International Journal of Innovative Computing, Information and Control
Volume **17**, Number **3**, June **2021**

ICIC International ©2021 ISSN 1349-4198
pp. **1041–1053**

A SURVEY ON INSTANCE SEGMENTATION: RECENT ADVANCES AND CHALLENGES

Huiyan Zhang[1], Hao Sun[2,*], Wengang Ao[1] and Georgi Dimirovski[3]

[1]National Research Base of Intelligent Manufacturing Service
Chongqing Technology and Business University
No. 19, Xuefu Avenue, Nan'an District, Chongqing 400067, P. R. China
{ huiyanzhang; aowg }@ctbu.edu.cn

[2]School of Computer Science and Technology
Harbin Institute of Technology
HIT Campus of University Town of Shenzhen, Shenzhen 518055, P. R. China
*Corresponding author: sunhaohit0@gmail.com

[3]Faculty of Electrical Engineering and Information Technologies
Ss. Cyril and Methodius University of Skopje
blvd. Goce Delcev 9, 1000 Skopje, Republic of North Macedonia
dimir@feit.ukim.edu.mk

Received December 2020; revised April 2021

Abstract. *Instance segmentation is the inevitable result of object detection and semantic segmentation in developing from coarse- to fine-grained inference, which is a popular research field in computer vision. Many technical methods have been published to solve the instance segmentation problems from the proposal (SDS) to the present (BlendMask) in a timeline manner. For this case, this paper surveys these methods and discusses the existing research difficulties and key challenges for instance segmentation. Finally, we conclude this paper by discussing some open challenges and the possible solutions.*
Keywords: Instance segmentation, Semantic segmentation, Object detection

1. **Introduction.** Object detection or localization refers to the progressive development process of digital image inference (i.e., from coarse- to fine-grained inference). It provides not only classification of image objects, but also the coordinate positions of the classified image objects in the form of border or center coordinates [1]. Semantic segmentation provides category inference by predicting each pixel tag in the input image, while the pixels are marked separately according to the category of the object where they are located [2]. With the further research, it is found that instance segmentation can provide different labels for individual instances belonging to the same category of objects. Therefore, it can be said that instance segmentation is a technology to solve problems in both object detection and semantic segmentation. In recent years, the update frequency of instance segmentation method is getting higher, but there has been no comprehensive and detailed research context of instance segmentation, which triggered the writing of this paper. This paper sorts out the key methods of instance segmentation in recent years. After seeing a clear context, it also gives possible research directions based on this context prediction, hoping to provide some reference for those who intend to study this direction.

Instance segmentation comes from semantic segmentation and object detection, as shown in Figure 1. Semantic segmentation refers to the separation of different types of objects in an image, e.g., dividing the sky, ground, cats, dogs and pedestrians into different

