Developing a Deep Learning Based Grab Cut Target Image Automatic Segmentation Algorithm

Jianfeng Liang[1,2,a]
[1]Department of Computer Science University of the Cordilleras Baguio City, Philippines
[a]e-mail: 419955409@qq.com

Thelma D. Palaoag [1,b*]
[2]College of Electronic and Information Engineering, University of Beibu Gulf Qinzhou City, China
[b*]e-mail: tpalaoag@gmail.com

2024 4th International Conference on Machine Learning and Intelligent Systems Engineering (MLISE) | 979-8-3503-7507-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/MLISE62164.2024.10674591

***Abstract*—In response to the problem that traditional Grab Cut algorithms cannot achieve automatic segmentation of foreground targets, this paper proposes an automatic segmentation algorithm that combines deep learning and graph cutting. Firstly, the YOLOv4 model is trained on a common dataset to achieve automatic recognition of multiple targets, and the coordinate parameters of the automatic recognition box are converted into vertex coordinate parameters for automatic annotation. Then, the Grab Cut algorithm is iterated to complete image segmentation. At the same time, in order to further improve the accuracy of image segmentation, the regression box loss function of the YOLOv4 model has been improved. The experimental results show that the automatic segmentation algorithm proposed in this paper has better segmentation performance than the traditional unsupervised MeanShift algorithm and is close to the Grab Cut algorithm. After improvement, the automatic segmentation algorithm has improved in both IoU and PA values.**

Keywords-Image segmentation; Deep learning; YOLOv4; Grab Cut

I. INTRODUCTION

With the rapid development of computer and Internet technology, as well as the popularity of camera equipment, the amount of image information data also grows rapidly. In practical applications, image information has enormous application value in multiple fields. For example, in the medical field, it can provide strong support for doctors to identify lesion areas, diagnose and treat diseases; In the field of autonomous driving, it can provide key information such as roads, vehicles, pedestrians, etc; In the field of security monitoring, it can provide information sources for tasks such as target detection, tracking, and behavior analysis. Therefore, how to accurately and quickly segment useful information from massive image information has important research value and significance.

Image segmentation plays a crucial role in technologies such as target recognition, image analysis, and image understanding [1]. After years of development, researchers have proposed many excellent research plans. Xu Yicheng et al. [2] proposed a region based 3D tracking method to improve the robustness, accuracy, and contour edge segmentation of image segmentation. Li Na et al.[3] proposed an improved K-means image clustering segmentation algorithm based on grayscale difference, which constructs an optimization problem through the criterion function of the K-means algorithm to achieve target segmentation of coal mine images. Zhang Lingshun et al. [4] proposed a Graph cuts label fusion method that utilizes generative model constraints to quickly and accurately segment the human brain and horse body structure. The above methods mainly use traditional image segmentation methods. Although these methods have achieved good results in image target segmentation, they still have problems such as weak anti-interference ability, low degree of automation, and over cutting and under cutting. With the widespread application of deep learning models, deep learning methods have also been applied in the field of image segmentation. Ali, M et al.[5] proposed a technique that directly combines 3D CNN and U-Net to automatically segment brain tumors from multimodal MR images. Liu Xia et al.[6] first used the U-Net model to obtain coarse segmentation results, and then used Grab cut to refine segmentation, thereby improving the precision of target segmentation. Deep learning based image segmentation methods can effectively solve the problems of traditional methods, but their drawbacks are high computational complexity and insufficient real-time performance.

In response to the above issues, this article proposes a deep learning based method for automatic segmentation of Grab Cut target images. Firstly, a deep learning model is trained using a universal dataset to achieve automatic recognition of the segmentation target, generate recognition box coordinates for the segmentation target, and convert the recognition box coordinates into the vertex coordinates and length/width parameters required for Grab Cut algorithm foreground segmentation. This method effectively compensates for the shortcomings of traditional methods and deep learning methods in image segmentation.

