Real-Time Instance Segmentation Method Based on Location Attention

Li Liu[1*], and Yuqi Kong[2]
[1] School of Automation and Electrical Engineering, University of Science and Technology Beijing
Beijing 100083, China
[e-mail: liuli@ustb.edu.cn]
[2] Shunde Innovation School, University of Science and Technology Beijing
Fushan 528399, China
[e-mail: m202221544@xs.ustb.edu.cn]
[*]Corresponding author: Li Liu

Received August 14, 2023; revised April 3, 2024; accepted September 8, 2024; published September 30, 2024

Abstract

Instance segmentation is a challenging research in the field of computer vision, which combines the prediction results of object detection and semantic segmentation to provide richer image feature information. Focusing on the instance segmentation in the street scene, the real-time instance segmentation method based on SOLOv2 is proposed in this paper. First, a cross-stage fusion backbone network based on position attention is designed to increase the model accuracy and reduce the computational effort. Then, the loss of shallow location information is decreased by integrating two-way feature pyramid networks. Meanwhile, cross-stage mask feature fusion is designed to resolve the small objects missed segmentation. Finally, the adaptive minimum loss matching method is proposed to decrease the loss of segmentation accuracy due to object occlusion in the image. Compared with other mainstream methods, our method meets the real-time segmentation requirements and achieves competitive performance in segmentation accuracy.

Keywords: Deep learning, Instance segmentation, Attention mechanism, Adaptive loss function

1. Introduction

Image segmentation is a pixel-level classification task, i.e., the division of graphic regions, which mainly consists of semantic segmentation, instance segmentation, and panoptic segmentation. Instance segmentation is a combination of object detection and semantic segmentation tasks, which requires not only the classification of pixels in an image but also the localization of different instances with the same semantic category in the foreground. Therefore, instance segmentation can provide richer and more accurate image feature information and is widely used in satellite remote sensing image processing [1,2,3], defect detection [4,5,6], intelligent vehicles [7,8], medical image processing [9,10,11], and other fields. The instance segmentation methods based on deep learning are mainly divided into two-stage, single-stage, and multi-stage.

According to the sequence of detection and segmentation, two-stage methods are classified as top-down and bottom-up. Top-down methods, such as PANet [12], and MS R-CNN [13], first identify the candidate regions of interest using an object detector, and then segment in each candidate region. However, the accuracy of object detection is not high enough, which will directly affect the subsequent segmentation results, resulting in lower detection accuracy. Another classical two-stage method Mask R-CNN [14], proposes a new mask representation to improve the model resolution, but the processing of fine details is still limited. The bottom-up methods PointRend [15] and Condlnst [16] newly introduce the clustering of neighboring pixels according to certain rules and combine the clustering results with semantic information to identify the object instances. However, because of the long processing time and high computational cost, bottom-up methods only have advantages in some specific complex scene detection tasks.

Under the unified framework of fully convolutional networks, one-stage instance segmentation methods are divided into two categories: anchor-based and anchor-free. The anchor-based one-stage methods complete instance segment tasks by breaking them into two parallel branches: generating a set of prototype masks and predicting per-instance mask coefficients. Then they produce instance masks by linearly combining the prototypes with the mask coefficients [17], such as YOLACT [18] and RTSS [19]. Anchor-based modeling can infer quickly but lacks the ability to describe object voids [20]. Unlike the anchor-based one-stage methods, the anchor-free one-stage methods add a mask branch to the detector or integrate the location information into a one-stage segmentation framework, such as CenterMask [21], BlendMask [22], Polarmask [23], SOLO [24], and LSNet [25]. The SOLOv2 [26] uses spatial correspondence between semantic and mask branches to achieve a well-balanced speed and accuracy, but the prediction results can be ambiguous when facing the presence of multiple objects in the same grid.

The multi-stage instance segmentation based on cascade structure, such as the Cascade R-CNN [27], is beneficial to train better features and further improve the instance segmentation performance. Query-based models like SOTR [28] and QueryInst [29], also the Transformer embedding method as K-Net [30] and MaskFormer [31] are the latest representatives of multi-stage methods which are the highest segmentation accuracy at present. However, the multi-stage design brings a costly computational effort. On this basis, some transformer-based methods like Mask2Former [32], Mask Transfiner [33], and Fastlnst [34] bring new improvements. Mask2Former designs masked attention, which extracts localized features by constraining cross-attention within predicted mask regions. Mask Transfiner proposes an incoherent regions detection mechanism, which decomposes and represents the image regions as a Quadtree, then corrects errors with the transformer autonomously. FastInst follows

the meta-architecture of Mask2Former, which incorporates instance activation-guided queries, a dual-path update strategy, and ground truth mask-guided learning to allow the model to utilize lighter pixel decoders and fewer Transformer decoder layers. In addition, PatchDCT [35], as a newly proposed multi-stage real-time instance segmentation method, partitions the mask derived from a DCT vector into multiple patches and enhances each patch using a specially designed classifier and regressor. Although these improvements improve the detection speed and the detection accuracy of edge details, they still have limitations in solving the problem of detecting small objects and handling object occlusion.

