Proceedings of CONF-MLA 2025 Symposium: Applied Artificial Intelligence Research
DOI: 10.54254/2755-2721/2025.BJ27326

An Introduction to Deep Learning-Based Object Segmentation Methods

Siyu Lu

School of Computer Science and Technology, Dongguan University of Technology, Dongguan, China
l2459934669@gmail.com

Abstract. Object segmentation is a challenging task that requires both instance-level and pixel-level predictions, and it has widespread applications in autonomous driving, video analysis, scene understanding, and other fields. This paper first categorizes object segmentation algorithms, then conducts an in-depth analysis of some representative algorithms and their improved versions, discusses the advantages and disadvantages of these algorithms, and finally provides an outlook on the future development of object segmentation methods. This paper It is found that the top-down detection algorithm represented by MaskR-CNN has the disadvantages of large computation and poor real-time performance; Bottom-up methods have a large gap in SSAP dependent post-processing and so on, but some breakthroughs in some new algorithms, such as Deep snake, etc.; the direct segmentation methods represented by the SOLO series achieve the accuracy and speed of the balance, while methods such as YOLACT make up for the shortcomings in sample segmentation. This paper also identifies current research gaps and outlines future directions.

Keywords: deep learning, object instance segmentation, pixel segmentation, computer vision

1. Introduction

Instance segmentation is a core problem in computer vision research, combining features of object detection and semantic segmentation. It can effectively identify each pixel and distinguish objects of the same type, with important applications in autonomous driving, medical imaging, and robot vision.

Current real-time object segmentation methods include three approaches: one is a two-step segmentation algorithm based on MaskR-CNN, which builds upon Fast R-CNN. Based on Transformer algorithms such as Detection Transformer (DETR) and MaskFormer, the segmentation problem is unified under mask classification, representing an innovative approach. Additionally, new algorithms like PointRend use point sampling to handle edge regions. In terms of performance, Mask2Former has achieved leading international standards in multiple evaluations and unified multiple segmentation tasks. This paper specifically explores the typical algorithmic principles, improvements, and experimental comparisons of three types of instance segmentation methods based on deep learning, and analyzes research shortcomings and future directions. By reviewing the structures and improvements of various algorithms and comparing their performance using metrics

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.BJ27326

such as AP and fps, this study can provide references for related industry applications, clarifying the advantages and disadvantages of algorithms and their applicable scenarios.

2. Top-down detection-based instance segmentation methods

2.1. Mask R-CNN

FasterR-CNN is a multi-object recognition method based on the Region Proposal Network (RPN) [1]. Based on FasterR-CNN, a mask branch is added to FasterR-CNN, enabling precise recognition of segmentation templates. In terms of integer errors, MaskR-CNN adopts the RoIAlign hierarchy to eliminate quantization errors in the RoIPool algorithm; it uses a bilinear interpolation algorithm to more accurately identify the corresponding features of each block, ensuring that the extracted features align with the input.

2.2. YOLACT

YOLACT is a real-time object segmentation algorithm that adds a mask branch to a single-level detector to achieve precise sample segmentation. It has significant advantages over algorithms like Single Shot MultiBox Detector (SSD) in image segmentation. Its main feature is adding a mask branch to a single-level detector to achieve precise sample segmentation [2].

YOLACT establishes k templates for each image and predicts k templates. The k template coefficients in each template correspond to the template, and then the positive and negative mask coefficients are linearly combined to obtain the final model prediction result [3].

2.3. BlendMask

The BlendMask architecture consists of a detection network and task branches, with the network structure shown in Figure 1. The tasks are divided as follows: the bottom-layer module extracts low-level detail features and predicts score maps. The top layer is primarily used to predict instance attributes, extracting certain instance-level information to guide the score template in capturing key position information and suppressing external regions, serving as a guide for instance perception; the boundary block integrates scores and attributes, fusing features to generate high-quality region templates, reducing the number of required channels, and producing high-quality masks without being constrained by the resolution of the top-layer sampling [4].

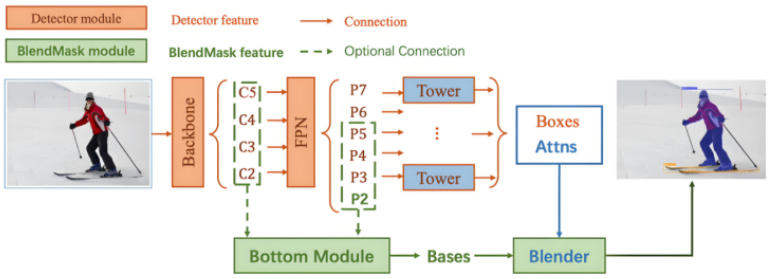

Figure 1. BlendMask network structure

DOI: 10.54254/2755-2721/2025.BJ27326

3. Bottom-up semantic segmentation based approach

3.1. Sequential grouping network of SGN for instance segmentation

To study image segmentation, two methods are used: one is based on multi-layer neural networks to achieve segmentation; the other is to use the watershed method, first do horizontal and vertical scanning of the image, and then divide the image into horizontal and vertical levels, respectively, to scan again and complete the segmentation. The LineNet network splices between the same labelled samples to form a complete sample template; for the segmentation of the samples caused by the occlusion, samples are spliced into the target samples, thus breaking through the difficulty of single sample decomposition [5]. The network structure is shown in Figure 2.

