Fully densely linked and strongly correlated road scene instance segmentation

Hao Wang
Wuhan University of Technology
Ying Shi (✉ a_laly@163.com)
Wuhan University of Technology
Changjun Xie
Wuhan University of Technology
Chaojun Lin
Wuhan University of Technology
Hui Hou
Wuhan University of Technology
Jie Hua
Wuhan University of Technology

Research Article

Keywords: Road scene, Automated driving, Instance segmentation, Multi-scale Feature enhancement, Correlation extraction

Posted Date: September 19th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2073024/v1

Fully densely linked and strongly correlated road scene instance segmentation

Hao Wang[1] · Ying Shi[1] · Changjun Xie[1] · Chaojun Lin[1] · Hui Hou[1] · Jie Hua[1]

Received: date / Accepted: date

Abstract Unlike conventional indoor or outdoor scenes, road images in driverless scenes usually have the characteristics of high resolution, large width (indicating a large span and containing many targets), and variable background. As a mainstream algorithm for the instance segmentation task, Polar-Mask can balance segmentation accuracy and real-time performance to some extent. However, confronted with the road scene images, its feature extraction is inadequate, and the regression branch and the classification branch in its network structure are disjoint, ignoring the potential correlation between instance contour and instance category. To overcome this shortcoming, a Polar-Mask-based fully densely linked and strongly correlated instance segmentation network (FCSIS-Polar) is proposed. Specifically, the original cascaded convolutional layers in Polar-Mask are densely connected to enhance the feature extraction of the residual network. In addition, the category features are co-encoded with the original mask prediction results as a priori information to establish a contour-category correlation. The experiment results based on the Cityscapes dataset corroborate its performance, which can achieve a segmentation accuracy of 26.4% and a segmentation speed of 14.2 FPS even in small images.

This work was supported by National Natural Science Foundation of China (Grant numbers [52177110])

✉ Ying Shi
E-mail: a_laly@163.com

✉ Chaojun Lin
E-mail: 242906@whut.edu.cn

[1] the Department of Automation, the Wuhan University of Technology, Wuhan, China

Keywords Road scene · Automated driving · Instance segmentation · Multi-scale Feature enhancement · Correlation extraction

1 Introduction

As hardware-software systems that can execute dynamic driving tasks (DDT) on a sustainable basis [1], the widespread deployment of Automated Driving Systems (ADSs) can reduce the societal loss caused by erroneous human behavior such as distraction, driving under influence and speeding [2]. It is essential for ADSs to gather necessary data for safe navigation and perceive their surroundings with object detection, semantic segmentation, and instance segmentation [3].

Object detection can achieve target localization in road scenes and accomplish instance-level category prediction tasks [3]. However, many important objects, such as roadways, traffic lines, walkways, and buildings, are poorly delineated by bounding boxes. Semantic segmentation segments these objects through pixel-by-pixel category prediction and can better describe their contours [4]. However, it is difficult for autonomous vehicles to classify significantly different instances of the same category and then to determine the target-level safety. Profiting from the advantages of object detection and semantic segmentation, instance segmentation can discriminate objects with various trajectories and behaviors, providing precise contours of the targets in the vision field [5].

Early instance segmentation algorithms usually use the underlying information of pixel points in images, such as gray value or morphological characteristics. Although training is not required, but they have poor robustness and weak generalization ability and therefore cannot be applied to road images with complex backgrounds. A full convolutional network (FCN) was firstly proposed in [6] to achieve pixel-by-pixel category prediction by deconvolution and then the application of neural networks was extended from object detection to instance segmentation. Numerous outstanding instance segmentation algorithms have arisen since the advancement of deep learning techniques [7–12]. These algorithms may be broadly classified into two categories: two-stage and one-stage. Typically, these two types of algorithms employ binary classification to tackle the segmentation problem at the mask level. But, such pixel-to-pixel correspondence prediction is luxurious, especially in the single-shot methods [12].

To solve this problem, Xie et al. [12] proposed Polar-Mask, formulating instance segmentation as instance center classification and dense distance regression in a polar coordinate. Inspired by fully convolutional one-stage (FCOS) [13], Polar-Mask employs n ray distances to determine the centerness for instance segmentation. Its advantages are three-fold, 1) it realizes the integration of object detection and instance segmentation within the prediction branch. 2) it is easy to locate the points in polar coordinates and then to connect them to form a whole contour. 3) it is applicable to different detection frameworks with minor adjustments.

In this paper, the fundamental of Polar-Mask is investigated and two drawbacks of its application in road scenes are analyzed. The first drawback is its poor fit to the mask borders of different multi-scale potential targets. This is due to the high target overlap and huge scale span in road scenes, and its poor understanding of head structure on meaningful pixels in the feature map. The second one is that the contour and the category of any object has not been correlated. In order to benefit from the advantages of Polar-Mask and to properly solve these issues, a fully densely linked and strongly correlated instance segmentation (FCSIS-Polar) network is proposed. Our contributions are as follows. 1) a dense connection scheme is adopted to connect the cascaded convolutional layers in the head, so that the head structure can better capture the edge features of complex contour. 2) a correlation extraction strategy of ray distance based on category is proposed to augment its generalization ability.

