JST: Engineering and Technology for Sustainable Development
Volume 31, Issue 3, July 2021, 097-104

Improve of Mask R-CNN in Edge Segmentation

Cải thiện Mask R-CNN cho phân đoạn vùng biên vật thể

Hoang Hong Hai*, Tran Bao Long
School of Mechanical Engineering, Hanoi University of Science and Technology, Hanoi, Vietnam
**Email: hai.hoanghong@hust.edu.vn*

Abstract

Nowadays, grasping robot plays an important role in many automatic systems in the industrial environment. An excellent grasping robot can detect, localize, and pick objects accurately but to perfectly achieve these tasks, it is still a challenge in the computer vision field. Especially, segmentation task, which is understood as both detection and localization, is the hardest problem. To deal with this problem, the state-of-the-art Mask Region Convolution Neural Network (Mask R-CNN) was introduced and obtained an exceptional result. But this superb model does not certainly perform well when working with harsh locations of objects. The edge and border regions are usually misunderstood as the background, this leads to the failure in localizing objects to submit a good grasping plan. Thus, in this paper, we introduce a novel method that combines the original Mask R-CNN pipeline and 3D algorithms branch to preserve and classify the edge region. This results from the improvement of the performance of Mask R-CNN in detailed segmentation. Concretely, the significant improvement practiced in harsh situations of object location was obviously discussed in the experimental result section. Both IoU and mAP indicators are increased. Specifically, mAP, which directly reflects the semantic segmentation ability of a model, raised from 0.39 to 0.46. This approach opens a better way to determine the object location and grasping plan.

Keywords: detection, edge, 3D segmentation, Mask R-CNN.

Tóm tắt

Ngày nay, rô-bốt gắp vật đóng một vai trò quan trọng trong nhiều hệ thống tự động hóa trong môi trường công nghiệp. Một hệ thống rô-bốt tối ưu có khả năng xác định, định vị và gắp đối tượng chính xác. Tuy nhiên, để tối ưu những nhiệm vụ này, nhiệm vụ trong xử lý ảnh vẫn là một thử thách lớn. Đặc biệt là trong những nhiệm vụ phân tách đối tượng khỏi bức ảnh là vấn đề khó nhất. Để xử lý vấn đề này, mô hình Mask R-CNN được giới thiệu và đạt được những kết quả nhất định. Nhưng mô hình này không thực sự tốt trong bài toán nhận diện vị trí đối tượng trong môi trường khó. Cạnh và đường bao thường bị nhầm thành nền, do đó nó dẫn đến sai số trong định vị vị trí của đối tượng để đưa ra chiến lược gắp vật. Trong bài báo này, chúng tôi đưa ra một phương pháp tiếp cận mới phối hợp mạng Mask R-CNN gốc và giải thuật 3D để tối ưu và phân loại cạnh. Điều này đã giúp thu được sự cải thiện trong khả năng phân loại chi tiết của Mask R-CNN. Cụ thể, thực nghiệm đã cải thiện đáng kể kết quả phân đoạn vật trong các trường hợp khắc nghiệt về vị trí đặt của vật thể. Các kết quả thí nghiệm được làm rõ hơn trong chương kết quả thí nghiệm. Cả hai chỉ số IoU và mAP đã được cải thiện. Đặc biệt, chỉ số mAP, chỉ số phản ánh khả năng dự đoán của mô hình trong nhiệm vụ phân đoạn vật thể đã tăng từ 0.39 tới 0.46. Cách tiếp cận này đưa ra một phương pháp tốt hơn để xác định vị trí của vật và chiến lược gắp vật.

Từ khóa: nhận dạng, cạnh, phân đoạn 3D, Mạng nơ-ron Mask-CNN

1. Introduction

Instance segmentation is an important and challenging problem in computer vision tasks. But with the rapid development of Convolution Neural Network (CNNs), this difficult task is almost perfectly achieved. To understand the development of Mask R-CNN model, back in the releasing of The Region based Convolution Neural Network (R-CNN) [1], which was designed to deal with object detection. It got a good result, but it was not enough for our demand. Then, Fast R-CNN [2] did not only improve the accuracy in detection, but also improve the computation cost. Dramatical performance of Faster R-CNN [3] then reaches the excellent in this task by modifying the architecture of Fast R-CNN.

