Review of object instance segmentation based on deep learning

Di Tian,[a] Yi Han,[a,*] Biyao Wang,[a] Tian Guan,[a] Hengzhi Gu,[a] and Wei Wei[b]

[a]Chang'an University, School of Automobile, Xi'an, China
[b]Xi'an University of Technology, School of Computer Science and Engineering, Xi'an, China

Abstract. As a challenging task in computer vision, instance segmentation has attracted extensive attention in recent years. Able to obtain very rich and refined object information, this technology shows important application value in many fields, such as intelligent driving, medical health, and remote sensing detection. Instance segmentation technology should not only identify the positions of objects but should also accurately mark the boundary of any single instance, which can be defined as solving object detection and semantic segmentation at the same time. Our study gives a detailed introduction to the background of instance segmentation technology, its development and the common datasets in this field, and further deeply discusses key issues appearing in the development of this field, with the future development direction of instance segmentation technology proposed. Our study provides an important reference for future research on this technology © *2021 SPIE and IS&T* [DOI: 10.1117/1.JEI.31.4.041205]

Keywords: instance segmentation; image segmentation; deep learning; computer vision.

Paper 210471SSV received Jul. 29, 2021; accepted for publication Nov. 19, 2021; published online Dec. 6, 2021.

1 Introduction

Computer vision has four basic tasks: image classification,[1–4] object detection,[5–9] semantic segmentation,[10–13] and instance segmentation.[14,15] These technologies form the basis for addressing many practical problems. The main purpose of image classification is to find a classification label from the classification label set and then to assign the label to the input image. The task of object detection is not only to assign classification labels to images but also to mark the locations of specific objects. Both image classification and object detection are the foundation for solving complex problems in computer vision, such as object tracking,[16–19] image segmentation, and scene interpretation. The task of semantic segmentation refers to assigning a category label to each pixel in an image, so as to segment and detect specific categories and finely label the boundaries. While distinguishing each category, instance segmentation also provides different labels for individual instances in the same type of objects. It can be regarded as delivering the tasks of object detection and semantic segmentation at the same time, which usually need to deal with multiple overlapping objects and complex backgrounds. This is the most challenging problem of the four basic tasks of computer vision and has important application value in the fields of intelligent driving,[20–22] remote sensing detection,[23–26] and medical health.[27–30] The four basic tasks of computer vision are shown in Fig. 1. With breakthroughs in computer vision technology, as well as progress in wireless sensor network[31–33] and multiple communications,[34–39] especially in cloud computing[40–42] and other techniques to lower the hardware requirements, it is possible to solve these practical problems.[43–47] Currently, neural network technology is widely deployed.[48–51] Instance segmentation can provide richer information than other computer vision tasks can, but it is also a more difficult job. Although the development of deep learning has greatly promoted the progress of instance segmentation tasks over recent years, there are still few related reviews, without a comprehensive summary of its latest development and a deep

*Address all correspondence to Yi Han, hany@chd.edu.cn

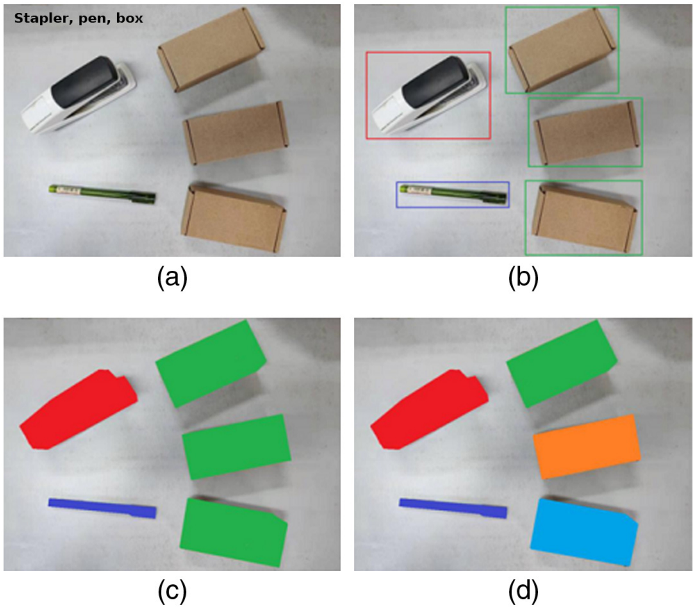


Fig. 1 Diagram of the four basic tasks of computer vision: (a) image classification, (b) object detection, (c) semantic segmentation, and (d) instance segmentation.

discussion of its current difficulties. This study will make an in-depth discussion of instance segmentation on the basis of image segmentation.

The task of instance segmentation is highly related to the task of object detection, and further segmenting the pixels of objects based on results of object detection is an instance segmentation task. Over recent years, the rapid development of object detection technology has also driven the progress in instance segmentation technology, delivering continuous breakthroughs in its accuracy and speed. Instance segmentation can be traced back to traditional image segmentation technology, which divides images into disjoint but meaningful subregions. Pixels in the same region have certain correlations, but some differences exist between pixels of different regions. Traditional image segmentation techniques mainly include segmentation methods based on thresholds,[52–54] edges,[55–57] and clustering.[58–60] The principle of the threshold-based image segmentation method lies in classifying the pixels of an image through different gray thresholds. Pixels in the same grayscale range are classified into the same category and regarded as having similar properties. This method is suitable for images with uniform grayscale distribution and clear grayscale distinctions between objects and their backgrounds; however, this method is susceptible to noise. The edge-based image segmentation method is mainly used to detect the pixels at the boundary of objects and then connect the pixels to form the edge contours of images. Common methods for it include Roberts,[61] Prewitt,[62] Sobel,[63] and Canny.[64] The principle of the cluster-based image segmentation method is to merge adjacent pixels with similar characteristics into an identical category, and finally gather all the pixels into several different categories, so as to divide image areas. Common methods for it include K-means,[65] FCM,[66] and SLIC.[67] Traditional image segmentation methods can deliver better processing effects for simple machine vision problems and can complete the required image segmentation tasks if reasonably implemented in actual scenes.

However, with wide deployments of computer vision in such complex scenes as intelligent driving and security monitoring, traditional image segmentation technology has no longer been able to meet the requirements of object segmentation in such scenes. As deep learning is adopted in many fields,[48,68,69] image segmentation technology has made tremendous progress. Such image segmentation algorithms as FCN and Mask R-CNN based on deep learning have pushed image segmentation technology to a new height by delivering the possibility of applying instance segmentation technology in complex environments. Nowadays, instance segmentation technology based on deep learning can realize accurate segmentation of objects in common scenes. However, the real-world environment is much more complex and changeable, with many external interfering factors. Therefore, the instance segmentation technology has to be further enhanced.

The remainder of this paper is arranged as follows. Section 2 will introduce the current mainstream instance segmentation algorithm in detail. Section 3 will discuss the common datasets, evaluation methods, and data augmentation methods in the field of instance segmentation. In Sec. 4, the key issues affecting the in-depth development of this field will be discussed deeply. Finally, conclusions and outlooks are given in Sec. 5.

2 Object Instance Segmentation Techniques

The current object instance segmentation technology has three types of solutions in terms of processing ideas. Instance segmentation task can be defined as a collection of object detection and semantic segmentation, so there are two solutions from these two aspects. One of them is detection-based method, which is similar to the object detection task. This method first detects the area of each instance and then divides the instance mask in each area. The other is a method based on pixel clustering. This method first predicts the category label of each pixel and then uses the clustering method to group them to form instance segmentation results. The above two solutions are both two-stage segmentation methods, which have attracted the attention of a large number of researchers. In addition, in recent years, a single-stage instance segmentation method has been developed, which has a greater improvement in processing speed compared to the above two segmentation methods. Next, this paper will discuss in detail the common instance segmentation methods. The classification of the instance segmentation methods based on deep learning is shown in Fig. 2.