The rest of this paper is organized as follows. First, related work on instance segmentation is briefly reviewed in Sect. 2. Then, the basic framework of Polar-Mask is introduced in Sect. 3, and its defects under road scenes are analyzed. Sect. 4 makes targeted improvements to address the defects proposed above. Sect. 5 verifies the effectiveness of the improved strategies, respectively; Sect. 6 concludes our work.

2 Related works

Detect then Segment Detect then Segment is often referred to as a two-stage algorithm and can be classified into top-down and bottom-up. The top-down instance segmentation algorithm locates its region through instance-by-instance target identification and then executes pixel-by-pixel semantic segmentation. Early segmentation algorithms, such as MNC [14], separately output the results of target detection and instance segmentation, and then combine them to produce the segmentation results. Faster R-CNN [15] integrates feature extraction, proposal extraction, bounding box regression, and classification all in one network. Its successor, Mask R-CNN [7] segments instances by concatenating FCN branches and classification confidence. Mask Scoring R-CNN [16] adds an additional Mask head to Mask R-CNN and replaces the confidence of instance classification with the intersection over union (IOU) ratio of predicted Mask and real Mask, so that the IoU of mask and ground-truth can be predicted. However, when dealing with multiple instances in road images, Mask Scoring R-CNN relies on target detection branches to generate a lot of proposed regions and is therefore significantly time-consuming. By clustering the pixels into each instance in an image, the bottom-up instance segmentation algorithm creates masks and circumvents the restriction of bounding boxes for the subsequent segmentation. Deep watershed transform [17] uses FCN networks to directly learn the energy of the watershed transform, and each energy region represents an instance. BAIS [18] predicts instances over semantic segmentation features via an extra instance mask prediction network to rectify errors during the object candidate generation process. SGN [19] splits instance segmentation tasks into multiple subtasks, forming a segmentation sub-network from points to lines and to components, and finally realizes the instance segmentation through a clustering network. All these algorithms are constrained within the “detect then segment” framework and their real-time performance is not suitable for automated driving.

Detect and Segment The representatives of Detect and Segment, e.g. SOLO [8], YOLACT [10], etc., are implemented by adding segmentation branches to one-stage object detection with a lightweight network topology and excellent real-time performance. SOLO [8] presents the concept of instance category and converts instance segmentation into a single-shot classification-solvable issue based on RetinaNet [20]. On the basis of YOLO [21–23], YOLACT [10] generates instance masks by predicting a set of instance-specific masks. These masks assign different coefficients to the common global mask features so that the RoI Align is not needed to generate local feature maps. Tensormask [9] uses a structured 4D tensor to represent the mask in the spatial domain and introduces the dense sliding-window paradigm for the instance segmentation. The aforementioned methods generally do not allow for direct target modeling, so it is difficult to optimize specific instances. Polar-Mask proposes a new ray modeling approach and uses the polar coordinate to model irregular targets, transforming the pixel-by-pixel instance segmentation into instance centroid categorization and ray distance regression. Two parallel branches are accordingly constructed to implement the classification and regression, but the potential correlation between the category and the mask is ignored.

3 Polar-Mask Instance Segmentation algorithms

This section further describes the Polar-Mask algorithm and analyzes its drawbacks when applied in automated driving. Polar-Mask is based on FCOS with parallel branches on head and uses four convolutional layers for classification and regression after feature mapping in the backbone network. In addition, a new contour modeling method is proposed, which uses n rays emanating from the center point to construct the target contour.

3.1 Feature extraction structure

Polar-Mask is built on FCOS and includes a backbone network [24], feature pyramid networks (FPN) [25], and two or more mission-specific heads, as is shown in Fig.1. Due to some inevitable overlaps between ground truth and bounding boxes, the full convolutional network (FCN) [6] introduces low recall and confusing data. FPN solves this problem with multi-scale prediction. The prediction in each scale corresponds to a head composed of two parallel sub-networks, respectively responsible for categorization and ray distance regression. The categorization sub-network is a tiny FCN whose parameters are identical for all scales. After obtaining an input feature map with C channels in a given scale layer, the categorization sub-network applies four 3×3 convolutional layers to form a cascaded convolutional layer for feature extraction. Then, a ReLU layer and a 3×3 convolutional layer are sequentially added to generate C category channels.

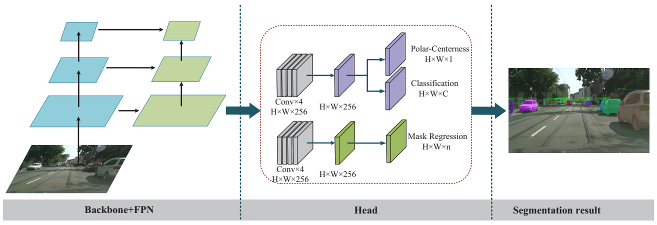


Fig. 1 The overall structure of the Polar-Mask

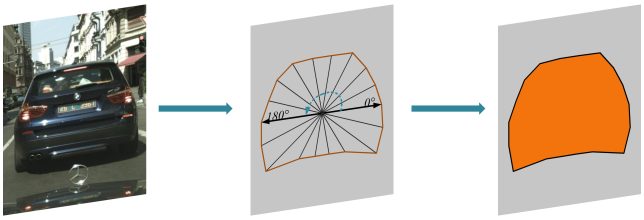


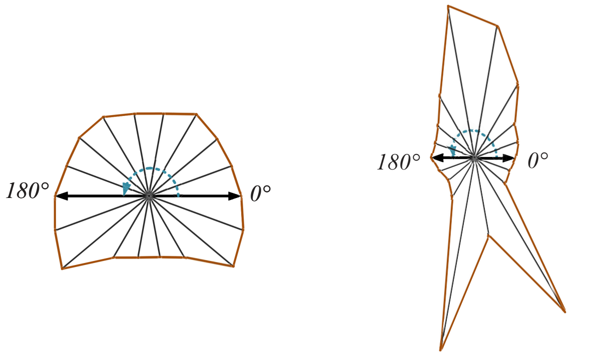
Fig. 2 The process of object profile modeling based on distance regression of rays