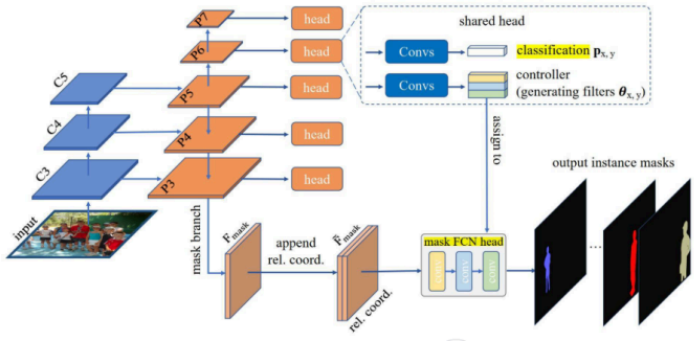

Figure 2. Sequential packet network

3.2. SSAP

The semantic segmentation-based no-proposal algorithm will take the similarity between examples as the basis and entry point, aiming at solving the 'bounding box' problem existing in the existing algorithms [6]. Gao et al. propose a single-stage no-proposal example segmentation method, SSAP [7]. First, a similarity pyramid structure based on similarity points is studied and the probability distribution of two pixels having the same indication is calculated; secondly, a combination of semantic segmentation and similarity pyramid is used to establish a multi-scale example prediction model; finally, a cascade graph is used to dissect into two steps, and the two steps are fused into a single one, and the network structure is shown in Figure 3.

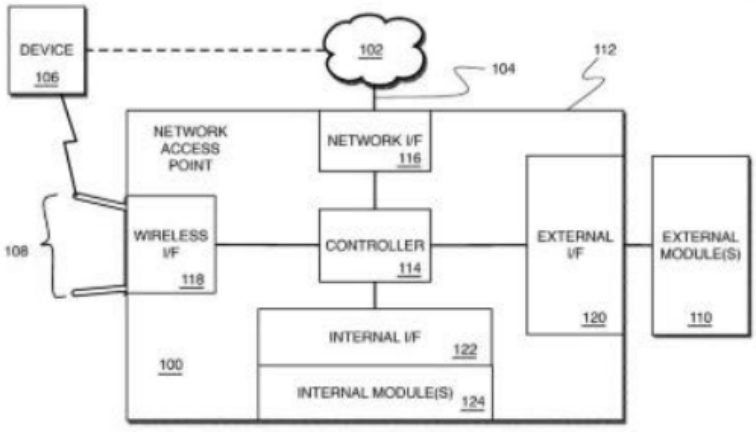

Figure 3. SSAP network structure

DOI: 10.54254/2755-2721/2025.BJ27326

4. Direct instance segmentation method

4.1. SOLO algorithm

The SOLO algorithm segments the image into S×S grids, and when the centre of gravity of the target falls into a particular grid, this grid will take on the following two tasks: a) classification of the target's objectives; and b) mask branches are used for the prediction of example masks of the object. In Figure 4, after SOLO uses the FPN for the backbone network, each layer of the FPN follows the two parallel branches in front of it for classification and localisation prediction; the number of grids in each branch varies according to the size of the samples, with smaller samples corresponding to a larger number of grids, and merging all the data gives a sample segmentation result for this object [8].

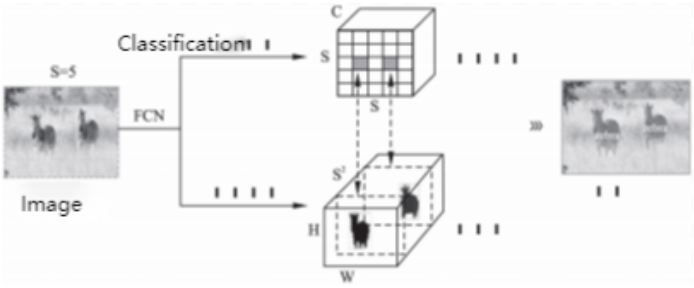

Figure 4. Framework of the SOLO model

4.2. SOLOv2

SOLOv2 makes two improvements to the SOLO algorithm: mask branching, and replacing NMS in SOLO with NMS matrix [9]. Therefore, SOLOv2 proposes to divide the task header into two parts, core and feature, in which the structure of the task header is shown in Figure 5: the core branch and feature branch extract features by convolution to form the mask layer, and at the maximum, at each layer of the grid there is There is a target, which can effectively filter out the purposeless grids, thus reducing the subsequent computation; The intersection and concatenation ratios between the masks are calculated by the maximum suppression matrix method, and the intersection and concatenation ratio matrix is used to exclude the amount of masks with higher intersection ratios, thus speeding up the computation of the maximum suppression algorithm [10].