2.1 *Method Based on Detection*

With the application of convolutional neural networks (CNN) in the field of object detection, it has greatly promoted the development of object detection. In 2014, Girshick proposed the RCNN framework, which uses selective search[70] to generate a region proposal boxes on the image and then uses CNN for feature extraction. Finally, train the support vector machine (SVM)[71,72] classifier to predict the result. The appearance of this algorithm has greatly promoted the development of object detection technology. Subsequently, the author improved RCNN and further proposed Fast RCNN and Faster RCNN. It solves problems such as repeated convolution calculations and effectively improves the effect of object detection technology.

The detection-based instance segmentation method first finds the region of the instance through object detection method and then performs mask prediction in the detection area. In the end, each prediction result is output as a different instance. This method is closely related to object detection. With the rapid development of object detection technology, it can directly promote the progress of this type of instance segmentation technology. In terms of algorithm research, it has significant advantages over other types of instance segmentation technologies.

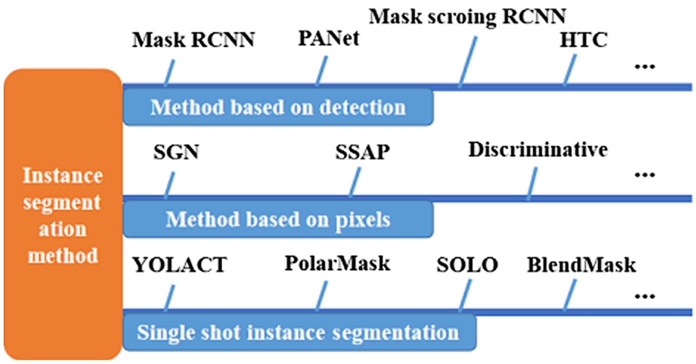


Fig. 2 Classification of instance segmentation.

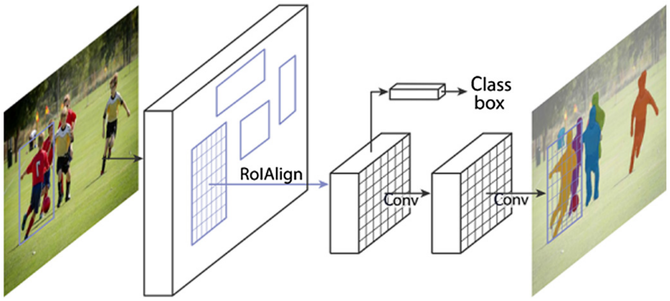


Fig. 3 Mask RCNN framework for instance segmentation.

In 2017, He et al. proposed the Mask RCNN algorithm based on Faster RCNN. The algorithm adds a mask branch to predict the instance mask on the basis of Faster RCNN. It is a classic instance segmentation algorithm based on object detection. It has obtained a good instance segmentation effect and strongly promoted the development of related technologies. The framework of Mask RCNN is shown in Fig. 3.

In 2016, Dai et al.[73] proposed a multi-task learning framework for the problem that semantic segmentation algorithms cannot be applied to instance segmentation. The model is composed of three parts: distinguishing instances, estimating masks, and classifying objects. These parts are designed as a cascade structure. The algorithm achieved the most advanced instance segmentation accuracy at the time on the VOC dataset. In 2018, Liu et al.[74] proposed PANet. Through the bottom-up path enhancement method, precise positioning signals are used at a lower level to enhance the entire feature hierarchy. The information path between the lower level and the uppermost level is shortened. And an adaptive feature pool is proposed, so that the useful information of each layer is directly transmitted to other suggestion subnets, and a complementary branch that captures different views for each suggestion is created, which further improves the mask prediction result. In 2019, Chen et al.[75] proposed a new hybrid task cascade framework HTC, which introduced cascade in instance segmentation. The framework interweaves detection and segmentation for joint multi-stage processing and uses full convolution to distinguish difficult samples, which can gradually learn more distinguishing features. At the same time, the complementary features are integrated at each stage, which effectively improves the effect of instance segmentation.

2.2 *Method Based on Pixel*

Pixel-based methods first predict the category label of each pixel and then group them to form instance segmentation results through methods such as clustering and metric learning.[76–78] Compared with detection-based methods, this type of algorithm is generally less accurate, and because it needs to predict each pixel, it puts a higher demand on the computing power of the computer.

Gao et al.[79] proposed a pixel-based instance segmentation method SSAP, which learns the probability that two pixels belong to the same instance by learning the affinity pyramid of pixel pairs. Among them, the affinity pyramid learns the short-distance affinity from the higher resolution image, learns the long-distance affinity from the lower resolution image, and then generates the multi-scale affinity pyramid. Use semantic segmentation and affinity pyramid joint learning to generate multi-scale instance predictions. Through the cascading graph division module, the running speed is effectively improved. It reached the most advanced level at the time on the Cityscapes dataset. The frame diagram of SSAP is shown in Fig. 4.

Brabandere et al.[80] proposed a pixel-level discriminative loss function to deal with instance segmentation tasks. The loss function maps each pixel in the network to a point in the feature space, and this makes the pixels belonging to the same instance very close, while the distance between different instances is very far. The loss function includes pulling force, thrust force, and regularization, where the pulling force is to reduce the distance between all elements in the same instance and their average value. The thrust force pushes the center point of each cluster farther.

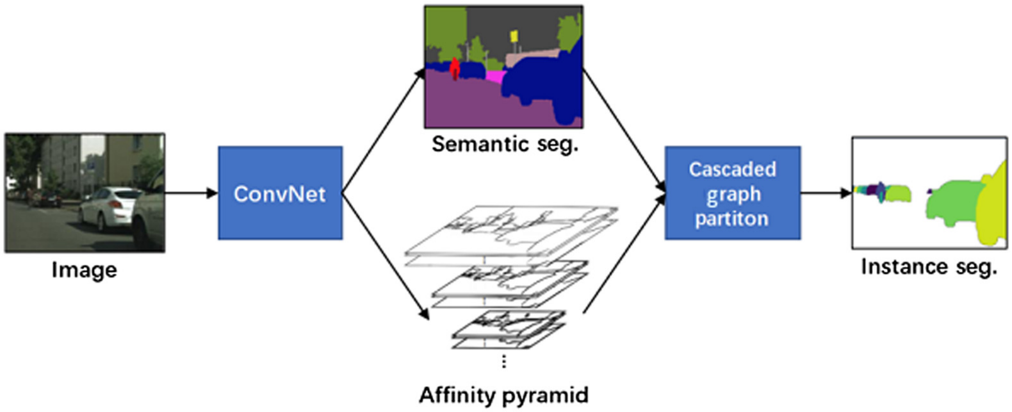


Fig. 4 SSAP instance segmentation algorithm framework.

Regularization means that the center point of each cluster is as close to the origin as possible. In addition, Bai and Urtasun[81] proposed a simple end-to-end CNN to complete instance segmentation. They combined semantic segmentation with the traditional watershed algorithm to generate an energy map and then segmented each instance of the energy map, which can complete a very fast and accurate estimation.

Kirillov et al.[82] defined the instance segmentation problem as two outputs of instance-agnostic semantic segmentation and all instance boundaries. The edge detector is trained using the annotation of instance segmentation, and then the semantically segmented region is segmented by the edge detector, forming a new multi-cutting form. Arnab and Torr[83] proposed an instance segmentation system based on an initial semantic segmentation module, which inputs the semantic segmentation result into an instance subnetwork for instance prediction and achieves instance segmentation with higher precision. Liu et al.[84] proposed the SGN network to decompose the instance segmentation task into a series of subtasks, then the final segmentation mask prediction is performed by the combination of these subtasks. The advantage of this method is to achieve task decomposition, but because each subtask is performed sequentially, it takes a long time.