3.2 Ray distance regression method

The architecture of the regression sub-network is identical to that of the classification subnetwork, with the exception that it culminates in n linear outputs at each spatial location, which correspond to n ray distances. Polar-Mask utilizes the concept of border regression in FCOS to obtain a better contour description in the following way. Firstly, for the input image, the network predicts n rays distances in the polar coordinate after getting the center point. Next, the contour coordinates are calculated based on the angles and ray distances. The ray angles are distributed uniformly in [0°, 360°]. Finally, the predicted coordinates are sequentially connected to form the contour of the target and the pixel points distributed therein represent the instance segmentation results. The process of contour modeling based on ray distance regression is shown in Fig.2. Like the pixel-by-pixel detection of FCOS, the center point is the pixel point in the feature layer. In instance segmentation, the category of the center point and the distances of n rays are predicted pixel by pixel to achieve fast segmentation. It can be observed that a larger n renders a more accurate predicted contour. Considering that the original centerness of FCOS is unable to evaluate the quality of irregular target contours, Polar-Mask suggests using Polar-Centerness in the following manner. Another parallel branch is created in the head to forecast Polar-Centerness.

$$Polar-Centerness=\sqrt{\frac{\min\left(\{d_1,d_2,\ldots\ldots,d_n\}\right)}{\max\left(\{d_1,d_2,\ldots\ldots,d_n\}\right)}} \quad (1)$$

where $\{d_1, d_2, \ldots\ldots, d_n\}$ respectively denotes the distance of n rays. The closer d_{min} and d_{max} are, the



(a)Modeling of vehicle contours

(b)Modeling of pedestrian profiles

Fig. 3 Ray modeling of objects

higher the probability of the candidate point selected as the center point.

3.3 Analysis of Polar-Mask defects in road scenes

Polar-Mask proposes a new instance segmentation modeling approach and can achieve a good balance between segmentation accuracy and real-time performance to some extent. However, when applied to road images, it has the following shortcomings:

(1)Insufficient feature extraction: Unlike conventional indoor or outdoor scenes, the road images under driverless scenes usually have high resolution, large width, and variable background. Although the backbone network within its structure is dedicated to the feature extraction, it is still difficult for Polar-Mask to extract features from these images.

(2)Isolation between task branches: there is a strong correlation between instance contour and instance category. For example, the contours of vehicle and pedestrian targets in road scenes are significantly different. Unlike the typical instance segmentation strategy of pixel-by-pixel category prediction, the ray distance regression is more sensitive to the target category, because the distribution of ray distances varies greatly with the target class. When the number of rays n=20, the ray models of vehicle and pedestrian targets are shown in Fig.3. The ray distance distributions of vehicles and pedestrians are obviously different. For vehicle, the difference in the ray distance is small. On the contrary, in the pedestrian model, the ray distance tends to change periodically according to its angle. It can be inferred that the target contour is strongly correlated with its class, which is neglected by Polar-Mask. If the classification branch and the ray distance regression branch in its structure are linked, then its performance is expected to be promoted.

4 Our method

In this section, the proposed Polar-Mask-based Fully densely linked and strongly correlated instance Segmentation network (FCSIS-Polar) is detailed and the improvement strategies according to the shortcomings of Mask-Polar in road scenes are respectively elaborated.

4.1 Architecture

The proposed FCSIS-Polar structure is illustrated in Fig4. It uses ResNet-50 as the backbone network to extract multi-scale features in the road scene image through the FPN, and then predicts the target contour through the head. Due to the complexity of the road scene, it is difficult to extract distinct features for this head, which leads to a weak instance segmentation. To address this problem, a feature enhancement strategy based on fully densely connection (FD-FE) is proposed. The cascaded convolutional layers are densely connected to enhance the feature extraction and to facilitate the subsequent prediction.

The instance contours present the features of different categories. In opposition to Mask-Polar, the proposed method exploits the strong correlation between contours and categories via an encoder-decoder structure based on ray distance. To link the classification branch with the ray distance prediction branch, a correlation extraction strategy of ray distance based on category prior (CP-CE) is proposed. The category features are co-encoded as a priori information with the ray distance prediction, and are then decoded through a convolution operation. The predicted contours are believed to better represent the corresponding target categories.

4.2 Feature Enhancement Strategy of Road Scene Based on Fully Densely Connection

In Polar-Mask, convolutional layers are simply cascaded to extract features which are not delicate enough in the case of complex road scenes and bring about low accuracy in the subsequent segmentation. To alleviate the gradient disappearance problem caused by network deepening, ResNet [24] gives shortcut connections between the front and back layers to enhance the feature fusion between convolutional layers. Its successor, densely connected convolutional networks (DenseNet) [26], can further achieve deep feature fusion between them.

