electronics

MDPI

Article

Improvement of Road Instance Segmentation Algorithm Based on the Modified Mask R-CNN

Chenxia Wan, Xianing Chang and Qinghui Zhang *

College of Information Science and Engineering, Henan University of Technology, Zhengzhou 450001, China; wanchenxia@haut.edu.cn (C.W.); 2023920196@stu.haut.edu.cn (X.C.)
* Correspondence: zqh131@haut.edu.cn

Abstract: Although the Mask region-based convolutional neural network (R-CNN) model possessed a dominant position for complex and variable road scene segmentation, some problems still existed, including insufficient feature expressive ability and low segmentation accuracy. To address these problems, a novel road scene segmentation algorithm based on the modified Mask R-CNN was proposed. The multi-scale backbone network, Res2Net, was utilized to replace the ResNet network, and aimed to improve the feature extraction capability. The soft non-maximum suppression algorithm with attenuation function (soft-NMS) was adopted to improve detection efficiency in the case of a higher overlap rate. The comparison analyses of partition accuracy for various models were performed on the adopted Cityscapes dataset. The results demonstrated that the modified Mask R-CNN effectively increased the segmentation accuracy, especially for small and highly overlapping objects. The adopted Res2Net and soft-NMS can effectively enhance the feature extraction and improve segmentation performance. The average accuracy of the modified Mask R-CNN model reached up to 0.321, and was 0.054 higher than Mask R-CNN. This work provides important guidance to design a more efficient road scene instance segmentation algorithm for further promoting the actual application in automatic driving systems.

Keywords: road scene segmentation; Modified Mask R-CNN; multi-scale backbone; soft non-maximum suppression algorithm; segmentation performance

Citation: Wan, C.; Chang, X.; Zhang, Q. Improvement of Road Instance Segmentation Algorithm Based on the Modified Mask R-CNN. *Electronics* **2023**, *12*, 4699. https://doi.org/10.3390/electronics12224699

Academic Editor: Petros Karvelis

Received: 24 September 2023
Revised: 11 November 2023
Accepted: 15 November 2023
Published: 18 November 2023

1. Introduction

With the increasing demand for intelligent transportation, autonomous driving, and internet of vehicles (IoV) technologies, deep learning has aroused widespread attention in the last few decades [1–4]. To improve the safety and reliability of intelligent vehicles, instance segmentation technology [5] for field-of-vision object detection has been introduced to obtain more accurate determination of shapes and attributes of traffic participants in the driving environment. Instance segmentation is a very challenging task that combines object detection [6–8] and semantic segmentation [9,10]. However, some problems currently exist in the segmentation task [11], including feature mapping inconsistencies with the actual read, insufficient feature expression, and poor segmentation performance, especially for small or occluded objects in the actual driving environment [12]. Robail [13] proposed an encoder–decoder model for semantic segmentation, which can operate under any lighting or weather conditions.

Mask R-CNN is the most adopted method for instance segmentation. It adds a mask branch [14,15] and feature pyramid network (FPN) [16] model, and uses ROIAlign [17] to reduce quantization loss and retain the accurate spatial position information, with the purpose of improving segmentation accuracy. However, the actual road scenes are complicated and changeable, and background lighting, as well as occlusions, may greatly affect the partition results. Based on Mask R-CNN dominance in the instance partition task, many scholars [18,19] have worked on improving Mask R-CNN. However, few researchers

have focused on considering small and occluded objects. To increase the accuracy of instance segmentation, Liu et al. [20] designed a PANet network based on Mask R-CNN that added a bottom-up path and used an adaptive feature pool to enrich ROI features, which improved the segmentation accuracy at the expense of time. Huang et al. [21] designed a mask score region model based on a convolutional neural network, which improved the quality of the mask segmentation. Hameed et al. [22] proposed a score-based mask edge improvement of Mask R-CNN to segment fruit and vegetable images in a supermarket environment. Bi et al. [23] proposed the information-enhanced Mask R-CNN, called IEMask R-CNN, which made significant gains of about 2.60%, 4.00%, and 3.17% over Mask R-CNN on MS COCO2017. In [24], a computer-vision-based Mask R-CNN associated with a modified residual network (ResNet) was employed as a hybrid method (Mask R-CNN-ResNet) for robust and accurate wood identification at the species level. Sahu et al. [25] proposed using ResNet101 as the backbone of Mask R-CNN to detect pedestrians more accurately in less time. However, low segmentation performance was still present, especially for small and occluded objects.

To address the existing problems of current approaches, a novel road scene segmentation algorithm based on the modified Mask R-CNN was designed. The improved backbone network was adopted for extracting feature maps. A new multi-scale backbone network, Res2Net [26], was utilized for detailed extraction of the effective feature. The Soft non-maximum suppression (Soft-NMS) algorithm with an attenuation function [14] was used to address the problem of occluded objects by filtering the candidate frames to reduce the missed detection rate of complex road scene objects, and to improve the mask prediction accuracy.

The structure was organized as follows: The original CenterMask model was analyzed, and the improved CenterMask model that utilized SCAG-Mask and regression strategy was designed in Section 2. The experimental configuration and data processing were conducted, and the subjective and objective evaluation analyses of segmentation performance were performed in Section 3. The conclusions are summarized in Section 4.

