Multi-scale Salient Instance Segmentation based on Encoder-Decoder

Houru Chen FRUITCHR@OUTLOOK.COM
Caijuan Shi SCJ-BLUE@163.COM
Wei Li LW@NCST.EDU.CN
Changyu Duan DUANCHANGYU@126.COM
Jinwei Yan YAN2696654544@163.COM
North China University of Science and Technology, Tangshan Hebei 063210, China

Editors: Vineeth N Balasubramanian and Ivor Tsang

Abstract

Salient instance segmentation refers to segmenting noticeable instance objects in images. In the face of multi-scale salient instances and overlapping instances, the existing salient instance segmentation methods have great limitations including inaccurate detection of large-scale instances, missing detection of small-scale instances, and wrong segmentation of overlapping in-stances. In order to solve these problems, a new multi-scale salient instance segmentation network (MSISNet) based on encoder-decoder is proposed. Firstly, we design a receptive field encoder (RFE), which adopts the serial dilated convolution instead of parallel dilated convolution and utilizes some common tricks to achieve better precision and speed. RFE can alleviate the problems of inaccurate detection of large-scale instances, missing detection of small-scale instances, and especially wrong segmentation of overlapping instances. Then, a pyramid decoder (PD) for the detection branch is designed to further alleviate the problem of inaccurate detection of large-scale instances and the difficulty in locating small-scale instances. Finally, a multi-stage decoder (MSD) is designed to improve the quality of the segmentation mask. In order to sufficiently evaluate the generalizability of our method, experiments are conducted not only on Salient Instance Segmentation-1K (SIS-1K) dataset, but also on Salient Objects in Clutter (SOC) dataset. The results show that the proposed method MSISNet is superior to the existing salient instance segmentation methods on $mAP^{0.5}$ and some recently proposed non-salient instance segmentation methods.

Keywords: Salient instance segmentation, Encoder-decoder, Multi-scale, Receptive field, Feature fusion.

1. Introduction

Visual saliency refers to focusing the most noticeable objects in the scene. Elazary et al. Elazary and Itti (2008) have researched on human visual system and have confirmed that the most attractive objects are more prominent in the visual system. Therefore, utilizing visual saliency to realize the object detection and instance segmentation, namely salient object detection (SOD) Wang et al. (2020a), Borji et al. (2019), Qin et al. (2019) and salient instance segmentation (SIS) Li et al. (2017a), Fan et al. (2019), can generate the salient maps only related to salient objects. This is in accord with the universal law of

human visual system Li et al. (2002). Although salient object detection has been widely studied in recent years and it can detect the most noticeable instances in the image scene, but it cannot distinguish the individual instances in the region. Therefore, salient instance segmentation has begun to get the attention and research, meanwhile, it has been applied to scene understanding of images and videos Anderson et al. (2018), assisted driving of intelligent vehicles Zeng et al. (2020), image media editing of humancomputer interaction Viazovetskyi et al. (2020), and robot perception system in industrial inspection Park et al. (2020).

The first work of salient instance segmentation MSRNet has been accomplished by Li et al. Li et al. (2017a) and it completes salient instance segmentation in a multi-stage way. However, MSRNet is timeconsuming and cannot meet the real-time application requirements. To alleviate the shortcoming of timeconsuming and improve the segmentation accuracy, S4Net Fan et al. (2019) has been proposed to segment instances by utilizing the relationship between instances and surrounding background. In the following work RDPNet Wu et al. (2021), regularized dense connections are used to optimize feature pyramid networks (FPN) Lin et al. (2017a) to further improve the performance of salient instance segmentation. Facing the multi-scale instances, the existing salient instance segmentation methods generally adopt FPN to realize the feature fusion. In addition, some work, such as PANet (Path Aggregation Network) Liu et al. (2018b) and TFPN Liang et al. (2019), combine the Residual Network (ResNet) He et al. (2016) with FPN to form ResNet-FPN encoder-decoder structure to accomplish multi-scale salient instance segmentation.

However, existing ResNet-FPN encoder-decoder structure still has great limitations: the bounding boxes of large-scale salient instances are not accurate enough, the small-scale salient instances are missing detected, and the overlapping salient instances are incorrectly segmented. The main reason is that the existing ResNet-FPN methods only pay attention to the information between adjacent scales but ignore the information between scales with large span. At the same time, FPN adopts the top-down connection, and each layer only contains the information of two consecutive layers so that the flow of information is limited. As a result, the high-level feature maps cannot access the spatial information in low-level features. Therefore, the boundaries of salient instances cannot be completely recovered leading to inaccurate detection of the large-scale instances. In addition, every feature fusion of FPN dilutes the information between nonadjacent scales, resulting in the lack of high-level semantic information in the low-level feature maps. This affects the location and segmentation of small-scale salient instances. What's more, there are overlapping among salient instances, so it is difficult for FPN to correctly generate bounding boxes and segment masks.