Regarding instance segmentation, Mask R-CNN [4] developed the state-of-the-art object detection model by adding a predicting mask branch. The mask covered the object region and distinguished them. The model achieved exceptional masks, which covered almost the object though working under multi-classes situations. Nonetheless, the masks were not always accomplished mask, the edge and boundary regions were misclassified and treated as the background. It was driven from the obstacle and closed location of objects. The boundary region made the hesitation in considering the classifier. To handle

ISSN: 2734-9381
https://doi.org/10.51316/jst.151.etsd.2021.31.3.17
Received: February 15, 2021; accepted: July 07, 2021

this problem, some recent proposed models as Mask Scoring R-CNN (MS R-CNN) [5] or Mask Refined R-CNN (MR R-CNN) [6] enhanced the mask quality. MS R-CNN added a scoring path to supervise the predicted mask. The confidence score and mask quality would be more satisfied. MR R-CNN upgraded the mask quality by combining the features layers, which focused on the global and detailed information.

Meanwhile, 3D algorithms could easily preserve the edge region but hard for instance segmentation task. The performance of 2D models is relied on the image brightness or object obstacle, while 3D approaches are painful to deal with instance segmentation requires. Hence, different from other improving models, we introduce a novel method that combining original Mask R-CNN and 3D algorithms branch. Original Mask R-CNN plays its role in generating the rough mask while 3D process branch saves the full object by applying Difference of Normals-Based Segmentation [7]. The edge and boundary region then will be distinguished completely by the difference in original mask, which does not contain border region and the consequence of the 3D algorithm. The important mission that cannot be completed by 2D approaches is considering the detailed edge region. Therefore, our model is necessary for classifying these regions. The spatial relationship between original masks and edge regions is considered as the reference for classifying. Concretely, this process is achieved by applying Euclidean Cluster Extraction [8]. This general development opens an opportunity for improving the performance of other 2D principles. The exceptional consequence of our method delivers a full acknowledgment of object to propose a better grasping plan.

This paper contains the following contexts: Section 2 details the related research to our goal. The detailed methodology is presented in section 3. Section 4 gives the experiments and results of our method. Finally, the discussion and conclusion are in last two sections.

2. Related Work

Dealing with instance segmentation, both 2D and 3D approaches have their own way, but they have specific advantages and disadvantages. Regarding 2D approach, many recent models based on Mask R-CNN achieve an excellent result in detailed segmentation.

- By combining with Grabcut [9], Improvement of the Mask R-CNN object segmentation algorithm [10] was introduced to get the misclassified regions. But Grabcut is incentive with the contrast of image or the lack of brightness. Moreover, the consequence did not categorize the boundary region.
- Modifying original Mask R-CNN by proposing Side Fusion feature pyramid network (SF-FPN) and replacing initial Resnet-101 backbone with Resnet-86 are the main contributions of Fast Vehicle and Pedestrian Detection Using Improved Mask R-CNN [11], which increased the accuracy of detailed information.
- To supervise the quality of the mask, which is usually poor while getting a high confidence score of original Mask R-CNN, Mask Scoring R-CNN (MS R-CNN) learns the quality of predicted mask by a network block. This development satisfies the relationship between instance mask and its confidence.
- A method was proposed in Mask Refined R-CNN to focus on global and detailed information. This task is achieved through a new semantic segmentation layer, which plays a role in learning feature fusion. This layer constructs a feature pyramid network and summing the transmissions of the same resolution. Comparing to Path Aggregation Network (PAN) [12], MR R-CNN significantly outperforms the prediction in large objects.

In terms of dealing with 3D data, PointNet [13] was introduced to deal with several missions such as object detection, part segmentation, and scene semantic parsing, while other 3D architecture considers the input data as 3D voxel grid or collections of images. This model directly works with not only three coordinates (x, y, z), but also computing normal and other global and local features. The novel architecture exceeds the result on standard benchmarks. Nonetheless, almost 3D architectures are heavy and need a large amount of computation space. It uses an irregular format of input data, which leads to adversity in implementation and customization. 2D architectures promptly achieve the segmentation goal but they are usually weak at detailed information or boundary region. Combining the 2D architectures and 3D algorithms is a comfortable way to satisfy the disadvantage while gaining the advantage of both two styles of approaching.

3. Methodology

3.1. Mask R-CNN Pipeline

With the advent of Region-Based Convolution Network (R-CNN), the object detection task is obtained with excellent accuracy. The entire image is fed to Selective Search [14] algorithm to extract proposal regions, which potentially contain object. These proposals are then separately put into a single convolution neural network (CNN) [15] to compute the feature. Although these regions have a number of the same parts the non-sharing studying feature of R-CNN makes the time-consuming of inference time.

