IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 12 November 2023, accepted 15 November 2023, date of publication 23 November 2023,
date of current version 29 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3335947

RESEARCH ARTICLE

Interactive Image Segmentation Technique Based on Improved Residual Network

FENG YANG[1] **AND DAN GENG**[2]
[1]Computer Center, Anshan Normal University, Anshan 114005, China
[2]Science and Technology Division, Anshan Normal University, Anshan 114005, China

Corresponding author: Feng Yang (yangfeng0800@163.com)

This work was supported in part by the Liaoning Provincial Natural Science Foundation, Simulation of Hyperspectral Image Dimension Reduction Algorithm Based on PCA Transform, under Grant 20190360.

ABSTRACT Interactive image segmentation offers useful guidance to users and can be applied in practical settings for production and daily life purposes. Nonetheless, the technology's intrinsic limitations, including complicated interaction methods and high error rates, have impeded its further advancement with the development of computer vision. To address the issue, this study introduces a method to determine the extreme point of center point prediction. The target center serves as the "symmetric center" of the extreme point, which facilitates the search for other extreme points. Furthermore, the Canny algorithm is combined to achieve edge image detection. Moreover, the residual network is enhanced through embedding a pre-activation step, introducing a BatchNorm layer, and adding a pyramid scene parsing network. Finally, the performance of this method is verified by analyzing its Intersection over Union, segmentation accuracy, efficiency, F1 value, and other indicators. The results show that on the Pascalvoc2012 dataset, the segmentation accuracy obtained through the extreme point method can reach over 90%. The addition of the pyramid scene analysis network stabilizes its accuracy on urban landscape datasets between 92% and 96%. When the proposed image segmentation method is applied to the Grabcut dataset, its union intersection can reach 88.7%. On the self-generated complex daily scenery dataset, this method achieves segmentation accuracy of 95% with superior stability and precision. This provides a fresh methodological reference for optimizing interactive image segmentation technology further.

INDEX TERMS Residual networks, extreme point, Canny, interactive, image segmentation.

I. INTRODUCTION

With the continuous progress of computer technology, the information society has also emerged, and human life has been filled with various information technologies [1]. In this regard, computer vision technology has entered humanity, occupying a crucial role. In terms of information acquisition, methods such as hearing and smelling can be used to obtain information [2]. Among these methods, the information obtained through visual perception is more intuitive and vivid compared to other approaches, but cannot be quantified. Furthermore, as image processing technology rapidly advances, users' demand for information volume and accuracy is continuously growing. Simultaneously, the demand for image processing technology has risen, and numerous experts have suggested cutting-edge image processing technologies, which has significantly advanced the field of image processing and analysis. Image segmentation is a crucial technology in image processing that holds significant importance in the field. It involves separating, extracting, and analyzing objects and backgrounds from complex environments.

Image segmentation is a technology with broad application prospects, which constantly affects people's daily lives. Image segmentation is a vital technology in the field of image processing with significant importance. It entails the separation, extraction, and analysis of objects and backgrounds from intricate surroundings. In the process of medical image processing, image segmentation technology is often used. In autonomous driving technology, experts use image segmentation technology to segment, recognize, and detect

The associate editor coordinating the review of this manuscript and approving it for publication was Felix Albu.

objects such as pedestrians and obstacles on the road. This enables the vehicle to drive and ensures road safety [3]. In biometric recognition, technologies such as fingerprint unlocking of mobile phones, fingerprint payment, and facial recognition all rely on the segmentation of parts of biological features as the basis for subsequent recognition analysis. By integrating the above applications, the important role of image segmentation in image processing technology can be discovered. Image segmentation uses the inherent characteristics of the image, predominantly low-dimensional features, to divide regions with a certain degree of similarity into the same class based on established similarity criteria. The goal is to segment the target area from the original image with accuracy and precision. The essence of image segmentation is to strip the required image from the original image, thereby providing a foundation for subsequent image analysis.

There are several techniques for segmenting images, which can be categorized as Interactive Image Segmentation (IIS) and Automatic Image Segmentation, depending on whether human intervention is involved during the segmentation process. Among them, IIS can overcome the problems of insufficient or excessive segmentation in automatic image segmentation. Moreover, IIS can enhance the outcomes of poor segmentation by utilizing point tracking and redrawing, thus making it a highly competitive method for practical applications. The characteristic of interactive segmentation is to achieve segmentation through user interaction guidance, but it is prone to excessive interaction times. At the same time, when facing the same interaction experiment, different people often achieve different results, even if the same person produces different results [4]. In addition, the complex interaction methods and high labeling error rate severely limit the improvement of interactive image segmentation performance as user interaction time increases.

With the widespread application of deep neural networks, interactive image segmentation has ushered in new development ideas. Optimizing traditional neural networks using residual network structures has emerged as a recent research area [5], particularly in the field of image segmentation. However, interactive image segmentation currently necessitates repeated operations due to the requirement of annotating specific seed points in both the foreground and background for precise segmentation. Although image segmentation is applicable to many scenarios, there are also problems such as complex interactions, difficult user operations, and low segmentation accuracy. The optimization of point-based interactive image segmentation is undergoing continuous improvement. However, some issues still persist, notably inadequate and absent image data, leading to suboptimal segmentation results.

Therefore, in response to the problems of complex interactive methods and high error rates in current interactive image segmentation, people have creatively proposed a method for determining extreme points based on center point prediction. This method takes the target center as the symmetric center of the extreme point, thereby reducing the difficulty of finding other extreme points. Using Canny algorithm for edge detection facilitates the identification of the ultimate value point. This study improves the technology on the basis of existing traditional networks, further improving the accuracy of the algorithm in image segmentation. At the same time, a new network model is introduced to improve the accuracy of dataset segmentation and confirmation using interactive prediction and extreme point confirmation methods. According to the method proposed in the study, the first step is to greatly shorten the time for determining the extreme point and improve efficiency. Improving the quality of image segmentation helps to obtain image regions of interest to users. This can reduce the amount of image data, facilitate analysis and understanding, and facilitate the application of images. In addition, the proposed image segmentation technology has improved the effectiveness of image segmentation technology in different fields such as medicine, transportation, and biometrics. This has driven the development of image engineering.

The paper is mainly divided into four parts. The first part is a literature review on the application of interactive residual technology in image segmentation. The second part has two sections. The first section is the proposed extreme point determination method based on center point prediction. The second section is based on the first section and proposes an image segmentation technology based on an improved residual network. The third part is the performance testing and actual effect verification of image segmentation technology. The fourth part summarizes the entire content and verifies the efficient performance advantages of the proposed method. Simultaneously, the present extreme point method for interactive image segmentation has been refined by substituting the conventional residual module, ultimately enhancing image data segmentation.

II. RELATED WORK

In recent years, the application of residual networks to image segmentation has attracted the attention of many experts. A number of research results have been achieved. To optimize the convolution layer and filter, J Miao's team proposed a human body segmentation algorithm based on the compressed depth Convolutional neural network, and adopted a two-stage global filter level pruning strategy. The results showed that the segmentation speed of this method is increased by 2.4 times [6]. Chu J et al. proposed an instance segmentation method that integrates non maximum suppression algorithms for semantic segmentation of bounding boxes returned by detectors. The results showed that it can stably improve the accuracy of instance segmentation [7]. Y Zhang et al. developed a masked region convolutional neural network, incorporating mask refinement to enhance its prediction capability for instance segmentation. They also modified the region of interest alignment step size. The results showed that the average accuracy of the method on large-scale cases reaches 56.6%, which has astrong generalization ability [8]. The neural network (NN) model proposed by Cornelio J et al. had been applied to yield prediction error learning and was able to effectively predict the error between parameters [9]. Deng H's team conceived a deep learning method built on improved adversarial networks to segment hippocampal images. In their experiments, they used different convolutional configurations for capturing and segmenting the acquired information. 130 Alzheimer's patients were used as subjects for the study: the model was able to effectively achieve segmentation of the hippocampus and correctly diagnose the affected individuals [10].