With this motivation, DenseNet is used to replace the cascaded convolutional layer in the head to realize a

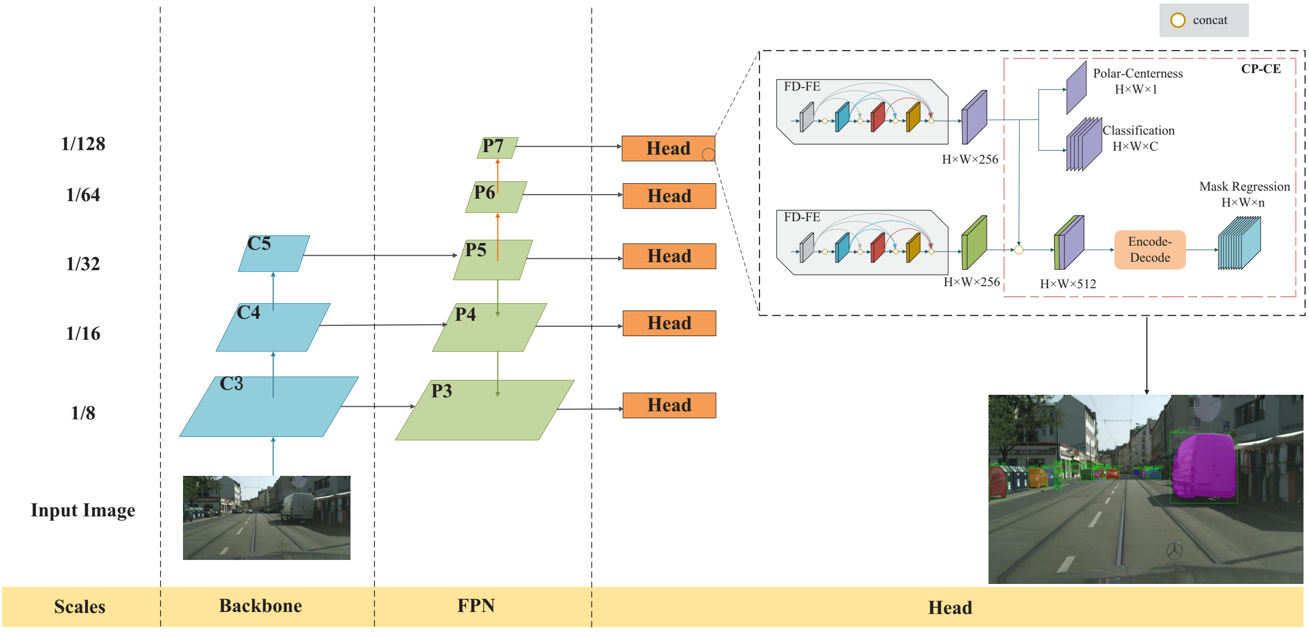


Fig. 4 The overall structure of the FCSIS-Polar

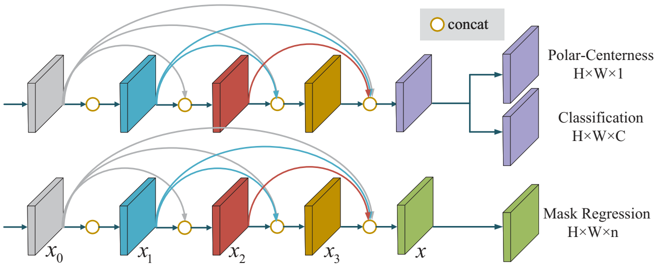


Fig. 5 Improved fully densely connected head structure

dense connection both in the classification branch and ray distance prediction branch, as is shown in Fig.5. Specifically, the features in each layer are fused with the features in all previous layers. The concat symbol represents the splicing of the features in the preceding layers. x_i represents the features in the densely connected convolutional layer and is calculated as

$$x_i = G\left([x_0, x_1, \ldots\ldots, x_{i-1}]\right) \tag{2}$$

where $[x_0, x_1, \ldots\ldots, x_{i-1}]$ denotes the splicing result of all features before the i-th layer. $G\left(\right)$ denotes the feature fusion and dimensionality reduction, respectively using 3×3 convolution and 1×1 convolution. This structure not only significantly enhances the feature extraction but also effectively alleviates the gradient disappearance problem.

4.3 Correlation Extraction Strategy of Ray Distance Based on Category Prior

When performing instance segmentation, if the category prior knowledge of the target is incorporated into the ray distance prediction, the optimization problem concerned with the ray distance regression can be further constrained. The proposed CP-CE strategy is based on this idea to improve the target contour prediction, as shown in Fig.6. Specifically, the connection of the classification branch is first added to the ray distance regression branch to splice the features used for category prediction with the ray distance regression features. Secondly, an encoder-decoder structure is added to receive the spliced features. The encoding module consists of a channel attention mechanism and a 3×3 convolutional layer. The first part weights the features of each branch and the second part encodes the number of feature channels from 512 to 1024. The channel attention module in CBAM [27]is used to compress the spatial dimensionality of the feature map with maximum pooling and average pooling of the spliced features, respectively. Then, two spatial context information F_{max} and F_{avg} are generated and fed into a shared multi-layer perceptron. After its output feature is implemented with an element-wise summation and the sigmoid function, the channel attention feature map M_c is obtained. The decoding module consists of k 3×3 convolutional layers, it decodes the 1024-dimensional encoded features and completes the final ray distance prediction.