Therefore, in order to solve the above problems, a new Multi-scale Salient Instance Segmentation Network (MSISNet) based on encoder-decoder framework is designed in this paper. In this framework, one encoder and two decoders are designed, namely Receptive Field Encoder (RFE), Pyramid Decoder (PD) and Multi-Stage Decoder (MSD). RFE adopts the serial dilated convolution and some common tricks to effectively fuse multi-scale features for obtaining the receptive fields that cover salient in-stances of multiple scales. Thereby RFE can alleviate the problems of inaccurate detection of large-scale instances, missed detection of small-scale instances and especially wrong segmentation of overlapping instances. Different from the parallel arrangement of dilated convolution in RFBNet Liu

et al. (2018c), RFE adopts the serial arrangement of dilated convolution to enhance the feature map for alleviating the problem of hard segmentation caused by overlapping instances. PD obtains feature maps of different scales by decoding up-sampling and down-sampling with different magnifications. In addition, PD adopts the top-down connection and the skip connection to further alleviate the problems of inaccurate detection of large-scale instances and difficult location of small-scale instances. MSD utilizes feature maps of different stages to realize element-wise addition, so the spatial edge information can be added to the obtained salient feature maps for improving the quality of the segmented masks. In addition, in order to increase the robustness against the complex background and non-salient objects, an attention module Convolutional Block Attention Module (CBAM) Woo et al. (2018) is introduced into MSISNet for standardizing features, improving salient information and suppressing non-salient information.

The designed encoder-decoder framework MSISNet can improve the accuracy of salient instance segmentation while increasing the complexity of the network structure in a small extent. Experiments have been conducted on the current salient instance segmentation dataset SIS-1K and instance segmentation dataset SOC. The results indicate that the proposed MSISNet outperforms other salient in-stance segmentation methods and some recently proposed non-salient instance segmentation methods.

Generally speaking, the main contributions of this paper are as follows:

- A receptive field encoder RFE with the serial arrangement of dilated convolution is designed to alleviate the problems of inaccurate detection of large-scale instances, missing detection of small-scale instances, especially difficult segmentation of overlapping instances.
- A pyramid decoder PD is designed to further alleviate the problems of inaccurate detection of large-scale instances and difficult location of small-scale instances. A multi-stage decoder MSD is designed to improve the quality of the segmented masks.
- A multi-scale salient instance segmentation network MSISNet is proposed to effectively alleviate the problems faced by multi-scale instance segmentation. Experiments show that the proposed MSISNet has better performance than existing salient instance segmentation methods.

The rest of this paper is arranged as follows. The second section introduces the related work, the third section introduces the proposed MSISNet in detail, the fourth section is experiments and analysis, and the last section is conclusion.

2. Related work

2.1. Salient Object Detection

Traditional salient object detection methods often rely on features extracted manually Yang et al. (2013), Cheng et al. (2014), Qin et al. (2015), resulting in inefficient and low precision. In recent years, due to the good feature extraction performance of CNN, the SOD method based on deep learning has achieved satisfactory results. SOD based on deep learning avoids the time-consuming and cost-effective feature selection, and at the same

time, feature extraction is efficient and robust. Therefore, a large number of SOD methods based on deep learning have emerged, such as Wang et al. (2018), Liu et al. (2019), Chen et al. (2018), and Liu et al. (2018a).

2.2. Instance Segmentation

The original instance segmentation method is to add the segmented mask to the framework of object detection. Inspired by the two-stage object detection framework R-CNN Girshick (2015), Hariharan et al. have proposed a simultaneous detection and segmentation (SDS) Hariharan et al. (2014). With the continuous development of object detection algorithms, the performance of instance segmentation is also improved with the performance improvement of object detection. After the object detection framework Faster R-CNN Ren et al. (2015) has been put forward, He et al. have added a mask branch in Faster R-CNN and have proposed the instance segmentation model Mask R-CNN He et al. (2017), which has become the basic framework of many instance segmentation tasks. In 2018, Liu et al. have proposed the PANet model, which effectively fuses the information contained in different levels of features to improve the performance of instance segmentation. Wang et al. have considered the interaction between object detection and instance segmentation, and have proposed the RDSNet Wang et al. (2020b) model, which makes full use of the information interaction between object detection and instance segmentation to improve the instance segmentation performance. Li et al. have put forward the FCIS Li et al. (2017b) model by using location information to distinguish instance objects and improve instance segmentation efficiency.

3. Framework

In this paper, a multi-scale salient instance segmentation network MSISNet based on encoder-decoder is proposed and the receptive field encoder RFE, the pyramid decoder PD and the multi-stage decoder MSD are particularly designed in MSISNet. The proposed framework is described in detail below.