Wu B et al. constructed a seismic impedance inversion system using a fully convolutional residual network (FCRN) with a migration learning algorithm. The method was able to accurately perform seismic inversion experiments [11]. To enhance the performance of conventional deep learning algorithms, Seo et al. introduced an attention channel residual network that relies solely on significant differences. They utilized it in experimental full reference image quality evaluations. The performance of this method on large datasets was significantly superior to other algorithms [12]. Feng X et al. put forward an improved multi-scale fractal residual network algorithm (MSFRN) to solve the problem of low image reconstruction performance in traditional models. This method can effectively improve subjective visual effects and achieve optimal overall performance [13]. Zaeemzadeh A et al. analyzed the impact of jump connections and interpreted the ResNet architecture. They had demonstrated through experiments that jumping and connecting residual network modules is beneficial for maintaining gradient norms and generating stable back propagation [14]. Akhenia P et al. designed a single image generation countermeasure network (SinGAN) data enhancement technology to overcome the data set problem in bearing fault diagnosis. Experimental data showed that this technology can diagnose faults with high classification accuracy, and its performance is significantly excellent than other algorithms [15].

In summary, numerous scholars have conducted extensive research and analysis on image segmentation techniques for diverse fields. Moreover, significant advancements have been made in the improvement and implementation of residual networks. Nevertheless, interactive image segmentation presently encounters challenging issues such as intricate interaction methods and low efficiency, hindering it from meeting computer vision's application necessities. Thus, this study proposes an improved residual network (IRN) method and applies its fusion to IIS technology. This is expected to further promote large-scale application and promotion of image segmentation techniques in other fields and provide some reference for subsequent research.

III. IIS BASED ON IMPROVED RESIDUAL NETWORK

A. DETERMINATION OF EXTREME VALUE POINTS BASED ON CENTROID PREDICTION

Target detection tasks can be converted into target key point estimation to avoid complex anchor frame design [16], [17]. The residual network will serve as the backbone network to achieve central point prediction of the target bounding box. The study's proposed method utilizes a convolutional network to extract a peak response spectrum and locate the local maximum as the center point. Regression operations based on the image characteristics of the center point then generate the bounding box of the target. Consequently, by inputting images into the residual network, the peak value of the output heat map serves as the corresponding target's central point. The target's size is obtained through image feature regression at the central point position, allowing for the formation of its bounding box. The training process employs supervised learning, while the inference stage solely uses the forward propagation of the network. Subsequently, the prediction of the central point is reconstructed into a key point estimation problem. Once the extreme point is identified, it will be transmitted to the segmentation network along with the input. This process delivers the necessary input for obtaining the required information. Figure 1 highlights the proposed IIS basic process, focusing on extreme point determination.

All real key points in category C during the training process are calculated to obtain low resolution key points. The Gaussian kernel is then used to distribute the actual key points into the corresponding heat map, as shown in Equation (1).

$$Y_{x,y,c} = \exp(-\frac{(x-\tilde{p}x)^2+(y-\tilde{p}y)^2}{2\sigma_p^2}) \tag{1}$$

In Equation (1), $\tilde{p}x$ represents the key point of ground resolution. $\tilde{p}y$ is the adaptive standard deviation of the object size. σ_p^2 refers to the value of the heat map obtained by the Gaussian kernel. When two Gaussian functions overlap in the same category, the maximum value of the element is taken. The objective function for training is Equation (2).

$$L_k = \frac{-1}{N}\sum_{xyc}\begin{cases}(1-Y_{xyc})^{\beta}(\hat{Y}_{xyc})^{\alpha}\log(1-\hat{Y}_{xyc}), & others\\ (1-\hat{Y}_{xyc})^{\alpha}\log(\hat{Y}_{xyc}), & Y_{x,y,c}=1\end{cases} \tag{2}$$

Both α and β in Equation (2) represent the hyper-parameters in focus loss, and their function is to normalize all focus losses. N is the number of key points in the image. Due to the fact that the data itself is discrete during down sampling, there will be deviations from the true key points. Therefore, this study adds an additional branch to predict the center point position offset. The added branch will output two values, i.e. the offset in the center point x and y directions. All categories will have the same coordinate offset value. The process of training coordinate offset is listed in Equation (3).

$$Loff = \frac{1}{N}\sum p\left|\hat{Q}_{\tilde{p}} - (\frac{p}{R}-\tilde{p})\right| \tag{3}$$

In Equation (3), $\hat{Q}$ is the offset of the central point position. This supervision solely impacts the location of crucial points on the heat map, without affecting non-key points. These key points are then utilized to estimate the network and predict all central points. Monitoring information is acquired from the

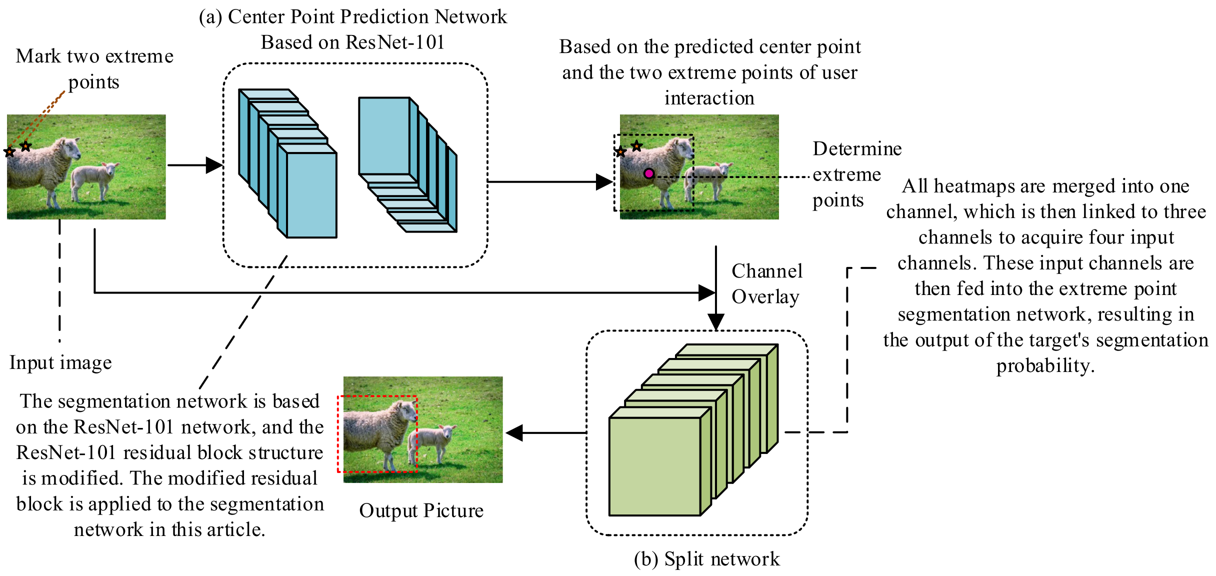

(a) Center Point Prediction Network Based on ResNet-101
Mark two extreme points
Based on the predicted center point and the two extreme points of user interaction
Determine extreme points
All heatmaps are merged into one channel, which is then linked to three channels to acquire four input channels. These input channels are then fed into the extreme point segmentation network, resulting in the output of the target's segmentation probability.
Channel Overlay
Input image
The segmentation network is based on the ResNet-101 network, and the ResNet-101 residual block structure is modified. The modified residual block is applied to the segmentation network in this article.
Output Picture
(b) Split network

