Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 3096961, 8 pages
https://doi.org/10.1155/2022/3096961

*Research Article*

# 2PN: A Unified Panoptic Segmentation Network with Attention Module

**Jianwen Wang and Zhiqin Liu**

*Southwest University of Science and Technology, Mianyang, China*

Correspondence should be addressed to Zhiqin Liu; lzq@swust.edu.cn

Received 26 January 2022; Revised 6 March 2022; Accepted 14 March 2022; Published 30 March 2022

Academic Editor: Yan Huo



Comprehensive and accurate surveillance of the environment forms the basis of secure Internet of things (IoTs), the threats can be observed, and the AI services of IoT systems can be preserved. Panoptic segmentation is an efficient and popular approach for environmental surveillance based on images captured by smart sensing devices. This approach can jointly detect stuffs and things within an image and feed subsequent tasks like image detection. So far, there are many methods for panoptic segmentation which focus on extracting sophisticated visual features for segmentation. However, these efforts are both heavy on their workload and cannot clearly distinguish essential features useful for surveillance in an open environment. Therefore, this paper proposes a novel deep learning model 2PN for panoptic segmentation. The model includes a 2-way pyramid network and an attention module to learn in a more concentrated and reasonable way which enhances the feature extraction part. It strikes a balance between the computing complexity and the power of model capability. Finally, 2PN (2-way pyramid network) results are reflected on the Cityscapes dataset.

## 1. Introduction

Securing the functionalities and services of the Internet of things (IoT for short) systems usually request a clear awareness of the environment, such that potential threats can be observed and the whole system can be guarded. Recently, AI-powered IoTs proposed both novel services and approach-secured IoTs; IoT systems can supply the multimedia data collected by intelligent sensing devices to perform environmental surveillance. Among these pioneering attempts, image segmentation is believed to be an essential and basic aspect of surveillance and also acts as an important research direction of computer vision. The panoptic segmentation [1], which is a combination of semantic segmentation and instance segmentation, is considered a novel frontier of image segmentation. Each pixel of the image must be obtained with a semantic label or an instance label, which may jointly contribute to the understanding of the environment. This segmentation method can bring new opportunities and challenges to computer vision, especially when dealing with complicated open environments.

Generally, the scene of image segmentation consists of "stuff" and "thing." "stuff" usually defines uncountable objects or an object without a fixed shape such as sky and building. At the same time, "thing" usually defines countable objects such as cars, bikes, and pedestrians. The main object of panoptic segmentation is to jointly and wisely detect and distinguish both parts as they are usually correlated. A panoptic segmentation image is shown in Figure 1.

Current trends of panoptic segmentation usually follow the paradigms of deep convolutional networks, which can be divided into three parts: feature extraction, semantic and instance segmentation branch, and subtask fusion. Feature extraction is the head part of panoptic segmentation which receives the input image and provides information for subsequent panoptic segmentation tasks. However, current methods for feature extraction may lose some information, resulting in poor results. Fortunately, the feature

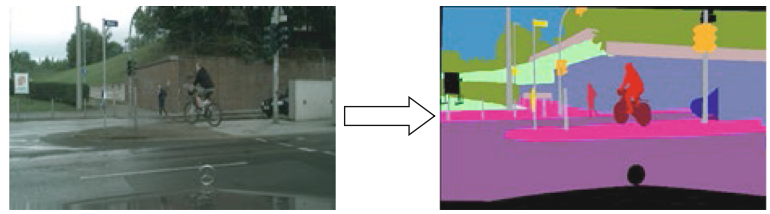

Figure 1: A panoptic segmentation image: (a) is the original image and (b) is the panoptic image. Background without fixed shapes such as sky, traffic light, and wall is "stuff." Pedestrians, cars, and bikes in this figure are "thing."

pyramid network (FPN) [2] is a method used in feature extraction. Current panoptic segmentation models such as BBFNet, PCV, and Panoptic-FPN [3–5] all adopt FPN to mitigate feature extraction. The function of FPN is to extract multiscale features, which can improve the effect of panoptic segmentation.

However, this method still has limitations. Due to the increase in the complexity of the image, only a single FPN cannot obtain more effective features. The FPN is a one-way network, which will lose local information and affect the accuracy of the thing detection part. Moreover, panoptic segmentation will change the branches of semantic and instance segmentation into different directions, and when the transmission is two-way, information will be lost due to different emphases of semantic and instance segmentation.