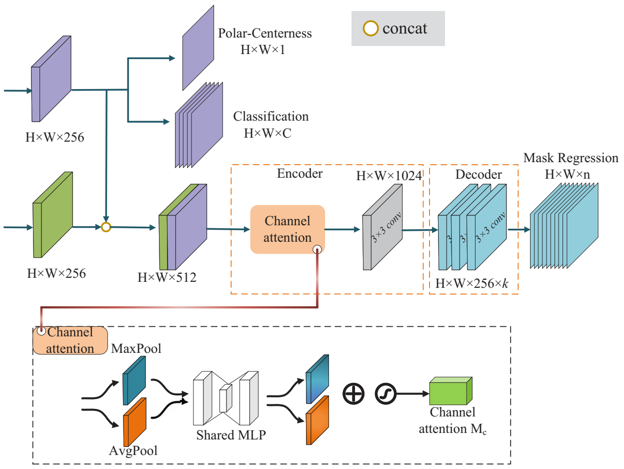


Fig. 6 The pipeline of CP-CE

5 Experiments and Discussion

5.1 Dataset and Evaluation

The current mainstream segmentation datasets are MS COCO [28], CamVid [29], Cityscapes [30], etc. MS COCO dataset has the largest number of images, which are mainly from daily life scenes but few from road scenes. The CamVid dataset is from city street scenes and contains pixel-level semantic annotation information, but does not contain instance-level object annotation. The Cityscapes dataset is also from city street scenes and contains a large amount of data with two kinds of annotation information. Therefore, Cityscapes dataset is chosen for the experiments.

The Cityscapes dataset contains road scene images from 50 regions, including 2975 images for training, 500 images for validation, and 1525 images as testing, with 2048×1024 resolution. This dataset contains 19 categories of pixel-level semantic annotation information and 8 categories of instance-level object annotation. In order to validate the instance segmentation algorithm, 8 categories with both pixel-level and instance-level annotation information are selected, namely car, pedestrian, truck, bus, rider, train, motorcycle, and bicycle.

Mask-mAP and Frame Per Second (FPS) are selected as the primary evaluation indexes to respectively measure the accuracy and the instance segmentation speed. In conformity to Microsoft COCO [28], by changing the IoU threshold varies from 0.5 to 0.95 with an increment of 0.05, the average precision (AP) for each IoU threshold is calculated, and after averaging, the instance segmentation accuracy Mask-mAP is obtained. AP is the ratio of the correct segmentation number to the total number of predicted segmentation. In addition, Mask-mAP50 for an overlap value of 50% are supplemented as a secondary evaluation index.

Table 1 The results of FD-FE strategy validation experiment

method	Mask-mAP(%)	FPS
Baseline	24.3	15.7
+FD-FE	25.1	14.5

5.2 Experimental Settings

(1) (1) Runtime environment: all the experiments are conducted in the Python environment, with an operating system of 64-bit Ubuntu 18.04 and a Pytorch1.5 deep learning framework; the hardware configuration is CPU Intel(R) i7-10700k; memory 32G; GPU NVIDIA GeForce RTX 2080Ti.

(2) Training details: the ResNet-50 based on ImageNet pre-training is used as the backbone. During training, all the images in the Cityscapes dataset are resized to 1280x768. In an iteration of 190400 times, a stochastic gradient descent (SGD) optimizer is used with the learning rate initially set to 0.00125 and reduced to 0.000125 at the 166600th iteration. The momentum of weight update and the batch size are respectively set to 0.9 and 0.1. The temporary models during the training are saved every 2975 iterations, and the model with the highest Mask-mAP value is selected for the following experiments.

5.3 Ablation Study

In this section, the effectiveness of FD-FE and CP-CE are respectively verified, then the ablation experiments on these two strategies are conducted.

(1) Verification of FD-FE strategy

Polar-Mask is used as the baseline algorithm and the FD-FE strategy is adopted, the comparison results are shown in Table 1.

As can be seen from Table 1, Mask-mAP is augmented by 0.8% with the FD-FE strategy. It is indicated that this strategy indeed reinforces the extraction of multi-scale features in the head. The instance segmentation speed is slightly decreased by 1.2 FPS due to an extra computation time consumed by dimensionality reduction.

Its positive effect on the segmentation accuracy can be seen from the comparison in Table 2. It can be seen that the segmentation performance is significantly improved with the FD-FE strategy for all 8 categories