2. Models and Methods

2.1. Mask R-CNN Model

The Mask R-CNN model was adopted in this paper to detect and segment the multiple objects in the road scene. Figure 1 illustrates the overall structure of Mask R-CNN. It mainly includes the Backbone, Region Proposal Network (RPN), RoIAlign, and Output branch.

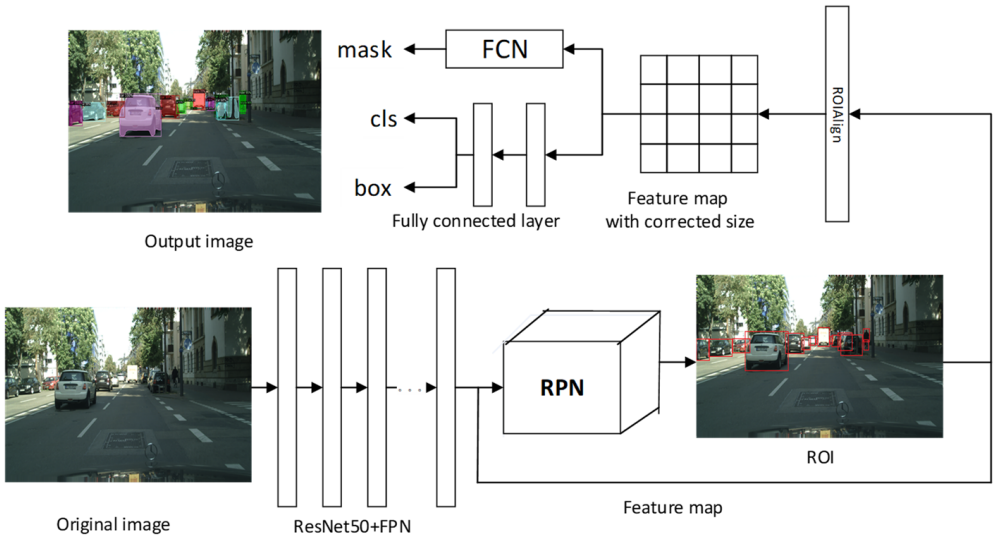


Figure 1. Mask R-CNN model.

(1) Backbone

The learning and extraction of image features are crucial in object detection and segmentation models, which determines the detection and segmentation accuracy. To obtain richer and more expressive features, the backbone network, Res2Net, was utilized. Res2Net was inspired by ResNet, which constructs the hierarchical residual class connection for obtaining more fine-grained features. The FPN was integrated with the backbone, which incorporated three channels and extracted feature information of various sizes.

(2) RPN

After the backbone, the obtained features were input into the RPN [5]. The RPN first generated the anchor boxes of different sizes and length-to-width ratios. The softmax method was adopted to determine the anchor boxes [27]. A bounding box regression was performed to obtain the initial regression, and then the regression prediction value was used to modify the anchor box, sort the foreground anchor box results according to the score, and use a non-maximum suppression algorithm to delete boxes below the threshold or with a high overlap rate.

(3) RoIAlign

The feature maps and the remaining anchor boxes in the RPN were sent to RoIAlign. RoIAlign removed the quantization process and utilized a bilinear interpolation algorithm to map the remaining anchor boxes in the RPN to the feature map, which effectively avoided the pixel mismatch problem and performed the pixel-level instance segmentation task.

(4) Output branch

The pixel-aligned feature map from RoIAlign was sent to two branches. The first was the mask branch network, which includes a full convolution network (FCN) [28] that is used to predict and generate an object segmentation mask. The second was the detection branch network, which includes fully connected layers for object classification.

2.2. Model Improvement

(1) Res2Net backbone network

Gao et al. [12] designed a new multi-scale backbone model called Res2Net. Res2Net can improve the feature extraction performance without increasing the computational load, which is very important in object detection, semantic segmentation, and instance segmentation tasks. Res2Net constructed the hierarchical residual class connection, extracted more detailed features, and improved the partition accuracy [22]. The structural diagram of Res2Net is shown in Figure 2. The left side depicts ResNet's basic CNN structure, and the right side represents Res2Net in Figure 2.

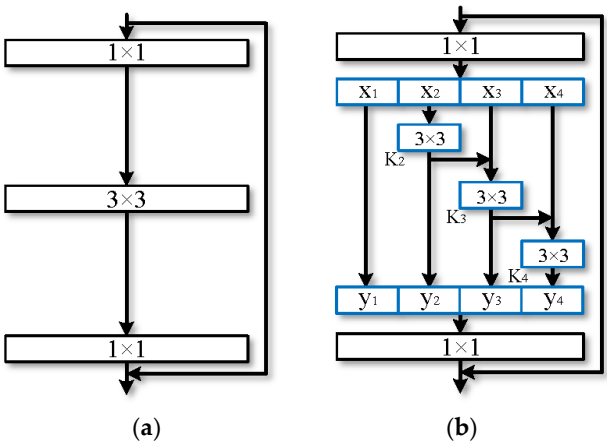


Figure 2. Comparison of (**a**) ResNet and (**b**) Res2Net.