This paper proposes a two-way pyramid network to solve this problem. The two-way pyramid network will carry out two-way propagation of information. Compared with regular FPN, the feature pyramid model in the upsampling direction is added, which can reduce the information loss caused by the convolution of the network and improve the segmentation effect of the "thing" part. In addition, due to the bidirectional propagation of information, the two-way pyramid network can integrate multiscale features better than FPN, which will also improve the effect of the "stuff" part. Specifically, in order to collect multiscale context information, we use Atrous Spatial Pyramid Pooling (ASPP) [6].

Moreover, the distribution and importance of various features are different in image segmentation. The attention module can play an important role in panoptic segmentation. The attention module has not been applied in these panoptic segmentation models. In this work, we include the attention-based methods to enhance feature extraction and get context information based on Panoptic-DeepLab [7]. Because the feature distribution of the image is unequal, the attention module focuses on more significant features. As a result, we get 60.4%PQ on Cityscapes with the ResNet-50 backbone, getting better performance better than the baseline Panoptic-DeepLab with the ResNet-50 backbone [8].

In summary, the main contribution of the paper is as follows:

(i) A two-way pyramid network is introduced, and a novel panoptic segmentation network is designed, by which reasonable and comprehensive visual features can be extracted and applied

(ii) An attention module is designed for getting multiscale context information concentrating on pivotal parts useful for environmental surveillance

(iii) Experimental results on benchmarks show the advancement of the proposed model in an open environment

## 2. Materials and Methods

*2.1. Related Work.* Panoptic segmentation is a concept proposed by Kirillov et al. [1]. It combines the characteristics of semantic segmentation and instance segmentation. In recent years, many methods have been proposed to improve the results of panoptic segmentation.

*Semantic segmentation*: semantic segmentation distinguishes the regions of different categories of the input image by distinguishing the category of each pixel. Early segmentation algorithms usually used traditional algorithms such as the conditional random field and random forest. The Fully Convolutional Network (FCN) proposed by Long et al. [9] is the semantic segmentation network based on CNN. FCN replaces the full connection layer with the convolution layer. U-Net proposed by Ronneberger et al. [10] is based on FCN and effectively obtains multiscale features through the encoder-decoder structure. Zhao et al. propose PSPNet [11] network structures, which adopt the Pyramid Pooling Module (PPM). The pyramid pooling structure uses four layers of pooling, which is easier to aggregate context information than a single pooling layer. Chen et al. [6] propose the Atrous Spatial Pyramid Pooling (ASPP) module. It samples the given input in parallel at different sampling rates, and the effect is to obtain the context information of the image in different scales. Chen et al. also propose DeepLabv3+ [12] to get better semantic segmentation performance; DeepLabv3+ takes an encoder-decoder structure as a whole, which can obtain more context feature information.

*Instance segmentation*: instance segmentation includes object detection and semantic segmentation. Instance segmentation is proposed by Hariharan et al. [13]. Instance segmentation generates segmentation results and then detects the segmentation results. Girshick et al. [14] propose the regional convolutional neural network (R-CNN), which first makes regional candidates and then classifies objects in the selected region. Fast R-CNN [15] has greatly improved the training speed of R-CNN.

Ren et al. propose Faster R-CNN [16]. As a continuation of Fast R-CNN, this method proposes the region proposal

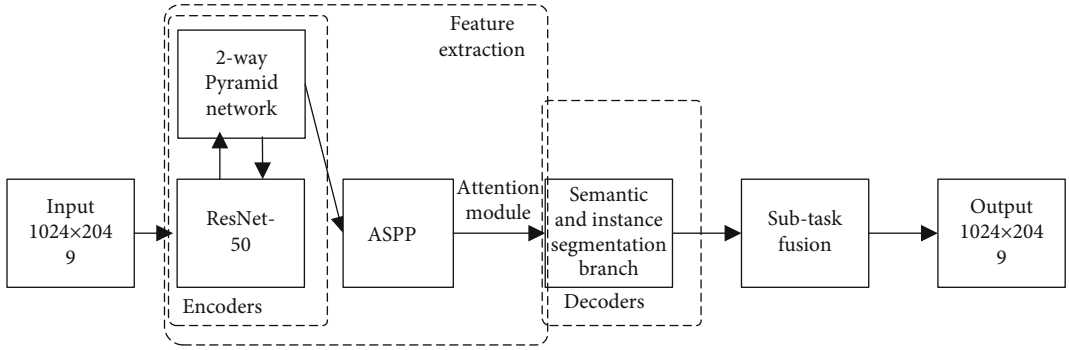


Figure 2: Overview of our architecture: ResNet-50, 2-way pyramid network, and ASPP consist of the feature extraction part of the attention module. The 2-way pyramid network as encoders transforms feature information to decoders in the semantic and instance segmentation branch.

network, which functions similar to the attention mechanism, generates target candidate boxes, and optimizes target detection. Then, Mask R-CNN proposed by He et al. [17] adds the mask mechanism on the basis of Faster R-CNN to perform parallel computing with Faster R-CNN added with FPN. Because of its accuracy and speed, this model is often used in the instance segmentation branch of panoptic segmentation. [18, 19] also introduce the semantic segmentation branch, so that each pixel can be marked. This branch of semantic segmentation is also very similar to the later semantic segmentation branch of panoptic segmentation.