2.3 *Single-Stage Instance Segmentation*

Compared with the above two types of instance segmentation algorithms, the single-stage instance segmentation technology discussed in this section has better computational efficiency, and the existing technology can already meet the requirements of real time in practical applications. In 2019, Bolya et al.[85] proposed the YOLACT algorithm. It decomposes the instance segmentation task into two subtasks: generating a set of prototype masks and predicting the mask coefficient of each instance. Then an instance mask is generated by linearly combining the two subtasks. In addition, in order to improve the running speed of the branch, the algorithm proposes a Fast non maximum suppression (NMS) algorithm to replace the NMS algorithm to achieve real-time instance segmentation speed. The structure diagram of YOLACT is shown in Fig. 5. In order to further improve the segmentation accuracy of YOLACT, the author proposes the YOLACT++ algorithm.[86] By introducing deformable convolution into the backbone network, the segmentation accuracy is effectively improved without affecting the efficiency of the algorithm.

In 2020, Xie et al.[87] proposed a single-stage instance segmentation method PolarMask. This method transforms instance segmentation into two tasks of instance center point classification and dense distance regression to jointly predict instance contours. In addition, two effective methods are proposed to process high-quality center samples and optimize dense distance regression, which can significantly improve performance and simplify the training process. This framework effectively simplifies the complexity of the instance segmentation task and has a higher segmentation accuracy. In the same year, Wang et al.[88] proposed a single-stage instance segmentation method SOLO, which directly predicts the object, predicts the instance category according to the position of the object center and the object size, and does not rely on suggestion box

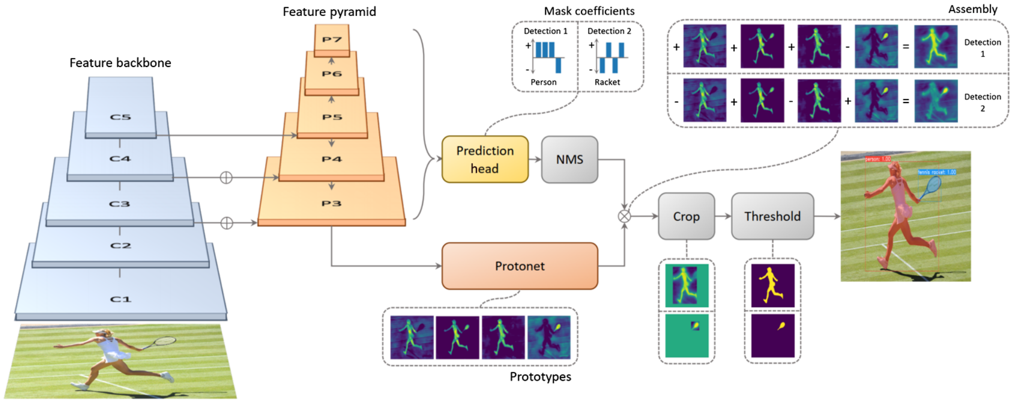


Fig. 5 YOLACT algorithm structure diagram.

extraction and postprocessing algorithms. The algorithm idea is to divide the picture into grids. If the center of the object falls into a certain grid, the grid uses the classification branch to predict the semantic category of the object, and the mask branch predicts the instance mask of the object. This processing algorithm is simple and efficient and has the same processing accuracy as Mask RCNN. It is better than all previous single-stage processing algorithms. Later, the SOLOv2[89] algorithm was developed on the basis of SOLO, which decomposes the object mask generation into kernel branches and feature branches, reducing subsequent calculations. In addition, the Matrix NMS method was proposed to execute NMS operations in parallel, and better results were obtained, and breakthroughs were made in segmentation accuracy and calculation speed.

The instance segmentation model based on deep learning has two typical processing ideas: top-down and bottom-up. The BlendMask algorithm[90] combines the advantages of the two ideas. The FCOS object detection algorithm[91] is used as the main structure, and it combines the global semantic information provided by the higher-level features with the location information provided by the lower-level features. The network learns richer feature information and achieves an excellent instance segmentation effect.

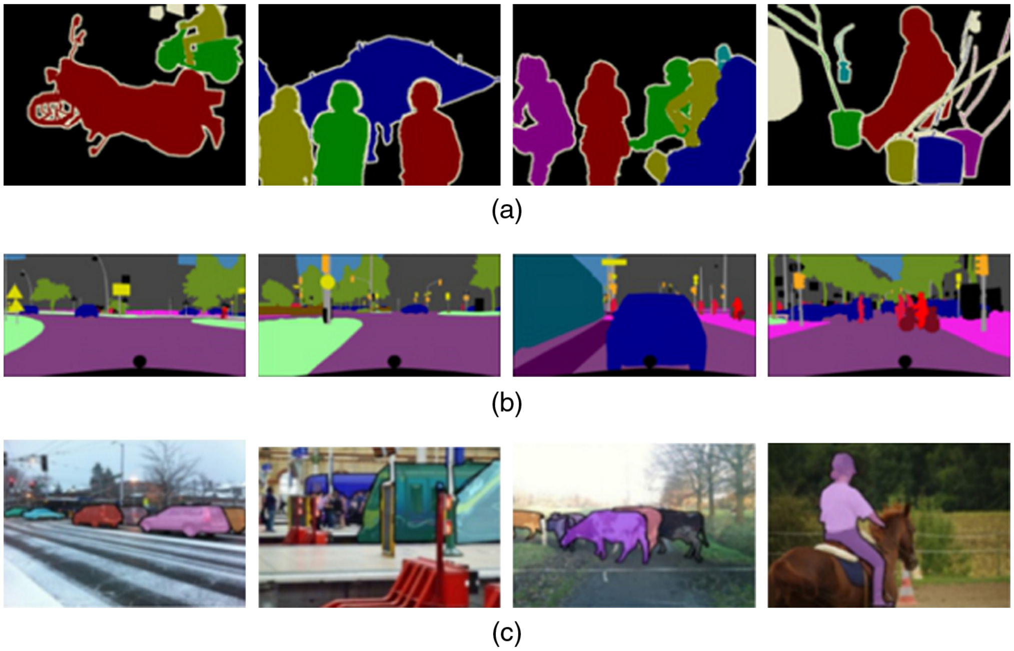
Among the above-mentioned mainstream methods of instance segmentation, both detection-based and pixel-based methods are two-stage processing methods, it generally has better segmentation accuracy than single-stage processing methods. Among them, detection-based processing methods are closely related to object detection tasks, and the rapid development of object detection tasks can directly promote the development of such algorithms, which has significant advantages. The single-stage instance segmentation algorithm is generally more concise and efficient than the two-stage algorithm and has a faster segmentation speed. In addition, the latest SOLO algorithm and BlendMask algorithm have also reached the segmentation accuracy of the two-stage algorithm, which is a recent research hotspot in the field of instance segmentation. Fields such as intelligent driving and security monitoring put forward higher requirements for the real-time performance of algorithms. In these fields, the current single-stage instance segmentation method with fast segmentation speed is undoubtedly the future development direction.

3 Datasets and Data Augmentation Methods

3.1 *Datasets*

Model training and testing are inseparable from a large amount of reasonable data. Datasets play a vital role in the development of computer vision technology. In addition, some public datasets also provide benchmarks for performance comparisons between different algorithms. The criteria for measuring the quality of a dataset include data richness, scene diversity, and label completeness. This section introduces common datasets in the field of instance segmentation.

The Pascal VOC dataset[92] is a dataset used in the PASCAL VOC Challenge between 2005 and 2012. It can be used for tasks such as object detection, semantic segmentation, and instance segmentation. It contains 11,540 pictures in 20 categories, and each picture has detailed


(a)

(b)

(c)

Fig. 6 Some example images in (a) Pascal VOC, (b) Cityscapes, and (c) Microsoft COCO.

annotation information. It has been the most important algorithm comparison benchmark before the large-scale application of the Microsoft COCO dataset.

The Cityscapes[93] dataset is an image dataset collected from more than 50 different city streets, with 5000 finely labeled images and 20,000 roughly labeled images. The dataset contains a total of 30 object categories, which are divided into 8 major categories related to urban scenes. It can be used for semantic segmentation and instance segmentation tasks and are mainly used to improve the algorithm's urban scene understanding performance.