II. METHODOLOGY

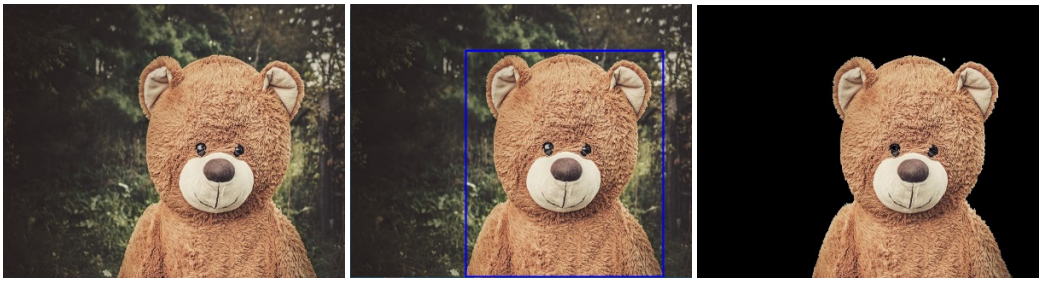
A. Grab Cut algorithm

The Grab Cut algorithm was proposed by Rother and Kolmogorov et al. [7] in 2004, and is an improved algorithm of Graph cuts. The Grab cut algorithm simplifies the user interaction process. Users only need to mark the foreground target by drawing a rectangular box, which is simpler than the manual selection of foreground and background pixels as seed points in Graph cuts. At the same time, the Grab cut algorithm utilizes multiple iterations to adjust parameters until the algorithm converges, replacing the initial energy minimization process of Graph cuts, in order to achieve better target segmentation results.

The expression for the energy function of the Grab cut algorithm is:

$$E(L,k,\theta,x) = U(L,k,\theta,x) + \lambda V(L,x) \qquad (1)$$

Among them, is the set of classification labels, is the Gaussian component, is the color feature of pixels, is the region term, is the data term, is the balance parameter between the region term and the data term, and is the parameter of GMM.



(a) Original picture (b) User marks foreground target (c) Segmentation results

Figure 1 Grab cut algorithm for image segmentation

The core of the Grab cut algorithm is to find a balance point through graph cutting technology, which can not only preserve important features in the image, but also ensure the smoothness and continuity of the segmentation results, and obtain the best segmentation of the image. The algorithm implementation process is as follows:

(1) The user annotates the foreground target using a rectangular box through the interactive interface, as shown in the blue box in Figure 1 (b).

(2) Use K-means clustering algorithm to initialize foreground and background GMM.

(3) Substitute the feature values into the Gaussian component, calculate the matching degree, and obtain the Gaussian component with the highest matching degree.

(4) Optimize GMM and use the maximum flow minimum cut algorithm to solve segmentation.

(5) Repeat (3) and (4) until the algorithm converges, and the segmentation result is shown in Figure 1 (c).

B. YOLOv4 network model

Although the Grab cut algorithm has achieved good results in image target segmentation, it still requires manual annotation by users and cannot achieve automatic segmentation of image targets. This paper studies the use of the YOLOv4 deep learning model to automatically recognize and annotate foreground targets instead of manual annotation, achieving automatic annotation of image foreground targets.

The current deep learning based object detection methods are mainly divided into two categories: one stage and two stage. The former belongs to regression based object detection and recognition algorithms with better detection speed than the latter, while the latter belongs to candidate region object detection and recognition algorithms with more advantages in detection accuracy. As a first-order detector, YOLOv4 not only maintains the detection speed of the first-order detector but also has higher recognition accuracy. The YOLOv4 network structure is shown in Figure 2, which is composed of three parts: the main feature extraction network (Backbone), the feature fusion network (Neck), and the prediction output network (Head).