DOI: 10.24507/ijicic.17.03.1041

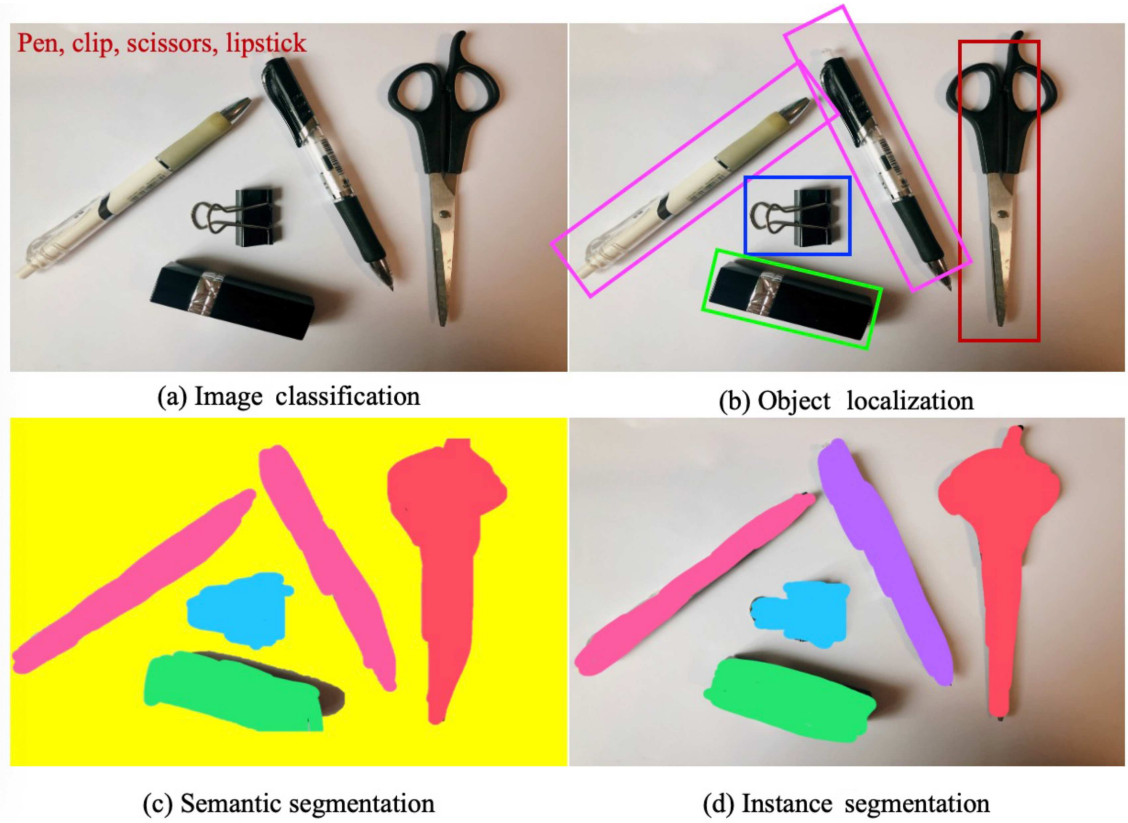


FIGURE 1. Object recognition evolution

areas, with a pixel level of accuracy [3]. However, each individual instance among multiple pedestrians is not distinguished [4]. The pixel level of accuracy means to distinguish each pixel point in the whole image into different objects. Object detection refers to finding out the objects defined by the current task in an image [5]. For example, when it needs to detect cats, dogs and people in an image, the detection results can only be an individual or a combination of cats, dogs and people, and there will be no objects out of the task such as pandas, monkeys and other tasks. Meanwhile, object detection also requires bounding box to frame the target objects separately, and provides the location information of these objects in the form of polar or center coordinates [6]. Instance segmentation is a method that further performs segmentation for different individuals belonging to the same class of objects on the basis of semantic segmentation. Compared with the bounding box for object detection, the profile of instance segmentation can be accurate to the edge of the instance. As opposed to semantic segmentation, instance segmentation can distinguish each individual. That is to say, instance segmentation is the inevitable result of object detection and semantic segmentation in developing from coarse- to fine-grained inference.

The organization of this survey is summarized as follows. Section 2 discusses some issues on instance segmentation: difficulties and challenges. Section 3 briefly overviews the evolution of instance segmentation. Section 4 discusses the instance segmentation method state and authors conclude this survey in Section 5.

2. **Issues on Instance Segmentation.** Instance segmentation is a popular research field in computer vision, which is used to predict the mask and the category for each object in a given image. According to the traditional definition, instance segmentation methods can be divided into two types: candidate-box-based approach and no-candidate-box-based approach. The former one is to perform coarse localization in advance with

the candidate box in an image, and then carry out the pixel level of mask segmentation to get object instance based on the objects framed [7]. A typical candidate-box-based instance segmentation method (Faster R-CNN) is shown in Figure 3(b). What should be mentioned is that, there are two shortcomings on this candidate-box-based approach: on the one hand, the image feature learning will be restricted by the range and position of candidate box due to the fact that the image mask is usually generated within the candidate area; on the other hand, the obtained mask has a coarser edge due to the fact that the lower mask bypass resolution. Therefore, enabling pixels with semantic affinity will have a significant influence on the segmentation results when performing an instance segmentation task. In addition, with the continuous improvement of object detection methods and effects, the results of instance segmentation have been further improved. However, in addition to instance object detection, how to accurately segment foreground and background pixels information in a given candidate box is still a direction worthy of challenge. Meanwhile, in addition to enabling pixels to be segmented with semantic affinity [8], the discussion on other approaches to better guide the execution of instance segmentation is also a direction of further research in the future.

While the latter one first adopts a multi-layer neural network to perform the pixel-level prediction for the instance object information, and then obtain the instance results through pixel clustering [9]. For this no-candidate-box-based segmentation approach, how to acquire the appropriate instance object information has always been the focus and difficulty for the research. In addition, in terms of data set of simple scenarios, this approach based on no-candidate-box has higher accuracy of segmentation than the former one based on candidate-box [10]. However, this approach is not only inferior to the former, but also has obvious disadvantages compared with other methods under a scenario with a large number of object categories and more complex data sets [11]. Therefore, how to learn more robust image features on data with limited space and complex objects is a problem well worth studying.

In addition, although convolutional neural network (CNN) [12, 13, 14] has been able to meet the requirements of most instance segmentation tasks, the effectiveness of network training is an unavoidable problem when the same class of instance objects are infinitely adjacent [15]. Meanwhile, although the obvious non-smooth boundary contours of the mask generated due to the fact that the fixed resolution of image features can be optimized with an approach based on image fusion, the improvement effect is not obvious in more complex scenarios. Therefore, the development of a new optimization method for instance segmentation results is also a worthy direction of research. Moreover, the early research on instance segmentation generally adopted the fully-supervised training for segmentation [16]. With the development of the Internet and big data technology, it is easy to collect massive data; however, there will be a huge labelling workload and a sharply increasing labelling cost, making the data labelling very difficult. Hence, how to train the model on data sets with limited or even no image labelling and propose a semi-supervised or unsupervised training approach to generate better instance segmentation results is a very challenging task.

On the other hand, with the continuous updating of visual information acquisition equipment, more and more data models are developed. How to obtain other key information using smart camera, depth camera, multi-view camera and other new data acquisition methods other than pixels to provide necessary and beneficial help for the current segmentation algorithm, is also a research direction with potentially hot application scenarios. Meanwhile, the cross-modal instance segmentation method [17] has also attached more and more attention from researchers. In addition, the data distribution under synthetic data labelling that is relatively easy to be acquired is not the same as that in real scenario.

So how to perform network training with the data in different modes is an issue worth studying. Furthermore, in addition to synthesized data, the data in other models can also be used as the target of cross-modal instance segmentation research. In addition, 3D data is also a popular direction currently, causing instance segmentation based on point cloud and voxel having more profound significance under scenarios such as piloted driving [18].

3. **The Evolution of Instance Segmentation.** The instance segmentation method has developed rapidly for recent years, and the increasing accuracy and speed are the driving forces for development. At the same time, in the process of development, the definition of instance segmentation is also changing, sometimes even directly overturn the problem, redefine. This section will discuss the development of instance segmentation in detail from the earliest instance segmentation method, and provide reference for the researchers who prepare to study instance segmentation. Instance segmentation was developed from semantic segmentation and object detection in a finer manner. Its development can be summarized in Figure 2.