The Microsoft COCO[94] dataset is collected from many different daily scenes, with a total of 330,000 pictures, including 80 categories and 1.5 million object instances. It can be used for a variety of different vision tasks, including object detection, semantic segmentation, instance segmentation, and panoptic segmentation. Due to the large amount of data and diverse scenarios, the Microsoft COCO dataset is currently recognized as an authoritative dataset in many computer vision fields and is often used as a benchmark for comparing the pros and cons of algorithms. The pictures in the Pascal VOC, Cityscapes, and Microsoft COCO datasets are shown in Fig. 6.

The Pascal VOC dataset is rich in types and has detailed annotation information. Compared with the COCO dataset, it is more suitable for algorithm research on personal terminals, and its appropriate amount of data can effectively reflect the pros and cons of the model. Compared with the Pascal VOC dataset, the COCO dataset has a richer amount of data. It can obtain better experimental results with abundant computing resources, which helps to improve the final results of the model. Compared with the above two comprehensive datasets, the Cityscapes dataset contains a large amount of data related to urban scenes, which has important application value in professional fields such as intelligent driving.

In addition to the above three common datasets in the field of instance segmentation, there are datasets such as MVD,[95] Kins,[96] and SBD.[97] The relevant information of these datasets is shown in Table 1.

3.2 *Evaluation Method*

Evaluation methods can be used to test the pros and cons of algorithms. Using the same evaluation method on the same dataset can compare the performance of different algorithms. At present, the performance of instance segmentation algorithms can be evaluated from many aspects, among which the most important evaluation indicators are segmentation accuracy and running speed.

Table 1 Dataset related information.

Dataset name	Category	Images of trainval	Images of test	Size	Characteristic
Pascal VOC	20	11,540	10,991	—	Multiple categories, very important early datasets
Cityscapes	8	3475	1525	1024 × 2048	Five thousand finely annotated images
COCO	80	123,287	40,670	—	Multiple categories, large-scale datasets
MVD	66	25,000	—	—	Images taken around the world in different weather, seasons, and daytime
Kins	8	7474	7517	—	Fine pixel level annotation
SBD	20	8498	2820	500 × 500	Object external boundary dataset

In terms of accuracy, average precision (AP) is generally used as an evaluation index, which represents the accuracy of instance segmentation. In addition, because common datasets usually have multiple categories, mAP is often used as an evaluation indicator to represent the average accuracy of multiple categories, the result is to take the mean value of each category AP. Intersection over Union (IoU) represents the intersection ratio between the algorithm segmentation result and the ground truth box. In the general definition, when IoU is greater than 0.5, the segmentation is considered successful. After 2014, due to the widespread use of COCO datasets, researchers began to pay more attention to accuracy. In COCO, a fixed IoU threshold is not used. Instead, multiple IoUs are averaged between 0.5 (coarse positioning) and 0.95 (perfect positioning). This metric change promotes more accurate object positioning.

In terms of running speed, processing time and frames per second are generally used as evaluation indicators. The processing time represents the time required to process a standard resolution image, and the number of frames per second represents the number of images that the algorithm can process in one second. When comparing the operating efficiency of different algorithms, in addition to ensuring the consistency of the dataset, it is also necessary to ensure the consistency of the hardware platform.

The performance of excellent instance segmentation algorithms in recent years is shown in Table 2.

Table 2 Algorithm performance information.

Algorithm	Dataset	AP (%)	Year
Mask RCNN	COCO	37.1	2017
SGN	Cityscapes	25.0	2017
PANet	COCO	40.0	2018
SSAP	Cityscapes	32.7	2019
TensorMask[98]	COCO	37.3	2019
YOLACT	COCO	29.8	2019
YOLACT++	COCO	34.6	2019
SOLO	COCO	37.8	2020
SOLOv2	COCO	39.7	2020
PolarMask	COCO	32.9	2020
BlendMask	COCO	41.3	2020
Deep snake[99]	Cityscapes	31.7	2020

3.3 *Data Augmentation*

At present, with the continuous development of deep learning technology, although the network performance is getting stronger and stronger, it also leads to the deepening and complexity of the network structure. The earliest Lenet5[100] only had five layers, but nowadays common networks such as ResNet can easily reach hundreds of layers. The complex network puts forward higher requirements on the amount of data, and it is easy to cause problems such as overfitting when the amount of data is insufficient.

For many specific application areas, there are not enough datasets available. The several datasets mentioned above have a data volume of several thousand to several tens of thousands, but they are often divided into many categories. For example, the VOC dataset is divided into 20 categories, the MVD dataset is divided into 66 categories, and the COCO dataset is divided into 80 categories. When focusing on the case segmentation research of a certain fine category, there is often a serious lack of corresponding data volume, thus developed a series of data augmentation methods. Through the data augmentation method, the amount of data can be appropriately expanded without affecting the content of the image expression, which is an effective method to increase the diversity of data.

Common data augmentation methods include multi-scale scaling, flipping, and cropping. Multi-scale instance segmentation has always been a tricky problem. Different sizes of the same object can help the network learn its features better. Multi-scale scaling is to randomly scale the original image to a specified set of sizes. In the process of multi-scale scaling, it is generally equal-scale scaling of the original image to avoid object distortion. Flip is generally divided into horizontal flip and vertical flip, which uses the center of the image as the center of rotation for symmetrical mapping of the image. The flipped image can help the network learn information about objects in different positions of the image. The cropping operation generally cuts part of the image randomly, so that objects appear in different positions of the image in different proportions, which can increase the data diversity to a certain extent and reduce the sensitivity of the model to the object position.

In addition, a data enhancement method called Mosaic is mentioned in YOLOv4.[101] This method randomly uses four pictures to perform random scaling. Randomly distributed splicing is then carried out, which can effectively enrich the detection dataset. In particular, this data enhancement method adds many small-scale objects to make the network more robust. Mosaic data enhancement is shown in Fig. 7.

4 Discussion on Key Issues in Instance Segmentation Technology

4.1 *Instance Segmentation Framework*

From the perspective of the development of instance segmentation technology, a large number of studies have focused on the two-stage instance segmentation framework. Including classic algorithms such as Mask RCNN that detects first and then segmentation and also includes algorithms such as SSAP that first pixel-by-pixel prediction and then clustering. They have achieved good segmentation results in the test. In recent years, inspired by detectors such as YOLO, some single-stage segmentation methods have also appeared in the field of instance segmentation, such as YOLACT and PolarMask algorithms.

In the technical development process of instance segmentation, compared with the single-stage processing method, the two-stage method has many advantages, such as better segmentation accuracy and easier promotion. Moreover, the continuous improvement of object detection technology can also promote the development of detection-based two-stage instance segmentation technology, so a large amount of research focuses on the two-stage method. However, the two-stage method is generally more cumbersome to train and requires more training time and higher hardware resources. In the test, the detection speed is often slower than the single-stage method. The single-stage segmentation algorithm has a simple structure and a fast running speed, which can meet the requirements of real-time applications. When this type of algorithm first appeared, the segmentation accuracy was low. However, after continuous research and