FIGURE 1. The proposed basic process of interactive image segmentation based on determining extreme points.

actual bounding box coordinates to obtain the actual center point. To reduce computational burden, the study employs the same prediction for all categories, and monitors it with Equation (4).

$$L_{size} = \frac{1}{N}\sum_{k=1}^{N}\left|\hat{S}_{pk} - S_k\right| \tag{4}$$

In Equation (4), L_{size} refers to the loss. $\hat{S}_{pk}$ is the same prediction. S_k refers to the size of the regression object. Equation (5) is the overall training goal.

$$L_{\text{det}} = L_k + \lambda_{size}L_{size} + \lambda_{off}L_{off} \tag{5}$$

According to the predicted central point of the bounding box, the final bounding box can be obtained. To further determine the coordinates of extreme points, image edge detection is required. Canny edge detection can provide critical information as an aid in determining extreme point positions. First, a smoothing process is performed on the image to filter out noise. To achieve this goal, a Gaussian filter is used to expand the convolution operation. The template for the Gaussian function is shown in Figure 2.

The formula for the convolution operation of Gaussian filter kernel expansion is expressed as Equation (6).

$$h(x, y, \sigma) = \frac{1}{2\pi\sigma^2}\exp(-\frac{x^2+y^2}{2\sigma^2}) \tag{6}$$

Then, the transformation pattern of the original image obtained by Gaussian smoothing is defined as Equation (7).

$$g(x, y) = h(x, y, \sigma) \times f(x, y) \tag{7}$$

In Equation (7), × is a convolution operation. $g(x, y)$ refers to the Gaussian smoothed image, it is conducted by using high and low threshold [18], [19].

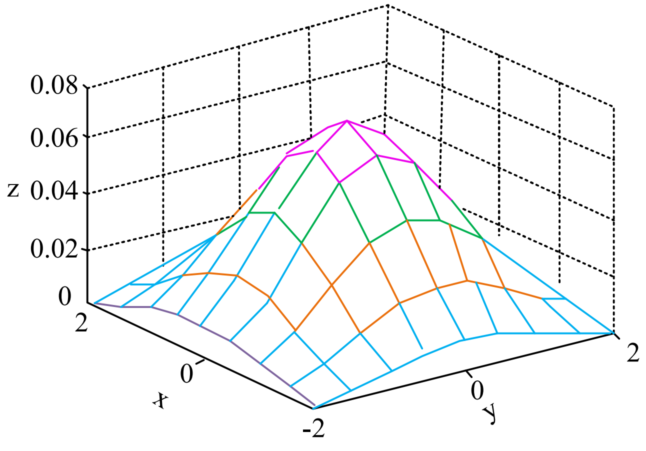


FIGURE 2. Template for a Gaussian filter function.

B. IMAGE SEGMENTATION BASED ON IMPROVED RESIDUAL NETWORK

After determining the extreme point using the proposed central point prediction and Canny edge detection, the extreme point and the original image are input together into the segmentation network for image segmentation. The residual network shows superb proficiency in deep learning and possesses a robust capability in image segmentation, enabling it to tackle gradient explosion and other probable hurdles. However, in the initial stage of gradient decline, the traditional residual structure has a very slow loss rate, which seriously restricts the improvement of neural network performance [20], [21]. Thus, the paper first improves the training ability of the residual network by adjusting the position of the unit modules. Equation (8) expresses the residual module in the residual network.

$$\begin{cases} y_l = h(x_l) + F(x_l, W_l) \\ x_{1+l} = f(y_l) \end{cases} \tag{8}$$

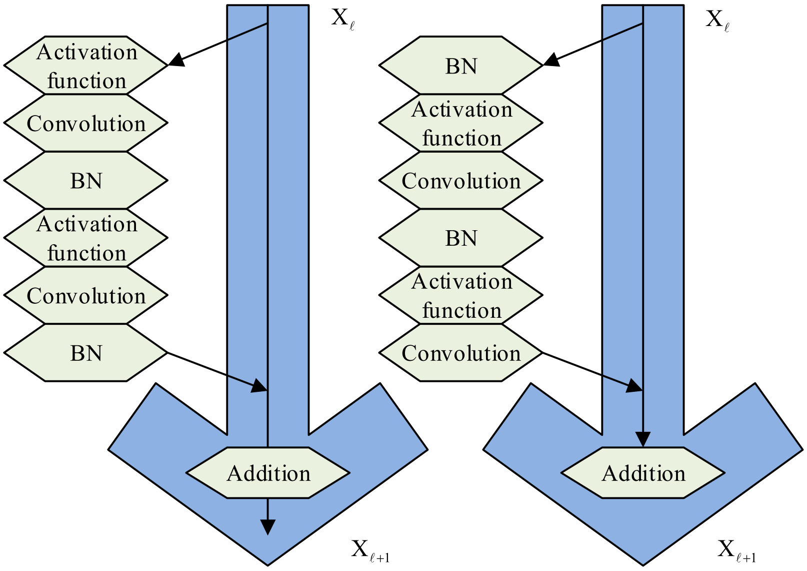


FIGURE 3. Two pre activated residual structures.

In Equation (8), x_l represents the residual unit input. x_{1+l} is the residual unit output. F refers to the residual function and f refers to the activation function RELU. This paper introduces the idea of pre activation and adjusts the position of the activation function to obtain a more efficient new residual structure. The traditional residual unit is $h(x_l) = x_l$. $f(y_l) = y_l$ is used to obtain $x_{l+1} = y_l$. The content shown in Equation (9) is obtained.

$$x_{1+l} = F(x_l, W_l) \tag{9}$$

Then, all networks are respectively substituted into Equation (9) to obtain the content of Equation (10).

$$x_L = \sum\nolimits_{i=1}^{L-1} F(x_i, W_i) \tag{10}$$

Whether it is forward or reverse propagation, network information can be transmitted between any layers. This procedure is accomplished by adjusting the positions of the BatchNorm (BN) layer and activation function and rendering the RELU activation function as an equation. To let the activation function f be an asymmetric function with $f(y_l) = y_l$, $f(y_l)$ is then re-written in function F to obtain $\hat{f}(y_l)$, thereby obtaining Equation (11).

$$y_{l+1} = F(\hat{f}(y_l), \omega_{l+1}) + f(y_l) \tag{11}$$

The equation in the shortcut section can be obtained by first activating a portion of the residual function using the activation function and then performing a convolution. From this, the two residual network structures shown in Figure 3 are obtained.

The above graphs on the left in Figure 3 were not processed in advance in the BN layer, and were first obtained through pre activation before the convolution layer [22], [23]. The structure on the right in Figure 4 moves the BN layer to process real-time data before activating the function. Using BN layer to regularize data can prevent the occurrence of over fitting phenomenon. Traditional neural networks have a tendency to lose context information due to convolutional kernels, leading to imprecise image segmentation [24], [25]. Subsequently, a context self-calibrating convolution module was designed to address this issue. Dividing the convolution core into multiple blocks enabled each block to perform different functions, thus acquiring various spatial receptive fields. As a result, the amount of context information obtained increases, and ultimately the enhancement of the spatial positioning ability of the neural network is realized. The convolutional neural network is shown in Figure 4.