*Panoptic segmentation*: BlitzNet proposed by Dvornik et al. [20] is considered to be the prototype of a single-stage panoptic segmentation model. It cascades object detection and semantic segmentation. Deeperlab proposed by Yang et al. [21] uses the bottom-up method and uses three parts of the mainstream panoptic segmentation. Then, Panoptic-DeepLab proposed by Cheng et al. [7] gets the best performance of panoptic segmentation. The structure of ASPP is added before Panoptic-DeepLab's semantic and instance segmentation branch, and Panoptic-DeepLab has strong expansibility. The performance of the model can be further improved by modifying the semantic and instance segmentation branch of feature extraction. FPSNet [22] uses a heuristic algorithm to make the model simpler and easier to implement. These methods are one-stage methods without using RPN. Some panoptic segmentation methods which use RPN are called two-stage methods. JSIS-Net proposed by De Geus et al. [23] uses a shared feature extractor to provide features for semantic and instance segmentation branches. TASC-Net proposed by Li et al. [24] reduces the fusion loss by adding a mask mechanism to align the "thing" categories of the semantic and instance segmentation branch. Panoptic-FPN proposed by Kirillov et al. [5] adds the feature pyramid network to help extract context information. Our model proposes two opposite FPNs to get more features from the segmentation input part.

*Attention module*: the spatial attention module and channel attention module are the two most commonly used modules. The channel attention module enables the neural network to automatically determine which channel is important or unimportant and then assign appropriate weight. SE (Squeeze-and-Excitation) [25] is based on the channel attention module. The spatial attention module is to find the most important part of the network for processing. Our attention module combines the spatial attention module and channel attention.

*AI-empowered IOT*: some research in different fields on the Internet of things focuses on datasets. [26] presents an out-of-core 3D segmentation method for large-scale image datasets on medical service. [26] introduces the novel concept of $\epsilon$-Kernel Dataset on Wireless Sensor Networks (WSNs) and designs a distributed algorithm to satisfy the $\epsilon$ requirement. [27–29] also propose algorithms for WSNs, while our approach focuses on the Cityscapes dataset, which is the representative of the open environment of street scenes.

## 3. Architecture

This section first introduces the overview of our proposed model. Then, each component of the model is separately introduced in each part.

As is shown in Figure 2, the size of our model's input from the Cityscapes is $1024 \times 2049$. Our model focuses on the improvement of feature extraction and consists of the following parts: a ResNet-50 backbone. The feature will be passed into a 2-way pyramid network, which sends large-scale feature images to decoders from semantic and instance segmentation branches and produces feature maps for the ASPP part. ASPP is used for getting multiscaled features. The attention module is used to get the most important features from ASPP. The semantic and instance segmentation branch and subtask fusion part are similar to Panoptic-DeepLab. Our semantic segmentation branch and instance branch are similar to DeepLabv3+ [12]. Subtask fusion obtains the loss function and gets the results of the model.

*3.1. 2-Way Pyramid Network.* Figure 3 is the 2-way pyramid network, which will adopt 4 or 8x downsampling for the input large-scale feature image to obtain more detailed information, and 16 or 32x downsampling for small-scale