3.1. Receptive Field Encoder

In order to solve the problems of inaccurate detection of large-scale instances, missing detection of small-scale instances, and especially wrong segmentation of overlapping instances, inspired by RFBNet and ResNet residual blocks, a receptive field encoder RFE is designed in this paper. The structure of RFE is shown in Fig. 1. RFE enhances the receptive field of feature maps with residual block structure including dilated convolution by fusing multiple scale features at the same time. This encoding structure not only eliminates the information dilution problem in traditional FPN, but also alleviates the problems of inaccurate detection large-scale instances and missing detection of small-scale instances by superimposing receptive fields in feature maps. In addition, RFE adopts serial arrangement of dilated convolution to enhance the feature maps, so as to alleviate the problem that overlapping instances are difficult to be segmented. Specifically, in the first step, the multi-scale features $\{P_3\text{-}P_5\}$ generated by FPN are upsampled by bilinear interpolation at different magnifications and they are merged into a feature map through element-wise addition to form the

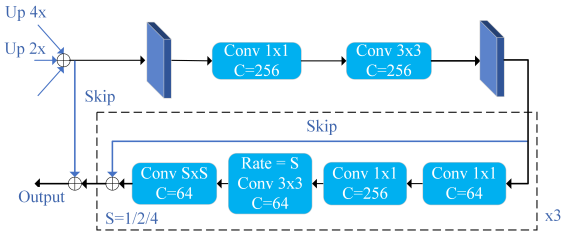

Figure 1: Detailed structures of the Receptive Field Encoder.

inputs of the receptive field encoder. Second, one 1x1 convolution layer and one 3x3 convolution layer same as those of FPN are adopted to process the input features for refining the context semantic information. Third, feature maps continuously pass through three groups of 3x3 dilated convolution with rate value S. At the same time, for dilated convolution with different rates, ordinary convolution with SxS convolution kernel is used in the previous layer to prevent pixel loss when the receptive field is expanded. In the last layer, 1x1 convolution is also used to recover the number of channels. Finally, skip connection is introduced into the receptive field encoder for adding the original inputs and the outputs of the residual block with dilated convolution, and then to generate a feature map containing multi-scale receptive fields for improving the detection rate of salient instances.

3.2. Decoder

Existing salient instance segmentation methods lack the design of the decoder in depth. This results in incomplete feature information contained in the decoded feature maps and unbalanced information among features in different scales. So it is difficult to correctly generate bounding boxes and complete mask segmentation among overlapping salient instances. Similarly, the feature maps used in the segment branch also lack low-level spatial information resulting in the incomplete mask segmentation map of salient instances. Therefore, a pyramid decoder PD is designed for the detection branch and a multi-stage decoder MSD is designed for the segmentation branch respectively in this paper. PD obtains the feature maps of different scales through the up-sampling and down-sampling with different magnification, and then the top-down connection and the skip connection are adopted to alleviate the problems of inaccurate detection of large-scale instances and difficult location of small-scale instances. MSD makes use of multi-stage inputs to complement the spatial edge information, so as to improve the quality of segmented masks.

3.2.1. Pyramid Decoder (PD)

In order to further alleviate the inaccurate detection of large-scale instances and the difficulty in locating small-scale instances, the pyramid decoder (PD) is designed by adopting the feature pyramid structure to produce a multi-scale feature map to predict the bounding box in the detection branch. The structure of PD is shown in Fig. 2. Through up-sampling and down-sampling, the outputs of RFE are resized into four-scale feature maps, each of which has multiple receptive fields to assist the detection branch correctly in detecting

multi-scale instances. Meanwhile, skip connection is used to add the feature maps {P_3, P_4, P_5} generated by FPN and the feature maps {N_3, N_4, N_5} decoded by PD to supplement the original information of salient instances. For large-scale salient instances, the feature maps with enhanced receptive fields can achieve more accurate frames. For large-scale salient instances, the feature maps with enhanced receptive fields can achieve more accurate frames. For small-scale salient instances, PD can provide small-scale feature maps with the low-level spatial features to solve the problem of missing detection.

In addition, because the feature map enhanced by RFE already contains enough instance information, the low-level scale layer N_2 is discarded when constructing the pyramid decoder, and only the high-level layers {N_3, N_4, N_5, N_6} is used for detection in order to balance the time and the accuracy. The specific operation is that the feature map output by

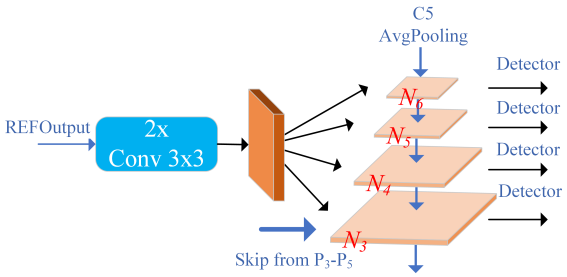

Figure 2: Detailed structures of the Pyramid Decoder.