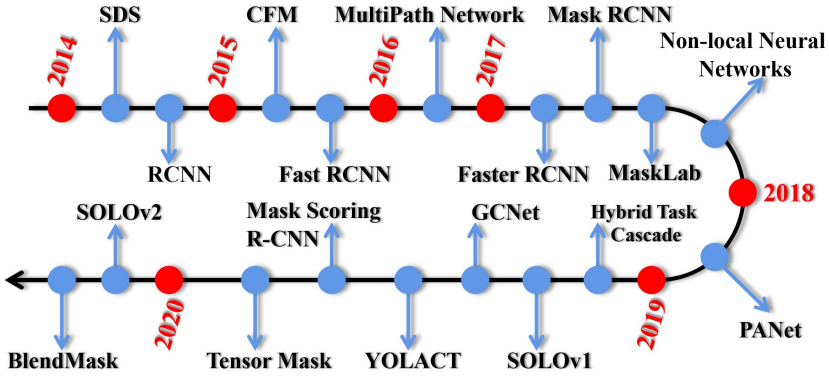


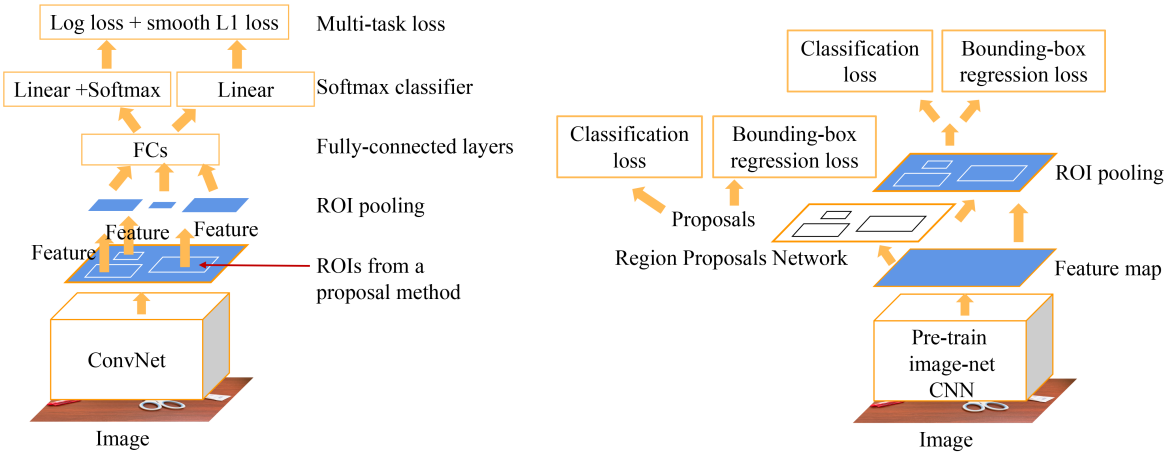
FIGURE 2. The key technology during the development process of instance segmentation

3.1. **SDS and R-CNN.** In 2014, Hariharan et al. proposed the earliest instance segmentation algorithm SDS [19] (simultaneous detection and segmentation) algorithm that realized the combination of object detection and semantic segmentation. Specifically, it realized coarse localization for instance objects by object detection and classification of pixels by semantic segmentation. Although there is a big gap between SDS algorithm and existing advanced algorithms in segmentation accuracy, it provides the most original idea for instance segmentation algorithm, and also lays a foundation for the subsequent research on instance segmentation [20]. Hariharan et al. then made improvements on the basis of SDS algorithm and proposed the Hyper Columns algorithm [21]. This algorithm can make finer inference to the details through fusion of low level features and high level ones, thus improving the classification accuracy.

Inspired by CNNs' breakthrough in image classification and the suggestion of selective search technology in manually generated feature areas, Girshick et al. first explored the influence of CNNs on instance segmentation in the same year. They developed the R-CNN [23] technology that integrated with AlexNet and applied the selective search solution. Despite the higher object detection quality, R-CNN also had some obvious defects. For example, the multi-stage training was slow and difficult, since it is necessary to conduct separate training for each stage. In addition, separate training on SVM classifier and BBox regression also required more resources and time. This led to the creation of improved detection frameworks such as Fast R-CNN [24] (in 2016) and Faster R-CNN [25] (in 2017).

3.2. **CFM and Fast R-CNN.** In 2015, Dai et al. proposed the CFM (convolutional feature masking) algorithm [26], which introduced the concept of mask into instance segmentation for the first time [20]. Image masking is a method of changing the image processing range by covering specified areas in an image with image blocks. The CFM algorithm generates image feature masks using bounding boxes, and fixed-size features in any area to facilitate the next segmentation. Dai et al. then proposed a new approach of instance segmentation-MNC (multi-task network cascades) [16]. Just as its name implies, MNC realizes multi-task cascading by sharing features, and integrates the three tasks of bounding box prediction for instance objects, instance mask segmentation and object classification into an end-to-end network framework for instance segmentation through cascading, thus making instance segmentation relatively efficient.

In 2015, Girshick proposed the object detection based on Fast R-CNN method [23], which takes an entire image and a set of object proposals as input [24]. In this method, a convolutional (conv) feature map is obtained by using several conv and max pooling layers for the entire image. Then, a fixed-length feature vector for each object proposal is extracted from the region of interest (ROI) pooling layer of the obtained feature map above. Finally, the obtained feature vector for each object proposal is fed into a set of fully connected layers, which consists of two output layers (one is to produce softmax probability and the other is to output the bounding-box positions). The model structure of Fast R-CNN is shown in Figure 3(a).