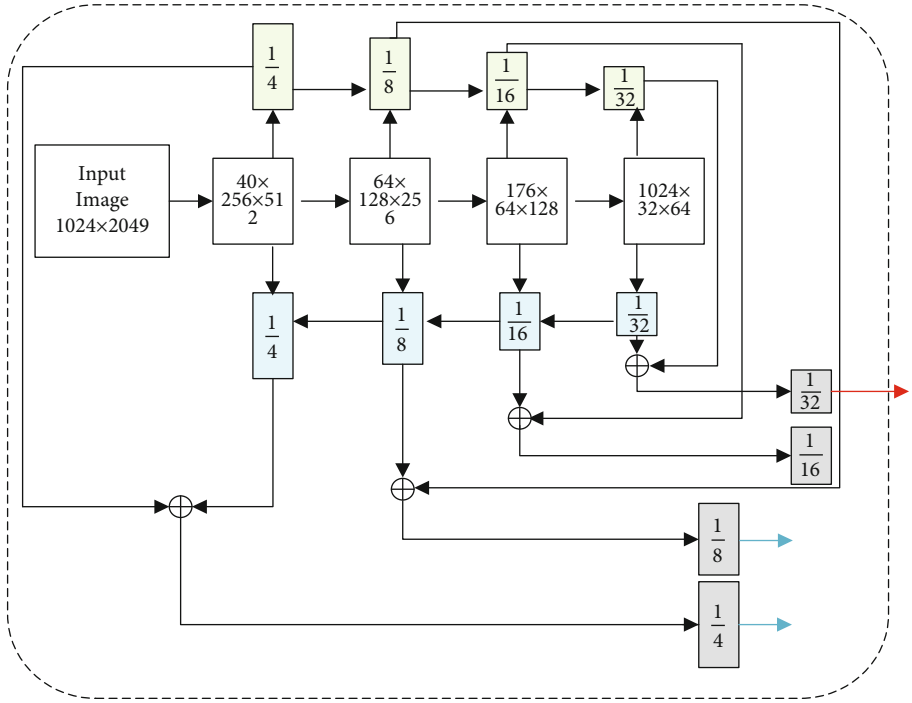


Figure 3: The 2-way pyramid network's information transmission is bidirectional. Four outputs are got from addition by two-way feature blocks.

features. The low-resolution features will be upsampled to the high-resolution features for fast operation and reduce the amount of calculation. The output of the last two branches corresponds to the sum and calculates through 3 × 3 separable convolutions with 256 output channels. After calculating, the results of 4, 8, 16, and 32x downsampling are obtained.

The 2-way pyramid network combined with the backbone network is divided into two parts, the light gray part for forward propagation and the light blue part for back propagation. The backbone network of the white block inputs the information into the two-part pyramid network with 4, 8, 16, and 32x downsampling, respectively, in which the light gray part propagates downward from the features with larger size and the light blue part propagates upward from the features with smaller size. At the same time, the gray and blue blocks are also fused with each other. In this way, the features of the obtained white blocks combine forward propagation and back propagation information. Among them, 32x of the output is used for Atrous Spatial Pyramid Pooling (ASPP), and 8x and 4x of the output will be sent to the decoder part.

*3.2. Atrous Spatial Pyramid Pooling.* Figure 4 is the architecture of ASPP. Four kinds of atrous convolutions with sampling rates will be input for sampling, which are atrous convolutions with the rate of 1, 6, 12, and 18, respectively. When the interval is 6, 12, and 18, 3 is adopted ×3. If the rate is too large, the context information obtained will be too rare and will not help feature extraction. If the rate is too small, too much feature information obtained will lead to a significant decrease in computing speed. Therefore, the rate of 6, 12, and 18 is the best combination of speed and precision. When the interval is 1, 3 × 3 kernels will become 1 × 1 because there is no rate. This method directly extracts the corresponding features. A total of four atrous convolutions and one pooling form the ASPP model. Finally, the two ASPP structures extract semantic, instance, and multiscale context information, respectively.

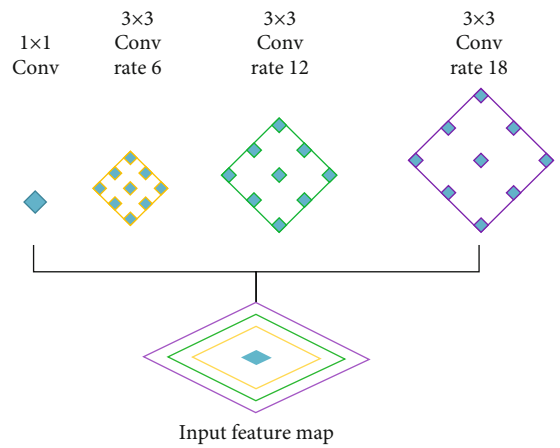


Figure 4: A feature map is divided into four parts. These four parts will be combined into an input feature map which is sent to the attention module.