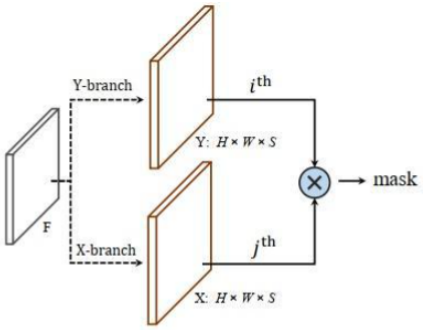

Figure 5. SOLOv2 mask header

DOI: 10.54254/2755-2721/2025.BJ27326

In SOLOv2, the real-time adjustment of multiple mask ratios is achieved through the simultaneous processing of multiple mask quantities, and the mAP of SOLOv2 in the ResNet-101 backbone network is 39.7%.

5. Experimental comparison of instance segmentation algorithms

The effect of instance segmentation algorithms in practical applications is mainly reflected in several aspects, such as accuracy, speed, and storage space. Measured by AP (AP), AP (AP) represents the average precision of the mask (AP), the mean value of the IoU threshold (AP), and different subscripts (AP50, AP75, AP75, AP75, AP75, small target, APM, APL) represent the computation results of the small, medium, and large targets [11]. Table 1 uses fps to measure the number of frames per second of the algorithm to segment the target. The memory footprint is also important if the model is used in devices with limited memory capacity.

Table 1. Comparison of segmentation speed of different algorithms

Classification of methods	methodologies	feature extraction	test set	AP(%)	hardware	ratio conversion	Time(ms)	frames per second (fps)
Top-down approach	Mask R-CNN	R-101-FPN	COCO	35.7	Titan Xp	86.3%	116.3	8.6
	MS R-CNN	R-101-FPN	COCO	38.3	Titan Xp	86.3%	116.3	8.6
	PANet	R-50-FPN	COCO	36.6	Titan Xp	86.3%	212.8	4.7
	FCIS	R-101-C5	COCO	29.5	Titan Xp	86.3%	151.5	6.6
	PolarMask	RX-101	COCO	32.9	2080Ti	103.2%	140.9	7.1
	YOLACT	R-101-FPN	COCO	29.8	Titan Xp	86.3%	29.8	33.5
	RetinaMask	R-101-FPN	COCO	34.7	Titan Xp	86.3%	166.7	6.0
	BlendMask-RT	R-101	COCO	36.8	1080Ti	84.5%	47.6	25
	Geometric Factors	R-101-FPN	COCO	33.2	1080Ti	84.5%	36.9	27.1
	TensorMask	R-101	COCO	37.3	1080Ti	84.5%	380.3	2.6
	EOLO	MobileNetv	COCO	11.7	1080Ti	84.5%	33.4	30
	Box2Pix	GoogLeNet	Cityscapes	13.1	1080Ti	84.5%	91.75	10.9
	FourierNet	R-50	COCO	24.3	2080Ti	103.2%	37.6	26.6

DOI: 10.54254/2755-2721/2025.BJ27326

Table 1. (continued)

Bottom-up approach	SGN	VGG16	Cityscapes	25.0	1080Ti	84.5%	1666.7	0.6
	Discriminative	R-38	Cityscapes	17.5	1080Ti	84.5%	200	5
	SSAP	R-50	Cityscapes	31.5	1080Ti	84.5%	263.1	3.8
	Deep snake	CenterNet	Pascal Voc	31.7	1080 Ti	84.5%	31.0	32.3
Direct segmentation method	EmbedMask	R-101-FPN	COCO	35.2	V100	100%	46.1	21.7
	ShapeMask	R-101-FPN	COCO	37.4	V100	100%	149	6.7
	Jointly Optimizin	ERFNet	Cityscapes	27.6	1080 Ti	84.5%	91	11
	SOLO	R-50-FPN	COCO	34.2	V100	100%	44.5	22.5
	SOLOv2	R-50-FPN	COCO	37.1	V100	100%	31.9	31.3

Table 1 lists the accuracy and speed of several typical algorithms. After comparison, it can be seen that the top-down method is more accurate in segmentation, but there is a big difference in the speed of segmentation. The speed of the test is influenced by the hardware facilities. Therefore, V100 can be used as a reference and its performance ratio with other graphic cards can be converted so that the reader can understand the effect of hardware on speed. It can be found that algorithms such as MaskR-CNN and its improved algorithms, TensorMask and RetinaMask have good segmentation accuracy but poor real-time performance [12]; algorithms such as YOLACT and BlendMask improve the accuracy of computation while ensuring real-time performance of the network; and algorithms such as Box2Pix and EOLO use simple, lightweight feature extraction network to accelerate segmentation, but did not achieve effective segmentation [13].