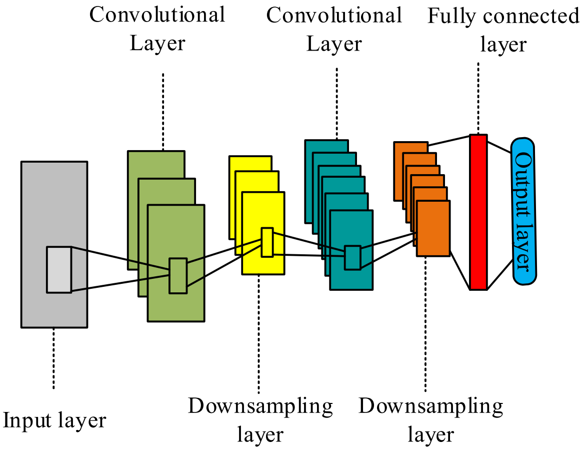


FIGURE 4. Basic architecture of convolutional neural networks.

Table 1 shows the basic parameters of the Convolutional neural network used in this study.

Prior to this, a cavity convolutional layer was included in the last two stages of the network with an attention module to ensure a consistent receptive field size and capture vital

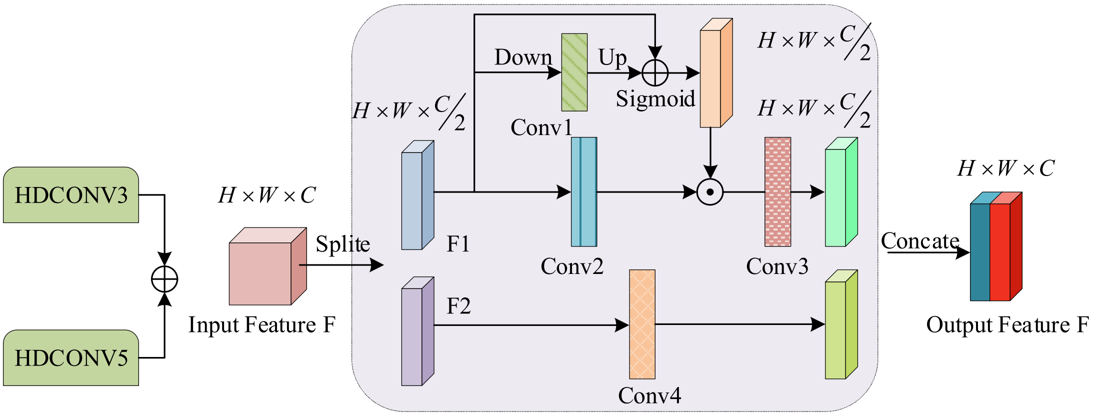


FIGURE 5. The basic structure of context self calibration convolution designed.

TABLE 1. Basic environment of experiment.

Storey	Input	Output	Kernel/Step Size
Conv1	224×224×1	224×224×16	5×5，1
Pool1	224×224×16	112×112×16	3×3，2
Conv2	112×112×16	112×112×32	5×5，1
Pool2	112×112×32	56×56×32	3×3，2
Conv3	56×56×32	56×56×64	5×5，1
Pool3	56×56×64	28×28×64	3×3，2
Conv4	28×28×64	28×28×128	5×5，1
Pool4	28×28×128	14×14×128	3×3，2
Conv5	14×14×128	14×14×128	5×5，1
Pool5	14×14×128	7×7×192	3×3，2
FC1	7×7×192	1024	/
FC2	1024	2	/
Activation function	RELU		

internal data structure features. Figure 5 depicts the fundamental framework of the context self-calibrating convolution design.

The image feature in Figure 5 obtains feature F after passing through a hole convolution group. Then four convolution operations, Conv1, Conv2, Conv3, and Conv4, are defined. The corresponding filters are C/2, C/2, k_h, and k_w. The first two correspond to input and output feature channels, while k_h and k_w are the height and width of the convolution sum [26], [27]. The four convolutional filters have different functions. Feature F is divided into F_1 and F_2, which are input into different paths to obtain different contextual space information. The path above the structure is path 1, and the path below is path 2. The convolution operations Conv1, Conv2, and Conv3 included in path 1 perform a self calibration operation on F_1 to obtain F_1'. In the other path, path 2, a general convolution operation is first performed on F_2, as shown in Equation (12).

$$F_2' = Conv4(F_2) = f_4(F_2) \tag{12}$$

F_2' in Equation (12) is a feature obtained by convolution of F_2, which can preserve the context information contained in the original feature. Then using the Contact operation is taken to fuse and splice the resulting F_1' and F_2' to obtain the output result, as listed in Equation (13).

$$F' = Concate\left\{F_1'; F_2'\right\} \tag{13}$$

In Equation (13), F' represents the result of fusion stitching. In the context of self-aligning structure, the first pathway is the crucial component. The context self calibration convolution output obtained from path 1 is equation (14).

$$F' = f_3[\sigma(Up(f_1(Down(F_1))) \oplus F_1) \otimes f_2(F_1)] \tag{14}$$

Image information undergoes contextual self calibration structure operations to obtain deeper features, including a large number of spatial and semantic features [28], [29]. However, this approach solely extracts contextual features, resulting in incomplete information features, leading to significant feature information waste, ultimately compromising the network's output. To rectify this issue, a study is performed to enhance the residual network by incorporating a pyramid scenario analysis network. It can perform fusion operations on feature information from deep and shallow layers. The result is image information that integrates more feature information. The pyramid scenario parsing network is Figure 6.

The pyramid scenario parsing network contains a total of four lines of operations. The initial row executes a singular, aggregated output operation on the inclusive qualities of the image to uphold its unimpaired integrity. The remaining three rows encompass a series of features, predominantly positional features within the image. The technical deployment process involves the mapping and division of the features from these three rows into distinct regions. The main function of each region obtained is to win features. In the resolution

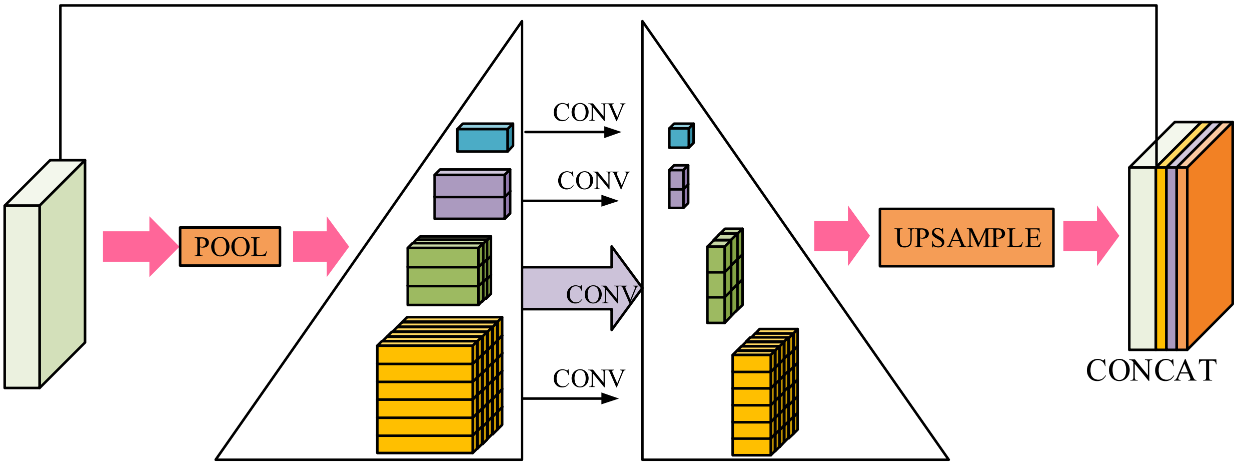


FIGURE 6. 6 Basic structure of pyramid scene analysis network.

