic feature pyramid network, and the extracted feature map is used as the input of semantic segmentation branch and instance segmentation branch.

Then the feature information flow of semantic segmentation is shared through the information sharing network between the two branches. Finally, a spatial sorting module with integrated output is used to deal with the occlusion problem between instances, and the multi-task learning method is used to self-adapt the loss function of the two subtasks, and the total loss function is obtained.

2.2 Spatial - Ranking Module

By comparing the scores of the panoramic quality index and the average category accuracy index of the instance segmentation, the prediction accuracy of some categories of objects in the instance segmentation is obviously higher than that in the panoramic segmentation.

These categories are blocked or covered by other categories in the panoramic segmentation. Therefore, these categories are taken as higher priority categories to solve the instance occlusion problem. After the score based on spatial hierarchy is obtained, the corresponding sub-branch prediction fusion algorithm is used for panoramic segmentation prediction. This spatial hierarchical sorting algorithm based on artificial prior information has many misjudgment problems, so this paper uses the spatial sorting module based on convolution. Its network structure is shown in Figure 5.

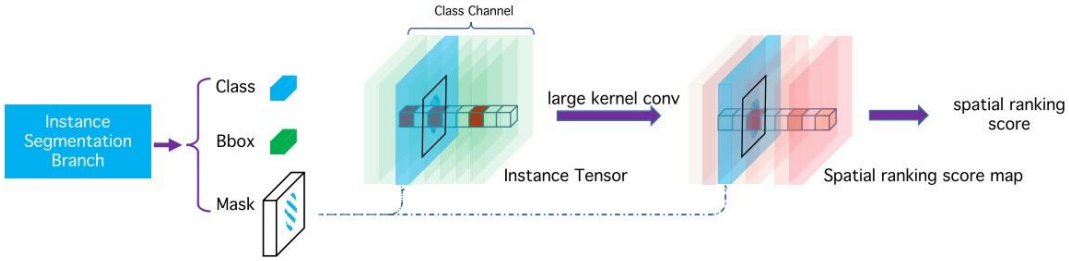


Figure 5. Spatial - Ranking module

2.3 SE Attention

The innovation point of SENet network is to focus on the relationship between channels, hoping that the model can automatically learn the importance of different channel characteristics. To do so, SENet has proposed the Squeeze-and-Excitation (SE) module.

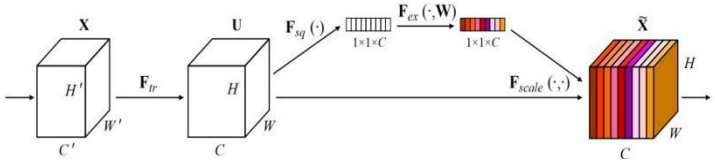


Figure 6. SE Attention module

The SE module first Squeeze the feature map obtained by convolution to get the channel level global features. Then, the global features are Excitation to learn the relationship between all channels and to get the weights of different channels. Finally, multiply the original feature map to get the final features.

Essentially, the SE module provides attention or gating for the channel dimension. This attention mechanism allows the model to focus on the most informative channel features while suppressing the less important channel features. Another point is that the SE module is generic, which means it can be embedded into existing network architectures. Figure 6 shows the SE module.

2.4 Loss Function

The feature fusion of semantic segmentation and instance segmentation uses the multi-task learning method, so that the model can adaptively learn the relative weights of different losses. The multi-task loss function obtained is:

$$L_{total} = L(W, \sigma_1, \sigma_2) = \frac{1}{2\sigma_1} L_1(w) + \frac{1}{2\sigma_2^2} L_2(w) + \log \sigma_1^2 \sigma_2^2$$

L_{total} is the total loss function, σ_1 andσ_2 are the noise parameters of semantic segmentation and instance segmentation, respectively. $L_1(w)$ is the loss function of semantic segmentation, and $L_2(w)$ is the sum of instance segmentation L_{cls}, L_{bboxs}, L_{mask} . Where L_{cls} represents the classification loss of candidate frame, L_{bboxs}represents the coordinate regression loss of candidate frame, andL_{mask}represents the segmentation loss of foreground and background in candidate frame.