In general, two-stage and multi-stage instance segmentation methods are widely used with high segmentation accuracy, but these models are difficult to deal with real-time problems. One-stage instance segmentation methods achieve a relative balance of segmentation accuracy and inference speed by performing detection and segmentation.

As an important part of the field of intelligent transportation, instance segmentation in the street scenes has a wide range of application prospects in real-time road condition detection, traffic planning optimization, and autonomous driving. Instance segmentation based on street scenes is the process of segmenting various objects such as various vehicles, road signs, buildings and so on at pixel level using computer vision techniques. Street scene instance segmentation requires high timeliness of system feedback, but most current instance segmentation algorithms cannot meet the requirements of segmentation accuracy and real-time at the same time. We propose a real-time instance segmentation network in the street scenes. The overall segmentation accuracy of the model is further improved by reducing the computation of the backbone network, also the small object miss segmentation and object occlusion problems are solved. Ablation experiments are conducted on MS COCO and Cityscapes datasets, comparing with other mainstream methods, our method achieves more competitive performance in both segmentation speed and segmentation accuracy.

In summary, our main contributions can be summarized as follows:

- We propose a cross-stage fusion backbone network based on location attention to fully extract image features and reduce computational effort.
- Decreasing the loss of shallow location information by integrating two-way feature pyramid networks; meanwhile, designing cross-stage mask feature fusion to resolve the problem of missing segmentation of small objects.
- The adaptive minimum loss matching method is proposed to reduce the occlusion caused by multiple object centers in an image falling into the same grid.

The remainder of the paper is organized as follows. Section 2 presents the real-time instance segmentation method based on SOLOv2. Extensive experiments are implemented and the performance of our method is evaluated compared with others in Section 3. Finally, the conclusion is presented in Section 4.

2. Instance Segmentation Methods

The cross-stage fusion backbone network is proposed based on location attention, feature information fusion structure design, and adaptive minimum loss function matching method accordingly. The general framework is shown in **Fig. 1**, which contains feature extraction, feature fusion, semantic branching, and mask branching.

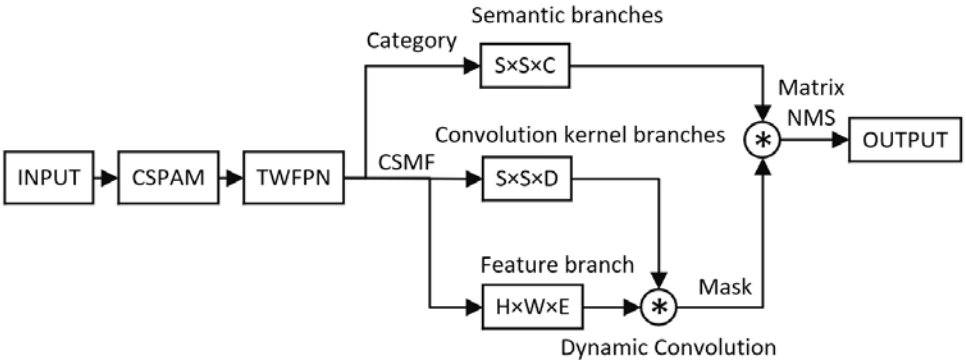

Fig. 1. The general framework of the instance partitioning.

In **Fig. 1**, CSPAM represents a cross-stage fusion backbone network based on location attention, TWFPN represents a two-way feature pyramid network, and CSMF represents cross-stage feature information fusion.

2.1 Cross-Stage Fusion Backbone Network

First, the cross-stage fusion structure is used to improve the original backbone network ResNet of the SOLOv2 model, and the number of parameters of the convolutional layers is reduced, which can decrease memory consumption and improve the training efficiency of the model. Then, the Relu activation function is replaced considering several advantages of the Mish activation function such as training stability, non-monotonicity, no upper limit, and smoothing. The structure of the improved network is shown in **Fig. 2**, where BN represents Batch Normalization, CONV represents Convolution, and Concat represents Dimensional Splicing.