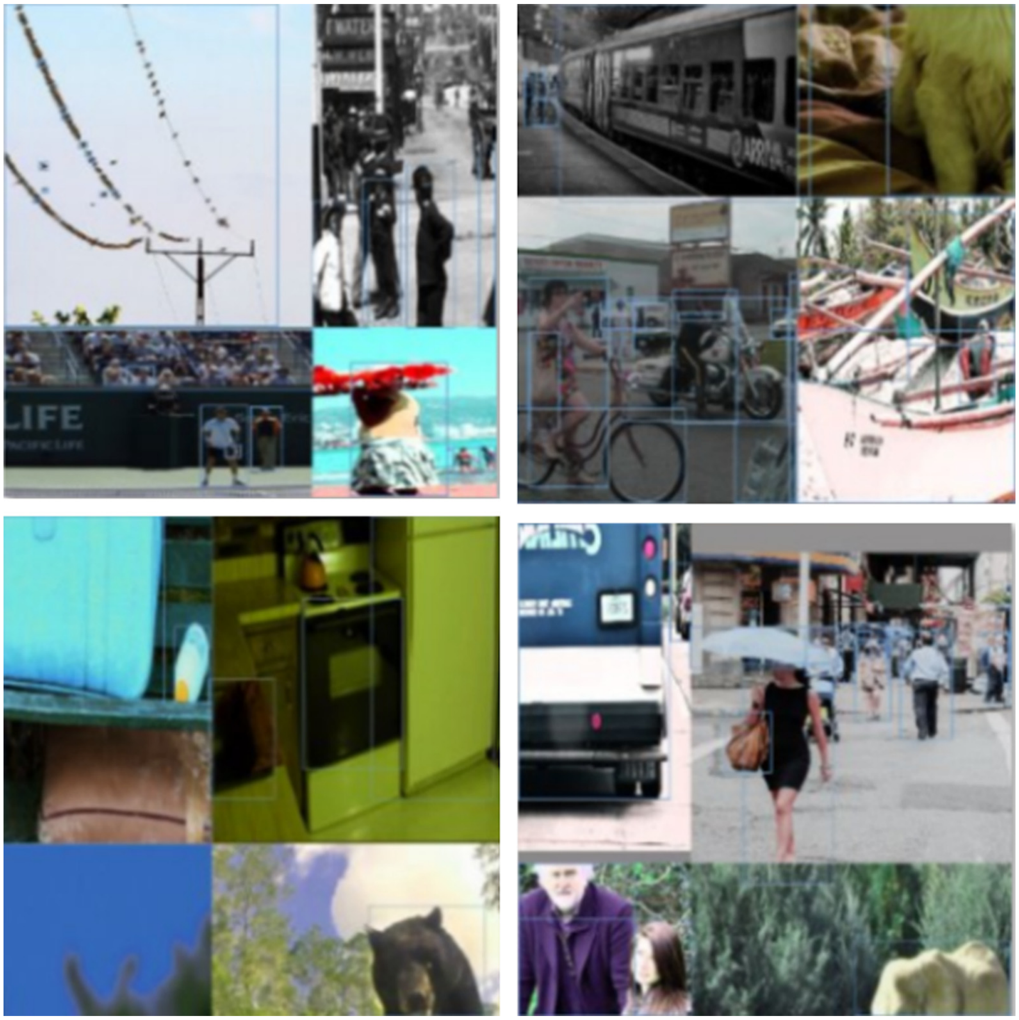


Fig. 7 Mosaic data enhancement results.

development, considerable progress has been made the latest SOLO, and BlendMask can reach the segmentation accuracy of the two-stage algorithm Mask RCNN.

In the current research process of instance segmentation technology, due to the advantages of single-stage frame segmentation speed, there are more and more researches on related algorithms, which strongly promote the development of single-stage instance segmentation technology. However, on the whole, the two-stage algorithm is easier to obtain excellent segmentation accuracy, and it still has important application value in medical applications and other occasions where real-time requirements are not high. In the research and application of instance segmentation technology, the specific frame selection should be closely combined with the application scenario to obtain a more suitable segmentation effect.

4.2 *Small Object Image Instance Segmentation*

Small object instance segmentation has important application value in many fields, such as remote sensing image instance segmentation, medical image object segmentation, and intelligent driving obstacle segmentation. Accurate instance segmentation results can provide reliable information for further in-depth analysis. At present, instance segmentation for small objects has low accuracy and poor effect and often has problems of missing segmentation and wrong segmentation. On the other hand, small objects that are successfully segmented often have problems such as low IoU with the real object and blurred segmentation boundaries. The small object image in the remote sensing field is shown in Fig. 8.

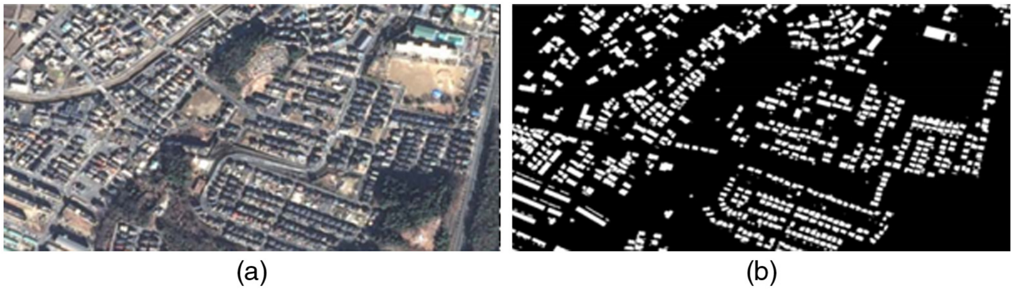


Fig. 8 Small object images in the field of remote sensing (a) satellite imagery and (b) ground truth for buildings.

After the CNN undergoes multi-layer feature extraction, the output layer generally contains only semantic information and lacks the detailed information needed to detect small objects. Although the detail information can be recovered to a certain extent through upsampling, it will cause serious loss of detail information after multi-layer convolution, which makes it difficult to segment small objects. Aiming at the difficulty of small object segmentation, Pang et al.[102] proposed a part-aware segmentation method, which can clearly detect the semantic part and retain relevant information during segmentation, which effectively improves the segmentation problem of small objects. Hamaguchi et al. found that the specific difficulties of remote sensing tasks were not considered in the segmentation of remote sensing images, such as small and dense objects in remote sensing images. To solve this problem, they proposed the LFE module,[103] which aggregated local features by reducing the expansion factor and achieved good results in the dataset test. For the problem of small object segmentation, Dijkstra et al.[104] proposed CentroidNetV2 on the basis of CentroidNet. Its loss function combines cross-entropy loss and Euclidean-distance loss to achieve high-quality object instance center detection and boundary segmentation, and effective improvements have been made in small object segmentation.

In the process of extracting object features using CNNs, generally only the semantic information needed to segment large objects is included at a high level. In order to accurately segment the small objects, the detailed information contained in the low-level feature map must be used, so the number of convolutional layers can be reduced to improve the small object segmentation ability. However, in practical applications, in order to retain sufficient large object detection capabilities, feature fusion is generally used to combine the information of multiple layers of features. Although the above method can improve the small object segmentation ability to a certain extent, the effect is still not ideal. How to segment small object images efficiently and accurately is a major difficulty in the field of instance segmentation.

4.3 *Instance Segmentation Edge Contour Optimization*

For some complex feature instances, the existing algorithms generally blur the segmentation of boundary regions. Although the segmentation result can complete the specific instance segmentation task, the overall visual effect is poor due to rough edges. Fine contour optimization can directly improve the visual effect of instance segmentation. The contour optimization effect comparison in the instance segmentation is shown in Fig. 9.

Aiming at the problem of fuzzy edge information of the segmentation result in the instance segmentation algorithm PolarMask, Zhang and Cao[105] accurately extracted the instance contour by predicting the angle offset and distance of the contour point. At the same time, the semantic segmentation subnetwork is used to further refine the edge of the instance, which improves the segmentation accuracy by 2.1% compared with the improvement before. Aiming at the problem that the edges of instances segmented by Mask RCNN are not fine enough, Liang et al. proposed to use PoolNet to process the detected images based on the segmentation results obtained by Mask RCNN. Use the results of PoolNet to optimize the edge of the mask image of the instance segmentation, thereby optimizing the edge contour of the instance segmentation.[106]

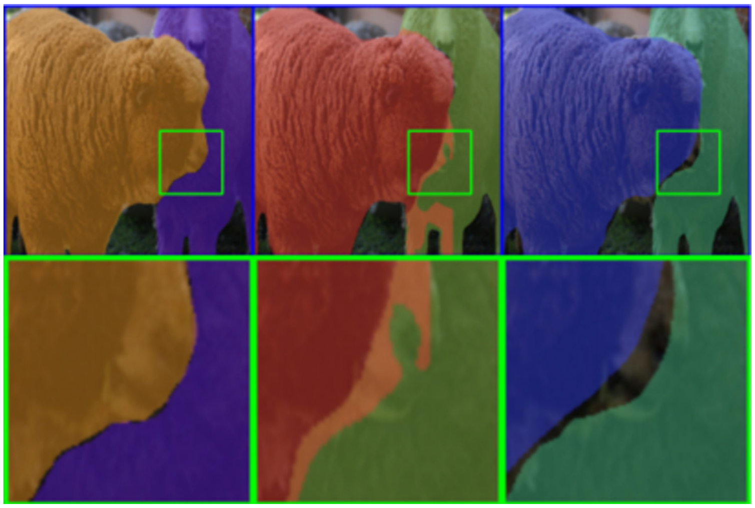


Fig. 9 Comparison of contour optimization effects.