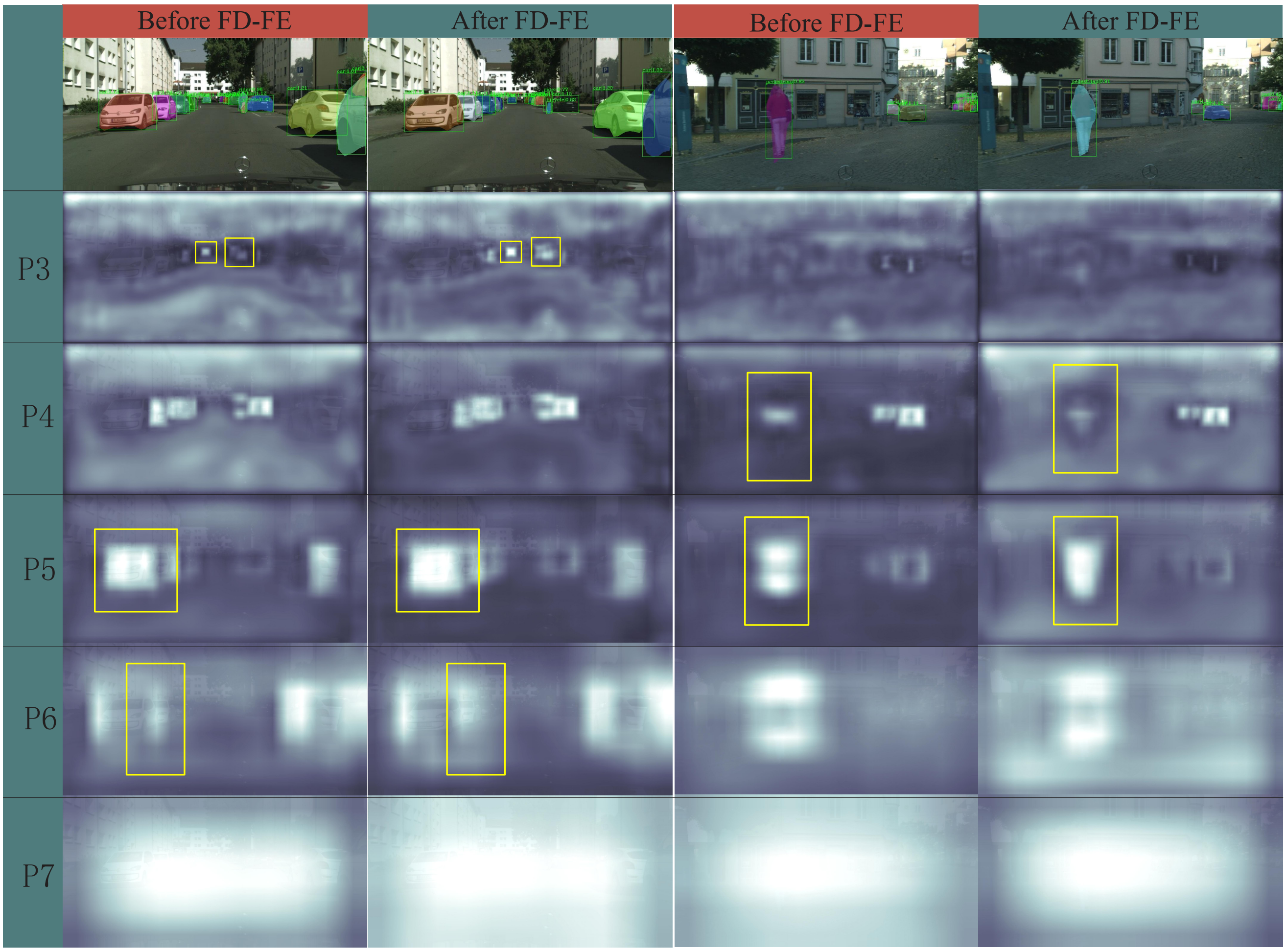


Fig. 7 Differences in features before and after applying FD-FE strategy

except for pedestrian and rider. One reason for this exception is that the rider contains the features of pedestrian and bicycle. Pedestrian tends to form local occlusion to bicycle, resulting in a partial loss of the rider. In addition, the features in each layer are superimposed with the features in its preceding layers. Therefore, the segmentation accuracy with respect to pedestrian and rider is lowered. Overall, the FD-FE strategy is effective.

The multi-scale features for category prediction are visualized and compared in Fig.7. By using the FD-FE strategy, the multi-scale features for prediction are more significant because the feature map corresponding to the ground truth object is fully activated. In addition, the features at different scales are all reinforced to some extent. For instance, for the farthest vehicles in the street, their feature maps in layer P3 are highlighted. Without this strategy, they are not adequately distinct. In layer P4, for the pedestrian in the urban road scene, before applying the FD-FE strategy, the feature map does not show her proper features, which are strengthened after applying the FD-FE strategy. In layer P5 and P6, the target features are also further activated while this strategy becomes less effective as the layer level increases.

(2) Validation of CP-EC strategy

Polar-Mask is used as the baseline algorithm and the CP-EC strategy is adopted, the comparison results are shown in Table 3. The encoder module consists of a channel attention mechanism and a 3×3 convolutional layer. The number of 3×3 convolutional layers in the decoder module is k. Its impact and the comparison results are shown in Table 4.

As can be seen from Table 3, Mask-mAP is augmented with the CP-EC strategy whatever the value of k. It is indicated that this strategy exploits the strong correlation between instance contour and instance category. On one hand, the network achieves the maximum Mask-mAP=25.7% when k=3. In comparison, the Mask-mAP of the baseline is 24.3%. On the other hand, the value of k has less impact on the instance segmentation speed. The number of feature channels in the

Table 2 Cityscapes instance segmentation class specific test set Mask-AP scores using metrics defined in Microsoft coco[28]

method	car	pedestrian	truck	bus	rider	train	motorcycl	bicycle
Baseline	42.9	22.1	25.6	42.7	14.8	22.6	9.9	13.7
+FD-FE	43.7	21.9	27.0	44.3	14.1	24.1	11.4	14.3

encode module is believed to impose a larger impact on the real-time performance.