Res2Net maintained a connection structure, which is similar to that of ResNet. Res2Net included the smaller filter banks connected in the form of hierarchical residuals and the increased number of scales for representing output features and extracting multiple scale features. Figure 2 demonstrates that Res2Net is convolved with the 1×1 block, and the features are divided into s subsets that are represented by x_i. Each subset feature x_i possessed the same spatial size, and the channel number was $1/s$. In addition to x_1, each x_i owned a 3×3 convolution, and is represented by $K_i()$. The output of $K_i()$ is represented by y_i, which equals sub-features x_i increased to the output of $K_{i-1}()$. Finally, all the blocks were spliced into another 1×1 convolution. The strategy of splicing after blocking enhanced the convolution and more efficiently processed features. To increase s and decrease the number of parameters, the 3×3 convolution kernel of x_1 was removed; this is shown as

$$y_i = \begin{cases} x_i & i = 1; \\ K(x_i) & i = 2; \\ K_i(x_i + y_{i-1}) & 2 < i \leq s. \end{cases} \tag{1}$$

Note that each 3×3 convolution operator $K_i()$ may obtain features from all feature splits $\{x_j, j \leq i\}$. The split is processed by using multiple scales in the Res2Net, which obtains the detailed features. A 1×1 convolution was used to connect all the splits to fuse the information from different scales. In this model, s is the parameter that controls the scale dimension, with larger values of s learning richer features. The calculation and memory overhead introduced by the cascading method is negligible. Therefore, using Res2Net instead of ResNet differs little in time while providing richer feature information. The input of rich feature maps to the subsequent network reduces the probability of false or missed detections, and thus improves the segmentation accuracy.

(2) Soft-NMS

The traditional NMS can easily miss detection for objects close to each other. Moreover, due to the complexity of the object scene in this paper, object vehicles and pedestrians in the road scene are easy to occlude. Therefore, it was necessary to improve the NMS algorithm [14]. The Soft-NMS algorithm was proposed in 2017, based on NMS. NMS was replaced by Soft-NMS in the model. The soft-NMS takes into account the possibility that there may be two objects in the overlapping part, and the direct deletion of the box may cause missing detection. Therefore, the Soft-NMS algorithm replaced the direct deletion of the box with an attenuation function, which has an important effect on improving the detection rate.

The rescoring function of NMS is written as

$$s_i = \begin{cases} s_i, iou(M, b_i) < N_t \\ 0, iou(M, b_i) \geq N_t \end{cases} \tag{2}$$

The hard threshold is utilized in NMS for obtaining the adjacent detection frame remains. The major disadvantage of NMS is that it makes the scores of adjacent detection frames zero. NMS cannot detect the overlapping area objects and decrease the performance.

Soft-NMS attenuates the detection score of adjacent detection frames that overlap with detection frame M. The more the detection frame overlaps with M, the more likely it is that false positive results could occur. In that case, the score attenuation is relatively larger. Therefore, the improved rescoring function of Soft-NMS is given as

$$s_i = \begin{cases} s_i & iou(M, b_i) < N_t \\ s_i(1 - iou(M, b_i)), iou(M, b_i) \geq N_t \end{cases} \tag{3}$$

When the overlapped value between the adjacent detection frame and M is higher than the overlap threshold N_t, the score of the detection frame is linearly attenuated. The detection frame near M is attenuated and the detection frame far from M is unaffected, which effectively improves the detection segmentation accuracy.

2.3. Dataset Preprocessing and Parameters Setting

(1) Dataset preprocessing

The pros and cons of the dataset selection occupy a very important position in the entire experimental process, which determines the upper limit of deep learning. To select a dataset that is closest to the complex road scenes in the real world, we first compared several commonly used road scene segmentation datasets in terms of region, number of labeled pictures, and data label density. Table 1 lists the comparative analyses of various datasets. The Cityscapes dataset only counts the labeling data.

Table 1. Comparative analyses of various datasets.

Dataset	Year	Area	Condition	Categories	Pictures	Label Density (%)
Cityscapes	2016	Germany	Season	30	5000	97.1
DUS	2013	Heidelberg	Day	5	500	63
KITTI	2012	Karlsruhe	Day	16	700	88.9
CamVid	2007	Cambridge	Day	32	700	96.2

As shown in Table 1, the fine labeling data in the Cityscapes dataset [29] had a high label density, and the labeling information was more delicate and had more annotation pixels. In addition, the Cityscapes dataset had a richer environment, a wider area, and a larger number of labeled pictures, which can provide better picture information that is closer to the real road scene. Therefore, we adopted the Cityscapes dataset in the following analyses.

The Cityscapes dataset used in this paper was obtained through the subsequent processing of the stereo video sequence of the on-board camera. While driving the vehicle, there will be blurred images caused by vehicle vibration and problematic images, such as excessively bright images when exposed to direct sunlight. Moreover, the number of pictures was relatively less, which was not conducive to training the instance segmentation model. Therefore, the dataset required preprocessing. First, the data was preprocessed to ensure that all data information was readable and operable; image sharpening operations were used to enhance the contrast of details in the picture and strengthen the edge line information. To make the contour of the vehicle edge clearer and improve the segmentation rates of small objects and blurred vehicles in the distance, the brightness of the over-exposed images was processed. Two data enhancement methods of horizontal flip and random cropping were adopted to expand the dataset. The diversity of the model reduced the overfitting and met the training requirements of the neural network model.