The researchers consistently develop the previous model to improve its performance day by day. As the result, Fast R-CNN is then proposed to decrease the computation cost of R-CNN by studying features of the entire image by a single CNN before surpassing the Selective Search algorithm. But it is still not enough for the demand, therefore, Faster R-CNN is then introduced to get a significant improvement. Instead of using Selective Search to propose the potential region, Faster R-CNN uses a network called Region Proposal Network. This network extracts fewer proposal regions leading to less computation cost. The object detection task is perfectly handled by this model. Taking this advantage, Mask R-CNN upgraded the architecture for instance segmentation task. A mask predict branch is added into Faster R-CNN to predict the mask of object at the same time of regression branch. Moreover, ROI Align is used to guard the mask accuracy. Nonetheless, the effort of Mask R-CNN cannot deal with pixels at edge region. Especially, when the experiments are practiced under harsh circumstances. Therefore, we propose a novel method that enhances the accuracy of the mask based on spatial relationship, which cannot be implemented by 2D approach. By applying our method, the pixel at the edge or boundary can be preserved and classified leading to a good plan for grasping robot [16-18]. A brief view of our research can be given in Fig. 1 below.

3.2. Method

For segmenting the edge regions and preserving the misclassified region, two applied particular 3D segmentation algorithms are Difference of Normals (DoN) Based Segmentation and Euclidean Cluster Extraction. The big hurdle suspending the performance of Mask R-CNN is the disorganized, overlapping location of objects in robot grasping field. Besides, 3D camera used in this field can acquire the distance from object to itself, this helpful data promotes the awareness of entire objects. Encouraged by this advantage, 3D segmentation algorithms are chosen for improving the accuracy of Mask R-CNN. Computation with 3D data is generally time consuming so that the down sampling step is necessary for this approach. To roughly decline the background, we used Random Sample Consensus (RANSAC) [19]. All objects are located on a plane inside the workspace of robot. RANSAC can easily deny the place, which objects are located on. Meanwhile, as a similar objective, Grabcut neglects the background based on the color model of an input image, which is hard to reach by 3D camera. DoN only depends on the difference of normals of points, so it does not abandon any piece of object. DoN compares at each point **p** the responses of the operator across two different radii $r_s < r_l$. Fig. 2. illustrates the effect of support radius on estimated surface normal for a point cloud. Formally, the Difference of Normals operator of any point **p** in a point cloud is defined as follows [7, 20].

$$\Delta_{\hat{n}}(p, r_s, r_l) = \frac{\hat{n}(p, r_s) - \hat{n}(p, r_l)}{2} \quad (1)$$

where $r_l, r_s \in \mathbb{R}$, $r_s < r_l$, and $\hat{n}(p, r)$ is the surface normal estimate at point **p**, at the given *r*.

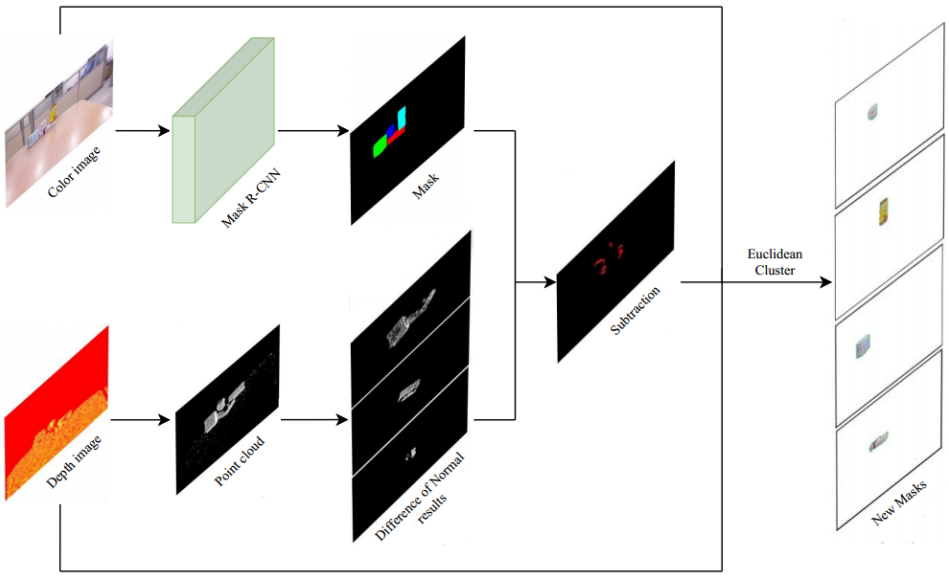
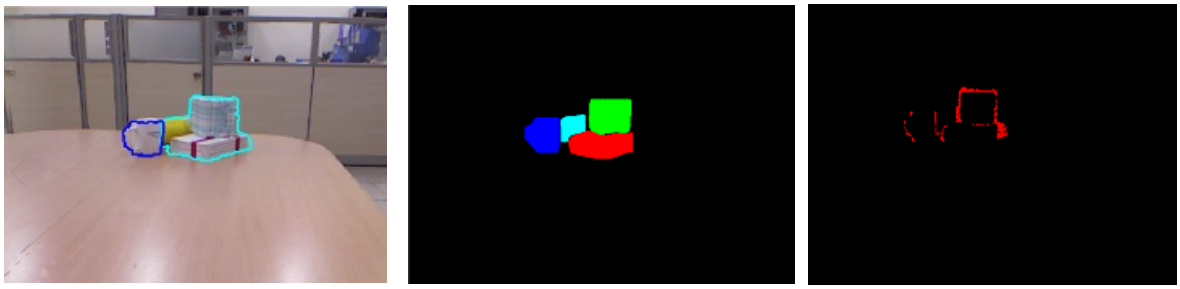