RFE is firstly refined by using two 3x3 convolutions continuously, and then four scale feature maps {N_3, N_4, N_5, N_6} are generated by up-sampling and down-sampling with different magnifications in the detection branch. In order to further supplement the information, the C_5 feature map of the residual network is input to the decoder after AvgPooling. This feature map is processed by bilinear interpolation in descending order, and then it is merged and transferred after up-sampling. This idea is consistent with that of top to down feature transmission of PANet.

3.2.2. Multi-Stage Decoder (MSD)

In this paper, a multi-stage decoder (MSD) is designed for the segmentation branch and its structure is shown in Fig. 3. MSD adopts a multi-stage mode to utilize inputs of different stages to generate the feature maps including more low-level spatial information. Therefore, MSD can effectively improve the quality of the segmentation mask and then to achieve good segmentation results. The inputs of MSD include three parts: spatial edge inputs, RFE inputs and supplementary inputs. First, the spatial edge inputs are to double down-sample the feature map P_2 of FPN in Fig. 3 to obtain the feature map S_1. Then, S_1 adds the RFE inputs of RFE by element-wise addition to obtain the feature map S_2, which contains high-level semantic information and low-level spatial information to make salient instance segmentation more easily in the segmentation branch. At last, supplementary inputs, which come from the last layer of PD, namely N_3 in Fig. 2, and added with S_2 to obtain the feature map S_3. This multi-stage way can play a role similar to the skip connection in residual network to supplement the original information for the final feature map.

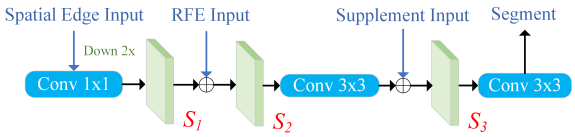

Figure 3: Detailed structures of the Multi-Stage Decoder.

The feature maps generated after fusion in each stage all adopt a 3x3 convolution to refine the semantic information. By multi-stage decoding and corresponding convolution operations, the generated feature maps can effectively improve the mask accuracy of salient instances.

3.3. Multi-scale Salient Instance Segmentation Network based on Encoder Decoder Framework (MSISNet)

The overall structure of the multi-scale salient instance segmentation network based on encoder-decoder MSISNet is shown in Fig. 4, and both the encoder and the decoders are encapsulated in a Mask R-CNN structure. The specific process of salient instance segmentation includes feature extraction, feature fusion, bounding box regression and mask segmentation.

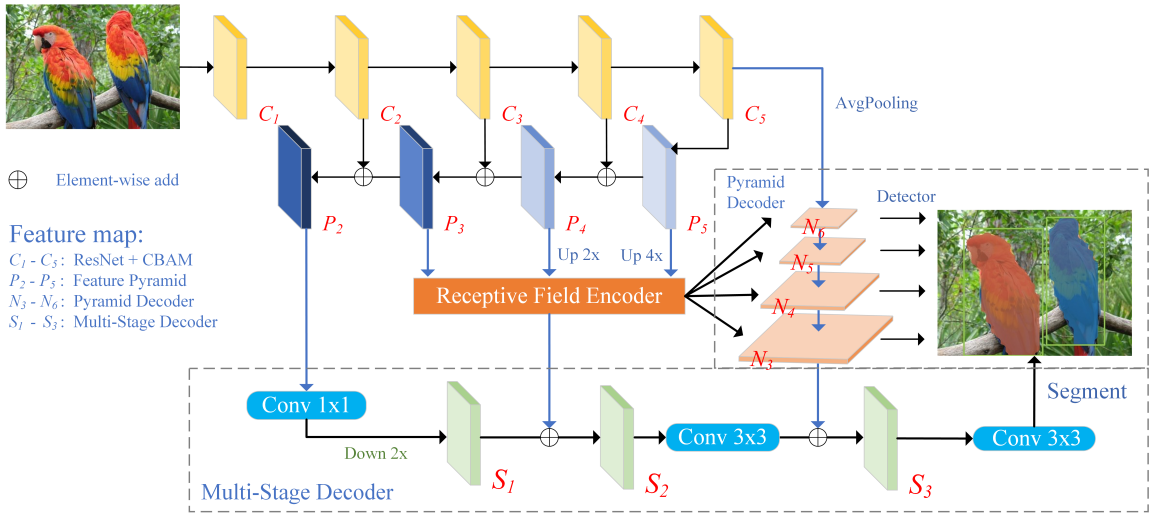

Figure 4: Overall structure of Multi-scale Salient Instance Segmentation Network.

3.3.1. Feature Extraction

The proposed MSISNet uses the basic ResNet as the encoder for feature extraction. In addition, in order to eliminate the interference of complex background and non-salient objects

in images, the plug-and-play conventional block attention module (CBAM) is embedded into the residual network to better standardize features and improve salient information.

3.3.2. Feature Fusion

Firstly, the FPN decoder is connected with ResNet embedded with CBAM to roughly fuse the features extracted from the residual network. Then, the receptive field encoder RFE designed in this paper is used to fuse multi-scale features. Specifically, the feature maps $\{P_3\text{-}P_5\}$ of three scales are fused first, and then the receptive fields of the feature maps are expanded by multiple scales and they are superimposed together through multiple dilated convolution and skip connection.