Table 3 The effects of CP-EC strategy with different k values

method	Mask-mAP(%)	FPS
Baseline	24.3	15.7
+ CP-EC (k = 1)	24.9	15.3
+ CP-EC (k = 2)	25.5	15.3
+ CP-EC (k = 3)	25.7	15.3
+ CP-EC (k = 4)	25.7	15.2
+ CP-EC (k = 5)	25.6	15.2

(3) Ablation analysis on FD-FE and CP-EC

The above experiments respectively verify the effectiveness of FD-FE and CP-EC and the following ablation experiment on these two strategies is conducted. The results are shown in Table 4. Compared to using one strategy, the average increase in Mask-mAP when two strategies are adopted is 1.0%. On the contrary, the real-time performance is further degraded, because each strategy will draw in additional convolutional computations. The FD-FE strategy has a greater impact on the real-time performance than the CP-EC strategy. Overall, this loss is acceptable.

Table 4 Ablation analysis for each improvement strategy

FD-FE	CP-EC	Mask-mAP(%)	FPS
		24.3	15.7
✓		25.1	14.5
	✓	25.7	15.3
✓	✓	26.7	14.2

5.4 Performance Comparison

In this section, the proposed FCSIS-Polar is compared with some representative instance segmentation algorithms, and the experiment results are shown in Table 5.

FCSIS–Polar reaches the highest segmentation accuracy in smaller images and it is known that the smaller the image size, the lower the segmentation accuracy. Compared with the baseline, FCSIS-Polar respectively achieves 2.1% and 6.7% improvements in Mask-mAP and Mask-mAP50, by sacrificing only 1.5 FPS. Considering the safety brought by a higher segmentation accuracy, this price is acceptable. Even in smaller images, the Mask-mAP of FCSIS-Polar is about 50% higher than that of MNC and BAIS. Moreover, FCSIS-Polar outperforms YOLACT [10] and Mask R-CNN [7] in both the segmentation accuracy and the instance segmentation speed. Therefore, the proposed algorithm achieves a better balance between these two evaluation indexes and is more applicable to the instance segmentation in road scenes.

6 Conclusion

In this paper, a Polar-Mask-based fully densely linked and strongly correlated instance segmentation network is proposed for automated driving and two strategies are used to modify the network structure of Polar-Mask. To reinforces the extraction of multi-scale features in the case of complex road scenes, DenseNet is used to replace the cascaded convolutional layer to realize a dense connection both in the classification branch and ray distance prediction branch. In order to exploit the strong contour-category correlation, a connection between these two branches and an encoder-decoder structure are adopted to incorporate the category prior knowledge into the ray distance regression prediction. The effectiveness of these two strategies are respectively verified by the experiment results on the Cityscapes dataset. When they are simultaneously adopted in FCSIS-Polar, its Mask-mAP and segmentation speed respectively reaches 26.4% and 14.2 FPS. In comparison with other prevalent algorithms in instance segmentation, FCSIS-Polar achieves a better balance between the segmentation accuracy and the real-time performance which are both critical to automated driving.

Table 5 Comparison (%) of our FCSIS–Polar and other advanced methods about segmentation on Cityscapes

method	size	Mask-mAP(%)	Mask-mAP50(%)	FPS
MNC[14]	2048×102	15.6	30.0	2.8
BAIS[18]	2048×1024	17.4	36.7	1.3
DTW[17]	2048×1024	19.4	35.3	–
SGN[19]	2048×1024	25.0	44.9	–
YOLACT[10]	2048×1024	21.4	40.5	8.8
Mask R-CNN[7]	1280×800	26.2	49.9	6.7
Polar-Mask(Baseline)	1280×768	24.3	44.1	15.7
FCSIS-Polar(Ours)	1280×768	26.4	50.8	14.2