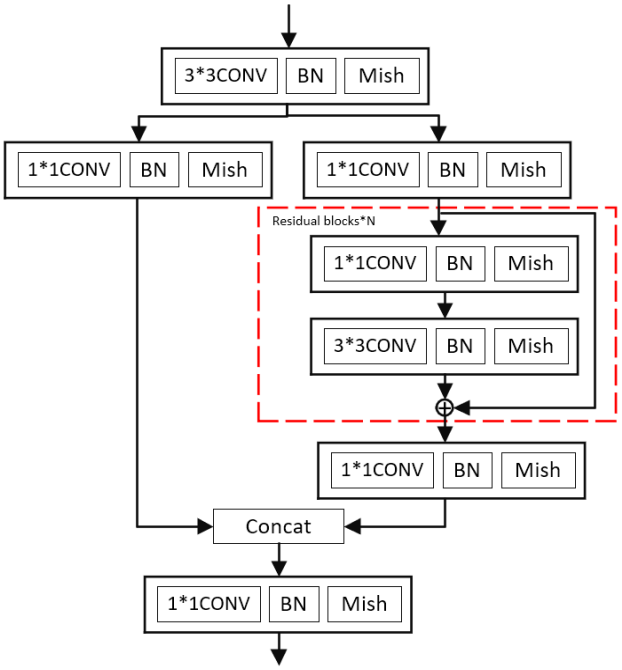

Fig. 2. Improved cross-stage feature extraction network.

In order to fully extract the position information of the bottom feature map, the Position Attention Module is applied between the residual blocks, as shown in **Fig. 3**. The input features C × H × W are first convolved by 1 × 1 to reduce the channel dimension C to 1/8 of the original one to get the feature map C1 × H × W, which decrease the computation. Then the H × W dimensions are combined by reshape operation to obtain three C1 × N 2D feature maps P, A, and M. Where, P is transposed and multiplied with A, after softmax processing, then multiplied with M to get the feature map PAM. Afterwards 1 × 1 convolution and reshape operation, there is the C2 × H × W feature map T2.

Finally, the reduced-dimensional feature map T1 is multiplied by the adaptive weight factor θ, and then the Concat operation of the channel dimension is performed with the feature map T2 to reduce the output to the same dimension as the input. Where θ is set to 0 at the beginning and the optimal hyper parameter value is obtained according to the training process. The final output of the module T3 is a concatenation of high and low-level information, which makes full use of the location and semantic information and increases the local feature accuracy of the network.

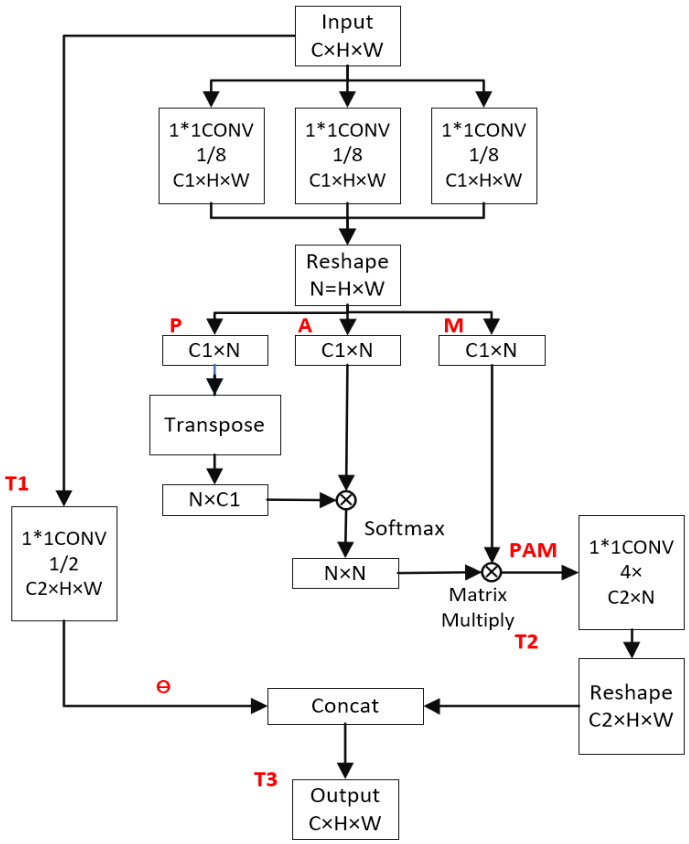

Fig. 3. Position attention mechanism.

2.2 Feature Fusion Structure

The FPN [36] does not make effective use of the shallow location information, resulting in the loss of location information during the delivery of feature maps, which is especially obvious in the segmentation of small objects. We introduce the bottom-up information flow and design

the two-way feature pyramid network TWFPN to supplement the missing location information. Specifically, the lateral connection inputs of the FPN are the output features of layers 2, 3, 4, and 5 of the backbone network, after 50–100 layers of convolution, the shallow location information is seriously lost, so the bottom-up location information enhancement is an effective solution.