(a) The model structure of Fast R-CNN

(b) The model structure of Faster R-CNN

FIGURE 3. The model structure of Fast R-CNN and Faster R-CNN

3.3. **MultiPath Network.** In 2016, Facebook AI proposed the MultiPath Network method through three improvements including multiple layers of fusion in object detection, context linking, and object function integration, with better detection effects on COCO data sets. Compared with Fast R-CNN, it improved the experimental effect by about 66% and the effect in detection of small scale objects by about 4 times. This method is regarded as an extension of the Fast R-CNN, and the loss function of the Fast R-CNN method is given as the following:

$$L(p, k^*, t, t^*) = L_{cls}(p, k^*) + \lambda[k^* \geq 1]L_{loc}(t, t^*),$$

where p is the predicted probability of a certain kind, k^* is the real category, t is the predicted border position, t^* is the actual border position, L_{cls} and L_{loc} are the same defined in [24]. The first term on the right side represents the classification loss, and the second

term represents the position loss. Take the threshold value as the evaluation criterion of 50 as an example: when the coincidence degree between the detected border and the real bounding box is greater than 50, it indicates that the prediction is correct; otherwise, $k^* = 0$. However, the first item on the right side has the following disadvantages: all thresholds greater than 50 are equal, for example, 100% coincidence and 50% coincidence are both correct. The authors proposed an improved method, that is, a higher coincidence degree should have more scores, and the improved classification loss function is given as follows:

$$\int_{50}^{100} L_{cls}(p, k_u^*)du,$$

where k_u^* is the corresponding value of different threshold values. Since the formula above is a continuous integral, the authors use the sum of $du = 5$, and the modified objective function is

$$L(p, k^*, t, t^*) = \frac{1}{n}\sum_n [L_{cls}(p, k^*) + \lambda[k^* \geq 1]L_{loc}(t, t^*)].$$

3.4. **Faster R-CNN, Mask R-CNN, MaskLab and Non-local Neural Networks.** Faster R-CNN is a target detection algorithm proposed by He et al. in 2015 [25], which won many first prizes in ILSVRV and COCO contests in 2015. Based on Fast R-CNN, RPN candidate box generation algorithm is proposed in this algorithm, which greatly improves the target detection speed [25]. And the model structure of Faster R-CNN is given in Figure 3(b).

In 2017, He et al. proposed the Mask R-CNN [27] detection algorithm. Mask R-CNN is an instance segmentation algorithm most widely used with the highest efficiency in the current, as shown in Figure 4. By increasing branches of mask segmentation on the basis of Faster R-CNN, this algorithm not only can achieve excellent results of instance segmentation, but also is highly extensible for further application in other fields such as human body feature point detection. Despite being the best in the field of instance segmentation at that time, it is still not as good as semantic segmentation in segmentation accuracy [20]. In any event, as the authors demonstrate, this model outperformed the other most advanced models on every task in the COCO data set challenge in 2016. During the training, the multi-task loss on each sampling area of interest is defined as

$$L = L_{cls} + L_{box} + L_{mask},$$

where the classification loss L_{cls} and the boundary box loss L_{box} are defined the same as these in Fast R-CNN, and L_{mask} is the average binary cross entropy loss.

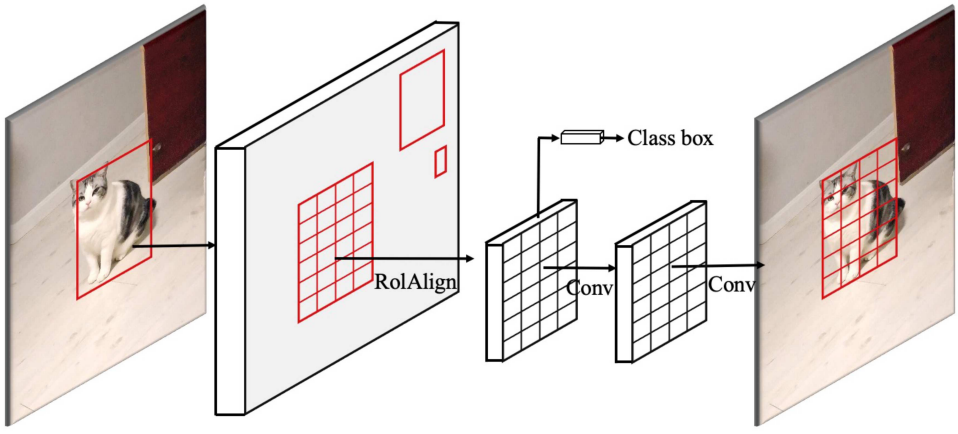


FIGURE 4. The Mask R-CNN framework for instance segmentation [27]

Furthermore, Chen et al. also improved Faster R-CNN and proposed the MaskLab [28] algorithm that generated two additional outputs, namely semantic segmentation and instance-centered direction. The predictive frame given by Faster R-CNN brought object instances with different scales into a standard scale for foreground and background segmentation by MaskLab in each predictive frame using semantic segmentation and direction prediction at the same time. To distinguish objects in different semantic categories, this algorithm adopted the prediction based on semantic segmentation to perform the pixel-level encoding for the classified data, thus eliminating the repeated background encoding.

Additionally, the Non-local Neural Networks [29] proposed by Wang et al. can be said to be the originator of Attention in images. It was the first to apply attention mechanism into the field of image, followed by the known SENet, SKNet, GCNet, Residual Attention Network, CAM, BAM, etc.