RFE, PD and ResNet-FPN are used as the inputs of the decoder to provide more effective feature map for bounding box regression and mask segmentation.

3.3.3. Bounding Box Regression and Mask Segmentation

In order to ensure that salient instances can be detected and segmented quickly, a single-stage RetinaNet Lin et al. (2017b) is used to replace the two-stage Faster R-CNN object detection network including the RPN regression in Mask R-CNN to ensure the real-time performance. The new encoder-decoder structure MSISNet proposed in this paper makes each level feature map generated by the detection branch have sufficient salient instance features. In addition, in order to further improve the speed, the P_2 feature map with the largest scale is abandoned and only the feature maps with smaller scales are used to save the calculation cost.

For the mask generation of salient instances, the decoder plays an important role in the one-stage segmentation circumstances of S4Net. The multi-stage decoder ensures that the spatial edge information of salient instances is retained in the feature map P3 used for segmentation to improve the mask accuracy.

4. Experiments

In this section, the proposed MSISNet is verified through experiments. Firstly, the datasets, evaluation metrics and related settings used in the experiments are introduced in section 4.1. The effectiveness of the proposed method MSISNet is verified in section 4.2. The comparison between MSISNet with existing salient instance segmentation methods and some instance segmentation methods is analyzed in Section 4.3.

4.1. Datasets, Evaluation Metrics and Experimental Settings

So far, there is only one proprietary salient instance segmentation dataset SIS-1K designed by Li et al. SIS-1K contains 1000 images with annotations of salient instance. In our experiments, 500 images are randomly selected as the training set, another 200 images are randomly selected as the validation set, and the rest 300 images as the testing set. The number of images is increased by flipping images horizontally in the training process of network model. In addition, in order to sufficiently evaluate the generalizability of our method and improve the convincingness of experimental results, one instance segmentation dataset, i.e. Salient Objects in Clutter (SOC) dataset Fan et al. (2018) are used in our

experiments. SOC dataset includes 6k images, which are divided into 3.6k/1.2k/1.2k for training/verification/testing respectively. The initial training learning rate of the network is set to 0.002 with a total of 40,000 iterations, while the learning rate becomes 0.0002 after 20,000 iterations. Weight decay and momentum are respectively set to 0.9 and 0.0001.

We follow MS COCO evaluation metrics Lin et al. (2014) and adopt mAP @ {0.5: 0.05: 0.95} as the main metrics, because it can better reflect the detection quality than the evaluation metrics in PASCAL VOC Everingham et al. (2010). $mAP^{0.5}$ and $mAP^{0.7}$ are also reported for reference. mAP represents the average accuracy of ten thresholds after IoU increases by 0.05 in turn from 0.5 to 0.95. $mAP^{0.5}$ represents the average accuracy at the threshold of intersection over union ratio of 0.5, and $mAP^{0.7}$ represents the average accuracy at the threshold of intersection over union ratio of 0.7. The larger the mAP value is, the stronger the performance of the algorithm is.

In this paper, the experimental configuration is Ubuntu 18.04, Tensorflow 1.15, Python 3.6 experimental environment with a single 1080Ti GPU.

In the training phase, the IoU is used to determine whether a bounding box proposal is a positive sample or a negative sample in the detection branch. If IoU is higher than 0.5, the bounding box proposal is positive; Otherwise, it's negative. For the detection of salient instances, the bounding box generated during training may not be enough to reach the threshold of the confidence score for NMS, so ground truth boxes are added to help the training process. We use SGD optimizer. The weight decay, the momentum, the input Batchsize, and the initial learning rate are respectively set to $1x10^{-4}$, 0.9, 256, and 0.002. The learning rate is divided by 10 after 10K iterations. A total of 30K iterations are executed. Due to the small batch size, all BatchNorm layers of the backbone network are frozen during training.

4.2. Ablation Studies

In this section, the effectiveness of the proposed MSISNet and the performance of each module are verified by ablation studies. If there is no special instruction in the tables, ResNet-50 is used as the basic backbone network for feature extraction.

4.2.1. RFE and PD Ablation Studies

In this experiment, the encoder RFE and the decoder PD in MSISNet are disassembled to verify their effectiveness respectively and the results are given in Table 1. Here RFE adopts the serial arrangement of dilated convolution and the operation of dimension reduction, while PD utilizes the top-down connection and the skip connection. It can be seen from

Table 1: Ablation studies of RFE and PD

Models	$mAP/mAP^{0.5}/mAP^{0.7}$	Time/s
Backbone	52.4%/87.1%/64.0%	0.025
+RFE	53.9%/89.6%/66.5%	0.028
+PD	54.9%/88.9%/67.6%	0.029
+RFE+PD	**55.5**%/**90.2**%/**68.6**%	0.029

Table 1: 1) when RFE and DP are introduced into the proposed MSISNet respectively, the accuracies of salient instance segmentation are improved by more than 1%. This indicates that both RFE and PD can improve the performance of salient instance segmentation. 2) When RFE and DP are introduced into the proposed MSISNet simultaneously, the accuracy of salient instance segmentation can be improved by more than 3% with only an extra time cost of about 0.003 second. This indicates RFE and PD are effective in salient instance segmentation.