For the operation of FPN to enlarge the feature layers to the same size and then overlay, the cross-stage mask feature fusion structure is designed in two ways: 1) after 2X upsampling, the higher semantic layer P5 concat with the lower feature layer N4 in the channel dimension to reduce the computation; 2) P5 forms a global semantic guide to P2 by 8X upsampling, which improves the detection and segmentation capability for small objects. The feature information fusion structure is shown in **Fig. 4**, where CoordConv represents Coordinate Convolution.

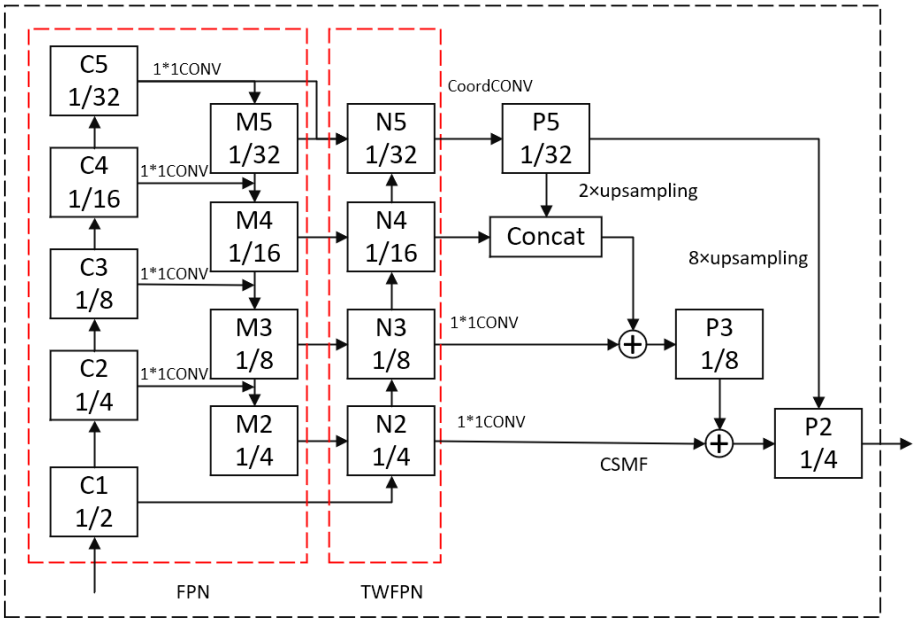

Fig. 4. Feature information fusion structure.

2.3 Adaptive Minimum Loss Function Matching Method

The SOLOv2 has the occlusion problem caused by multiple target centers falling into the same grid. To solve this problem, the influence factor ε is first introduced in the function L_{Mask}, which adaptively adjusts the weight information of the predicted *Mask* to increase the instance segmentation accuracy. Then, the category and mask losses are normalized by the sigmoid function, and the two losses are integrated by summation operation. Finally, the *Loss* is compared and the minimum loss is used for network training. L_{Mask}, ε, and *Loss* are defined as formulas 1, 2, and 3, respectively.

$$L_{\text{Mask}} = \frac{1}{N_{\text{pos}}} \sum_{i \in P} \left(1 - \frac{2 \sum m_i g_i + 1}{\sum m_i + \sum g_i + 1} + \varepsilon \right) \tag{1}$$

$$\varepsilon = \frac{\left(C_{PX} - C_{GX}\right)^2 + \left(C_{PY} - C_{GY}\right)^2}{\left|\sum m_i g_i\right| + 1} \tag{2}$$

$$Loss = sigmoid\left(L_{\text{Cate}}\right) + sigmoid\left(\lambda L_{\text{Mask}}\right) \tag{3}$$

where, L_{Cate} represents the category loss function calculated using focal loss, P is the set of positive samples, and N_{pos} is the number of positive samples. The m_i and g_i are the prediction mask and true mask corresponding to the ith feature point. C_P and C_G are the predicted *Mask* and the true labeled centroid position respectively. X and Y are respectively the horizontal and vertical coordinates.

3. Performance Evaluation

We perform ablation experiments and evaluate the performance of our method by comparing it with other instance segmentation methods on MS COCO and Cityscapes datasets. The Cityscapes dataset is transformed into the COCO dataset format and the original categories of the Cityscapes dataset are reduced, keeping the five categories of common objects in street scenes: cars, pedestrians, trucks, buses, and riders. The accuracy of different instance segmentation methods is evaluated by AP, the segmentation effect of small objects is evaluated in terms of AP_S, and FPS (frames per second) is the speed of different models.