3.5. **PANet.** In 2018, the path aggregation network (PANet) [30] proposed by Liu et al. established a basic framework for instance segmentation tasks, aiming at improving the flow of information graphs. This method improves the feature hierarchy of the deep network by using specific localization-related signals at the bottom. This process of improvement is known as bottom-up path enhancement, which makes the information path between bottom features and top features of the deep network shorter. In addition, this method also proposed a technology called adaptive feature pooling that realized the connection between feature mesh and all hierarchical features. Thanks to this technology, reliable category suggestions will be generated as the relevant information stream with features at each level flows to the subsequent subnetworks. The framework of the PANet is illustrated in Figure 5.

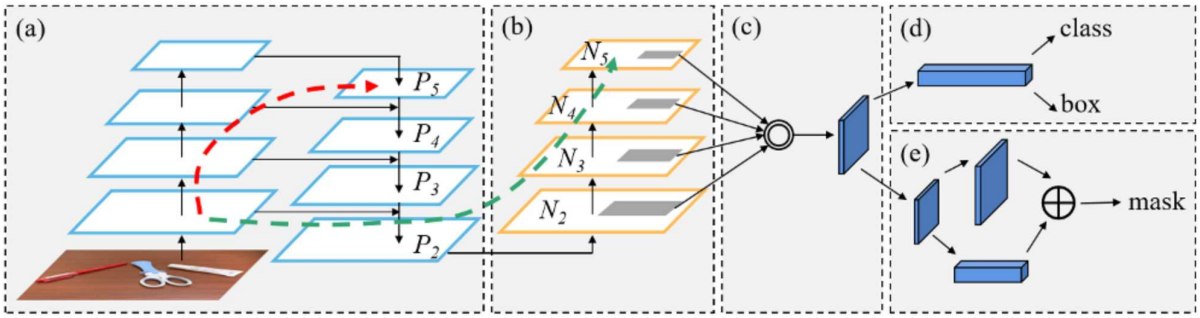


FIGURE 5. The framework of the PANet: (a) FPN backbone; (b) bottom-up path expansion; (c) adaptive feature pools; (d) box branches; (e) fusion of the full connection layer. For simplicity, the channel dimensions of the feature map are omitted in (a) and (b).

3.6. **HTC, GCNet, SOLOv1, YOLACT, Mask Scoring R-CNN and Tensor Mask.** In 2019, Chen et al. proposed the hybrid task cascade (HTC) [31] algorithm. The key for this algorithm to perform instance segmentation is to make the best use of the inverse relationship between object detection and object instance segmentation. In summary, HTC differs from traditional cascading in two important ways. First, HTC deals with object detection and object instance segmentation in multiple stages in a combination manner, instead of cascading them. Second, it uses a fully convolutional fragment to provide a spatial context to help distinguish foreground from background.

Moreover, Cao et al. also proposed a novel instance segmentation method, namely Global Context Network (GCNet) [32] this year. This approach captures remote dependencies by aggregating the global context for specific query at each query location in an image, and creates a simple network based on a stand-alone query formula that maintains the accuracy of the non-local network while having less computational cost. The authors for GCNet pointed out that their design is similar to the Squeeze-and-Excitation Networks (SENet) in structure, and suggested a universal three-step model to model the global context.

SOLOv1 is neither the first detection and then segmentation, nor the learning affinity method. It is a mask that directly segments instances. Comparing instance segmentation and semantic segmentation, the biggest difference in algorithm is that instance segmentation needs to deal with the problem of overlapping or adhesion of similar instances. SOLOv1 proposes a method that different instances correspond to different channels to solve this problem. SOLOv1, an end-to-end instance segmentation framework, achieves comparable results compared with Mask R-CNN.

YOLACT [10] proposed by Bolya et al. is a fast and simple instance segmentation model with a fully convolutional topology. It is mainly used for real-time instance segmentation and has the feature of the fastest real-time instance segmentation when it is introduced. When one Titan XP GPU is in use, it achieved a segmentation score of 29.8 in masking on COCO data sets at a speed of 33 frames per second, faster than other state-of-the-art methods at that time. Under the background of image segmentation, an important task is to make deep neural networks aware of their quality of prediction. For the purpose of instance segmentation, most methods adopt reliable estimation of instance classification as the quality score of the mask. What contradicted with this approach is the mask quality, which is not properly associated with the classification score as a quantified instance mask.

Huang et al. studied this problem, and then proposed Mask Scoring R-CNN [33]. The object detection model integrated with the sliding window technology had gained significant popularity and rapid development by generating predictive bounding boxes with dense and regular spatial meshes. For this reason, Chen et al. proposed a model called Tensor Mask [34] that performed instance segmentation with dense sliding windows. This is a field relatively undeveloped. In this work, the dense form of instance segmentation was implemented through prediction for four-dimensional tensor. This differs from the earlier work in instance segmentation with methods such as Deep Mask and Instance FCN. The latter adopted an unstructured 3-D tensor with segmentation masks encapsulated on the third channel axis. And the results suggested that the Tensor Mask framework/model can smooth the way for instance segmentation based on dense sliding windows in the future.

3.7. **BlendMask.** In 2020, Chen et al. proposed BlendMask [35], a one-stage dense instance segmentation approach, that integrated the ideas of top-down and bottom-up to extract low-level detail features by adding low-level models on the basis of FCOS and meanwhile predicting an object result at the instance level. It, based on the fusion of FCIS and YOLACT, proposed a hybrid module to better integrate the two features. As a result, both the accuracy (41.3AP) and speed (BlendMask-RT 34.2mAP, 25FPS on 1080ti) on COCO data sets surpassed that of Mask R-CNN. The comparisons of the BlendMask method with other methods on accuracy and efficiency are given in Tables 1 and 2. This method is mainly considered from top-down and bottom-up methods, combining instance information and semantic information to improve the accuracy of instance segmentation, and has achieved obvious results. So far, it is the latest effective segmentation method.