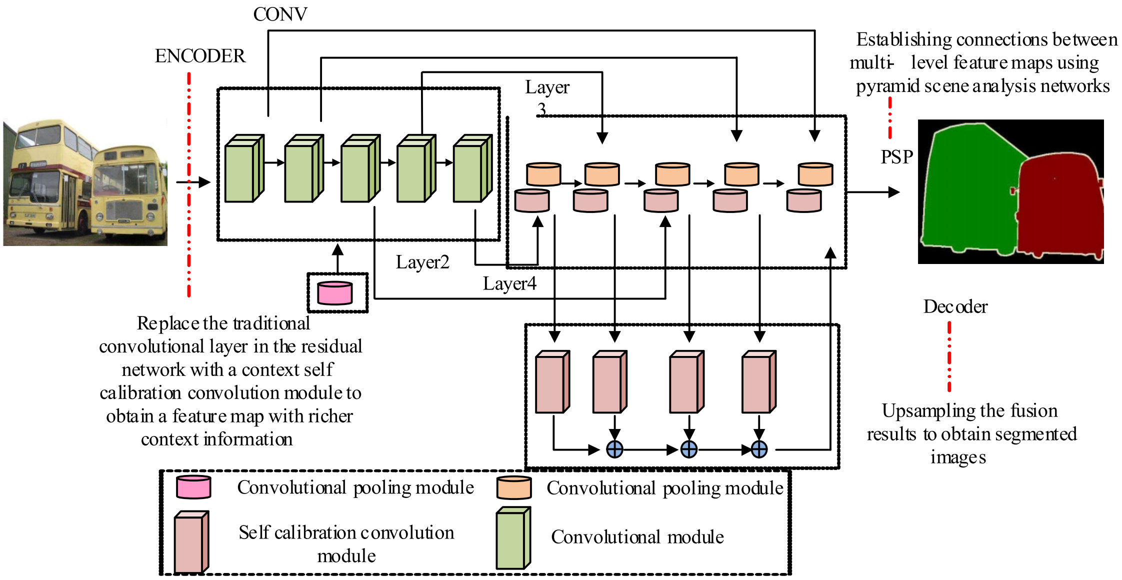

CONV
ENCODER
Layer
3
Layer2
Layer4
Establishing connections between multi- level feature maps using pyramid scene analysis networks
PSP
Decoder
Replace the traditional convolutional layer in the residual network with a context self calibration convolution module to obtain a feature map with richer context information
Upsampling the fusion results to obtain segmented images
Convolutional pooling module
Convolutional pooling module
Self calibration convolution module
Convolutional module

FIGURE 7. Overall network structure diagram.

pooling module, all feature maps have a one-to-one correspondence with their matching output size scales. To ensure the proportion of all branch features in the global feature, all branches pass the 1 × 1. The low-dimensional feature map is then processed by bilinear interpolation up-sampling. The result is a feature map with the same size as the original feature. Finally, the feature maps of all branches are fused to obtain the global features of pyramid pooling. The result obtained using this module is a probability map, which is the pixel of the object that ultimately needs to be segmented. The overall network framework of the research method is shown in Figure 7.

In Figure 7, the proposed interactive image segmentation network mainly consists of three parts, i.e. encoder, feature fusion layer and decoder. After the image is input, it first passes through the encoder structure. The encoder includes convolution layers, pooling layers, activation functions, and other features that extract advanced image characteristics. Its purpose is to progressively diminish the image's spatial resolution and increase the level of abstraction, resulting in the extraction of semantic information from the image. In Figure 8, the residual network's traditional convolutional layer is substituted with a context-self-calibrated convolutional module. This alteration generates a feature map with more contextual information, thereby altering the residual structure and enhancing the segmentation network's performance. The feature map produced by the encoder passes through the feature fusion layer, which is primarily responsible for merging features of different scales. The purpose of this is to synthesize different levels of feature information and improve the ability of the model to express image semantics. In Figure 7, the idea of analyzing the network by using the

TABLE 2. Basic environment of experiment.

Type	Configuration conditions
Internal storage	32G
GPU	Intel i9-11900k
Video storage	24G
Graphics card GPU	NVIDIA RTX 3090
CUDA	Cuda 11.0 with cudnn
Programming software	Vscode2017 Anaconda3
Operating system	Ubuntu 18
Programming language	Python 3.6
Learning rate	Le-8
Weight_decay	0.005
Batch_size	5
Momentum	0.9
Epoch	100
Tensorflow	1.14.0

pyramid scene is studied to establish the relationship between the multi-level feature maps and obtain a better feature representation. The output feature map of the feature fusion layer is passed to the decoder section. The decoder uses a series of de convolution layers, up sampling operations, etc., to gradually restore the spatial resolution of the image and convert the feature mapping into pixel-level predictions. The role of the decoder is to generate a predictive segmentation map of the same size as the input image. Throughout the entire process, the interactive image segmentation network obtains semantic information from the image via the encoder. The feature fusion layer combines the traits of different scales, and ultimately translates the feature mapping into pixel-level predictions via the decoder. In this way, the network can accurately segment the image, distinguish and label different objects or areas in the image.

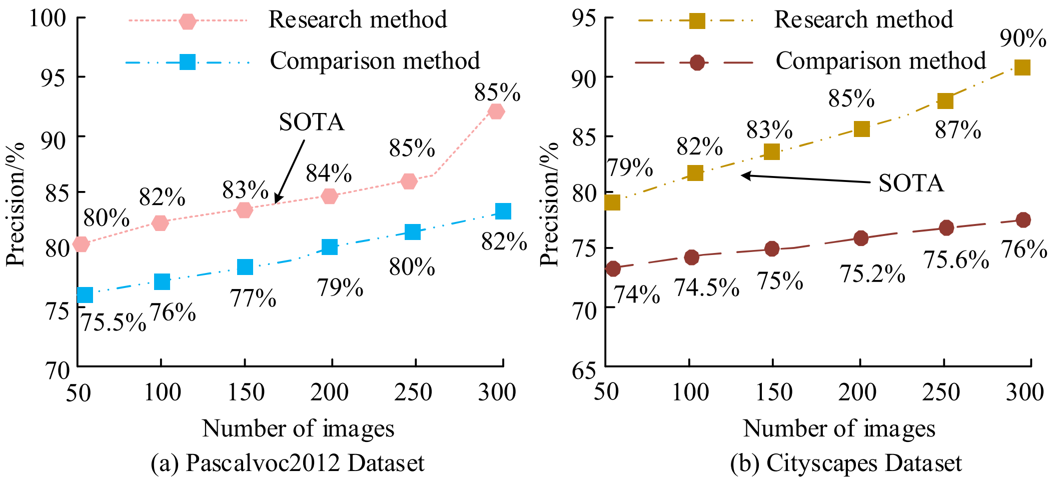
IV. ANALYSIS OF INTERACTIVE IMAGE SEGMENTATION EFFECT BASED ON IMPROVED RESIDUAL NETWORK