*3.3. Attention Module.* Figure 5 shows our attention module. We propose the attention module which combines the channel attention module and spatial attention module in Figure 4. The channel attention module performs the

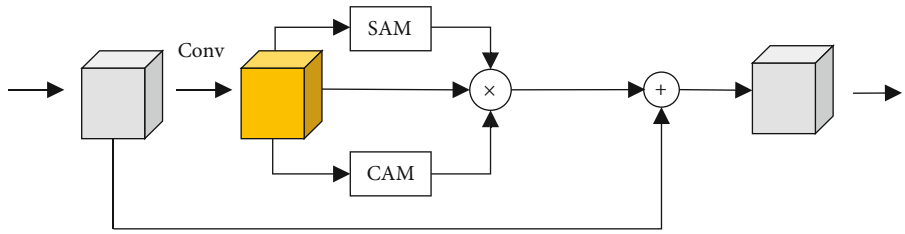


Figure 5: The attention module using spatial attention module (SAM) and channel attention module (CAM) focuses on the spatial and channel part which determines the weight of the feature.

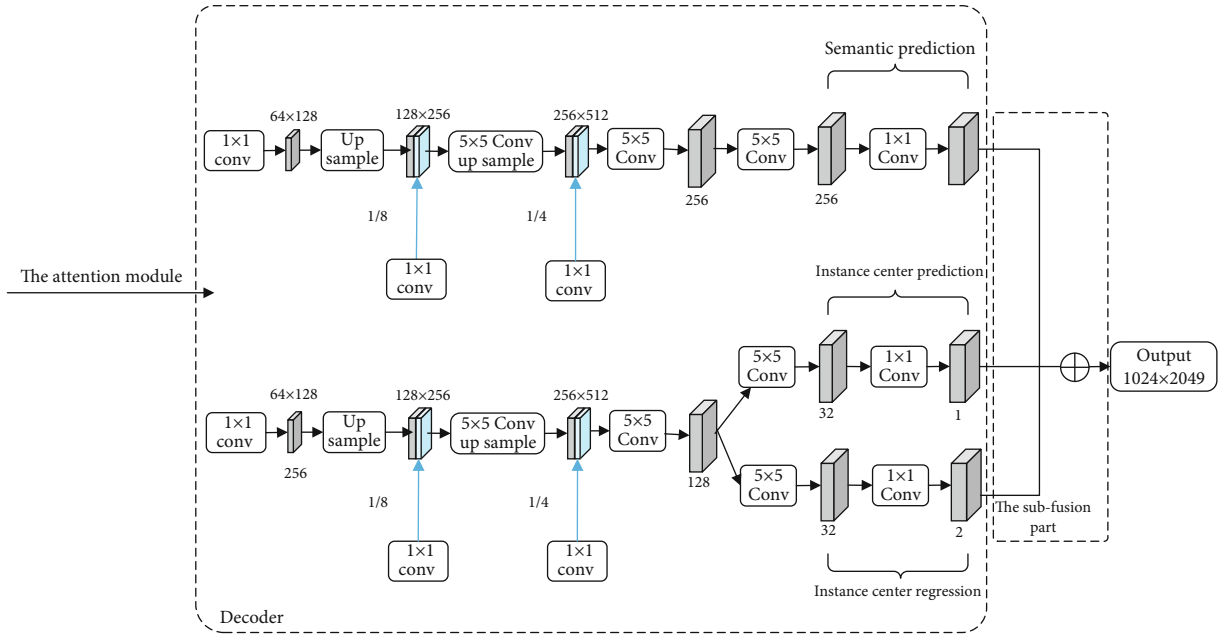


Figure 6: The structure of the image segmentation part. The semantic and instance segmentation branch is the decoder of our model. The subfusion part will obtain the output of semantic segmentation prediction, instance center prediction, and instance center regression.

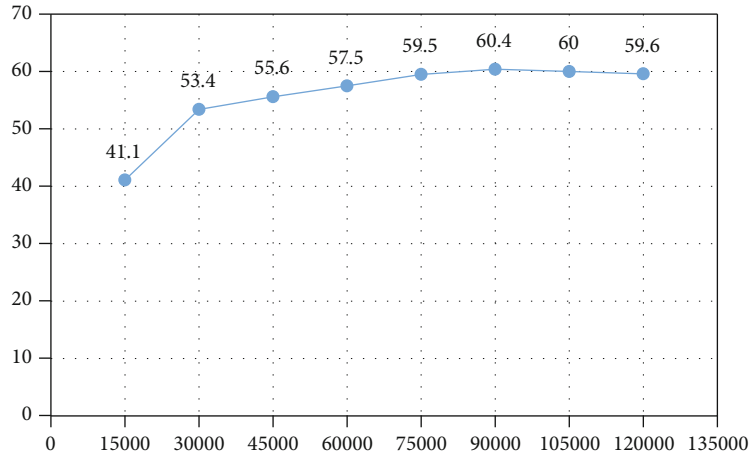


Figure 7: Comparison of panoptic segmentation experiments with different iterations.