In Table 1, Mask R-CNN* is the implementation detail of the modified Mask R-CNN on the TensorMask; the column "Aug." is multi-scale training with short margins [640, 800];

TABLE 1. The comparison of BlendMask with the R-CNN and TensorMask

Method	Backbone	Epochs	Aug.	T (ms)	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Mask R-CNN	R-50	12		97.0	34.6	56.5	36.6	15.4	36.3	49.7
Mask R-CNN*		72	√	97+	36.8	59.2	39.3	17.1	38.7	52.1
TensorMask		72	√	400+	35.5	57.3	37.4	16.6	37.0	49.1
BlendMask		12		78.5	34.3	55.4	36.6	14.9	36.4	48.9
BlendMask		36	√	78.5	37.0	58.9	39.7	17.3	39.4	52.5
BlendMask*		36	√	74.0	37.8	58.8	40.3	18.8	40.9	53.6
Mask R-CNN	R-101	12		118.1	36.2	58.6	38.4	16.4	38.4	52.1
Mask R-CNN*		36	√	118+	38.3	61.2	40.8	18.2	40.6	54.1
TensorMask		72	√	400+	37.3	59.5	39.5	17.5	39.3	51.6
SOLO		72	√	–	37.8	59.5	40.4	16.4	40.6	54.2
+deform convs		72	√	–	40.4	62.7	43.3	17.6	43.3	58.9
BlendMask		36	√	101.8	38.4	60.7	41.3	18.2	41.5	53.3
BlendMask*		36	√	94.1	39.6	61.6	42.6	22.4	42.2	51.4
+deform convs (interval = 3)		60	√	105.0	41.3	63.1	44.6	22.7	44.1	54.5

TABLE 2. The comparison of YOLACT real-time setting speed and accuracy indexes on COCO val2017 obtained through the official code and training model

Method	Backbone	NMS	Resolution	Time (ms)	AP^{SS}	AP	AP_{50}	AP_{75}
YOLACT	R-101	Fast	550×550	34.2	32.5	29.8	48.3	31.3
YOLACT		Fast	700×700	46.7	33.4	30.9	49.8	32.5
BlendMask-RT		Batched	$550 \times *$	47.6	41.6	36.8	61.2	42.4
Mask R-CNN	R-50	Batched	$550 \times *$	63.4	39.1	35.3	56.5	37.6
BlendMask-RT	R-50	Batched	$550 \times *$	36.0	39.3	35.1	55.5	37.1

Speed Mask R-CNN 1× and BlendMask are measured with MaskRCNN-Benchmark on a single 1080Ti GPU. BlendMask* is based on Detectron 2, and speed differences are caused by different measurement rules; '+deform convs (Interval = 3)' was convs obtained using deformable convolution over an interval of 3 backbone network.

In Table 2, among them, Mask R-CNN and BlendMask models use Detectron2 for training and measurement, with a resolution of 550 meaning a shorter edge 550 is used in reasoning. What can be seen is that the faster version of BlendMask is significantly better in precision than YOLACT's execution time.

3.8. **Discussions and future directions.** It can be seen from the above that there are more top-down instance segmentation studies, from the two-stage pursuit of accuracy to the single-stage pursuit of speed. The bottom-up instance segmentation method is also pursuing speed while minimizing post-processing. Bottom-up research is relatively rare, but it provides clever ideas and is instructive for the subsequent research. The emergence of direct methods is a new idea, which has a lot of room for development and exploration. Aiming at the problem of instance segmentation, problems remain unsolved are worthy of further exploration and optimization. Future research work can refer to the following aspects.

a) **The Problem of Object Splitting Due to Occlusion.** Due to occlusion, an object is divided into multiple parts, resulting in fragmentation of the instance. For example, in the Cityscapes dataset, it is very common that the car is blocked by a

pole and the car is divided. The current fragment merging method is computationally expensive, complex and time-consuming. The accuracy has not reached the expected effect, so for the problem of instance fragmentation caused by occlusion, how to improve the network's processing of instance fragmentation has further research value.

b) **Optimization of Edge Contour.** For some examples of features with complex contours, the segmentation of the boundary area is generally fuzzy and not fine enough. Alexander scholars have observed the edges of most objects in the inaccurate segmentation, and proposed the PointRend network (iv) to select the Top N most fuzzy in prediction of the points to restore the detailed segmentation on the finer grid, refine the segmentation of the edge contour. Although the edge only occupies a very small part of the entire object, optimizing the edge of the object is essential to improve the quality of the segmentation. The segmentation is still the goal pursued by instance segmentation, and fine edge contour segmentation is also one of the focuses of research.

c) **Performance Optimization for Single-Stage Target Detection Algorithm.** Since the instance segmentation method is based on the target detection algorithm, accurate target detection algorithm is conducive to the improvement of instance segmentation performance, so the improvement of target detection algorithm research has also promoted the development of instance segmentation to a certain extent. However, in recent years, the target detection network capacity of the method is smaller, the hyperparameters are less, the speed is faster, and the accuracy is higher. The selection and improvement of the instance segmentation method based on the anchor-free single-stage detector to achieve real-time performance is an important direction for subsequent development.

d) **Information Fusion.** BlendMask combines high-level rough instance information with low-level fine-grained information, so that the network learns a richer feature representation. The high-level feels wild and has rich semantic information, and the low-level has a lot of local information, which can provide more detailed information. Therefore, information fusion makes the segmentation result better. How to perform information fusion better and more concisely and let the network learn better features is worth further research.