(2) Training method

When training a huge network model, many labeled data must be provided to avoid overfitting. For some actual problems, obtaining sufficiently large amounts of training data was difficult, and manually labeling the unlabeled data was both time-consuming and energy-intensive. It was wise to adopt a pre-trained model to initialize one's training model. It resulted in obtaining better training results with smaller datasets, reducing the data requirements and mitigating the bottleneck caused by insufficient data.

The loss function used is a cross entropy loss function based on label smoothing, which is expressed as:

$$L(x, y) = -\frac{1-\alpha}{k-1}\sum_{i=1}^{k} \widetilde{y}_i \log(\hat{y}_i) \tag{4}$$

where $\hat{y}_i$ represents the output of the softmax function, which is obtained by taking the logarithm; $\widetilde{y}$ represents the weighted mixed label, and is written as:

$$\widetilde{y} = -(1-\alpha)y + \frac{\alpha}{k} \tag{5}$$

where y represents the real label, α represents the smoothing parameter, the value is set to 0.1, and k is the number of categories.

Adam optimization is considered a first-order optimization algorithm that can replace the traditional stochastic gradient descent process, which has the powerful advantages of decreasing the computation resource and speeding up the model convergence. It can update the weight of the neural network iteratively based on the training data, and an important feature of its updating rules is to choose the step size carefully.

We adopted the deep learning framework Pytorch 1.7. A computer with a core i7-6700 CPU and a GTX 1080 graphics card was utilized. The network framework adopted in this paper included three sub-network structures, which had their different functions and cooperated. Therefore, a combination of step training and end-to-end training can be used during training.

The designed network structure was only improved in the feature extraction sub-network. To save training time, step-by-step training was first performed, and only the changed feature extraction part was trained. When the loss was stable, end-to-end training was then performed.

End-to-end training ignores intermediate stages and directly trains the entire network. After step-by-step training, the parameters were adjusted more comprehensively through end-to-end training, which can improve the overall fit, optimize the model as a whole, and get a better training effect.

The pre-trained backbone classification network of the ImageNet data is set as the basis and adopts the method of combining step-by-step training and end-to-end training. It relieved the pressure of insufficient data, so that the model could obtain better training results on small datasets. In addition, using this method greatly reduced the training time and improved the efficiency of the road scene segmentation model.

3. Experimental Results and Analyses

3.1. Subjective Evaluation Analyses

The Mask R-CNN model and the modified Mask R-CNN model in the subjective evaluation experiment were used to demonstrate the partition accuracy. The comparison analyses of the test results were illustrated in Figure 3. Therein, (a-1), (a-2), and (a-3) are the original images of the testing dataset; (b-1), (b-2), and (b-3) refer to the segmentation results of the Mask R-CNN model; (c-1), (c-2), and (c-3) represent the segmentation results of the modified model.

As shown in Figure 3(a-1), two heavily occluded objects were included: the vehicle behind the tree on the left and the vehicle in front of the road. In Figure 3(b-1), Mask R-CNN did not segment these objects, and the wire above the house on the right was mistakenly detected as a bicycle. In Figure 3(c-1), the modified method correctly segmented all the objects. In Figure 3(b-2), Mask R-CNN missed the multiple pedestrians behind the right side of the image. However, the proposed method correctly detected the several overlapping pedestrians in Figure 3(c-2). In Figure 3(a-3), three heads can be seen behind the car on the left side. Mask R-CNN only detected one person in Figure 3(b-3). The modified Mask R-CNN segmented all three persons in Figure 3(c-3). Therefore, the modified Mask R-CNN also demonstrated better contour information.

To demonstrate the partition accuracy of various models, Figure 4 illustrates the detailed detection results of Mask R-CNN and modified Mask R-CNN.

Figure 4 demonstrates that the modified Mask R-CNN can detect more vehicles and persons, which demonstrates that it could get better segmentation results over the Mask R-CNN in high overlap objects. The accuracy of the modified Mask R-CNN also improved, as shown in Figure 4b.