Aiming at the problem that the quality of detection affects the integrity of the mask, Chen et al.[107] proposed a method that can learn the association between object features and bounding boxes. It can provide a more accurate edge contour for instance segmentation and has been effectively improved. Zhao et al.[108] proposed an instance segmentation model for the accuracy of segmentation contours, which used detection and segmentation as a multi-stage process to obtain accurate segmentation edges and improve the geometric regularity of the segmentation results.

The edge contour optimization of instance segmentation can not only improve the accuracy of segmentation but also greatly improve the recognition and trust of humans for computer vision instance segmentation. It is very important to improve the quality of segmentation and is a key factor that affects the final effect of instance segmentation.

5 Conclusion

Instance segmentation is an important problem of computer vision, and it is the most challenging research topic among the four basic tasks of computer vision. It has an important influence on the development of intelligent driving, security monitoring, medical health, etc. With the continuous development of CNNs, instance segmentation technology has become a current research hotspot and also a current research difficulty. This paper first reviews the traditional image segmentation methods, and on this basis, a comprehensive discussion of object instance segmentation based on deep learning. Subsequently, the common datasets in the field of instance segmentation and their respective characteristics are introduced in detail, and the key issues affecting the development of instance segmentation are discussed at the end.

The current instance segmentation technology mainly has problems with different algorithm frameworks, small object image instance segmentation, and edge contour blurring. Different algorithm frameworks have different advantages in object instance segmentation technology due to their different solution ideas. For example, detection-based methods generally have higher accuracy, and single-stage methods generally have faster speed. A reasonable selection of the algorithm framework according to the specific application field can effectively improve the segmentation results. Small object images often cause segmentation difficulties and inaccurate segmentation due to their object occupies fewer pixels in the image. For example, the segmentation ability of small objects in the field of remote sensing detection is particularly important. The solution of such problems can directly promote the wide application of object instance segmentation technology in related fields. Fuzzy edge contour is a common problem faced by object instance segmentation. Accurate edge contour can directly increase human trust in the object instance segmentation technology and can promote the application of this technology in different fields.

This paper proposes that the instance segmentation technology will mainly have the following development directions in the future.

(1) In the actual environment, it is often necessary to deal with the problem of multi-scale object segmentation. Multi-scale object segmentation requires accurate segmentation of large and small objects in the image at the same time, which is more difficult than the small object segmentation problem and has always been an important issue that affects the accuracy of instance segmentation technology. Improving the segmentation effect of small objects and the segmentation ability of multi-scale objects is an important way to improve the accuracy of the algorithm, and it is an important research direction in the field of instance segmentation.

(2) The accuracy of instance segmentation directly affects whether it can be applied in practice, and low-quality weather conditions such as rain, snow, and fog will inevitably be encountered in fields such as intelligent driving. The low-quality images collected in these weathers directly affect the effect of instance segmentation. How to combine cameras with sensors such as lidar to improve the robustness of the algorithm in various environments is an important development direction in the future.

(3) Because the segmentation accuracy of the overall instance segmentation technology is low, the main core problem at present is how to improve the accuracy of instance segmentation. As a result, the model is often more complicated, difficult to train, and difficult to deploy to the mobile terminal. How to simplify the model structure and reduce the hardware requirements without affecting the accuracy of the model is an important issue for its current application in a specific environment, and it is an important direction of current development.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant Nos. 2020YFB1313400 and 2020YFB1713302. This job was also supported by the Natural Science Foundation of Shaanxi Province of China (No. 2021JM-344) and Shaanxi Key Laboratory of Intelligent Processing for Big Energy Data (No. IPBED7). The authors declare no conflicts of interest.

References

1. A. Krizhevsky, L. Sutskever, and G. Hinton, “ImageNet classification with deep convolutional neural networks,” *Commun. ACM* **60**(6), 84–90 (2017).
2. K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arxiv.org/abs/1409.1556 (2014).
3. C. Szegedy et al., “Going deeper with convolutions,” in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 1–9 (2015).
4. K. He et al., “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 770–778 (2016).
5. R. Girshick et al., “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 580–587 (2014).
6. R. Girshick, “Fast R-CNN,” in *Proc. IEEE Int. Conf. Comput. Vision*, pp. 1440–1448 (2015).
7. S. Ren et al., “Faster R-CNN: towards real-time object detection with region proposal networks,” in *Adv. Neural Inf. Process. Syst.*, pp. 91–99 (2015).
8. J. Redmon et al., “You only look once: unified, real-time object detection,” in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 779–788 (2016).
9. W. Liu et al., “SSD: single shot multibox detector,” *Lect. Notes Comput. Sci.* **9905**, 21–37 (2016).

10. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, Boston, pp. 3431–3440 (2015).
11. L. Chen et al., "DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.* **40**(4), 834–848 (2018).
12. H. Zhoa et al., "Pyramid scene parsing network," in *Proc. 30th IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, Honolulu, pp. 6230–6239 (2017).
13. G. Lin et al., "RefineNet: multi-path refinement networks for high-resolution semantic segmentation," in *Proc. 30th IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, Honolulu, pp. 5168–5177 (2017).
14. K. He et al., "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vision*, pp. 2980–2988 (2017).
15. Z. Huang et al., "Mask scoring R-CNN," in *Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recognit.* (2019).
16. A. He et al., "A twofold Siamese network for real-time object tracking," in *Proc. 31st IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, Salt Lake City, pp. 4834–4843 (2018).
17. A. Rangesh and M. Trivedi, "No blind spots: full-surround multi-object tracking for autonomous vehicles using cameras and LiDARs," *IEEE Trans. Intell. Veh.* **4**(4), 588–599 (2019).
18. P. Zhang et al., "Non-rigid object tracking via deep multi-scale spatial-temporal discriminative saliency maps," *Pattern Recognit.* **100**, 107130 (2020).
19. M. Fiaz, A. Mahmood, and K. Soon, "Tracking noisy targets: a review of recent object tracking approaches," arxiv.org/abs/1802.03098 (2018).
20. D. Yang, J. Gan, and Y. Luo, "Urban road image segmentation algorithm based on statistical information," in *Proc. 26th Int. Conf. Geoinf.*, Kunming, China (2018).
21. Q. Wang and X. Liu, "Traffic sign segmentation in natural scenes based on color and shape features," in *Proc. IEEE Workshop Adv. Res. and Technol. Ind. Appl.*, Ottawa, Canada, pp. 374–377 (2014).
22. J. Nan and L. Bo, "Infrared object image instance segmentation based on improved mask-RCNN," *Proc. SPIE* **11187**, 111871E (2019).
23. H. Su, S. Wei, and S. Liu, "HQ-ISNet: high-quality instance segmentation for remote sensing imagery," *Remote Sens.* **12**(6), 989 (2020).
24. J. Lu et al., "An instance segmentation based framework for large-sized high-resolution remote sensing images registration," *Remote Sens.* **13**(6), 1657 (2021).
25. J. Ran, F. Yang, and C. Gao, "Adaptive fusion and mask refinement instance segmentation network for high resolution remote sensing images," in *Proc. IEEE Int. Geosci. and Remote Sensing Symp.*, Waikoloa, pp. 2843–2846 (2020).
26. I. Karakaya, B. Demirel, and O. Oztork, "HVLSeg: an ensemble model for instance segmentation on satellite images," in *Proc. 28th Signal Process. and Commun. Appl. Conf.*, Gaziantep, Turkey (2020).
27. G. Xu, Z. Song, and Z. Sun, "CAMEL: a weakly supervised learning framework for histopathology image segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vision*, Seoul, South Korea, pp. 10681–10690 (2019).
28. M. Rezaei, H. Yang, and C. Meinel, "Instance tumor segmentation using multitask convolutional neural network," in *Proc. Int. Joint Conf. Neural Networks*, Rio de Janeiro, Brazil (2018).
29. T. Konopczynski, R. Heiman, and P. Woznicki, "Instance segmentation of densely packed cells using a hybrid model of U-net and mask R-CNN," *Lect. Notes Comput. Sci.* **12415**, 626–635 (2020).
30. T. Prangemeier, C. Reich, and H. Koeppl, "Attention-based transformers for instance segmentation of cells in microstructures," in *Proc. IEEE Int. Conf. Bioinf. and Biomed.*, Seoul, South Korea, pp. 700–707 (2020).
31. H. Song et al., "Gradient-driven parking navigation using a continuous information potential field based on wireless sensor network," *Inf. Sci.* **408**(2), 100–114 (2017).