Fig. 1. Diagram of combination between original Mask R-CNN and 3D segmentation algorithms. A point cloud process brand goes together with Mask R-CNN to operate 3D segmentation algorithms. The result after applying our proposed method could get fully of objects.



(a) (b) (c)

Fig. 2. Illustration of subtraction between (**a**): contours generated by Diffrence of Normals, (**b**): contours generated by Mask R-CNN and (**c**) subtracted region. The misclassified regions of Mask R-CNN are then segmented by Euclidean Cluster Extraction.

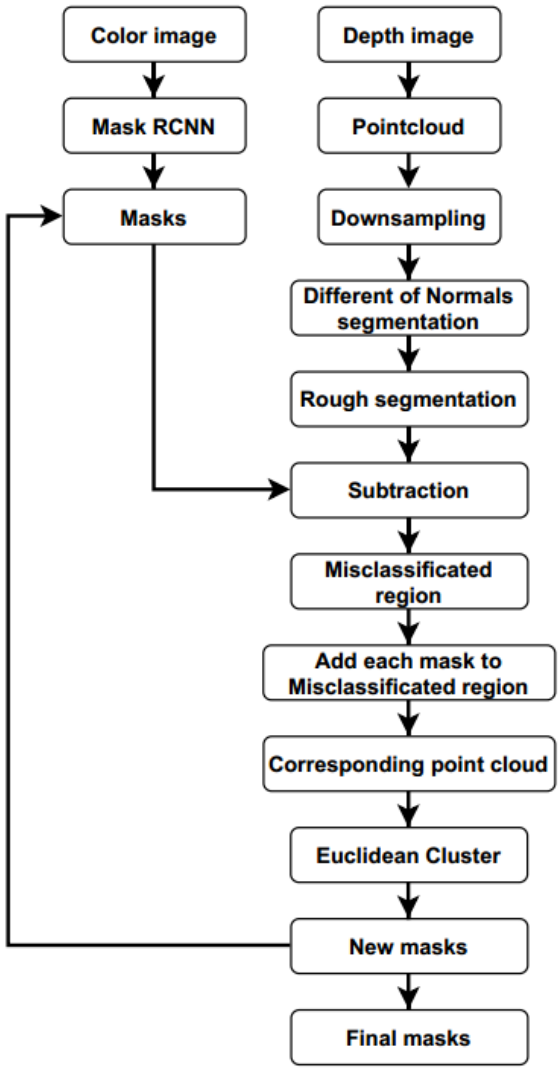


Fig. 3. Flow chart of our approach for improving edge region segmentation

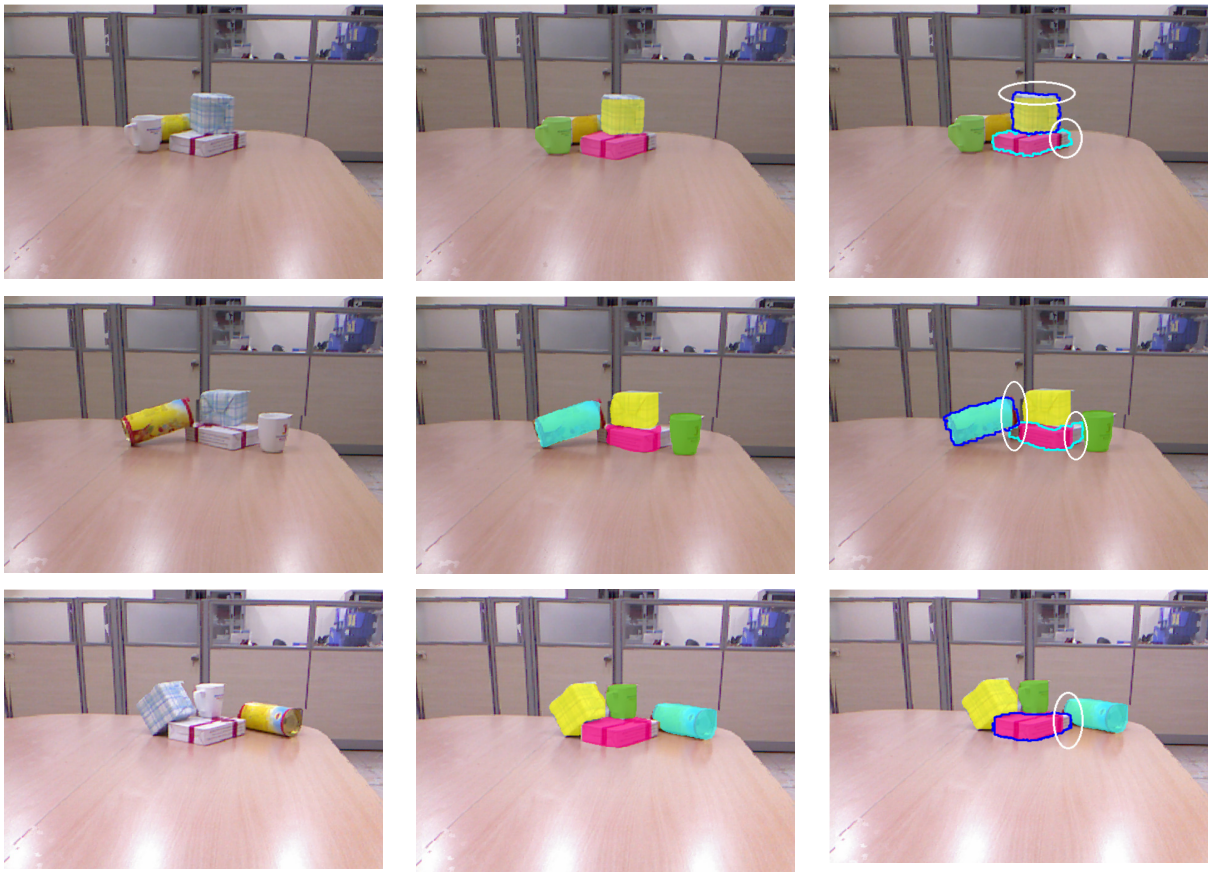
1. Estimate the normals for every point using a large support radius of r_l.
2. Estimate the normals for every point using the small support radius of r_s.
3. For every point the normalized difference of normals for every point, as defined above.
4. Filter the resulting vector field to isolate points belonging to the scale/region of interest.

Difference of Normals could gain the region, which contain almost the entire objects. However, the crucial task of this research is classifying the misunderstood region of Mask R-CNN. It is considered as categorizing the added pixel and take them to the corresponding Mask R-CNN masks. Fig.2 (a) shows the result of applying 3D segmentation algorithms. The foreground is bounded by the color polygon line. Whereas Fig.2 (b) illustrates the original mask generated by Mask R-CNN. To identify the edge regions, we subtract the original Mask R-CNN and result of Difference of Normals and have a result as Fig. 2 (c) illustrated above.

Fig. 3 explained the detailed process of our method. The next step, which categories the added region to objects. To complete the research, we matched each original mask to the subtracted area and get the corresponding depth map from depth image. Using both images as the elements to generate the point cloud, we then applied Euclidean Cluster Extraction [8] to find the edge of object. The added regions are totally the parts of object, so that using Euclidean Cluster Extraction can gain the border region based on the distance from original mask to the border in 3D-dimension space.

This work is achieved forcefully by following this procedure which is adapted from Point Cloud Library (PCL) tutorial that operated Euclidean Cluster Extraction.

1. Create a Kd-tree representation for the input point cloud dataset P.
2. Set up an empty list of clusters C, and a queue of the points that need to be checked Q.
3. Then for every $p_i \in P$, perform the following steps:
   - Add p_i to the current queue Q.
   - For every point $p_i \in Q$ do:
     + Search for the P_i^k set of point neighbors p_i in a sphere with radius $r < d_{th}$.
     + For every neighbor $p_i^k \in P_i^k$, check if the point has already been processed and if not add it to Q.
   - When the list of all points in Q has been processed, add Q to the list of clusters C, and reset Q to an empty list.
4. The algorithm terminates when all points $p_i \in P$ have been processed and are now part of the list of point clusters C.