References

1. SHADRIN S.S., IVANOVA A.A.: Analytical review of standard Sae J3016,taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles with latest updates. Avtomobil'. Doroga. Infrastruktura. (3 (21)):10 (2019)
2. MONTGOMERY W.D., MUDGE R., GROSHEN E.L., et al.: America's Workforce and the Self-Driving Future: Realizing Productivity Gains and Spurring Economic Growth. https://avworkforce.secureenergy.org/(2018).Accessed 06 June 2018
3. WU X., SAHOO D., HOI S.C.: Recent advances in deep learning for object detection. Neurocomputing. 396:39-64 (2020)
4. ASGARI TAGHANAKI S., ABHISHEK K., COHEN J.P., et al.: Deep semantic segmentation of natural and medical images: a review. Artificial Intelligence Review. 54(1):137-178 (2021)
5. HAFIZ A.M., BHAT G.M.: A survey on instance segmentation: state of the art. A survey on instance segmentation: state of the art. 9(3):171-189 (2020)
6. LONG J., SHELHAMER E., DARRELL T.: Fully convolutional networks for semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition. 3431-3440 (2015)
7. HE K., GKIOXARI G., DOLLáR P., et al.: Mask R-CNN. IEEE Transactions on Pattern Analysis & Machine Intelligence. pp. 2961-2969 (2017)
8. WANG X., KONG T., SHEN C., et al.: Solo: Segmenting objects by locations. European Conference on Computer Vision. 649-665 (2020)
9. CHEN X., GIRSHICK R., HE K., et al.: Tensormask: A foundation for dense object segmentation. Proceedings of the IEEE/CVF international conference on computer vision. 2061-2069 (2019)
10. BOLYA D., ZHOU C., XIAO F., et al.: Yolact: Real-time instance segmentation. Proceedings of the IEEE/CVF international conference on computer vision. 9157-9166 (2019)
11. CHEN H., SUN K., TIAN Z., et al.: Blendmask: Top-down meets bottom-up for instance segmentation. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 8573-8581 (2020)
12. XIE E., SUN P., SONG X., et al.: PolarMask: Single Shot Instance Segmentation With Polar Representation. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12193-12202 (2020)
13. TIAN Z., SHEN C., CHEN H., et al.: Fcos: Fully convolutional one-stage object detection. Proceedings of the IEEE/CVF international conference on computer vision. pp. 9627-9636 (2019)
14. DAI J., HE K., SUN J..: Instance-aware semantic segmentation via multi-task network cascades. Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3150-3158 (2019)
15. REN S., HE K., GIRSHICK R., et al.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems. 28 (2015)
16. HUANG Z., HUANG L., GONG Y., et al.: Mask Scoring R-CNN. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6409-6418 (2019)
17. BAI M., URTASUN R.: Deep watershed transform for instance segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5221-5229 (2017)
18. HAYDER Z., HE X., SALZMANN M.: Boundary-aware instance segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5696-5704 (2017)

19. LIU S., JIA J., FIDLER S., et al.: Sgn: Sequential grouping networks for instance segmentation. Proceedings of the IEEE International Conference on Computer Vision. pp. 3496-3504 (2017)
20. LIN T-Y., GOYAL P., GIRSHICK R., et al.: Focal loss for dense object detection. Proceedings of the IEEE international conference on computer vision. pp. 2980-2988 (2017)
21. REDMON J., DIVVALA S., GIRSHICK R., et al.: You only look once: Unified, real-time object detection. Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779-788 (2016)
22. REDMON J., FARHADI A.: YOLO9000: better, faster, stronger. Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7263-7271 (2017)
23. REDMON J, FARHADI A.: Yolov3: An incremental improvement. Computer Vision and Pattern Recognition. (2018) https://doi.org/10.48550/arXiv.1804.02767
24. HE K., ZHANG X., REN S., et al.: Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770-778 (2016)
25. LIN T-Y., DOLLáR P., GIRSHICK R., et al.: Feature pyramid networks for object detection. Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117-2125 (2017)
26. HUANG G, LIU Z, VAN DER MAATEN L, et al.: Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700-4708 (2017)
27. WOO S., PARK J., LEE J-Y., et al.: Cbam: Convolutional block attention module. Proceedings of the European conference on computer vision (ECCV). pp. 3-19 (2018)
28. LIN T-Y., MAIRE M., BELONGIE S., et al.: Microsoft coco: Common objects in context. European conference on computer vision. pp. 740-755 (2014)
29. BROSTOW G J, FAUQUEUR J, CIPOLLA R.: Semantic object classes in video: A high-definition ground truth database. Pattern Recognition Letters. Pura Appl. 30(2):88-97 (2009)
30. CORDTS M, OMRAN M, RAMOS S, et al.: The cityscapes dataset for semantic urban scene understanding. Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213-3223 (2016)

Publisher's Note Springer Nature remains neutral with regard tojurisdictional claims in published maps and institutional affiliations.

Hao Wang received the B.S. degree in automation from the Wuhan University of Technology (WHUT), Hubei, China, in 2021, where he is currently pursuing the masters degree in Electronic Information. His research interests include computer vision and deep learning.

Ying Shi received the Ph.D. degree in marine engineering from the Wuhan University of Technology (WHUT), Hubei, China, in 2006. She is currently a Professor of artificial intelligence with WHUT. She has published over 40 articles. Her research interests include environment perception technology for safe driving assistance systems and unmanned systems, digital image processing, 3D point cloud data processing, big data analysis, machine learning, and deep learning.

Changjun Xie (Member, IEEE) received the Ph.D. degree in vehicle engineering from WHUT, Wuhan, Hubei, China, in 2009. From 2012 to 2013, he was a Visiting Scholar with UC Davis, Davis, CA, USA. He is currently a Professor with the School of Automation, WHUT. He has published over 50 articles, of which more than 40 are indexed by SCI or EI. His research interests include battery management systems, control strategies of intelligent and connected vehicles, and vehicle control and optimization of new energy vehicles.

Chaojun Lin received the M.A. degree in control science and engineering form Wuhan University of Technology, Hubei, China, where he is currently studying for Ph.D. degree.

Hui Hou received the Ph.D. degreein electric power system and automation form Huazhong University of Science and Technology, Hubei,China.She is a Professor of automation with WuhanUniversity of Technology, Hubei, China. Her current research interests includebig data systems, and machine learning.

JIE HUA received the M.A. degree in control science and engineering from the Wuhan University of Technology (WHUT), Hubei, China, in2022. He is currently pursuing the Ph.D. degree in control science and engineering from Wuhan University. His research interests include computer vision and deep learning.