4. **Other Factors Affecting the Effect of Instance Segmentation.** Obviously, instance segmentation tasks are attracting the attention of more and more researchers. Until now, however, instance segmentation is still one of the most challenging tasks in computer vision. Instance segmentation needs separate segmentation and identification on different object instances in an image. However, the light conditions in the natural light scenario will directly affect the imaging, causing the judgment on instance category and boundary to be influenced by the too dark or bright condition. When the foreground object is similar to the background in colour, image segmentation will be very difficult [8]. Meanwhile, due to the influence of feature learning style, it is often difficult to acquire perceptive features for small objects, which greatly affects the detection and segmentation performance. In addition to this, the influence of factors such as angle, occlusion, and motion blur, in image acquisition also brings great challenges for the stability of the algorithm. Traditional methods usually rely on manual features extracted from the image; however, it is often difficult for such features to be enough robust under complex cases and meanwhile to be sufficient in generalization. With the continuous development of deep learning, enough reliable features can be learned for neural networks through massive data. However, it is quite difficult to cover all scenarios for data collection. Furthermore,

the labelling for instance segmentation tasks is also very tedious, which requires not only to be the pixel-level as defined in semantic segmentation, but also to distinguish different individual instances in the same class of detection objects, thus being quite costly. Essentially speaking, instance segmentation is also a feature learning task, similar to other tasks [20]. From the perspective of data, on the one hand, the more convenient acquisition of unlabelled data, simplified labelling process, pre-labelling with existing algorithms and technologies, and lower labelling cost with the advent of the era of big data, can help establish a larger and more comprehensive data set in the academic circle. On the other hand, exploring the connections between existing data and proposing new ways for data enhancement can also help the network to learn. From the perspective of algorithm, improving network design by improvements on network structure and meanwhile performing targeted optimization for instance segmentation tasks will enable networks to learn robust features more easily.

5. **Conclusion.** This paper sorts out the development of instance segmentation, and the emergence of this technique proves the development trend of image processing from coarse- to fine-grained inference. It has only been a few years since the initial proposal of instance segmentation. This paper summarizes and discusses a series of technical methods that have played a key role in the development of instance segmentation from its proposal (SDS) to the present (BlendMask) in a timeline manner, and meanwhile sums up several existing research difficulties and key challenges to be overcome next for instance segmentation, so as to provide some references for researchers in this field. Furthermore, various applications of instance segmentation of these methods, such as medical image process and industrial image process, will be considered in the future.

Acknowledgments. This paper is supported by National Natural Science Foundation of China (62003062), Science and Technology Research Project of Chongqing Municipal Education Commission (KJZD-M201900801, KJQN201900831), Chonqqing Natural Science Foundation of China (cstc2020jcyj-msxmX0077), High-level Talents Research Project of CTBU (1953013, 1956030, ZDPTTD201918), Key Platform Open Project of CTBU (KFJJ2019062, KFJJ2017075).

REFERENCES

[1] S. Wibirama, I. Ardiyanto, T. Satriya, T. B. Adji, N. A. Setiawan and M. T. Setiawan, An improved pupil localization technique for real-time video-oculography under extreme eyelid occlusion, *International Journal of Innovative Computing, Information and Control*, vol.15, no.4, pp.1547-1563, 2019.

[2] A. S. Agoes, Z. Hu and N. Matsunaga, LICODS: A CNN based, lightweight RGB-D semantic segmentation for outdoor scenes, *International Journal of Innovative Computing, Information and Control*, vol.15, no.5, pp.1935-1946, 2019.

[3] A. A. S. Gunawan, I. Arifiany and E. Irwansyah, Semantic segmentation of aerial imagery for road and building extraction with deep learning, *ICIC Express Letters*, vol.14, no.1, pp.43-51, 2020.

[4] Q. Li, H. Wang, J. Li, Y. Xiao and W. Hu, Deep hierarchical semantic segmentation algorithm based on image information entropy, *ICIC Express Letters, Part B: Applications*, vol.11, no.1, pp.25-32, 2020.

[5] A. Patrik, G. Utama, A. A. S. Gunawan, A. Chowanda, J. S. Suroso and W. Budiharto, Modeling and implementation of object detection and navigation system for quadcopter drone, *ICIC Express Letters*, vol.13, no.6, pp.461-468, 2019.

[6] J. Choi, M. Han and N. Kim, Object detection of cochlea in micro-computed tomography using faster region convolutional neural network, *ICIC Express Letters, Part B: Applications*, vol.10, no.7, pp.651-656, 2019.

[7] Y. Xu, Y. Li, Y. Wang, M. Liu, Y. Fan, M. Lai and E. Chang, Gland instance segmentation using deep multichannel neural networks, *IEEE Transactions on Biomedical Engineering*, vol.64, no.12, pp.2901-2912, 2017.

[8] J. Ahn and S. Kwak, Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2018)*, Salt Lake City, UT, USA, pp.4981-4990, 2018.

[9] M. Bai and R. Urtasun, Deep watershed transform for instance segmentation, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2017)*, Honolulu, HI, USA, pp.2858-2866, 2017.

[10] D. Bolya, C. Zhou, F. Xiao and Y. J. Lee, YOLACT: Real-time instance segmentation, *IEEE/CVF International Conference on Computer Vision (ICCV2019)*, Seoul, Korea, pp.9156-9165, 2019.

[11] S. Graham, H. Chen, J. Gamper, Q. Dou, P.-A. Heng, D. R. J. Snead, Y.-W. Tsang and N. M. Rajpoot, MILD-Net: Minimal information loss dilated network for gland instance segmentation in colon histology images, *Medical Image Analysis*, vol.52, pp.199-211, 2019.