3.1 Experimental Setting

Data enhancement (random flipping, size scaling, random cropping, etc.) and data normalization are used for data preprocessing, in line with the original SOLOv2 algorithm. To reduce the model computation, the training and test images were scaled to 1200 × 680 pixel size. After several experiments, the model parameters are set as follows: batch_size is set to 16, 8 graphics card × 2 images; the optimizer is selected as SGD, the initial learning rate is 0.01, the momentum is 0.9; the regularization weight decay coefficient is 0.0001; the epochs is 40; the decay rate of learning rate is 0.1 setting at the 10th time of model training.

3.2 Ablation Experiment

Table 1 shows the comparison results of 8 sets of ablation experiments, and the 3 structural designs for the SOLOv2 network effectively improve the model recognition and segmentation accuracy. Based on the original backbone network ResNet-50-FPN of the SOLOv2 model, CSPAM indicates cross-stage fusion based on location attention, TWFPN + CSMF indicates feature information fusion, and AML indicates the addition of adaptive minimum loss matching method. It can be seen that our improved SOLOv2 network combined simultaneously with CSPAM, TWFPN + CSMF and AML achieves the best accuracy metrics with 3% increase of AP and 2.6% increase of APs respectively.

Table 1. Results of ablation experiments on the COCO dataset

CSPAM	TWFPN + CSMF	AML	AP (%)	AP_S (%)
			32.6	12.2
✓			34.2	13.0
	✓		34.6	14.5
		✓	33.2	12.9
✓	✓		34.8	14.6
	✓	✓	34.3	14.3
✓		✓	34.7	14.5
✓	✓	✓	35.6	14.8

Fig. 5 shows the comparison of the experimental results of the original SOLOV2 algorithm and the improved algorithm of this paper for seven groups of instance segmentation. From the up-down comparison of group 1 and group 2, it can be seen that the original algorithm has defects in the segmentation of boundary details and our improved algorithm performs better in the segmentation of edge details; From the up-and-down comparison of group 3 and the left-right comparison of group 4, the original algorithm has labeling errors in the category differentiation of birds and trucks, and our improved algorithm gets the correct category labels; From the experimental comparison of group 4, group 5, group 6, and group 7, it can be seen that the original algorithm is prone to the problem of missed segmentation in the segmentation task of smaller objects, and our improved algorithm is better at improving the problem of missed segmentation of small objects such as people and cars in images.

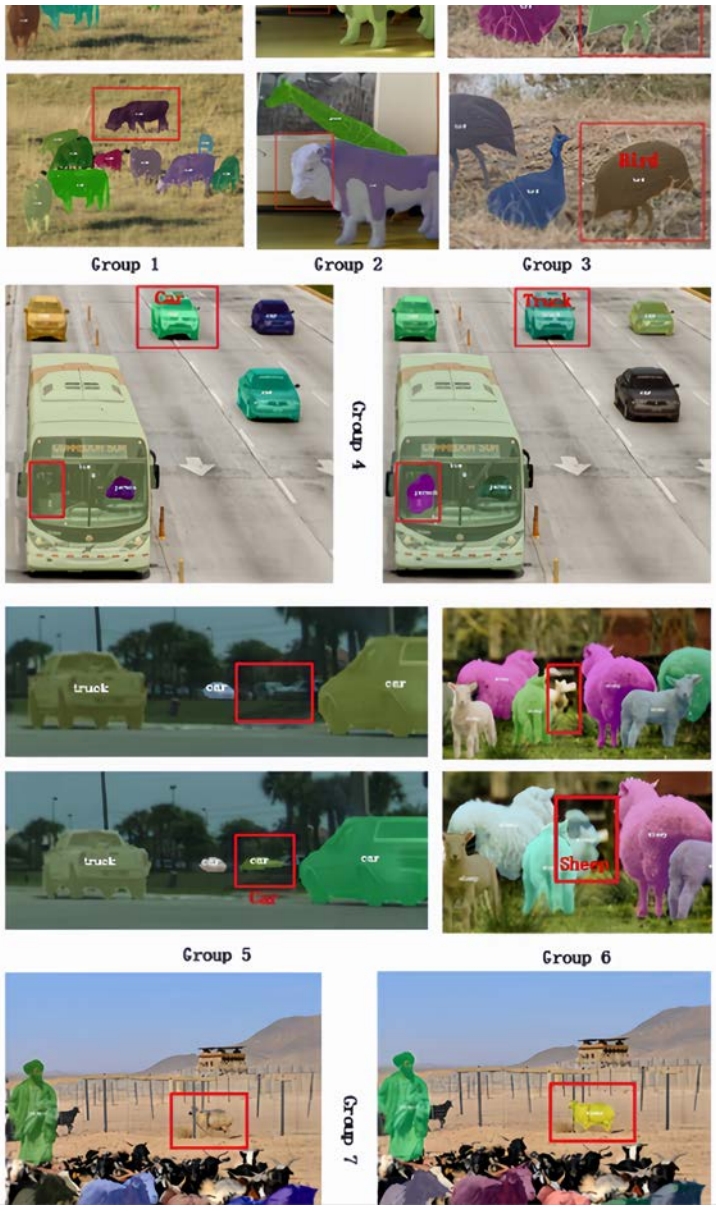

Fig. 5. Comparing experimental results on the COCO dataset.