32. Z. Sun et al., "Energy balance-based steerable arguments coverage method in WSNs," *IEEE Access* **6**, 33766–33773 (2018).
33. W. Wei et al., "Algorithm research of known-plaintext attack on double random phase mask based on WSNs," *J. Internet Technol.* **20**(1), 39–48 (2019).
34. Q. Xu et al., "GI/Geom/1 queue based on communication model for mesh networks," *Int. J. Commun. Syst.* **27**(11), 3013–3029 (2014).
35. J. Su, H. Song, and H. Wang, "CDMA-based anti-collision algorithm for EPC global C1 Gen2 systems," *Telecommun. Syst.* **67**(3), 63–71 (2018).
36. H. Song, H. Wang, and X. Fan, "Research and simulation of queue management algorithms in ad hoc networks under DDoS attack," *IEEE Access* **5**, 27810–27817 (2017).
37. X. Yang, P. Shen, and B. Zhou, "Holes detection in anisotropic sensornets: topological methods," *Int. J. Distrib. Sens. Netw.* **8**(10), 135054–135062 (2012).
38. X. Fan et al., "H infinity control of network control system for singular plant," *Inf. Technol. Control* **47**(1), 140–150 (2018).
39. Y. Qi, "Information potential fields navigation in wireless ad-hoc sensor networks," *Sensors* **11**(2), 4794–4807 (2011).
40. X. Xia et al., "Multi-sink distributed power control algorithm for cyber-physical-systems in coal mine tunnels," *Comput. Netw.* **161**, 210–219 (2019).
41. W. Li et al., "System modelling and performance evaluation of a three-tier cloud of things," *Future Gener. Comput. Syst.* **70**, 104–125 (2017).
42. W. Wei et al., "Imperfect information dynamic Stackelberg game based resource allocation using hidden Markov for cloud computing," *IEEE Trans. Serv. Comput.* **11**(1), 78–89 (2018).
43. B. Zhou, D. Polap, and M. Wozniak, "A regional adaptive variational PDE model for computed tomography image reconstruction," *Pattern Recognit.* **92**, 64–81 (2019).
44. Y. Qiang and J. Zhang, "A bijection between lattice-valued filters and lattice-valued congruences in residuated lattices," *Math. Probl. Eng.* **36**(8), 4218–4229 (2013).
45. X. Yang et al., "Combined energy minimization for image reconstruction from few views," *Math. Probl. Eng.* **2012**, 154630 (2012).
46. H. Srivastava et al., "A local fractional integral inequality on fractal space analogous to Anderson's inequality," *Abstract Appl. Anal.* **46**(8), 5218–5229 (2014).
47. X. Fan, H. Song, and H. Wang, "Video tamper detection based on multi-scale mutual information," *Multimedia Tools Appl.* **78**(19), 27109–27126 (2019).
48. G. Chen et al., "Fully convolutional neural network with augmented atrous spatial pyramid pool and fully connected fusion path for high resolution remote sensing image segmentation," *Appl. Sci.* **9**(9), 1816 (2019).
49. S. Liu, W. Li, and D. Du, "Fractal intelligent privacy protection in online social network using attribute-based encryption schemes," *IEEE Trans. Comput. Social Syst.* **5**(3), 736–747 (2018).
50. X. Li et al., "Study on remote sensing image vegetation classification method based on decision tree classifier," in *Proc. IEEE Symp. Ser. Comput. Intell.*, Bengaluru, India, pp. 2292–2297 (2018).
51. Q. Ke et al., "Accurate and fast URL phishing detector: a convolutional neural network approach," *Comput. Netw.* **178**, 107275 (2020).
52. N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst. Man Cybern.* **9**(1), 62–66 (1979).
53. J. Yen, F. Chang, and S. Chang, "A new criterion for automatic multilevel thresholding," *IEEE Trans. Image Process.* **4**(3), 370–378 (1995).
54. H. Ye and S. Yan, "Double threshold image segmentation algorithm based on adaptive filtering," in *Proc. IEEE 2nd Inf. Technol., Networking, Electron. and Autom. Control Conf.*, Chengdu, China, pp. 1008–1011 (2017).
55. F. Da and H. Zhang, "Sub-pixel edge detection based on an improved moment," *Image Vis. Comput.* **28**(12), 1645–1658 (2010).
56. M. Baydoun and M. Al-Alaoui, "Modified edge detection for segmentation," in *Proc. Int. Symp. Signals, Circuits and Syst.*, Iasi, Romania (2015).