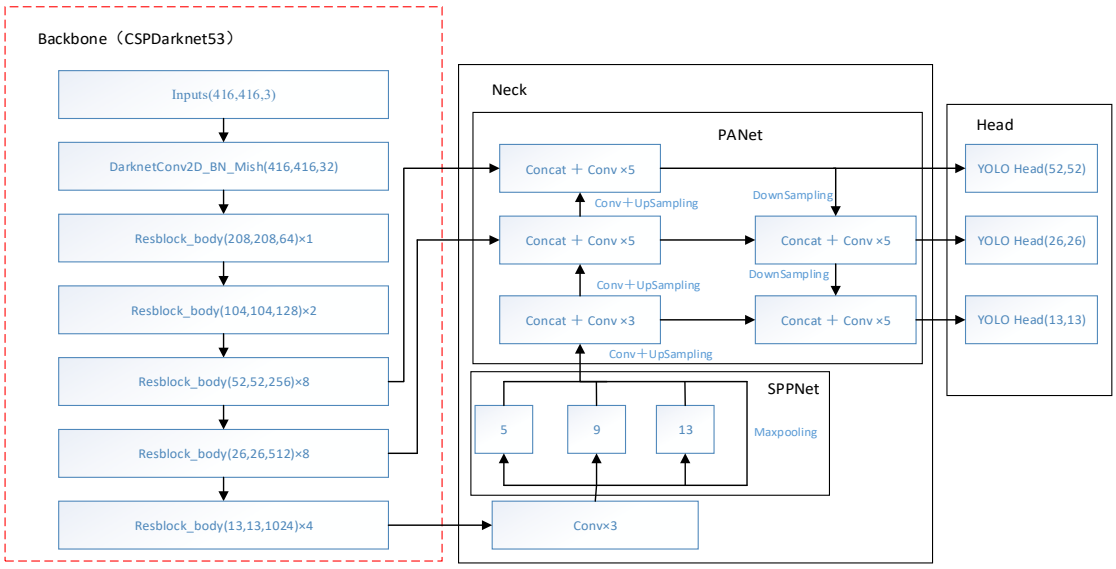


Figure 2 YOLOv4 Network Structure

In the YOLOv4 network structure, the main feature extraction network (Backbone) adopts a cascaded approach of multiple convolutional layers of different sizes to increase the dimensionality and number of channels of the network, enrich the receptive field of feature extraction. On the basis of Darkenet53, the CSPnet module is added to form the network structure of CSPMarkenet53, solving the problems of high computational load and memory consumption of Darkenet53. At the same time, the Resblock stacking method is used to increase the learning network depth and improve the detailed features of feature information. The feature fusion network (Neck) is composed of SPPnet and PANnet, and uses SPPnet to extract different feature information from the main feature extraction network. Maximizing pooling of size and stacking to expand the receptive field of feature information and enhance its spatial expression ability. Then, PANnet upsamples and Downsampling and concatenation processing are used to obtain composite feature map information of different sizes for large, medium, and small categories. This not only increases the semantic information of the target to be predicted, improves the fitting effect of the network, but also further enhances the recognition accuracy of small targets. Finally, the predicted output network (Head) predicts and recognizes the target to be measured.

1) Target box coordinate conversion

After automatic annotation by the YOLOv4 deep learning model, the position of the foreground target is determined by parameters such as the length, width, and center point coordinates of the recognized rectangular box. The required parameters for foreground segmentation by the Grab cut algorithm are length, width, and vertex coordinates. Therefore, coordinate transformation is required, and the transformation relationship is shown in Figure 3.

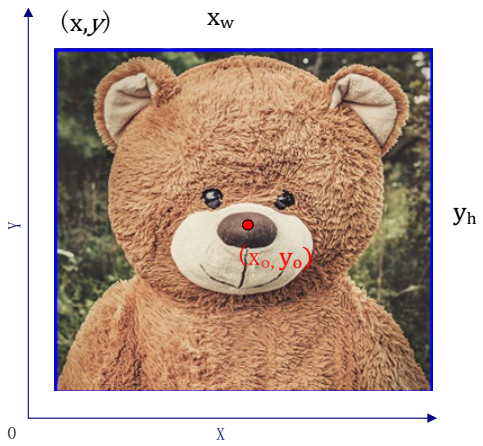


Figure 3 Coordinate Relationship Diagram

Assuming that the center coordinates of the foreground target box and its length and width parameters $(x_o, y_o, x_w, y_h)$are known, and the vertex coordinates are (x, y), the following can be obtained:

$$x = x_o - \frac{x_w}{2} \tag{2}$$

$$y = y_o + \frac{y_h}{2} \tag{3}$$

From this, the vertex coordinates and length/width of the foreground target box can be obtained as $(x_o - \frac{x_w}{2}, y_o + \frac{y_h}{2}, x_w, y_h)$.