(a) Original image (b) Mask R-CNN (c) Our approach

Fig. 4. The comparison image between (a) original (b) Mask R-CNN and (c) our approach.

4. Experimental Result

We practiced the experiment with the support of 3D camera Kinect V1 [21], an Intel Core i5 9th Gen computer and Nvidia GeForce 1650 graphic card. We experience with four classes of object: blue box, cup, cylinder, and white box. We set up objects from 0.4m to 3.5m far from the camera because the depth image of objects only can be captured in this space. With supporting of Kinect for Window SDK 1.8 and C# computer language, we developed an application to acquire both color and depth images to feed into our model. Before depth images are used as the material of 3D process branch, we calibrated it to fit with the color image. This is an important step since it guaranteed that the 3D information would be used in the right way. We collected 1000 images of all four classes prepared for the training with COCO [22] pre-trained. We set up a number of parameters for the training process such as 30 epochs with 100 steps each, 0.001 of the learning rate, 11000 iterations, 0.0001 of weight decay, 0.9 of learning momentum, and 5.0 of Gradient norm clipping rate. The detail of training process is described as Fig. 5.

As shown in Fig. 4 our approach could get and segment the edge region. The result was shown that it classified the edge region and added it to the corresponding mask generated by Mask R-CNN. All objects region is obtained completely and then can approve the grasping plan. Fig. 4(b) shows that original masks generated by Mask R-CNN are poor since these masks could not consist of all information of objects. This forgotten information is separated after the subtraction step as Fig. 4(c). As described above, for segmenting this information, we placed continuously each original mask to subtraction result and applied Euclidean Cluster Extraction. However, the previous mask was replaced by the new mask after segmentation not only reducing the computation cost because the additional region, which is analyzed by Euclidean Cluster Extraction will not be computed again but also increase the accuracy of the next calculation. Our consequence proves that our method significantly improves the poor mask of Mask R-CNN by the focus on edge and boundary regions. This enhances the acknowledge of the entire object, which is used to propose a good plan for robot manipulation.

Moreover, to clearly identify the improvement of our method in influencing edge segmentation, we evaluate our model with two indicators IoU (Intersection of Union) and mAP (mean Average Precision). Although all practices are implemented in harsh locations of object, our method obtains a significant performance in detailed segmentation. In comparing our method with orginal Mask R-CNN and applied 3D segmentation algorithms, DoN and Euclidean Cluster Extraction perform the worst result in IoU indicator because applying only disuse can not reach semantic segmentation, the IoU fluctuates from 0.195 to 0.358. Mask R-CNN performs well around 0.772 to 0.835 in each object but neglecting the edge region. Besides, the IoU is irregular and uneven, only the clear object gets high IoU, whereas, the object located behind gets severe results. This leads to the unqualified mask applying for robot manipulation. Considering throughout 4 objects as set up, our model upgrades 4 to 7% higher in IoU indicator than original model. Although IoU directly considers the covering ability of the mask on object truth ground, it does not mean getting a higher IoU as well as getting a high semantic segmentation. In terms of segmentation task, we analyze mAP indicator, which reflects the precision ability of model, throughout our approach and Mask R-CNN. Our approach significantly rises the mAP from 0.39 to 0.46. This improvement is led by the increase of IoU indicator. These improvements are certain evidences for the effective performance of our method.