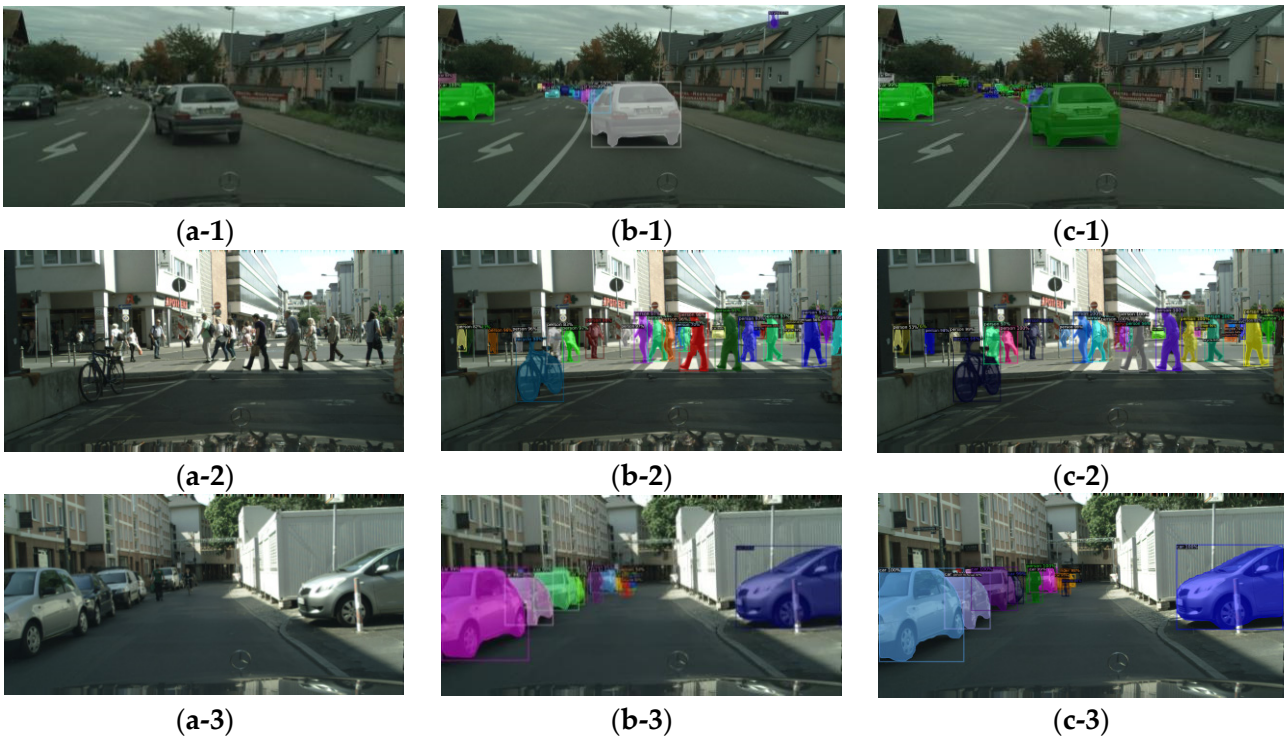
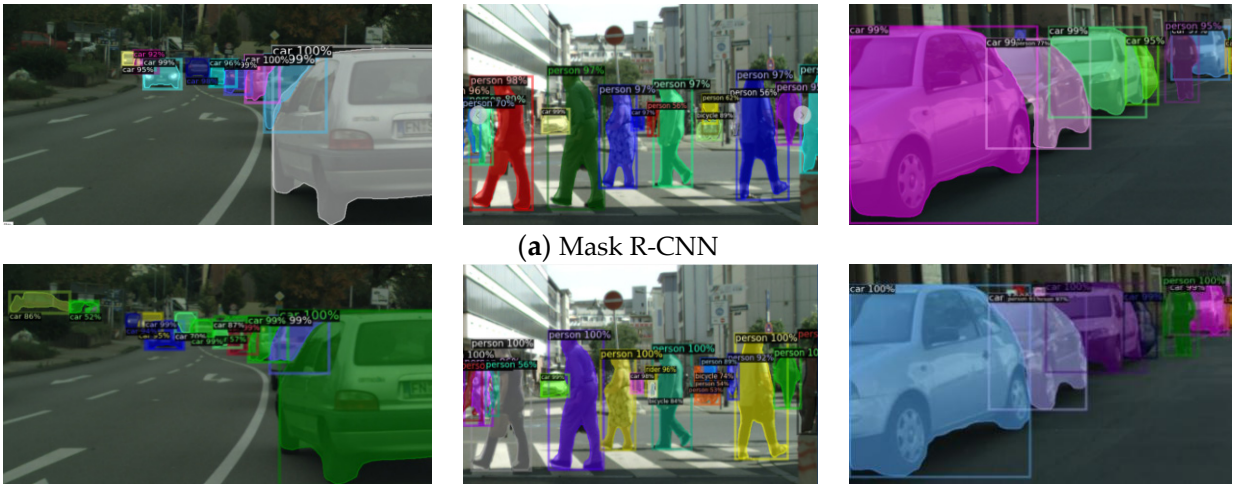


Figure 3. Prediction results of different models on the Cityscapes validation dataset: ((**a-1**)–(**a-3**)) Testing dataset; ((**b-1**)–(**b-3**)) Mask R-CNN; ((**c-1**)–(**c-3**)) Modified Mask R-CNN.



(**a**) Mask R-CNN

(**b**) Modified Mask R-CNN

Figure 4. Details segmentation results of various models.

3.2. Objective Evaluation Analyses

To demonstrate the training and learning process, Figure 5 illustrates the specific change process of the learning rate.