Table 1: Comparison of panoptic segmentation of modules.

| Model | PQ (%) | SQ (%) | RQ (%) | $PQ^{st}$ (%) | $PQ^{th}$ (%) |
|---|---|---|---|---|---|
| Original | 57.2 | 79.9 | 70.2 | 63.7 | 48.2 |
| Model+2-way pyramid network | 59.6 | 81.5 | 71.7 | 64.1 | 51.8 |
| Model+attention | 58.7 | 80.7 | 71.0 | 64.0 | 49.9 |
| Model+2-way pyramid network+attention | 60.4 | 82.4 | 72.6 | 64.5 | 53.2 |

Table 2: Comparison between our model and mainstream panoptic segmentation networks.

| Model | PQ (%) | SQ (%) | RQ (%) | $PQ^{st}$ (%) | $PQ^{th}$ (%) |
|---|---|---|---|---|---|
| Deeperlab [21] | 56.5 | — | — | — | — |
| Panoptic-FPN [5] | 58.1 | — | — | 62.5 | 52.0 |
| Panoptic-DeepLab +Res50 [7] | 58.0 | 80.2 | 70.7 | 64.3 | 48.5 |
| Ours | 60.4 | 82.4 | 72.6 | 64.5 | 53.2 |

maximum pooling and average pooling of the input feature map, respectively, and then puts it into the shared Multilayer Perceptron (MLP):

$$M_C(F) = \text{Sigmoid}(\text{MLP}(\text{AvgPool}(F)) + \text{MLP}(\text{MaxPool}(F))). \quad (1)$$

The output features of these two types perform the summation of the element-wise level first, and then, the final channel attention features are obtained through the operation of the sigmoid activation function. The spatial attention module takes the feature map output by the channel attention module as the feature map input by this module:

$$M_S(F) = \text{Sigmoid}\left(f^{7\times7}([\text{AvgPool}(F); \text{MaxPool}(F)])\right). \quad (2)$$

First, do the same operations of maximum pooling and average pooling to obtain two types of outputs and splice them together. Then, the dimension is reduced by convolution, and finally, the spatial attention feature is obtained by the operation of sigmoid activation function:

$$F_1 = M_C(F) \otimes M_S(F) \otimes F. \quad (3)$$

$M_C(F)$ is the output of CAM, while $M_S(F)$ is the output of SAM. Our method uses $F_1$ to get the output obtained by the cross-multiplication of CAM and SAM.

*3.4. Image Segmentation Part.* Figure 6 shows the structure of the image segmentation part, and the details of the picture are described below.

*Semantic segmentation*: this part adopts a method similar to DeepLabv3+, and after each upsampling operation, the $5 \times 5$ separable convolution will be used to improve the acquisition of context information. After the $1 \times 1$ kernel is finally used, weighted bootstrapped cross-entropy loss is also used in the semantic segmentation branch.

*Instance segmentation*: the part of the instance segmentation branch is similar to semantic segmentation, while the instance segmentation branch is divided into two parts. One is the instance center prediction, and the other is the instance center regression. $F_1$ loss is used in the instance center regression to minimize the distance between the predicted heating map and the ground truth heating map. The Mean Square Error (MSE) is the loss of instance segmentation regression.

*Fusion*: in the subfusion part, there are three kinds of losses, which are from instance center prediction, instance center regression, and semantic segmentation prediction. These three loss functions will be obtained in the form of accumulation:

$$L = \lambda_{\text{ICP}} L_{\text{ICP}} + \lambda_{\text{ICR}} L_{\text{ICR}} + \lambda_S L_S, \quad (4)$$

where $\lambda$ is the set superparameter, which will be adjusted according to the change of iteration time in training. $L$ is the total loss of our model. ICP represents the instance center prediction, ICR represents the instance center regression, and $S$ represents the semantic segmentation prediction.

## 4. Results and Discussion

*4.1. Experiments.* This section introduces the evaluation results of the 2PN model. The first part introduces the applied datasets and corresponding implementation details. The second part discusses the results and their comparison with baseline solutions.

*4.2. Datasets and System Settings. Cityscapes*: Cityscapes [30] is known as the urban street scene dataset. The images of the dataset are mainly from the street scenes provided by German companies. The dataset consists of 20000 weak annotation frames and 5000 high-quality annotation frames. The Cityscapes dataset has 19 categories, including 2975 pictures to form the training set, 500 pictures to form the Val set, and 1525 pictures to form the test set. The Cityscapes dataset focuses on street scenes with high image quality and fine annotation, which plays an important role in the understanding of street scenes. According to the development direction of panoptic segmentation in the future and in order to reduce the training time, the dataset selected in this experiment is Cityscapes, which focuses on street scenes and has a smaller scale than the Mapillary Vistas dataset [31].