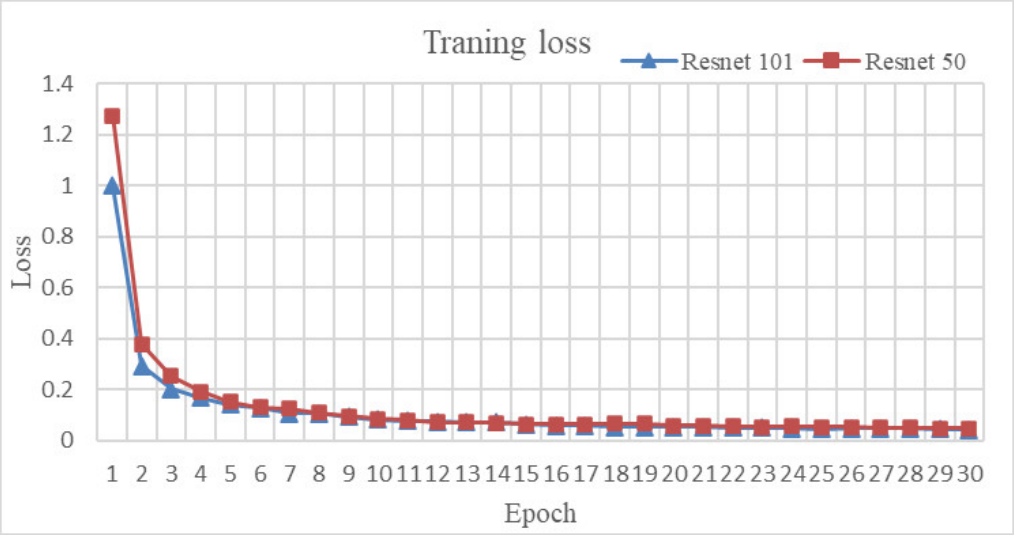


Fig. 5. Traning loss original network with Resnet 101 and Resnet 50

5. Discussion

Original masks of Mask R-CNN are oblivious of covering full objects. This constraint inspired us to submit a combination of Mask R-CNN and 3D traditional segmentation algorithms, that undoubtedly improve the coverage of the masks in border parts. To lead to this approach, we have some opinions as discussed now. The performance of 2D typical architectures mainly requires an enormous amount of dataset. Besides, the lightness of image or obstacle of object location deadly impact on the determination of boundary regions. It means it is difficult to achieve excellent results if considering 2D information without spatial pixel relationships. But the robustness in the instance segmentation of central region is assured. Meanwhile, 3D architectures for segmentation tasks need the unusual format input of datasets such as point cloud, 3D voxel grids, meshes, etc. This adversity makes the limitation on implementation and heavy computation. Although this type of architecture has advantages in dealing with edge pixels, it difficult to classify or segments the object region because it does not consider 2D features. Therefore, we combine analyzing 2D features and spatial location to achieve a remarkable improvement in detailed segmentation.

As shown in Fig. 4, the objective of our method is excellently achieved. All practices are implemented in harsh conditions of object location, where 2D architectures are usually stuck in getting exceptional segmentation such as closed located, obstacle of location, etc. This is principal evidence for the effectiveness of our method. It removes the disadvantages of both typical approaches. In aspects of time consuming, practiced in hardware as noticed earlier, it spends around 1s for finishing Mask R-CNN branch. The masks are used as the element of the classifying edge region process, so our method requires the complete operation of the original model. To get an exhaustive goal, our proposed method needs 3s. This result is far from real-time requirements but it does not mean our method is trivial in applying for a real system. In robot grasping field, the computer vision can operate while the robot manipulates. Thus, with the time consuming around 3s, it satisfies for the continuous operation of robot without any interruption as recent research [23, 24].

6. Conclusion

In this paper, we have proposed the combination of Mask R-CNN and 3D segmentation algorithms as Difference of Normals based segmentation, Euclidean Cluster Extraction to effectively segment the objects in Grasping environment. For robot grasping, the 2D images acquired by 3D camera are usually low resolution and there are many occlusions. Moreover, data collection and labeling data are time-consuming cause limited dataset size. All make Mask R-CNN does not work effectively. Our method is easy to implement and performs well at overcoming these constraints. It can not only get edge information but also categorize what object it belongs to. Our approach performs remarkably well in specifying the contours of objects. This can be observed by the increase in IoU and mAP indicator mentioned in the result section. This success makes a complete object understanding that will donate a proper plan for robot manipulation. However, the contours are still not smooth and have a bad effect on localization accuracy. So that, improving the localization accuracy is the most important assignment in future work.

Acknowledgements

This work was funded by Vietnam Ministry of Education and Training under project number B2020-BKA-02.

References

[1]. R. Girshick, J. Donahue, T. Darrell, J. Malik, Region-Based convolutional networks for accurate object detection, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2015, pages 142-158.

[2]. R. Girshick, Fast R-CNN, IEEE International in Conference on Computer Vision (ICCV), 2015. https://doi.org/10.1109/ICCV.2015.169

[3]. S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, in IEEE International Conference on Computer Vision, pages 2980–2988, 2017.

[4]. K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask rcnn, IEEE International Conference on Computer Vision, pages 2980–2988, 2017.