This study first proposed central point prediction and Canny algorithm edge detection to determine extreme points for IIS technology. The residual network structure has been improved, and the pyramid scene analysis module has been introduced to fuse context feature information. To verify the effectiveness of IIS technology, the study first designs two ablation experiments, including the extreme point determination method and the ablation experiment of the pyramid scene analysis module. TensorFlow, as a deep learning framework, is widely used in image processing, natural language processing and other fields due to its flexibility and portability. The study cites Srinivasu P N et al.'s real-time brain magnetic resonance image segmentation method that employs self-learning networks. This method accurately assesses impaired zones within the human brain through automated segmentation procedures, and it can learn from prior experimental outcomes. It is more efficient than other supervised learning strategies (such as convolutional neural network) in computation, and has 77% accuracy under the minimum training of the model. Therefore, the study utilizes the TensorFlow platform written in Python API to construct an image segmentation model. The resolution of the training image is 512 × 512, learning rate is 1.25e^{-4}, batch size is 32, momentum is set to 0.9, the iterations are 100, and weight attenuation coefficient is set to 10^{-4}. Table 2 shows the experimental environment.

First, the ablation experiment of the extreme point determination method is carried out. The proposed approach entails employing the interaction method as a bounding box, while also utilizing centre point prediction and Canny algorithm edge detection to detect extreme points in the initial stage. In the second case, during input, the fourth and fifth channels are dedicated to the segmentation of extreme value point information and background "label points". The results obtained by the two extreme point methods in the Pascalvoc2012 dataset and the Cityscapes dataset are shown in Figure 8. Figure 8 (a) shows the model accuracy results under two extreme point methods in the Pascalvoc2012 dataset. Figure 8 (b) shows the results obtained in the Cityscapes dataset. The PSACAL dataset is one of the benchmark data in object detection, image segmentation, and other technologies, containing 20 types of foreground and one type of background. The Grabcut dataset is often applied to image segmentation. In Figure 8 (a), the minimum accuracy obtained by the extreme point determination method proposed in the Pascalvoc2012 is 80%, and can ultimately reach over 90%. The method for comparison is only 82% at the highest and below 80% at the lowest, far lower than the proposed method. In the Cityscapes of Figure 8 (b), the proposed method has a maximum accuracy of 90% and a minimum accuracy of 79%, while the comparison methods are all below 77%. This indicates that the proposed extreme point determination method can effectively improve the accuracy of image segmentation.

In Figure 8, SOTA represents the curve with the best accuracy in this parameter set, among which the method used in this study has the best accuracy [2].

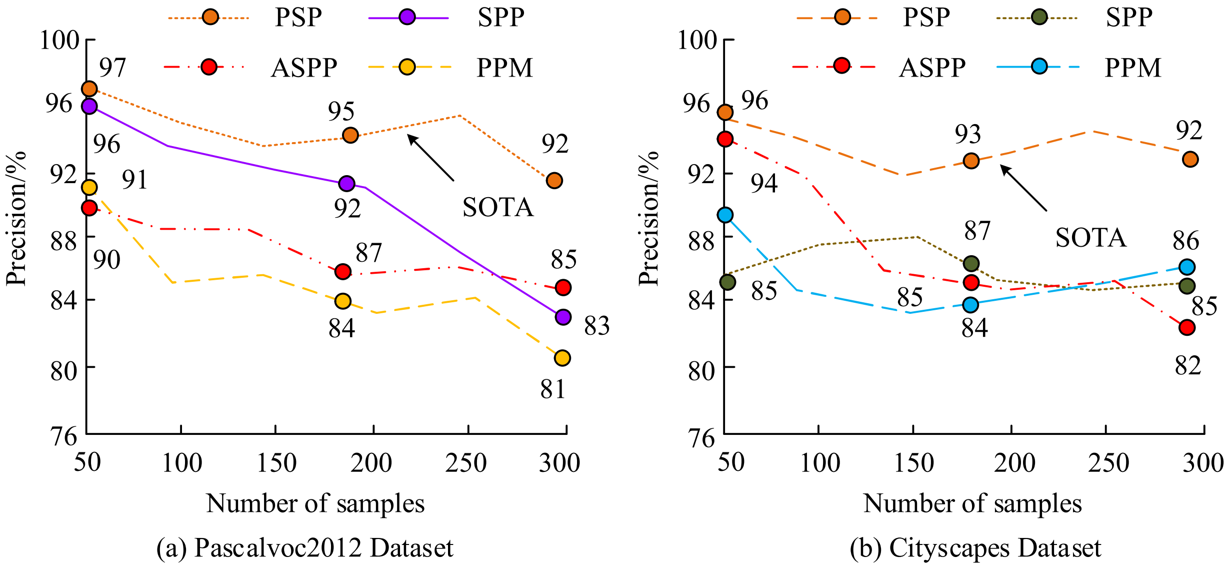
The ablation experiment of the pyramid scene parsing (PSP) module is conducted again. Comparing it with spatial pyramid pooling (SPP), around spatial pyramid pooling struc-



(a) Pascalvoc2012 Dataset

(b) Cityscapes Dataset

FIGURE 8. Image segmentation accuracy of two extreme point determination methods in Pascalvoc 2012 and Cityscapes datasets.



(a) Pascalvoc2012 Dataset

(b) Cityscapes Dataset

FIGURE 9. Experimental results of ablation of pyramid scene analysis module.

ture (ASPP), and pyramid pooling module structure (PPM), the results in Pascalvoc2012 and Cityscapes are listed in Figure 9. In Figure 9 (a) of Pascalvoc2012, the accuracy of ASPP and PPM models decreases significantly with an increase in the number of samples. They drop to around 81% and 83%, respectively, while the highest SPP accuracy is at 96%. The accuracy variation under the PSP module is small, with a maximum of 97% and a minimum of 92%, which is much higher than the other three models. Figure 9(b) of Cityscapes illustrates that SPP and PPM have similar accuracy with a minimum of 85% and 86%, respectively. Meanwhile, ASPP shows the largest decrease in accuracy, plummeting from 94% to 82% and displaying poor stability. In contrast, the PSP module exhibits consistent accuracy between 92% and 96%, reaching a maximum of 95%. The PSP module's accuracy is superior to the other three models.

In Figure 9, SOTA represents the curve with the best accuracy in different scene simulations, among which the PSP pyramid scene analysis module used in this study has the best accuracy, so it is selected as the SOTA optimal curve [3], [4].