*Settings*: we choose ResNet-50 [8] as our model's backbone. Our experiments use the same parameter setting as

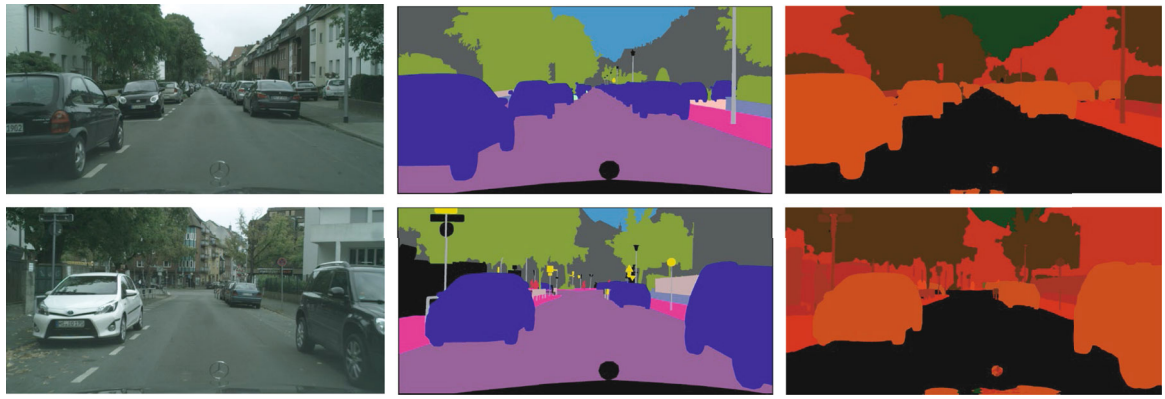

Figure 8: Our visualization results on Cityscapes. We show image, ground truth, prediction from left to right.

Panoptic-DeepLab [7]. Experiments are trained on one NVidia GeForce RTX 3090 with 24 GB video memory.

Our panoptic segmentation model is evaluated by panoptic quality (PQ), which is obtained by multiplying segmentation quality (SQ) and recognition quality (RQ) [1]. $PQ^{st}$ and $PQ^{th}$ represent the panoptic segmentation results of "stuff" and "thing."

## 5. Results

We first explore the impact of the number of iterations on the accuracy of the network. Figure 7 shows the comparison of panoptic segmentation experiments with different iterations.

As Figure 7 shows, when the number of iterations is 90000, the performance of the network reaches the maximum. Accuracy of the panoptic segmentation network will be reduced due to overfitting.

*5.1. Ablation Studies.* We analyze the impact of the two-way pyramid network and attention module on the accuracy of panoptic segmentation.

In Table 1, our model gets 3.2% better than the original model in Cityscapes. The model with the 2-way pyramid network works 0.9% better than the model with the attention module.

Due to the advantages of small object feature extraction, the two-way pyramid network and attention module have a great improvement in the "thing" part and less impact on the "stuff" part. The difference is that the two-way pyramid network affects more network layers and has two-way features. Therefore, the two-way pyramid network greatly improves the model. The attention module also improves the panoptic segmentation network, but its improvement range is obviously less than that of the two-way pyramid network, and the main improvement part is also in the "thing" part.

*Comparison*: compared with the above three models in Table 2, this model has some advantages. First, compared with the baseline, the PQ value in the thing part has obvious advantages. The model in this chapter enhances the feature extraction method and strengthens the extraction of small objects and the acquisition of context information. Meanwhile, compared with Panoptic-FPN, due to the use of empty feature pyramid pooling, it has been significantly improved in semantics and instances.

Finally, Figure 8 shows the results of the proposed model. Our prediction model works well in most areas. It can be observed that things on the roadside are detected, while the things and stuffs can also be correctly distinguished. However, the prediction in the black part of ground truth needs to be improved, where the detailed detection of background is still not clear.

## 6. Conclusion

This paper proposes a novel framework for panoptic segmentation towards surveillance in an open environment. The proposed network includes a light-weighted two-way pyramid network for better feature extraction, and an attention module is adapted to adjust the importance of extracted features. In this way, our model can obtain the context information features of the image. The attention module also optimizes the space and channel of small-scale features. Finally, the experimental tests show that the methods proposed in this chapter are effective and workable.