[5]. Z. Huang, L. Huang, Y. Gong, C. Huang, X. Wang, Mask-Scoring R-CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019. https://doi.org/10.1109/CVPR.2019.00657

[6]. Y. Zhang, J. Chu, L. Leng, J. Miao, Mask-Refined R-CNN: A Network for refining object details in instance segmentation. Sensors 2020, 20, 1010. https://doi.org/10.3390/s20041010

[7]. Y. Ioannou, B. Taati, R. Harrap, M. Greenspan, Difference of normals as a multi-scale operator in unorganized point clouds. In Proceedings of the 2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission, Zurich, Switzerland, 13–15 October 2012. https://doi.org/10.1109/3DIMPVT.2012.12

[8]. pcl.readthedocs.io. Available online: https://pcl.readthedocs.io/en/latest/cluster_extraction.html (accessed on 1 March 2021).

[9]. C. Rother, V. Kolmogorov, A. Blake, GrabCut: Interactive foreground extraction using iterated graph cuts. ACM Trans. Graph. 2004, 23, 309–314. https://doi.org/10.1145/1015706.1015720

[10]. X. Wu, S. Wen, Y. Xie, Improvement of Mask-RCNN object segmentation algorithm. In ICRIA

2019: Intelligent Robotics and Applications, Springer: Cham, Switzerland, 2019.

[11]. C. Xu, G. Wang, S. Yan, J. Yu, B. Zhang, S. Dai, Y. Li, L. Xu, Fast vehicle and pedestrian detection using improved Mask R-CNN. Math. Probl. Eng. 2020, 2020, 5761414. https://doi.org/10.1155/2020/5761414

[12]. S. Liu, L. Qi, H. Qin, J. S. J. Jia, Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, SU, USA, 18–22 June 2018. https://doi.org/10.1109/CVPR.2018.00913

[13]. C. R. Qi, H. Su, K. Mo, L. J. Guibas, PointNet: Deep learning on point sets for 3D classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.

[14]. J. R. R. Uijlings, K. E. A. Van de Sande, T. Gevers, A. W. M. Smeulders, Selective search for object recognition. Int. J. Comput. Vis. 2012, 104, 154–171. https://doi.org/10.1007/s11263-013-0620-5

[15]. S. Albawi, T. A. Mohammed, S. A. I-Zawi, Understanding of a convolutional neural network. In Proceedings of the International Conference on Engineering and Technology (ICET), Antalya, Turkey, 21–23 August 2017. https://doi.org/10.1109/ICEngTechnol.2017.8308186

[16]. S. Albawi, T. A. Mohammed, I-Zawi, S.A. Understanding of a convolutional neural network. In Proceedings of the International Conference on Engineering and Technology (ICET), Antalya, Turkey, 21–23 August 2017.

[17]. D. Rao, Q. V. Le, T. Phoka, M. Quigley, A. Sudsang, A. Y. Ng, Grasping novel objects with depth segmentation. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS), Taipei, Taiwan, 18–22 October 2020.

[18]. A. Uckermann, C. Elbrechter, R. Haschke, H. Ritter, 3D scene segmentation for autonomous robot grasping. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS), Vilamoura-Algarve, Portugal, 7–12 October 2012. https://doi.org/10.1109/IROS.2012.6385692

[19]. R. Kurban, F. Skuka, H. Bozpolat, Plane segmentation of Kinect point clouds using RANSAC. In Proceedings of the 2015 7th International Conference on information Technology, ICIT, Huangshan, China, 13–15 November 2015. https://doi.org/10.15849/icit.2015.0098

[20]. pcl.readthedocs.io. Available online: https://pcl.readthedocs.io/projects/tutorials/en/latest/don_segmentation.html (accessed on 1 March 2021).

[21]. H. Sarbolandi, D. Lefloch, A. Kolb, Kinect Range Sensing: Structured-Light versus Time-of-Flight Kinect. Comput. Vis. Image Underst. 2015, 139, 1–20. https://doi.org/10.1016/j.cviu.2015.05.006

[22]. T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft COCO: Common objects, in context. In ECCV; Springer: Cham, Switzerland, 2014.

[23]. J. Lundell, F. Verdoja, V. Kyrki, Beyond Top-grasps through scene completion. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 31 May–31 August 2020; pp. 545–551. https://doi.org/10.1109/ICRA40945.2020.9197320

[24]. Gualtieri, M.; Pas, A.t.; Saenko, K.; Platt, R. High precision grasp pose detection in dense clutter. In Proceedings of the International Conference on Intelligent Robots and Systems (IROS), Daejeon, Korea, 9–14 October 2016. https://doi.org/10.1109/IROS.2016.7759114