2) Improvement of Regression Box Loss Function

In the Grab cut algorithm, the accuracy of foreground target box annotation directly determines the effectiveness of image target segmentation. This article studies the use of the YOLOv4 deep learning model to automatically identify and annotate foreground targets instead of manual annotation. In deep learning based object detection methods, the regression box loss function directly determines the accuracy of target recognition box localization. The YOLOv4 regression box loss function uses CIoU Loss, which includes three elements: overlapping area, center point distance, and aspect ratio to determine the quality of the regression box. The formula is expressed as:

$$CIoU = 1 - IoU + \frac{\rho^2\left(d, d^{gt}\right)}{C^2} + \alpha\upsilon \tag{4}$$

Among them, $1-IoU$ represents the overlapping area between the predicted box and the real box, $\frac{\rho^2\left(d, d^{gt}\right)}{C^2}$ is the distance between the center points of the predicted box and the real box, and $\alpha\upsilon$ is the aspect ratio between the predicted box and the real box.

However, in practical applications, the aspect ratio cannot represent the true value of length and width, which can lead to a decrease in the accuracy of the target recognition box. Therefore, this article proposes to use EIoU Loss instead of CIoU Loss to improve the positioning accuracy of the target recognition box. The EIoU Loss formula is expressed as:

$$EIoU = 1 - IoU + \frac{\rho^2\left(d, d^{gt}\right)}{C^2} + \frac{\rho^2\left(w, w^{gt}\right)}{C_w^2} + \frac{\rho^2\left(h, h^{gt}\right)}{C_h^2} \tag{5}$$

From Eq (5) It can be seen that EIoU Loss directly uses the aspect ratio as a penalty term instead of the aspect ratio in CIoU Loss, solving the problem of reducing the accuracy of the localization box in CIoU Loss and effectively improving the accuracy of the target localization box.

III. RESULTS AND DISCUSSION

A. Data collection and parameter settings

The image data used for deep learning model training and image segmentation experiments in this article are all from the VOC2007 public dataset, which contains approximately 5000 images and covers 20 different object categories, including common items such as humans, animals, and vehicles. The data is divided into three parts: training set, validation set, and testing set.

The experimental environment is the Windows 11 operating system, with a deep learning framework of tensorflow gpu=2.0, CUDA version 10.1, CUDNN version 10.1, programming language using Python, and running hardware environment with Intel i9-13900HX CPU, 32GB of running memory, and NVIDIA GeForce RTX 4060 GPU. YOLOv4 model training related parameter settings: batch number 20, iteration number 2000, attenuation coefficient 0.0005, initial learning rate 0.0001, Grab cut minimum cut algorithm iteration number 12.

B. Analysis of Image Segmentation Results

To verify the effectiveness of the image target automatic segmentation algorithm proposed in this paper, comparative experiments were conducted on the VOC2007 public dataset with the Grab cut algorithm and MeanShift algorithm based on human-computer interaction. The experimental results are shown in Figure 4.