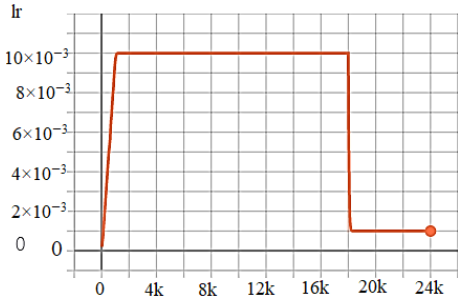
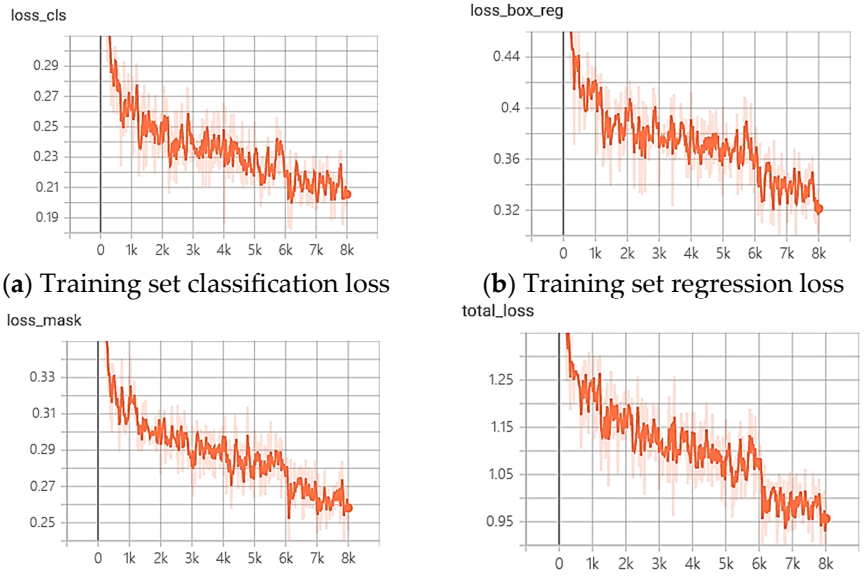


Figure 5. The process of learning rate change.

The initial learning rate was set as 0.01. Before 1000 steps, it reached the preset value through linear growth. At step 18,000, the learning rate dropped to 0.001.

The training process of the loss function in the optimization model is visualized with TensorBoard, as shown in Figure 6.



(**a**) Training set classification loss (**b**) Training set regression loss

(**c**) Training set mask segmentation loss (**d**) Training set loss

Figure 6. Loss function.

Figure 6 demonstrated that the loss function curve rapidly decreases during the training process. The classification loss, detection regression loss, mask loss, and total model loss on the training dataset were all in a convergent state, which indicates that the model is in multiple task branches of the training set. Therefore, both have a good learning effect.

To demonstrate the partition accuracy, the performance indicators of average precision (AP), accuracy (AC), recall (RE), and F1-score (F1) were adopted in this paper. The intersection of the union (IoU) required the calculation of AP. Therein, IoU refers to the intersection size between the object prediction area and the real tag area, or the ratio of the degree of intersection to the region of union. An evaluation AP by averaging the 10 thresholds range from 0.50 to 0.95 was obtained. AP_{50} indicates an AP value with an IoU threshold greater than 0.50.

To study the influence of different backbone networks on segmentation performance, three different backbone networks for feature extraction, including ResNet-50-FPN, ResNeXt-50-FPN, and Res2Net-50-FPN, were adopted. Table 2 lists the segmentation results of three different backbone network Mask R-CNN algorithms on the Cityscapes verification dataset.

Table 2. Segmentation performance of Mask R-CNN using different backbone networks.

Algorithm	Backbone	AP	AC	RE	F1	AP_{50}
Mask R-CNN	ResNet-50-FPN	0.267	0.713	0.697	0.681	0.525
Mask R-CNN	ResNeXt-50-FPN	0.281	0.749	0.738	0.713	0.541
Mask R-CNN	Res2Net-50-FPN	0.298	0.775	0.759	0.735	0.559

As shown in Table 2, different backbone networks had a significant influence on the segmentation performance of the Mask R-CNN model. The model with Res2Net-50-FPN had the best segmentation performance over others, which obtained the highest values of AP, AC, RE, F1, and AP_{50}. Compared with ResNet-50-FPN, ResNeXt-50-FPN uses a parallel stack of blocks to take the place of ResNet's three-layer convolution block. The Res2Net network provides more abundant extracted feature information with richer semantic features. This explains why the high AP and AP_{50} values can be obtained for Res2Net-50-FPN.

To evaluate the segmentation performance of each object, ResNet-50-FPN and Res2Net-50-FPN are used in this paper as backbone networks. The segmentation results on the Cityscapes verification set are listed in Table 3.

Table 3. ResNet-50-FPN and Res2Net-50-FPN segmentation performance.

Backbone Networks	ResNet-50-FPN		Res2Net-50-FPN	
Categories	**AP**	**AP_{50}**	**AP**	**AP_{50}**
person	0.342	0.670	0.359	0.692
rider	0.209	0.588	0.228	0.613
car	0.503	0.766	0.527	0.789
truck	0.210	0.324	0.263	0.380
bus	0.278	0.437	0.322	0.483
train	0.210	0.453	0.258	0.504
motorcycle	0.200	0.431	0.221	0.455
bicycle	0.187	0.532	0.206	0.556
Average	0.267	0.525	0.298	0.559

Table 3 showed that Mask R-CNN segmentation performance using the Res2Net-50-FPN backbone network was better than ResNet-50-FPN, and the AP values of all categories were higher than that. The algorithm using Res2Net had a significant improvement in the truck and train categories. The AP value was increased by 0.053 and 0.048, respectively, and the AP_{50} value was increased by 0.056 and 0.051, respectively; the small object person and rider categories also improved, and the AP value increased, respectively. In terms of the overall segmentation performance of several categories, the AP value of the Mask R-CNN algorithm using the Res2Net-50-FPN backbone network was larger than that using the ResNet-50-FPN backbone network. The model was 0.031, and the AP_{50} value was 0.034 higher than that. It shows that the Res2Net multi-scale backbone network can effectively extract objects of various sizes and enhance performance.