3.3 Performance Comparing

Table 2 shows the experimental results of our improved SOLOv2 algorithm compared with other classical instance segmentation algorithms on the COCO dataset. To ensure the validity of the performance comparison, the backbone networks of various instance segmentation methods are unified as ResNet-101-FPN, and the forward inference times are tested at 2080TiGPU. However, due to the differences in hardware devices and algorithm parameter settings, the reproduction results of some algorithms marked with * are somewhat different from the original paper.

Table 2. Comparing with other methods on the COCO dataset

Category	Method	AP	APs	Time	FPS
Two-stage	Mask R-CNN [14]	35.3	52.3	96.2	10.4
	PANet [12]	36.6	53.1	212.8	4.7
	MS R-CNN [13]	38.3	54.4	116.3	8.6
Multi-stage	K-Net [30]	40.1	58.8	61.7	16.2
	SOTR [28]	40.2	73.0	140.1	7.1
	QueryInst [29]	40.6	53.8	166.7	6.0
	MaskFormer [32]	42.4	63.8	119.1	8.4
One-stage	YOLACT [18]	28.7	47.2	26.2	38.1
	CenterMask [21]	38.3	54.5	71.9	13.9
	PolarMask [23]	30.4	42.8	81.3	12.3
	SOLOv2 [26]	38.8	56.6	29.9	33.5
	Ours	40.7	58.5	42.1	23.8

It can be seen from **Table 2**, that the two-stage and multi-stage instance segmentation algorithms are better in terms of segmentation accuracy, but the FPS is mostly below 10 FPS, which cannot reach the minimum requirement of real-time inference. The segmentation accuracy of the one-stage instance segmentation algorithm is generally lower than the two-stage and multi-stage methods, while the segmentation speed is faster. Among one-stage instance segmentation algorithms, the YOLACT and SOLOv2 algorithms reach the lower limit requirement of real-time instance segmentation. In this paper, the improved instance segmentation method with parallel inference is structured in such a way that the segmentation accuracy is the highest among the classical single-stage instance segmentation methods, also with the fastest inference time.

The performance comparing results of our improved SOLOv2 with other instance segmentation methods on the street scene dataset Cityscapes are shown in **Table 3**, where the backbone network is ResNet-50-FPN. Among the mainstream algorithms nowadays, most of the instance segmentation algorithms can't achieve real-time performance and their backbone networks are all relatively large. At present, the common algorithms in street scene applications mainly include Mask R-CNN, YOLACT, etc. Compared with these algorithms, the instance segmentation method proposed in this paper has a great improvement in the segmentation accuracy of different object categories.

Table 3. Comparing with other methods on the Cityscapes dataset

Methods	Backbone	AP	Car	Pedestrian	Truck	Bus	Rider
Mask R-CNN	ResNet-50-FPN	30.0	37.1	16.5	41.3	39.2	15.7
YOLACT	ResNet-50-FPN	26.8	34.1	12.7	37.2	37.6	12.2
SOLOv2	ResNet-50-FPN	28.3	35.1	13.5	39.3	39.6	13.8
Our method	CSPAM-50-TWFPN	31.5	38.5	16.0	42.4	44.3	15.8

4. Conclusion

In this work, we propose a real-time instance segmentation network applied to street scenes. The cross-stage fusion backbone network based on location attention improves model accuracy and reduces computational effort. The feature information fusion structure design enriches shallow location information and high-level semantic information to solve the problem of small objects' missed segmentation. The adaptive minimum loss matching method reduces the loss of segmentation accuracy due to objects occlusion in images. The evaluation results show that our proposed network structure achieves competitive performance compared with other mainstream methods, and the inference speed satisfies the real-time requirement.

Acknowledgment

This work is supported by the National Natural Science Foundation of China under Grants 12071025; Scientific and Technological Innovation Foundation of Foshan Municipal People's Government (BK20AE004).