[12] P. Shi, F. Li, L. Wu and C. C. Lim, Neural network-based passive filtering for delayed neutral-type semi-Markovian jump systems, *IEEE Transactions on Neural Networks and Learning Systems*, vol.28, no.9, pp.2101-2114, 2017.

[13] P. Shi, Y. Zhang, M. Chadli and R. Agarwal, Mixed H_∞ and passive filtering for discrete fuzzy neural networks with stochastic jumps and time delays, *IEEE Transactions on Neural Networks and Learning Systems*, vol.27, no.4, pp.903-909, 2016.

[14] P. Shi, Y. Zhang and R. Agarwal, Stochastic finite-time state estimation for discrete time-delay neural networks with Markovian jumps, *Neurocomputing*, vol.151, pp.168-174, 2015.

[15] Y. Liu, Y.-H. Wu, P.-S. Wen, Y.-J. Shi, Y. Qiu and M.-M. Cheng, Leveraging instance-, image- and dataset-level information for weakly supervised instance segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, doi: 10.1109/TPAMI.2020.3023152, 2020.

[16] J. Dai, K. He and J. Sun, Instance-aware semantic segmentation via multi-task network cascades, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2016)*, Las Vegas, NV, USA, pp.3150-3158, 2016.

[17] J. Pei, H. Tang, C. Liu and C. Chen, Salient instance segmentation via subitizing and clustering, *Neurocomputing*, vol.402, pp.423-436, 2020.

[18] D. Liu, D. Zhang, Y. Song, F. Zhang, L. O'Donnell, H. Huang, M. Chen and W. Cai, PDAM: A panoptic-level feature alignment framework for unsupervised domain adaptive instance segmentation in microscopy images, *IEEE Transactions on Medical Imaging*, doi: 10.1109/TMI.2020.3023466, 2020.

[19] B. Hariharan, P. A. Arbeláez, R. B. Girshick and J. Malik, Simultaneous detection and segmentation, *The 13th European Conference on Computer Vision (ECCV2014)*, Zurich, Switzerland, pp.297-312, 2014.

[20] H. Su, S. Wei, M. Yan, C. Wang, J. Shi and X. Zhang, Object detection and instance segmentation in remote sensing imagery based on precise mask R-CNN, *IEEE International Geoscience and Remote Sensing Symposium (IGARSS2019)*, Yokohama, Japan, pp.1454-1457, 2019.

[21] B. Hariharan, P. A. Arbeláez, R. B. Girshick and J. Malik, Hypercolumns for object segmentation and fine-grained localization, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2015)*, Boston, MA, USA, pp.447-456, 2015.

[22] Z. Zhang, A. G. Schwing, S. Fidler and R. Urtasun, Monocular object instance segmentation and depth ordering with CNNs, *IEEE International Conference on Computer Vision (ICCV2015)*, Santiago, Chile, pp.2614-2622, 2015.

[23] R. Girshick, Fast R-CNN, *arXiv.org*, arXiv: 1504.08083, 2015.

[24] K. Wang, Y. Dong, H. Bai, Y. Zhao and K. Hu, Use fast R-CNN and cascade structure for face detection, *Visual Communications and Image Processing (VCIP2016)*, Chengdu, China, pp.1-4, 2016.

[25] S. Ren, K. He, R. B. Girshick and J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.39, no.6, pp.1137-1149, 2017.

[26] J. Dai, K. He and J. Sun, Convolutional feature masking for joint object and stuff segmentation, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2015)*, Boston, MA, USA, pp.3992-4000, 2015.

[27] K. He, G. Gkioxari, P. Dollar and R. Girshick, Mask R-CNN, *IEEE International Conference on Computer Vision (ICCV2017)*, pp.2961-2969, 2017.

[28] L.-C. Chen, A. Hermans, G. Papandreou, F. Schroff, P. Wang and H. Adam, MaskLab: Instance segmentation by refining object detection with semantic and direction features, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2018)*, Salt Lake City, UT, USA, pp.4013-4022, 2018.

[29] X. Wang, R. B. Girshick, A. Gupta and K. He, Non-local neural networks, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2018)*, Salt Lake City, UT, USA, pp.7794-7803, 2018.

[30] S. Liu, L. Qi, H. Qin, J. Shi and J. Jia, Path aggregation network for instance segmentation, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2018)*, Salt Lake City, UT, USA, pp.8759-8768, 2018.

[31] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang, C. C. Loy and D. Lin, Hybrid task cascade for instance segmentation, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2019)*, Long Beach, CA, USA, pp.4974-4983, 2019.

[32] Y. Cao, J. Xu, S. Lin, F. Wei and H. Hu, GCNet: Non-local networks meet squeeze-excitation networks and beyond, *arXiv.org*, arXiv: 1904.11492., 2019.

[33] Z. Huang, L. Huang, Y. Gong, C. Huang and X. Wang, Mask Scoring R-CNN, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2019)*, Long Beach, CA, USA, pp.6409-6418, 2019.

[34] X. Chen, R. B. Girshick, K. He and P. Dollár, TensorMask: A foundation for dense object segmentation, *IEEE/CVF International Conference on Computer Vision (ICCV2019)*, Seoul, Korea, pp.2061-2069, 2019.

[35] H. Chen, K. Sun, Z. Tian, C. Shen, Y. Huang and Y. Yan, BlendMask: Top-down meets bottom-up for instance segmentation, *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR2020)*, Seattle, WA, USA, pp.8570-8578, 2020.