4.2.2. The influence of the arrangement of dilated convolution in RFE

In this paper, RFE is designed inspired by RFBNet, but it adopts serial arrangement instead of parallel arrangement of dilated convolution. In order to show the influence of the arrangement of dilated convolution, this experiment is conducted and the experiment results are shown in Table 2.

From Table 2, we know that the serial arrangement has 1.1% improvement on the total mAP compared with the parallel arrangement of dilated convolution. No matter under the case of loose IoU=0.5 or strict IoU=0.7, RFE has an accuracy improvement of nearly 2% in serial mode. Although the improvement is less after setting IoU to 0.7, it also has an improvement of 1.5%. Therefore, it can be concluded that the serial arrangement of dilated convolution in RFE can improve the salient instance segmentation accuracy more effectively than the parallel arrangement.

Table 2: Influence of the Arrangement of Dilated Convolution in RFE

Mode	$mAP/mAP^{0.5}/mAP^{0.7}$	Time/s
Parallel	54.4%/88.4%/67.1%	0.025
Serial	**55.5**%/**90.2**%/**68.6**%	0.029

4.2.3. Influence of PD

The pyramid decoder PD in the detection branch is a four-level pyramid with top-down connection and horizontal connection similar to FPN. In addition, in order to supplement the original information in the feature maps, the features generated by the backbone network are connected to the PD through the skip connection.

In this section, experiments are designed to verify the effectiveness of PD and the results are shown in Table 3. The first row represents that the pyramid decoder designed in this paper is not included, and the feature map generated by RFE is directly sent to the detection branch. The second row represents only top-down connection is included and the third row represents only skip connection is included. The fourth row represents the proposed method with the pyramid decoder, which adopts both the top-down connection and the skip connection. It can be seen from Table 3 that: 1) compared with the simple connectionless mode, mAP has the 1.6% improvement only with the top-down connection. When FPN feature map is connected with the detection pyramid with the Skip connection, mAP is improved by 0.5% without additional time overhead. Therefore, the PD designed in this paper can effectively improve the salient instance segmentation performance.

Table 3: Influence of Pyramid Decoder

Models	$mAP/mAP^{0.5}/mAP^{0.7}$	Time/s
Without PD	52.8%/88.0%/64.6%	0.026
+Top to Down	54.4%/88.1%/66.0%	0.029
+Skip	53.1%/89.0%/65.5%	0.026
+PD(Top to Down+ Skip)	**54.9**%/**88.9**%/**67.6**%	0.029

4.3. Comparison Experiment of Salient Instance Segmentation

4.3.1. Quantitative comparison

In this section, the proposed MSISNet is compared with the existing salient instance segmentation methods, i.e., MSRNet, S4Net and RDPNet. At the same time, some recently proposed non-salient instance segmentation methods, including HTC Chen et al. (2019), CenterMask Lee and Park (2020), BlendMask Chen et al. (2020) and DetectoRS Qiao et al. (2021), have been compared with MSISNet here. Table 4 shows the quantitative comparison results of these four methods with $mAP/mAP^{0.5}/mAP^{0.7}$. At the same time, some recently proposed non-salient instance segmentation methods have been compared in the section of quantitative comparison. As can be seen from Table 4 that: 1) Whether compared with two-stage RDPNet or single-stage S4Net, MSISNet has the best performance at $mAP^{0.5}$ among these salient instance segmentation methods; 2) The proposed MSISNet has 3.1% improvement on $mAP/mAP^{0.5}$ and 4.6% improvement on strict $mAP^{0.7}$ compared to the single-stage benchmark S4Net; 3) Compared with these non-salient instance segmentation methods, the proposed MSISNet also achieves excellent performance. The good performance of MSISNet is mainly attributed to the receptive field encoder, pyramid decoder and multi-stage decoder designed for the proposed method MSISNet.

Table 4: Quantitative comparison of different methods.