References

[1] X. Xu, Z. Feng, C. Cao, M. Li, J. Wu, Z. Wu, Y. Shang, and S. Ye, "An Improved Swin Transformer-Based Model for Remote Sensing Object Detection and Instance Segmentation," *Remote Sens.*, vol.13, no.23, 2021. Article (CrossRef Link)

[2] W. Zhao, J. Na, M. Li, and H. Ding, "Rotation-Aware Building Instance Segmentation From High-Resolution Remote Sensing Images," *IEEE Geosci. Remote Sens. Lett.*, vol.19, pp.1-5, 2022. Article (CrossRef Link)

[3] B. Chai, X. Nie, H. Gao, J. Jia, and Q. Qiao, "Remote Sensing Images Background Noise Processing Method for Ship Objects in Instance Segmentation," *J. Indian Soc. Remote Sens.*, vol.51, pp.647-659, 2023. Article (CrossRef Link)

[4] M. Ding, B. Wu, J. Xu, A. N. Kasule, and H. Zuo, "Visual inspection of aircraft skin: Automated pixel-level defect detection by instance segmentation," *Chin. J. Aeronaut.*, vol.35, no.10, pp.254-264, 2022. Article (CrossRef Link)

[5] E. Antwi-Bekoe, G. Liu, J.-P. Ainam, G. Sun, and X. Xie, "A deep learning approach for insulator instance segmentation and defect detection," *Neural Comput. Applic.*, vol.34, pp.7253-7269, 2022. Article (CrossRef Link)

[6] C. Zhang and X. Zhang, "Multi-target domain-based hierarchical dynamic instance segmentation method for steel defects detection," *Neural Comput. & Applic.*, vol.35, pp.7389-7406, 2023. Article (CrossRef Link)

[7] Y. Sun, J. Li, X. Xu, and Y. Shi, "Adaptive Multi-Lane Detection Based on Robust Instance Segmentation for Intelligent Vehicles," *IEEE Trans. Intell. Veh.*, vol.8, no.1, pp.888-899, 2023. Article (CrossRef Link)

[8] S. Du, Z. Chen, L. Li, H. Zhang, D. Cao, and L. Chen, "NFIS: A NMS-free FCOS Method for Instance Segmentation," in *Proc. of 2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pp.3150-3155, Indianapolis, IN, USA, Sep. 2021. Article (CrossRef Link)

[9] X. Zhang, Y. Yang, Y. Shen, P. Li, Y. Zhong, J. Zhou, K. Zhang, C. Shen, Y. Li, M. Zhang, L. Pan, L. Ma, and H. Liu, "SeUneter: Channel attentive U-Net for instance segmentation of the cervical spine MRI medical image," *Front. Physiol.*, vol.13, 2022. Article (CrossRef Link)

[10] X. Pang, Z. Zhao, Y. Wang, F. Li, and F. Chang, "LGMSU-Net: Local Features, Global Features, and Multi-Scale Features Fused the U-Shaped Network for Brain Tumor Segmentation," *Electronics*, vol.11, no.12, 2022. Article (CrossRef Link)

[11] D. N. H. Thanh, L. T. Thanh, U. Erkan, A. Khamparia, and V. B. Surya Prasath, "Dermoscopic image segmentation method based on convolutional neural networks," *Int. J. Comput. Appl. Technol.*, vol.66, no.2, pp.89-99, 2021. Article (CrossRef Link)
[12] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path Aggregation Network for Instance Segmentation," in *Proc. of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.8759-8768, Salt Lake City, UT, USA, Jun. 2018. Article (CrossRef Link)
[13] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, "Mask Scoring R-CNN," in *Proc. of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.6402-6411, Long Beach, CA, USA, Jun. 2019. Article (CrossRef Link)
[14] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. of 2017 IEEE International Conference on Computer Vision (ICCV)*, pp.2980-2988, Venice, Italy, Oct. 2017. Article (CrossRef Link)
[15] A. Kirillov, Y. Wu, K. He, and R. Girshick, "PointRend: Image Segmentation As Rendering," in *Proc. of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.9796-9805, Seattle, WA, USA, Jun. 2020. Article (CrossRef Link)
[16] Z. Tian, B. Zhang, H. Chen, and C. Shen, "Instance and Panoptic Segmentation Using Conditional Convolutions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol.45, no.1, pp.669-680, 2023. Article (CrossRef Link)
[17] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT++ Better Real-Time Instance Segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol.44, no.2, pp.1108-1121, 2022. Article (CrossRef Link)
[18] S. Koles, S. Karakas, A. P. Ndigande, and S. Ozer, "Using Different Loss Functions with YOLACT++ for Real-Time Instance Segmentation," in *Proc. of 2023 46th International Conference on Telecommunications and Signal Processing (TSP)*, pp.264-267, Prague, Czech Republic, Jul. 2023. Article (CrossRef Link)
[19] J. Cai and Y. Li, "Realtime single-stage instance segmentation network based on anchors," *Comput. Electr. Eng.*, vol.95, 2021. Article (CrossRef Link)
[20] E. Xie, P. Sun, X. Song, W. Wang, X. Liu, D. Liang, C. Shen, and P. Luo, "PolarMask: Single Shot Instance Segmentation With Polar Representation," in *Proc. of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.12190-12199, Seattle, WA, USA, Jun. 2020. Article (CrossRef Link)
[21] Y. Lee and J. Park, "CenterMask: Real-Time Anchor-Free Instance Segmentation," in *Proc. of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.13903-13912, Seattle, WA, USA, Jun. 2020. Article (CrossRef Link)
[22] H. Chen, K. Sun, Z. Tian, C. Shen, Y. Huang, and Y. Yan, "BlendMask: Top-Down Meets Bottom-Up for Instance Segmentation," in *Proc. of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.8570-8578, Seattle, WA, USA, Jun. 2020. Article (CrossRef Link)
[23] E. Xie, W. Wang, M. Ding, R. Zhang, and P. Luo, "PolarMask++: Enhanced Polar Representation for Single-Shot Instance Segmentation and Beyond," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol.44, no.9, pp.5385-5400, 2022. Article (CrossRef Link)
[24] X. Wang, T. Kong, C. Shen, Y. Jiang, and L. Li, "SOLO: Segmenting Objects by Locations," in *Proc. of 16th European Conference on Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, vol.12363, pp.649-665, Glasgow, UK, Dec. 2020. Article (CrossRef Link)
[25] K. Duan, L. Xie, H. Qi, S. Bai, Q. Huang, and Q. Tian, "Location-Sensitive Visual Recognition with Cross-IOU Loss," *arXiv preprint arXiv:2104.04899*, Apr. 2021. Article (CrossRef Link)
[26] X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen, "SOLOv2: Dynamic and fast instance segmentation," in *Proc. of NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, pp.17721-17732, Vancouver, BC, Canada, Dec. 2020. Article (CrossRef Link)
[27] Z. Cai and N. Vasconcelos, "Cascade R-CNN: High Quality Object Detection and Instance Segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol.43, no.5, pp.1483-1498, 2021. Article (CrossRef Link)