At the same time, to evaluate the effect of the effectiveness of Soft-NMS algorithm on the segmentation model, this section used the Soft-NMS algorithm to train and calculate the segmentation performance of Mask R-CNN with Res2Net-50-FPN in the Cityscapes dataset, as listed in Table 4.

Table 4. Segmentation results of NMS and Soft-NMS algorithms.

Backbone Networks	**Res2Net-50-FPN**		**Res2Net-50-FPN**	
Non-Maximum Suppression Algorithm	**NMS**		**Soft-NMS**	
Categories	**AP**	**AP_{50}**	**AP**	**AP_{50}**
person	0.359	0.692	0.390	0.725
rider	0.227	0.613	0.258	0.645
car	0.526	0.789	0.547	0.811
truck	0.263	0.380	0.275	0.393
bus	0.311	0.483	0.326	0.499
train	0.256	0.504	0.269	0.518
motorcycle	0.237	0.455	0.267	0.486
bicycle	0.205	0.556	0.236	0.589
Average	0.298	0.559	0.321	0.583

Table 4 demonstrated that the segmentation performance of the Mask R-CNN model using the Soft-NMS algorithm with attenuation function was slightly better than not using that algorithm with the function. In several categories with high occlusion overlap rates, the segmentation performance improved. The AP value of the person category increased by 0.031, AP_{50} rose to 0.033; the AP value of the rider category was 0.031, and the AP_{50} rose to 0.032; the AP value of the car category was 0.021, and the AP_{50} rose to 0.022; the AP value of the motorcycle category was 0.030, the AP_{50} rose to 0.031; the AP value of the bicycle category was 0.031, and the AP_{50} rose to 0.033. Therefore, the segmentation performance improved by adopting Soft-NMS.

In general, after the Mask R-CNN model using the Res2Net-50-FPN backbone network adopted the Soft-NMS algorithm with attenuation function, it can be seen from the test chart in the subjective evaluation and the data analysis in the objective evaluation. The segmentation performance was enhanced, especially in the case of severe occlusion. The smaller object can be segmented, which reduced the false detection rate and had a good segmentation effect.

To demonstrate the segmentation performance of the modified Mask R-CNN, Table 5 lists the comparison of other published algorithms on the Cityscapes dataset.

Table 5. Performance comparison of the modified and reported models.

Name	**AP**	**AP_{50}**
Mask R-CNN+Res2Net+Soft-NMS	0.321	0.583
Mask R-CNN+Res2Net	0.298	0.559
PANet [20]	0.318	0.571
Mask R-CNN	0.267	0.525

Table 5 demonstrated that AP value of the instance segmentation algorithm (Mask R-CNN+Res2Net) that only changes the backbone network was 0.031 higher than the original Mask R-CNN, and the AP_{50} value was 0.034 higher than the original Mask R-CNN; the modified model that changes the backbone network and adopts the maximum suppression algorithm (Mask R-CNN+Res2Net+Soft-NMS) was 0.321, which is better than PANet. On the whole, the Mask R-CNN instance segmentation algorithm that combines Res2Net and Soft-NMS had a certain competitiveness among the current advanced algorithms, which indicates that the modified Mask R-CNN possesses better segmentation performance. This work provided important guidance in improving instance segmentation performance and enhancing the actual applications in automatic driving systems.

4. Conclusions

An instance segmentation method for complex road scenes was proposed for effectively improving segmentation performance. A multiple scale backbone model, Res2Net, was introduced to extract more detailed features, and the Soft-NMS with an attenuation function was adopted to reduce the missed detection of the occluded objects. The results showed that the modified Mask R-CNN can correctly segment more occluded and small objects than Mask R-CNN in subjective evaluation analyses. The AP value of the adopted Res2Net algorithm increased to 0.031 higher than ResNet. The average accuracy of the modified Mask R-CNN reached 0.321, and was 0.054 higher than the Mask R-CNN. The modified model demonstrated the superior segmentation performance for many road objects. This work provided important guidance to design a more efficient road scene instance segmentation algorithm for further promoting the actual application in automatic driving systems.

Future research in complex road scene segmentation will focus on compressing the network parameters and reducing the running time to enhance generalization and robustness.

Author Contributions: Conceptualization, C.W.; Methodology, Q.Z.; Validation, C.W. and X.C.; Resources, X.C.; Data curation, C.W. and X.C.; Writing—original draft, C.W.; Writing—review & editing, Q.Z.; Supervision, Q.Z.; Funding acquisition, C.W. and Q.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was financially supported by the National Natural Science Foundation of China (Grant No. 62073123), and the High-Level Talent Research Start-up Fund Project of Henan University of Technology (2023BS040).