57. B. Sumengen and B. Manjunath, “Multi-scale edge detection and image segmentation,” in *Proc. 13th Eur. Signal Process. Conf.*, Antalya, Turkey, pp. 1–4 (2005).
58. W. Zhen and Y. Meng, “A fast clustering algorithm in image segmentation,” in *Proc. 2nd Int. Conf. Comput. Eng. and Technol.*, Chengdu, China, pp. 592–594 (2010).
59. Y. Shi, Z. Chen, and Z. Qi, “A novel clustering-based image segmentation via density peaks algorithm with mid-level feature,” *Neural Comput. Appl.* **28**, 29–39 (2017).
60. L. Cong et al., “Image segmentation algorithm based on superpixel clustering,” *IET Image Proc.* **12**(11), 2030–2035 (2018).
61. A. Rosenfeld, “The max Roberts operator is a Hueckel-type edge detector,” *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-3**(1), 101–103 (1981).
62. L. Yang et al., “An improved Prewitt algorithm for edge detection based on noised image,” in *Proc. 4th Int. Cong. Image and Signal Process.*, Shanghai, China, pp. 1197–1200 (2011).
63. W. Gao et al., “An improved Sobel edge detection,” in *Proc. 3rd IEEE Int. Conf. Comput. Sci. and Inf. Technol.*, Chengdu, China, pp. 67–71 (2010).
64. J. Canny, “A computational approach to edge detection,” *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-8**(6), 679–698 (1986).
65. M. Ayech and D. Ziou, “Segmentation of Terahertz imaging using k-means clustering based on ranked set sampling,” *Expert Syst. Appl.* **42**(6), 2959–2974 (2015).
66. J. Noordam, W. van den Broek, and L. Buydens, “Geometrically guided fuzzy C-means clustering for multivariate image segmentation,” in *Proc. 15th Int. Conf. Pattern Recognit.*, Barcelona, Spain, pp. 462–465 (2000).
67. R. Achanta et al., “SLIC superpixels compared to state-of-the-art superpixel methods,” *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(11), 2274–2282 (2012).
68. P. Shen et al., “A near-infrared face detection and recognition system using ASM and PCA + LDA,” *J. Networks* **9**(10), 2728–2733 (2014).
69. F. Yang et al., “DDTree: a hybrid deep learning model for real-time waterway depth prediction and smart navigation,” *Appl. Sci.* **10**(8), 2770 (2020).
70. K. E. Van de Sande et al., “Segmentation as selective search for object recognition,” in *Proc. IEEE Int. Conf. Comput. Vision*, Barcelona, Spain, pp. 1879–1886 (2011).
71. V. Sanchez A, “Advanced support vector machines and kernel methods,” *Neurocomputing* **55**(1), 5–20 (2003).
72. A. Patle and D. Chouhan, “SVM kernel functions for classification,” in *Proc. Int. Conf. Adv. Technol. and Eng.*, Mumbai, India (2013).
73. J. Dai, K. He, and J. Sun, “Instance-aware semantic segmentation via multi-task network cascades,” in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, WA (2016).
74. S. Liu et al., “Path aggregation network for instance segmentation,” in *Proc. 31st IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, Utah (2018).
75. K. Chen et al., “Hybrid task cascade for instance segmentation,” in *Proc. 32nd IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, Long Beach, CA (2019).
76. W. Li, Y. Shi, and W. Yang, “Interactive image segmentation via cascaded metric learning,” in *Proc. IEEE Int. Conf. Image Process.*, Quebec City, Canada, pp. 2900–2904 (2015).
77. B. Bue et al., “Metric learning for hyperspectral image segmentation,” in *Proc. 3rd Workshop Hyperspectral Image and Signal Process.: Evol. Remote Sens.*, Lisbon, Portugal (2011).
78. Y. Kong, D. Wang, and L. Shi, “Adaptive distance metric learning for diffusion tensor image segmentation,” *PLoS One* **9**(3), e92069 (2014).
79. N. Gao et al., “SSAP: single-shot instance segmentation with affinity pyramid,” in *Proc. IEEE/CVF Int. Conf. Comput. Vision*, Seoul, South Korea (2019).
80. B. Brabandere, D. Neven, and L. Gool, “Semantic instance segmentation with a discriminative loss function,” arXiv:1708.02551 (2017).
81. M. Bai and R. Urtasun, “Deep watershed transform for instance segmentation,” in *Proc. 30th IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, Honolulu, HI (2017).
82. A. Kirillov et al., “InstanceCut: from edges to instances with MultiCut,” in *Proc. 30th IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, Honolulu, HI (2017).

83. A. Arnab and P. Torr, "Pixelwise instance segmentation with a dynamically instantiated network," in *Proc. 30th IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, Honolulu, HI (2017).
84. S. Liu et al., "SGN: sequential grouping networks for instance segmentation," in *Proc. 16th IEEE Int. Conf. Comput. Vision*, Venice, Italy (2017).
85. D. Bolya et al., "YOLACT: real-time instance segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vision*, Seoul, South Korea (2019).
86. D. Bolya et al., "YOLACT++: better real-time instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.* (2019).
87. E. Xie et al., "PolarMask: single shot instance segmentation with polar representation," in *Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, Seattle (2020).
88. X. Wang et al., "SOLO: segmenting objects by locations," *Lect. Notes Comput. Sci.* **12363**, 649–665 (2020).
89. X. Wang et al., "SOLOv2: dynamic and fast instance segmentation," arxiv.org/abs/2003.10152 (2020).
90. H. Chen et al., "BlendMask: top-down meets bottom-up for instance segmentation," arxiv.org/abs/2001.00309 (2020).
91. Z. Tian et al, "FCOS: fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vision*, Seoul, South Korea (2019).
92. M. Everingham et al., "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.* **88**(2), 303–338 (2010).
93. M. Cordts et al., "The Cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, Seattle, pp. 3213–3223 (2016).
94. T. Lin et al., "Microsoft COCO: common objects in context," *Lect. Notes Comput. Sci.* **8693**, 740–755 (2014).
95. G. Neuhold et al., "The mapillary vistas dataset for semantic understanding of street scenes," in *Proc. 16th IEEE Int. Conf. Comput. Vision*, Venice, Italy (2017).
96. L. Qi et al., "Amodal instance segmentation with KINS dataset," in *Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, California (2019).
97. B. Hariharan et al., "Semantic contours from inverse detectors," in *Proc. IEEE Int. Conf. Comput. Vision*, Barcelona, Spain (2011).
98. X. Chen et al., "TensorMask: a foundation for dense object segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vision*, Seoul, South Korea, pp. 2061–2069 (2019).
99. S. Peng et al., "Deep snake for real-time instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, Seattle, pp. 8530–8539 (2020).
100. Y. Lecun, L. Bottou, and Y. Bengio, "Gradient-based learning applied to document recognition," *Proc. IEEE* **86**(11), 2278–2324 (2010).
101. A. Bochkovskiy, C. Wang, and H. Liao, "YOLOv4: optimal speed and accuracy of object detection," arxiv.org/abs/2004.10934 (2020).
102. C. Pang et al., "Exploring part-aware segmentation for fine-grained visual categorization," *Multimedia Tools Appl.* **77**(23), 30291–30310 (2018).
103. R. Hamaguchi, A. Fujita, and K. Nemoto, "Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery," in *Proc. 18th IEEE Winter Conf. Appl. Comput. Vision*, Nevada (2018).
104. K. Dijkstra et al., "CentroidNetV2: a hybrid deep neural network for small-object segmentation and counting," *Neurocomputing* **423**, 490–505 (2021).
105. X. Zhang and J. Cao, "Contour-point refined mask prediction for single-stage instance segmentation," *Acta Opt. Sin.* **40**(21), 2115001 (2020).
106. Z. Liang et al., "Research and implementation of instance segmentation and edge optimization algorithm," *J. Graphics* **41**(6), 939–946 (2020).
107. X. Chen et al., "Supervised edge attention network for accurate image instance segmentation," *Lect. Notes Comput. Sci.* **12372**, 617–631 (2020).
108. W. Zhao, C. Persello, and A. Stein, "Building instance segmentation and boundary regularization from high-resolution remote sensing images," in *Proc. IEEE Int. Geosci. and Remote Sens. Symp.*, Waikoloa, HI (2020).

Di Tian received his master's degree from Chang'an University. He is currently studying his PhD in vehicle engineering at Chang'an University. His research interests include computer vision and artificial intelligence.

Yi Han received his PhD from the Northwestern Polytechnical University, Xi'an, China, in 2003. He is currently a professor at the School of Automobile of Chang'an University. He has authored more than 20 papers and received more than 100 patents. His current research interests include vehicular networking and automatic driving technology.

Biyao Wang received her bachelor's degree in vehicle engineering from Xi'an University of Science and Technology, Xi'an, China. She is currently pursuing her PhD in vehicle engineering at the School of Automobile of Chang'an University. Her current research interests include automatic driving and hybrid vehicles.

Tian Guan received her bachelor's degree from Chang'an University, Xi'an, China, where she is currently pursuing her PhD in the Department of Vehicle Engineering. Her research interests are automatic driving and image processing.

Hengzhi Gu received his bachelor's degree in vehicle engineering from Hebei Agricultural University in 2019. He is currently a graduate student studying for his master's degree at the school of automobile of Chang'an University under the supervision of Prof. Han Yi. His research interests include artificial intelligence and its application in the field of automatic driving.

Wei Wei received his PhD in computer software and theory from Xi'an Jiaotong University in 2011. In 2009, he visited the University of Nebraska. In 2015, he completed postdoctoral research in electrical engineering at Xi'an University. In 2017, he visited and completed post-doctoral research at the University of Texas at Dallas. He is an associate professor and a senior member of IEEE. He has been studying the Internet of Things and big data.