[28] R. Guo, D. Niu, L. Qu, and Z. Li, "SOTR: Segmenting Objects with Transformers," in *Proc. of 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp.7137-7146, Montreal, QC, Canada, Oct. 2021. Article (CrossRef Link)
[29] Y. Fang, S. Yang, X. Wang, Y. Li, C. Fang, Y. Shan, B. Feng, and W. Liu, "Instances as Queries," in *Proc. of 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp.6890-6899, Montreal, QC, Canada, Oct. 2021. Article (CrossRef Link)
[30] W. Zhang, J. Pang, K. Chen, and C. C. Loy, "K-Net: towards unified image segmentation," in *Proc. of NIPS'21: Proceedings of the 35th International Conference on Neural Information Processing Systems*, pp.10326-10338, Jun. 2024. Article (CrossRef Link)
[31] B. Cheng, A. G. Schwing, and A. Kirillov, "Per-Pixel Classification is Not All You Need for Semantic Segmentation," *arXiv preprint arXiv:2107.06278*, Oct. 2021. Article (CrossRef Link)
[32] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention Mask Transformer for Universal Image Segmentation," in *Proc. of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.1280-1289, New Orleans, LA, USA, Jun. 2022. Aricle (CrossRef Link)
[33] L. Ke, M. Danelljan, X. Li, Y.-W. Tai, C.-K. Tang, and F. Yu, "Mask Transfiner for High-Quality Instance Segmentation," in *Proc. of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.4402-4411, New Orleans, LA, USA, Jun. 2022. Article (CrossRef Link)
[34] J. He, P. Li, Y. Geng, and X. Xie, "FastInst: A Simple Query-Based Model for Real-Time Instance Segmentation," in *Proc. of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.23663-23672, Vancouver, BC, Canada, Jun. 2023. Article (CrossRef Link)
[35] Q. Wen, J. Yang, X. Yang, and K. Liang, "PatchDCT: Patch Refinement for High Quality Instance Segmentation," in *Proc. of Eleventh International Conference on Learning Representations (ICLR)*, pp.1-15, Kigali, Rwanda, Feb. 2023. Article (CrossRef Link)
[36] M. Hu, Y. Li, L. Fang, and S. Wang, "A^2-FPN: Attention Aggregation based Feature Pyramid Network for Instance Segmentation," in *Proc. of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.15338-15347, Nashville, TN, USA, Jun. 2021. Article (CrossRef Link)

Li Liu is a professor at School of Automation and Electrical Engineering, University of Science and Technology Beijing. Her research focuses on deep learning, image instance segmentation, and data analysis.

Yuqi Kong is a master candidate at Shunde Innovation School, University of Science and Technology Beijing. Her research focuses on deep learning, image instance segmentation, and object detection.