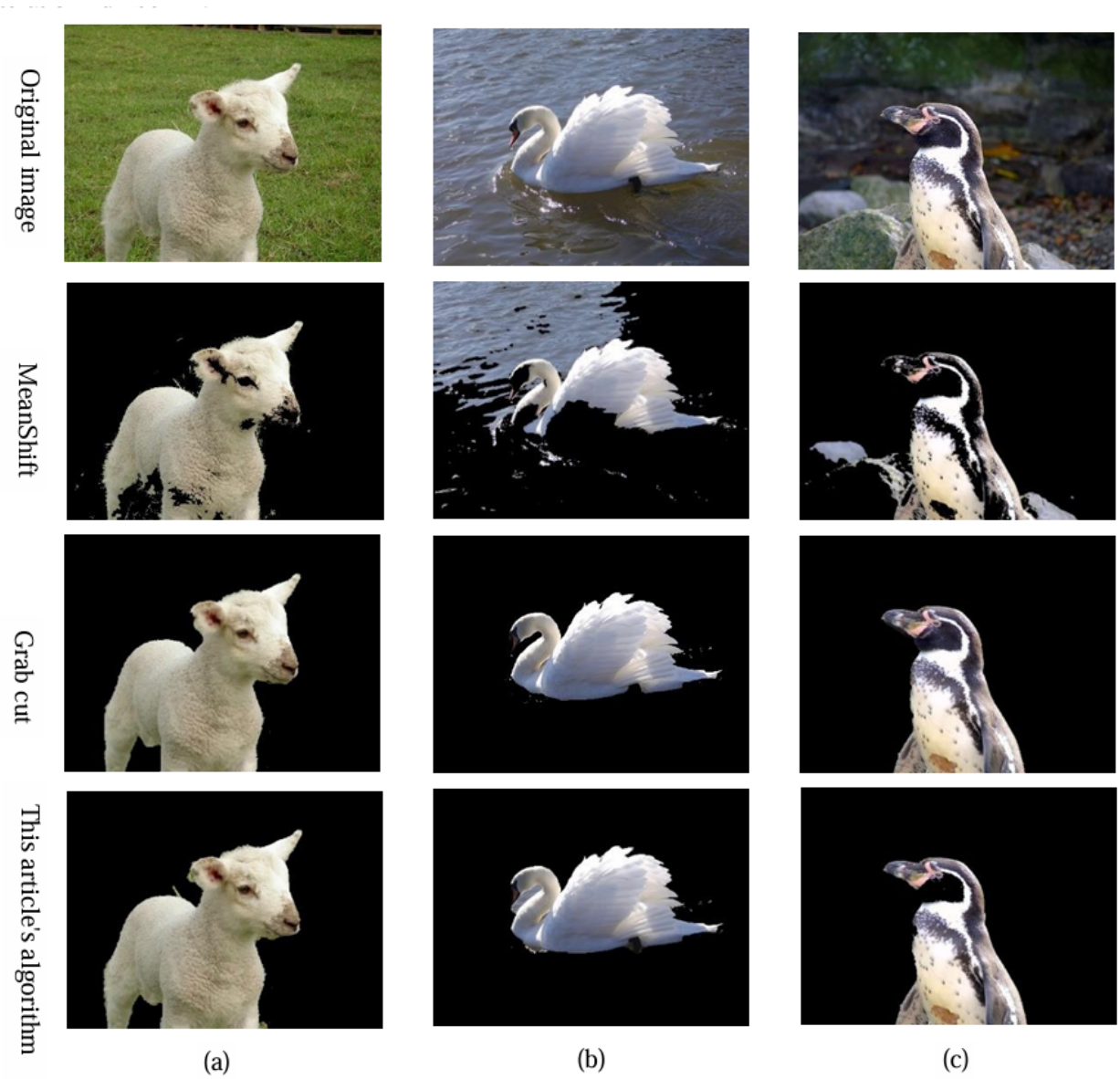


Figure 4 Different algorithm foreground target segmentation results

Figure 4 (a) shows a clear distinction between foreground target pixels and background pixels in the original image. The background image pixels are simple and the color distribution is continuous. Therefore, all three algorithms have good segmentation results for foreground targets. Figure 4 (b) shows that the color difference between the foreground target and background image pixels in the original image is not clear, and the Meanshift algorithm has problems with over cut, under cut, and a large number of noise points in the foreground target and background images, respectively. The human-computer interaction Grab cut algorithm can be manually adjusted, so the effect is the best. Although the algorithm in this article has a

small amount of background undercutting, the overall effect is similar to the Grab cut algorithm, and the advantage of the algorithm in this article is that it can achieve automatic segmentation. Figure 4 (c) shows phenomena such as complex background, rich texture, and unclear regional continuity in the original image. The Meanshift algorithm also suffers from over cutting and under cutting problems. Although the foreground target of this algorithm has a small amount of over cutting problems, the overall effect is also similar to that of the Grab cut algorithm.