Datasets		SIS-1K			SOC		
Evaluation Metrics		mAP	mAP0.5	mAP0.7	mAP	mAP0.5	mAP0.7
Non-salient method	HTC	45.4%	81.5%	55.9%	32.7%	57.6%	41.2%
	CenterMask	54.0%	87.2%	68.7%	23.8%	39.5%	29.9%
	BlendMask	53.6%	88.0%	67.4%	32.3%	56.2%	38.7%
	DetectoRS	50.4%	82.7%	63.7%	24.3%	49.1%	28.4%
salient method	MSRNet	-	65.3%	52.3%	-	-	-
	S4Net	52.4%	87.1%	64.0%	24.0%	51.8%	27.5%
	RDPNet	58.6%	88.9%	73.8%	37.7%	59.4%	48.4%
	MSISNet	55.5%	90.2%	68.6%	31.7%	62.7%	36.3%

4.3.2. Vision comparison

In order to intuitively show the results of salient instance segmentation, Fig. 5 shows the visual comparison between MSISNet and S4Net. Each row from top to bottom in Fig. 5 respectively represents the original image, the ground-truth, the results of S4Net and the

results of MSISNet. Column (a) and column (b) both contain a single large-scale salient instance, and the former has the simple background and the latter has the complex background. Column (c) and column (d) both contain salient instances with large-scale differences. The background of the column (c) is simple and the background of the column (d) is complex. In addition, the multiple scale instances contained in the column (e) and column (f) have overlapping in different degree. It can be clearly seen from Fig. 5 that the proposed

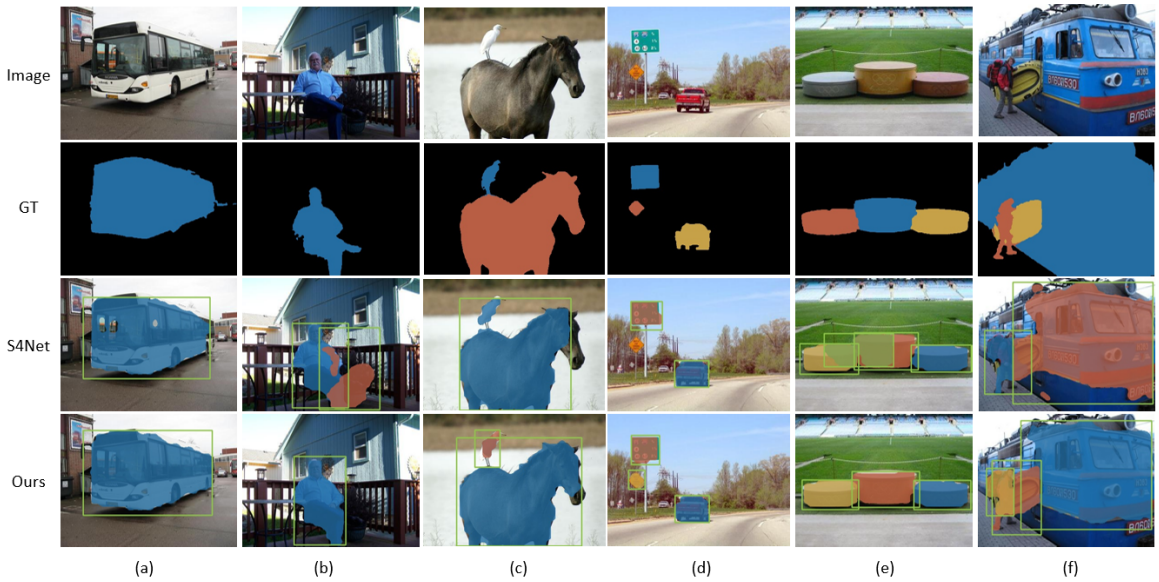

Figure 5: Vision comparison of the proposed MSISNet with S4Net.

MSISNet can better detect and segment salient instances than S4Net. 1) MSISNet can detect large-scale objects more accurate and generate the masks more comprehensive than S4Net. As shown in column (a), the bounding box produced by MSISNet can completely include the outline of the bus, and the mask does not have loopholes. Column (b) shows that the detection of MSISNet is not disturbed by complex background and MSISNet correctly identifies the old man's outline and generates the corresponding mask. This indicates MSISNet can well alleviate the problem of inaccurate detection of large-scale instances. 2) MSISNet can still work well in the case of different instance scales, even if the scales are small. Column (c) and column (d) show that MSISNet can detect and segment small-scale salient instances, such as birds and small road signs well. This indicates that MSISNet can overcome the problem of missing detection of small-scale instances. 3) In the case of overlapping of salient instances, the proposed method MSISNet can also successfully segment different instances when there is less overlapping of salient instances as shown in column (e). When the overlapping of instances is complicated, although the mask segmentation is inaccurate, MSISNet is able to segment three overlapping instances as shown in column (f). Therefore, the proposed method MSISNet can effectively alleviate the error segmentation problem caused by salient instance overlapping.

Through this experiment, it is intuitively and clearly verified that the proposed MSISNet has the best salient instance segmentation performance among the existing methods. This

is mainly attributed to the multi-scale salient instance segmentation network MSISNet, especially the receptive field encoder, the pyramid decoder and the multi-stage decoder.