3. Experiment and Result Analysis

3.1 Experimental Environment and Parameter Configuration

Experimental hardware configuration: 16G video memory, 64G memory, Intel core i7 CPU, NVDIA GeForce GTX1060 GPU.

Experimental software configuration: Windows 10 operating system, using Python 3.6 programming language and Pytorch deep learning framework.

The initial learning rate set in the experiment was$10^{-3}$and the number of iterations was2×10^4. Every iteration 1000 times, the learning rate decreases 10 times. The FPN of Imagenet was used as the pre-training model in the experiment.

3.2 Experimental Analysis

Table 1 shows the results of this method compared with the most advanced methods on the COCO dataset. Wherein this method It refers to a network structure model with information sharing module and spatial sorting module. It can be seen that the accuracy of this method is greatly improved compared with other methods, but SQ is 1.5% lower than AUNet network.

Table 1. Comparison of experimental results

	PQ	SQ	RQ	PQ^{th}	PQ^{st}
JSISNet	27.2	71.9	35.9	29.6	23.4
OANet	40.7	78.2	49.6	50.0	26.2
AUNet	46.5	81.0	56.1	55.8	32.5
Ours	51.5	79.5	65.8	57.6	35.5

(a) Input picture (b) Annotated picture (c) Forecast result

Figure 7. The visualization results of this method on the COCO dataset

The visualization results of this method on the COCO dataset are shown in Figure 7. (a) The input picture, (b) the real annotated picture, and (c) the prediction result of this method.

4. Conclusion

This chapter proposes a panoramic segmentation method based on attention mechanism. Firstly, the residual network ResNet-50 is used as the backbone network to extract the features of the input image, and the obtained feature maps are transferred to the semantic segmentation branch and the instance segmentation branch respectively. Semantic segmentation branch and instance segmentation branch respectively use FCN and Mask RCNN to get the final output, and introduce information flow between the two branches for information sharing. Finally, the multi-task learning method is used to

fuse the two branches and the space sorting module is used to solve the occlusion problem between instances.

The semantic segmentation network generates the mask coefficient of each category, the instance segmentation network generates the result of target detection and the mask coefficient, and the cross-layer attention fusion module aggregates the features of different scales into the prototype mask module, and then combines the prototype mask with the semantic branch and the instance branch respectively to get the final mask coefficient of each branch. Finally, the predictive results of semantic segmentation and instance segmentation are fused heuristically to obtain the output of panoramic segmentation.

References

[1] Mueller, Shane T, Hoffman, et al. Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publicati- ons, and Bibliography for Explainable AI[J]. Computer Science - Artificial Intelligence, 2019, 8:82-115.

[2] Dai J, He K, Sun J. Box Sup: Exploiting Bounding Boxes to Supervise Convolutional Networks for Semantic Segmentation[C]. Proceedings of the IEEE International Conference on Computer Vision, 2015: 1635-1643.

[3] Pinheiro P, Collobert R. From Image-Level to Pixel-Level Labeling with Convol- utional Networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015:1713-1721.

[4] He K, Gkioxari G, Piotr Dollár, et al. Mask R-CNN[J]. IEEE Transactions on Pattern Analysis Machine Intelligence, 2017, 42(2):386-397.

[5] Xie E, Sun P, Song X, et al. PolarMask: Single Shot Instance Segmentation with Polar Representation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020:12193-12202.

[6] Chen H, Sun K, Tian Z, et al. BlendMask: Top-Down Meets Bottom-Up for Instance Segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Patt- ern Recognition, 2020:8573-8581.

[7] De G, Daan, Panagiotis M, and Gijs D. Panoptic Segmentation with A Joint Semantic and Instance Segmentation Network[C]. Computer Vision and Pattern Recognition. 2018:1-6.

[8] Cheng B, Collins M D, Zhu Y, et al. Panoptic-DeepLab: A Simple, Strong, and Fast Baseline for Bottom-Up Panoptic Segmentation[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2020: 12472-12482.

[9] Jayasumana S, Ranasinghe K, Jayawardhana M, et al. Bipartite Conditional Random Fields for Panoptic Segmentation [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2020:6301-6312.