Experiments are carried out on the PASCAL and Grabcut datasets to compare the proposed and improved residual network image segmentation method with the loss function values of GraphCuts, Geodesic Matching. Random Walker, and RIS-Net for the specific task of method classification. The PASCAL dataset serves as a benchmark for visual AI tasks, specifically object detection and semantic segmentation. Likewise, the Grabcut dataset is also publicly available for image classification and segmentation purposes. The variation of loss function values on different data sets is shown in Figure 10. From Fig.10, as the number of iterations changes, Geodesic matching and RIS-Net begin to converge at the

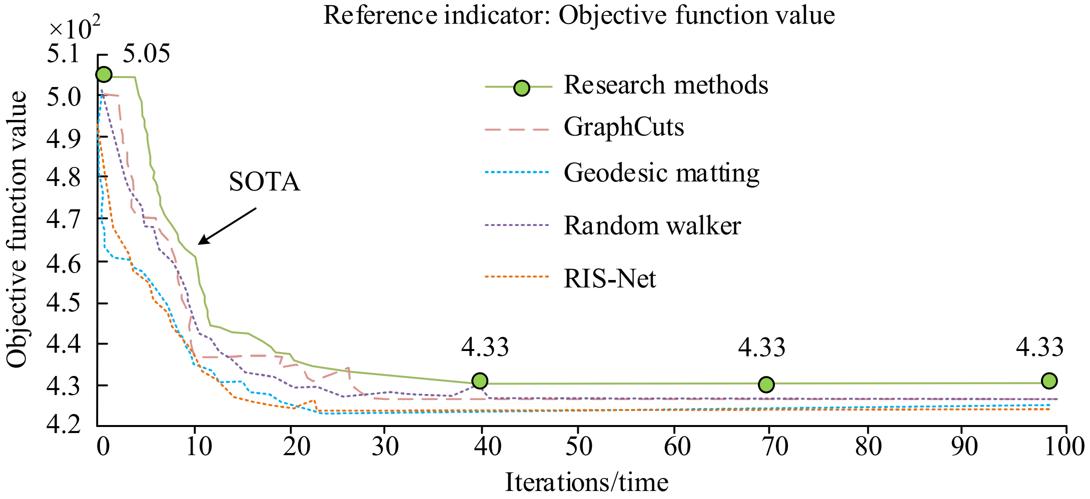


FIGURE 10. Comparison of convergence of different algorithms.

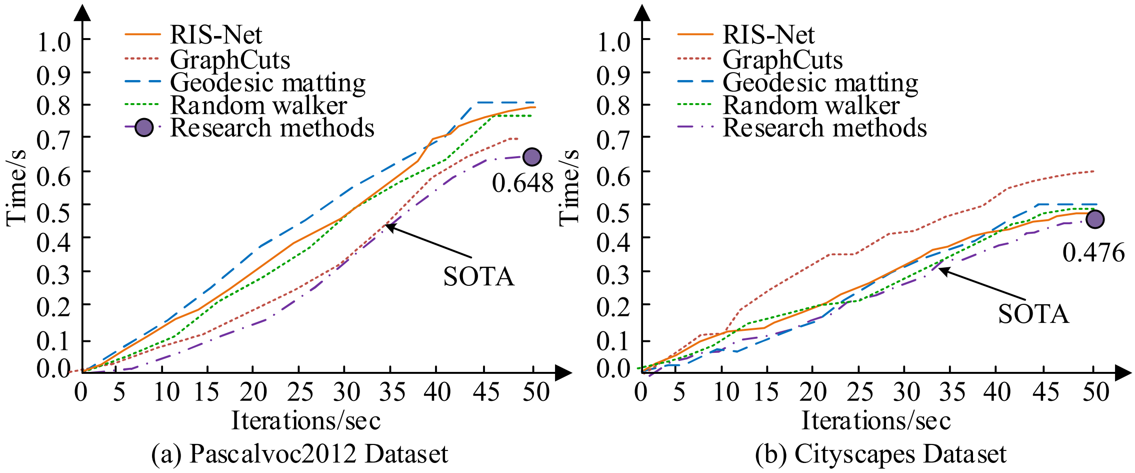


FIGURE 11. Time for different algorithms to reach a steady state.

20th iteration. It leads to the system's segmentation results falling into local optima and ultimately resulting in lower quality segmented images. Random walkers and GraphCuts do not begin to converge until about the 35th iteration. The four comparative algorithms all showed certain curve changes in the early stages of iteration. The research method began to converge at the 39th iteration and remained stable thereafter. From the comparison, it is evident that the research method does not rapidly enter the convergence state but rather stays stable during the later stages of convergence, leading to high convergence accuracy.

In Figure 10, SOTA represents the curve with the best accuracy among the stability change curves of different methods as the number of iterations increases. Among them, the method used in this study has the best stability, so it is selected as the optimal SOTA curve [7], [8].

The next step is to compare the time it takes for different algorithms to reach a stable state when they are run on two sets of data, as shown in Figure 11. From Fig.11, as the number of iterations increases, the time it takes for all five algorithms to reach a stable state increases, and the operation of the research algorithm is more complex. In Figure 11 (a), when the number of iterations reaches 45, the running time of the research method and RIS-Net, GraphCuts, Geodesic matching, and Random walker tends to stabilize, with times of 0.648s, 0.795s, 0.697s, 0.812s, and 0.774s, respectively. In Figure 11 (b), when the number of iterations reaches 50, the running time of the research method and RIS-Net, GraphCuts, Geodesic matching, and Random walker tends to stabilize, with times of 0.476s, 0.478s, 0.601s, 0.498s, and 0.482s, respectively. By comparison, the research method takes significantly less time to achieve stable operation than other algorithms, which can to some extent accelerate the successful efficiency of system image segmentation.

The SOTA in Figure 11 represents the shortest time variation curve among the time curves used by different methods

TABLE 3. IoU evaluation results of five models in Pascal and Grabcut datasets.

Model Type	Grabcut (IoU/%)	PASCAL (IoU/%)
GraphCuts	59.3	43.6
Geodesic matting	55.6	55.1
Random walker	56.3	75.2
RIS-Net	79.8	80.5
Research methods proposed	88.7 (SOTA)	87.2 (SOTA)

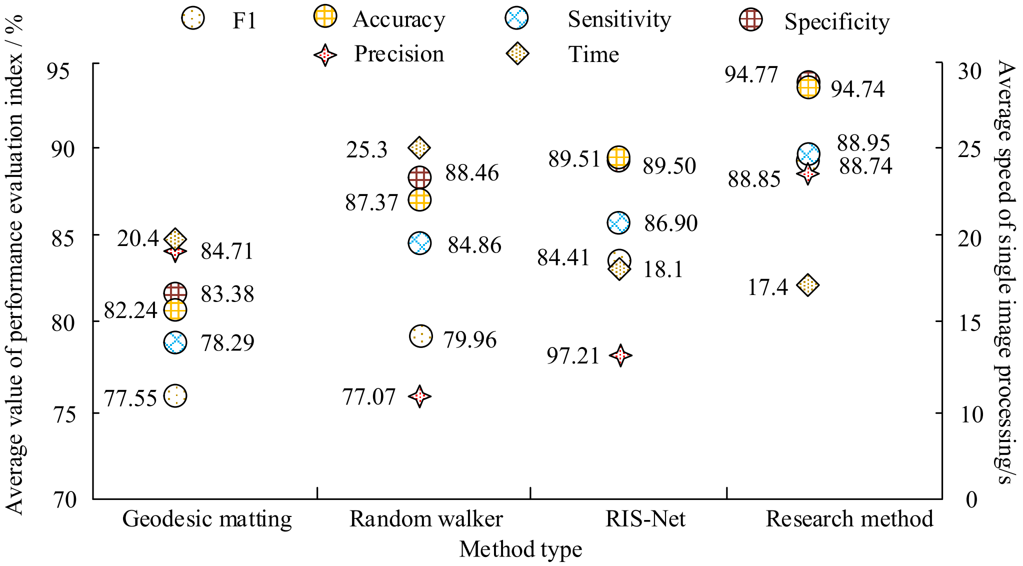


FIGURE 12. Average value of four models in the dataset PASCAL and Grabcut.