5. Conclusions

In this paper, we propose a multi-scale salient instance segmentation network based on encoder-decoder MSISNet, which mainly includes the receptive field encoder, the pyramid decoder and multi-stage decoder are mainly designed in MSISNet. Extensive experiments have been conducted on the SIS dataset and SOC dataset,and the results show that MSISNet can effectively alleviate the problems of inaccurate detection of large-scale salient instances, missing detection of small-scale salient instances, and error segmentation caused by overlapping of salient instances, at the same time, MSISNet can achieve a balance between precision and speed. In the future, we will explore the more efficient network structure to improve the accuracy of salient instance segmentation, especially to overcome the overlapping problem. While seeking the generalization of the salient instance segmentation model, application in actual scenes is also our heading towards.

Acknowledgments

This research was supported by Postgraduate Demonstration Project in Hebei Province (KCJSX2019097); Excellent Youth Foundation of North China University of Science and Technology Scientific Committee (JQ201715).

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

Ali Borji, Ming-Ming Cheng, Qibin Hou, Huaizu Jiang, and Jia Li. Salient object detection: A survey. *Computational visual media*, 5(2):117–150, 2019.

Hao Chen, Kunyang Sun, Zhi Tian, Chunhua Shen, Yongming Huang, and Youliang Yan. Blendmask: Top-down meets bottom-up for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8573–8581, 2020.

Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4983, 2019.

Shuhan Chen, Xiuli Tan, Ben Wang, and Xuelong Hu. Reverse attention for salient object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 234–250, 2018.

Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE transactions on pattern analysis and machine intelligence*, 37(3):569–582, 2014.

Lior Elazary and Laurent Itti. Interesting objects are visually salient. *Journal of vision*, 8 (3):3–3, 2008.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

Deng-Ping Fan, Ming-Ming Cheng, Jiang-Jiang Liu, Shang-Hua Gao, Qibin Hou, and Ali Borji. Salient objects in clutter: Bringing salient object detection to the foreground. In *European Conference on Computer Vision (ECCV)*. Springer, 2018.

Ruochen Fan, Ming-Ming Cheng, Qibin Hou, Tai-Jiang Mu, Jingdong Wang, and Shi-Min Hu. S4net: Single stage salient-instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6103–6112, 2019.

Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *European conference on computer vision*, pages 297–312. Springer, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

Youngwan Lee and Jongyoul Park. Centermask: Real-time anchor-free instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13906–13915, 2020.

Fei Fei Li, Rufin VanRullen, Christof Koch, and Pietro Perona. Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences*, 99(14):9596–9601, 2002.

Guanbin Li, Yuan Xie, Liang Lin, and Yizhou Yu. Instance-level salient object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2386–2395, 2017a.

Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2359–2367, 2017b.

Yi Liang, Wang Changjian, Li Fangzhao, Peng Yuxing, Lv Qin, Yuan Yuan, and Huang Zhen. Tfpn: twin feature pyramid networks for object detection. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1702–1707. IEEE, 2019.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017a.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017b.

Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. A simple pooling-based design for real-time salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3917–3926, 2019.

Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3089–3098, 2018a.

Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768, 2018b.

Songtao Liu, Di Huang, et al. Receptive field block net for accurate and fast object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 385–400, 2018c.

Dongwon Park, Yonghyeok Seo, Dongju Shin, Jaesik Choi, and Se Young Chun. A single multi-task deep neural network with post-processing for object detection with reasoning and robotic grasp detection. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7300–7306. IEEE, 2020.

Siyuan Qiao, Liang-Chieh Chen, and Alan Yuille. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10213–10224, 2021.

Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7479–7489, 2019.

Yao Qin, Huchuan Lu, Yiqun Xu, and He Wang. Saliency detection via cellular automata. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 110–119, 2015.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015.

Yuri Viazovetskyi, Vladimir Ivashkin, and Evgeny Kashin. Stylegan2 distillation for feed-forward image manipulation. In *European Conference on Computer Vision*, pages 170–186. Springer, 2020.

Bo Wang, Quan Chen, Min Zhou, Zhiqiang Zhang, Xiaogang Jin, and Kun Gai. Progressive feature polishing network for salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12128–12135, 2020a.

Linzhao Wang, Lijun Wang, Huchuan Lu, Pingping Zhang, and Xiang Ruan. Salient object detection with recurrent fully convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1734–1746, 2018.

Shaoru Wang, Yongchao Gong, Junliang Xing, Lichao Huang, Chang Huang, and Weiming Hu. Rdsnet: A new deep architecture forreciprocal object detection and instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12208–12215, 2020b.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

Yu-Huan Wu, Yun Liu, Le Zhang, Wang Gao, and Ming-Ming Cheng. Regularized densely-connected pyramid network for salient instance segmentation. *IEEE Transactions on Image Processing*, 30:3897–3907, 2021.

Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3166–3173, 2013.

Wenyuan Zeng, Shenlong Wang, Renjie Liao, Yun Chen, Bin Yang, and Raquel Urtasun. Dsdnet: Deep structured self-driving network. In *European conference on computer vision*, pages 156–172. Springer, 2020.