C. Analysis of the Improvement Effect of Regression Box Loss Function

In order to verify the effectiveness of the improved algorithm for regression box loss function, the improved YOLOv4 model and the original YOLOv4 model proposed in this paper were used to randomly select three sets of images from the VOC2007 common dataset for automatic segmentation experiments. The segmentation evaluation criteria were mainly measured using the following two methods:

(1) Intersection Over Union (IoU):

$$IoU = \frac{|A| \cap |B|}{|A| \cup |B|} \tag{6}$$

Among them, A and B represent the segmentation value and the true value. The degree of overlap between the segmentation value and the true value is measured by the ratio of the intersection and union of two values.

(2) Pixel Accuracy (PA):

$$PA = \frac{\sum_{i=0}^{n} p_{ii}}{\sum_{i=0}^{n} \sum_{j=0}^{n} p_{ij}} nij \tag{7}$$

The formula represents the proportion of correctly classified pixels to the total number of pixels, where n is the total number of categories, p_{ii} is the predicted total number of pixels in class i of real pixels, and p_{ij} is the total number of pixels predicted as class j in class i of real pixels.

The improved YOLOv4 model and the original YOLOv4 model proposed in this article provide three sets of image automatic segmentation evaluation index data as shown in Table 1. From the segmentation results, it can be seen that the improved YOLOv4 model has improved segmentation performance and accuracy compared to the YOLOv4 model on both IoU and PA data.

Table 1 Evaluation data of segmentation results for two models

Split image	Segmentation model	IoU	PA
1 group	YOLOv4	0.823	0.856
	Improve YOLOv4	0.939	0.956
2 group	YOLOv4	0.783	0.804
	Improve YOLOv4	0.892	0.911
3 group	YOLOv4	0.901	0.922
	Improve YOLOv4	0.963	0.971

IV. CONCLUSIONS

This article proposes a scheme for automatic segmentation of foreground targets by combining the deep learning YOLOv4 model and the image segmentation algorithm Grab cut. The YOLOv4 model is used to automatically recognize foreground targets and convert the recognized target box coordinates to achieve automatic annotation and segmentation of foreground targets. At the same time, the regression box loss function of the YOLOv4 model is improved to improve the accuracy of target segmentation. According to the experimental results, it can be seen that the automatic segmentation method proposed in this paper has more advantages in segmentation accuracy and noise processing compared to the traditional MeanShift algorithm, and the segmentation effect is close to the manually interactive Grab cut algorithm. Moreover, compared with the original YOLOv4 model automatic segmentation method, the improved YOLOv4 model automatic segmentation method proposed in this paper has improved IoU and PA values, higher segmentation accuracy, and closer to the true value effect.

ACKNOWLEDGMENTS

This work was supported by the 2024 Guangxi University Young and Middle aged Teachers Research Foundation Enhancement Project "Research on the Construction Method of Electromagnetic Maps in Marine Environment Based on Group Intelligence" (No.2024KY0440).

REFERENCES

[1] Jin,H.Y, Peng,J, Zhou,T et al. (2021) A binocular image segmentation method based on Graph Cuts multi feature selection. Computer Science, 48 (08): 150-156.

[2] Xu,Y.C, Li,P, Li,S, et al. (2024) Region based 3D tracking method for weakly textured parts Computer Integrated Manufacturing Systems: 1-21.

[3] Li,N, Xue,J.M, Jia,P.T. (2022) Application of improved K-means algorithm based on grayscale difference in coal mine image segmentation. Mining Research and Development, 42 (12): 180-185.

[4] Zhang,L.S, Ma,Y, Lu,Y, et al. (2020) Graph cuts label fusion algorithm based on generative model constraints. Computer Application Research, ,37 (06): 1910-1915.

[5] Ali, M., Gilani, S., Waris, A., Zafar, K., & Jamil, M. (2020) Brain Tumour Image Segmentation Using Deep Networks. *IEEE Access*, 8, 153589-153598.

[6] Liu,X, Yu,H.B, Li,B, et al. (2021) Spinal CT image segmentation method based on improved U-Net model. Journal of Harbin Institute of Technology, 26 (03): 58-64.

[7] Rother C, Kolmogorov V, Blake A. (2004) Grabcut: Interactive foreground extraction using iterated graph cuts. ACM Transactions on Graphics, 23: 309-314.