6. Conclusion

This paper focuses on deep learning-based object sample segmentation techniques, and through the study of various classifications, it clarifies the advantages and shortcomings of each, and gives an outlook on the future research direction. It is found that the top-down detection algorithm represented by MaskR-CNN has the disadvantages of large computation and poor real-time performance; Bottom-up methods have a large gap in SSAP dependent post-processing and so on, but some breakthroughs in some new algorithms, such as Deep snake, etc.; the direct segmentation methods represented by the SOLO series achieve the accuracy and speed of the balance, while methods such as YOLACT make up for the shortcomings in sample segmentation. In this thesis, the performance of instance segmentation algorithms in extreme scenarios such as low-light and severe occlusion is not thoroughly investigated, and the impact of different backbone networks on their performance is not quantitatively analysed. Future research can focus on exploring how the instance segmentation method can improve the robustness of complex scenes and optimise the effect of small target segmentation, exploring the lightweight model framework, and integrating with Transformer and other technologies to solve the bottlenecks of the algorithms in terms of computational accuracy and speed, and promote their wide application in more real-world scenes.

DOI: 10.54254/2755-2721/2025.BJ27326

References

[1] Charisis C, Wang M, Argyropoulos D.Instance segmentation of oyster mushroom datasets: A novel data sampling methodology for training and evaluation of deep learning models [J].Smart Agricultural Technology, 2025, 12101146-101146.
[2] Zhu Y, Wang Z, Ye Q, et al.SPA-Net: An Offset-Free Proposal Network for Individual Tree Segmentation from TLS Data [J].Remote Sensing, 2025, 17(13): 2292-2292.
[3] Nasim A, Anwar M, Owais M, et al.Regressor driven incremental instance segmentation for contraband items identification [J].Engineering Applications of Artificial Intelligence, 2025, 157111421-111421.
[4] Dhathri M, Mustafa M A, Prafulla K, et al.Enhanced Structural Damage Detection, Segmentation, and Quantification Using Computer Vision and Deep Learning [J].Journal of Computing in Civil Engineering, 2025, 39(5):
[5] Charisis C, Nuwayhid S, Argyropoulos D.A novel Mask R-CNN-based tracking pipeline for oyster mushroom cluster growth monitoring in time-lapse image datasets [J].Computers and Electronics in Agriculture, 2025, 237(PA): 110590-110590.
[6] Liu Y,Yang S,Li B,et al． Affinity derivation and graph merge for instance segmentation[C]/ / Proceedings of the European Confer- ence on Computer Vision,2018 : 686-703.
[7] Gao N,Shan Y,Wang Y,et al． SSAP: single-shot instance segmen- tation with affinity pyramid[C]/ / Proceedings of the IEEE Interna- tional Conference on Computer Vision,2019 : 642-651.
[8] Awan M, Whangbo K T, Shin J.Deep learning methods for autonomous driving scene understanding tasks: A review [J].Expert Systems With Applications, 2025, 287128098-128098.
[9] Akrouchi E M, Mhada M, Gracia R D, et al.Optimizing Mask R-CNN for enhanced quinoa panicle detection and segmentation in precision agriculture [J].Frontiers in Plant Science, 2025, 161472688-1472688.
[10] Huang D Y, Jadav N, Rutter G, et al.A generalist deep-learning volume segmentation tool for volume electron microscopy of biological samples. [J].Journal of structural biology, 2025, 217(3): 108214.
[11] Vogelgesang M, Kaczmarek V, Lopes P C D A, et al.Automated material flow characterization of WEEE in sorting plants using deep learning and regression models on RGB data. [J].Waste management(New York, N.Y.), 2025, 204114904.
[12] Ren X, Zhao P, Le X, et al.Intelligent detection of traveling ionospheric disturbances from dense GNSS TEC observations utilizing instance segmentation model Cascade Mask R-CNN [J].GPS Solutions, 2025, 29(3): 118-118.
[13] PetrovićE, SimonovićM, JovanovićD D, et al.Deep learning-based instance segmentation for detection of tire tread area [J].Progress in Engineering Science, 2025, 2(2): 100080-100080.