## Data Availability

The author's e-mail is 630211995@qq.com, and the code of ResNet-50 with 2PN is available.

## Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

## References

[1] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, "Panoptic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9404–9413, Long Beach California, 2019.

[2] T. Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2117–2125, Honolulu, 2017.

[3] U. Bonde, P. F. Alcantarilla, and S. Leutenegger, "Towards bounding-box free panoptic segmentation," in *DAGM German Conference on Pattern Recognition*, vol. 12544 of Lecture Notes in Computer Science, , pp. 316–330, Springer, Cham, 2021.

[4] H. Wang, R. Luo, M. Maire, and G. Shakhnarovich, "Pixel consensus voting for panoptic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9464–9473, Long Beach California, 2020.

[5] A. Kirillov, R. Girshick, K. He, and P. Dollár, "Panoptic feature pyramid networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6399–6408, Long Beach California, 2019.

[6] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[7] B. Cheng, M. D. Collins, Y. Zhu et al., "Panoptic-DeepLab: a simple, strong, and fast baseline for bottom-up panoptic segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12475–12485, 2020.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, Las Vegas, 2016.

[9] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, Boston, 2015.

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-assisted Intervention*, pp. 234–241, Springer, 2015.

[11] H. S. Zhao, J. P. Shi, X. J. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2881–2890, Honolulu, 2017.

[12] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semanticimage segmentation," in *Proceedings of the European Conference on ComputerVision (ECCV)*, pp. 801–818, Munich, 2018.

[13] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Simultaneous detection and segmentation," in *European Conference on Computer Vision*, pp. 297–312, Springer, Cham, 2014.

[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, Columbus, 2014.

[15] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1440–1448, Boston, 2015.

[16] S. Ren, K. He, R. Girshick, and R.-C. N. N. Faster, *Towards real-time object detection with region proposal networks*, 2015.

[17] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *Proceedings of the IEEE International Conference On Computer Vision*, 2017, pp. 2961–2969, Long Beach California, 2017.

[18] A. Fathi, Z. Wojna, V. Rathod et al., "Semantic instance segmentation via deep metric learning," https://arxiv.org/abs/1703.10277.

[19] D. Neven, B. D. Brabandere, M. Proesmans, and L. V. Gool, "Instance segmentation by jointly optimizing spatial embeddings and clustering bandwidth," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8837–8845, Long Beach California, 2019.

[20] N. Dvornik, K. Shmelkov, J. Mairal, and C. Schmid, "Blitznet: a real-time deep network for scene understanding," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4154–4162, Venice, 2017.

[21] T. J. Yang, M. D. Collins, Y. Zhu et al., "Deeperlab," https://arxiv.org/abs/1902.05093.

[22] D. De Geus, P. Meletis, and G. Dubbelman, "Fast panoptic segmentation network," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1742–1749, 2020.

[23] D. De Geus, M. Panagiotis, and G. Dubbelman, "Panoptic segmentation with a joint semantic and instance segmentation network," https://arxiv.org/abs/1809.02110.

[24] J. Li, A. Raventos, A. Bhargava, T. Tagawa, and A. Gaidon, "Learning to fuse things and stuff," https://arxiv.org/abs/1812.01192.

[25] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, Salt Lake City, 2019.

[26] K. Kwon and B. Shin, "3D segmentation for high-resolution image datasets using a commercial editing tool in the IoT environment," *Journal of Information Processing Systems*, vol. 13, no. 5, pp. 1126–1134, 2017.

[27] S. Cheng, Z. Cai, and J. Li, "Curve query processing in wireless sensor networks," *IEEE Transactions on Vehicular Technology.*, vol. 64, no. 11, pp. 5198–5209, 2015.

[28] Z. He, Z. Cai, S. Cheng, and X. Wang, "Approximate aggregation for tracking quantiles and range countings in wireless sensor networks," *Theoretical Computer Science*, vol. 607, pp. 381–390, 2015.

[29] J. Li, S. Cheng, H. Gao, and Z. Cai, "Approximate physical world reconstruction algorithms in sensor networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 12, pp. 3099–3110, 2014.

[30] M. Cordts, M. Omran, S. Ramos et al., "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3213–3223, Las Vegas, 2016.

[31] G. Neuhold, T. Ollmann, S. R. Bulo, and P. Kontschieder, "The Mapillary Vistas Dataset for semantic understanding of street scenes," in *IEEE International Conference on Computer Vision*, pp. 4990–4999, Venice, 2017.