Data Availability Statement: Data are available on request to the authors.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yang, Z. A Novel Brain Image Segmentation Method Using an Improved 3D U-Net Model. *Sci. Program.* **2021**, *2021*, 4801077. [CrossRef]
2. Li, Y.; Du, T.; Zhu, L.; Qu, S. An Efficient Minimal Text Segmentation Method for URL Domain Names. *Sci. Program.* **2021**, *2021*, 9946729. [CrossRef]
3. Gona, A.; Subramoniam, M. Convolutional neural network with improved feature ranking for robust multi-modal biometric system. *Comput. Electr. Eng.* **2022**, *101*, 108096. [CrossRef]
4. Izadi, S.; Ahmadi, M.; Nikbazm, R. Network traffic classification using convolutional neural network and ant-lion optimization. *Comput. Electr. Eng.* **2022**, *101*, 108024. [CrossRef]
5. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
6. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
7. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving into High Quality Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
8. Chen, H.; Sun, K.; Tian, Z.; Shen, C.; Huang, Y.; Yan, Y. BlendMask: Top-Down Meets Bottom-Up for Instance Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Seattle, WA, USA, 13–19 June 2020; pp. 8570–8578.
9. Wang, Y.; Lee, D.; Heo, J.; Park, J. One-shot summary prototypical network toward accurate unpaved road semantic segmentation. *IEEE Signal Process. Lett.* **2021**, *28*, 1200–1204. [CrossRef]
10. Chen, X.; Girshick, R.; He, K.; Dollar, P. TensorMask: A Foundation for Dense Object Segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2061–2069.
11. Lee, Y.; Park, J. CenterMask: Real-Time Anchor-Free Instance Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, Seattle, WA, USA, 13–19 June 2020; pp. 13903–13912.
12. Gao, S.-H.; Cheng, M.-M.; Zhao, K.; Zhang, X.-Y.; Yang, M.-H.; Torr, P. Res2net: A new multi-scale backbone architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 652–662. [CrossRef] [PubMed]
13. Yasrab, R. ECRU: An Encoder-Decoder Based Convolution Neural Network (CNN) for Road-Scene Understanding. *J. Imaging* **2018**, *4*, 116. [CrossRef]

14. Bodla, N.; Singh, B.; Chellappa, R.; Davis, L.S. Soft-NMS—Improving Object Detection with One Line of Code. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), IEEE, Venice, Italy, 22–29 October 2017; pp. 5562–5570.
15. Nie, S.; Jiang, Z.; Zhang, H.; Cai, B.; Yao, Y. Inshore ship detection based on mask R-CNN. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, IEEE, Valencia, Spain, 22–27 July 2018; pp. 693–696.
16. Tayara, H.; Chong, K.T. Object detection in very high-resolution aerial images using one-stage densely connected feature pyramid network. *Sensors* **2018**, *18*, 3341. [CrossRef] [PubMed]
17. He, K.; Gkioxari, G.; Dollar, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), IEEE, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
18. Liao, B.; Xu, J.; Lv, J.; Zhou, S. An Image Retrieval Method for Binary Images Based on DBN and Softmax Classifier. *IETE Tech. Rev.* **2015**, *32*, 294–303. [CrossRef]
19. Yang, X.; Sun, H.; Sun, X.; Yan, M.; Guo, Z.; Fu, K. Position detection and direction prediction for arbitrary-oriented ships via multitask rotation region convolutional neural network. *IEEE Access* **2018**, *6*, 50839–50849. [CrossRef]
20. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
21. Huang, Z.; Huang, L.; Gong, Y.; Huang, C.; Wang, X. Mask Scoring R-CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Long Beach, CA, USA, 15–20 June 2019; pp. 6409–6418.
22. Hameed, K.; Chai, D.; Rassau, A. Score-based mask edge improvement of Mask-RCNN for segmentation of fruit and vegetables. *Expert Syst. Appl.* **2022**, *190*, 116205. [CrossRef]
23. Bi, X.; Hu, J.; Xiao, B.; Li, W.; Gao, X. IEMask R-CNN: Information-Enhanced Mask R-CNN. *IEEE Trans. Big Data* **2022**, *9*, 688–700. [CrossRef]
24. Bello, R.W. Wood Species Identification Using Mask RCNN-Residual Network. *Pro Ligno* **2023**, *19*, 41–51.
25. Sahu, S.; Sahu, S.P.; Dewangan, D.K. Pedestrian detection using ResNet-101 based Mask R-CNN. *AIP Conf. Proc.* **2023**, *2705*, 020008.
26. Li, B.; Yan, Q.-R.; Wang, Y.-F.; Yang, Y.-B.; Wang, Y.-H. A binary sampling Res2net reconstruction network for single-pixel imaging. *Rev. Sci. Instrum.* **2020**, *91*, 033709. [CrossRef] [PubMed]
27. Wang, F.; Cheng, J.; Liu, W.; Liu, H. Additive margin softmax for face verification. *IEEE Signal Process. Lett.* **2018**, *25*, 926–930. [CrossRef]
28. Fu, G.; Liu, C.; Zhou, R.; Sun, T.; Zhang, Q. Classification for high resolution remote sensing imagery using a fully convolutional network. *Remote Sens.* **2017**, *9*, 498. [CrossRef]
29. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The Cityscapes Dataset for Semantic Urban Scene Understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 3213–3223.