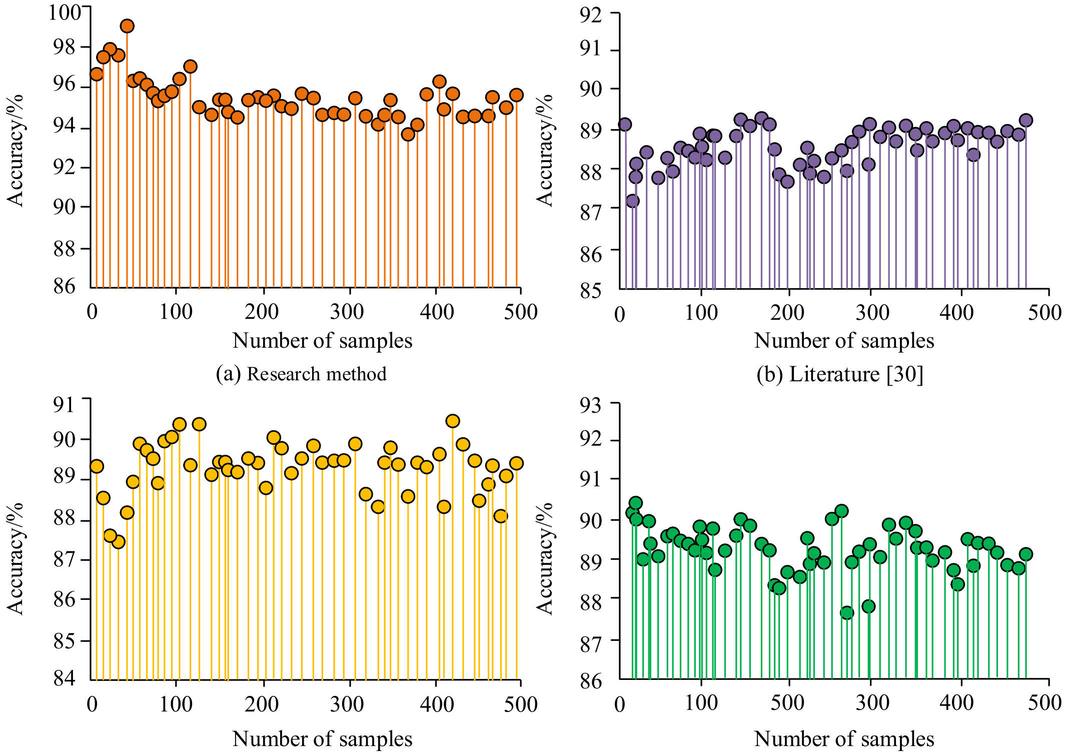
to reach a stable state as the number of iterations increases. Among them, the method used in this study took the shortest time to reach stability, so it is selected as the optimal SOTA curve [11], [12], [13].

The Intersection over Union (IoU) evaluation results of five models in the Pascal and Grabcut datasets are shown in Table 3. IoU is a standard for measuring the accuracy of detecting corresponding objects in a specific dataset. From Table 3, in the dataset Grabcut, the IoU of the GraphCuts, Geodesic matching, Random walker, and RIS-Net models are 59.3%, 55.6%, 56.3%, and 79.8%, respectively. However, the proposed method can achieve 88.7%, which is significantly higher than the other four models. In the PASCAL dataset, the IoU of GraphCuts and Geodesic matching models in Grabcut is below 60%, while the Random Walker and RIS-Net models are 75.2% and 80.5%, respectively. The proposed model is 5.3% and 12% higher than the Random Walker and RIS-Net models, indicating better performance.

In Table 3, SOTA represents the method with the most excellent cross joint cross value performance among several methods in the dataset. Among them, the method used in the experiment has the highest cross joint cross value, so SOTA is chosen [17].

Due to the poor IoU of GraphCuts, the study only compared Geodesic matching, Random walker, and RIS-Net with IRN in terms of F1 score, precision, specificity, sensitivity, and accuracy. Figure 12 shows the average values of the four models in PASCAL and Grabcut. The data in Figure 12 shows that the average value of the performance evaluation indicators of the IRN image segmentation method is better than the other three models. Its specificity and accuracy reach 94.77% and 94.74%, respectively. The corresponding values of F1 score, precision, and sensitivity are 88.74%, 88.85%, and 88.95%, all above 88%. Geodesic matching is below 85%, and the five indicators of Random Walker and RIS-Net are below 89% and 90%, respectively. In the comparison of average processing speeds, Geodesic matting, Random walker, and RIS-Net are 20.4s, 25.3s, and 18.1s, respectively, and IRN is 17.4s. Although IRN has not taken a significant lead, the combination of its five evaluation indexes has demonstrated the superior performance of this method.

To further validate the interactive image segmentation performance of the proposed method, it is applied to a self-assembled dataset that collected a significant amount of data in complex everyday environments with a wide range of variations. The segmentation accuracy of the four models in the self-built dataset is shown in Figure 13. In Figure 13, (a)~(d) correspond to the segmentation accuracy of the proposed method, Literature [30], RIS-Net, and Geodesic matching, respectively. The horizontal axis represents the number of image samples, while the vertical axis represents the accuracy. According to the results in Figure 13, the segmentation accuracy of Literature [30] fluctuates between 87% and 90%, mostly around 88%. Geodesic



(a) Research method

(b) Literature [30]

(c) RIS-Net

(d) Geodesic matting

FIGURE 13. Comparison results of the segmentation accuracy of the four models in the self-built datasets.

matching has the highest image segmentation accuracy of 90.5%, mostly within the range of 88% to 90%. RIS-Net models are all above 87%, but most fluctuate slightly below 90%. The image segmentation methods proposed by the improved residual network are all above 90%, and the vast majority are stable at 95%. They have higher segmentation accuracy, stronger stability, and better practical application results.

V. CONCLUSION

Image segmentation is an important foundation for image analysis and image processing. In line with the use of center point prediction and Canny algorithm for edge detection, this study determines the extreme points of the image, and improves the residual network.

Eventually, the image segmentation effect of the proposed method was verified. In the ablation experiment using the extreme point determination method, the segmentation accuracy of this method in the dataset Pascalvoc 2012 and Cityscapes could reach a maximum of over 90%, and a minimum of over 78%.

In the ablation experiment of PSP blocks, the module could achieve a maximum of 97% and a minimum of 92% in Pascalvoc 2012. Meanwhile, the segmentation accuracy of this module in Cityscapes was up to 95%, while the accuracy of SPP and PPM models was down to 85% and 86%, respectively. ASPP exhibited the greatest variation in accuracy, ranging from 94% to 82%, with inadequate stability. In the comparison of IoU indicators, in the dataset Grabcut, the IoU of the four models GraphCuts, Geodesic matching, Random walker, and RIS-Net were 59.3%, 55.6%, 56.3%, and 79.8%, respectively. However, the proposed method could achieve 88.7%, which was significantly higher than the other four models. In the PASCAL dataset, the proposed model was 5.3% and 12% higher than the Random Walker and RIS-Net models, respectively. In Grabcut, the IRN could reach 88.7%, significantly higher than the other four models. In the comparison of average processing speeds, Geodesic matting, Random walker, and RIS-Net were 20.4s, 25.3s, and 18.1s, respectively, and IRN was 17.4s.

When the IIS was tested on a self-generated dataset, the accuracy rate exceeded 90%. This suggests that the enhanced IIS yields better results, and outperforms previous methods in terms of performance and segmentation. Based on the findings, the proposed method presented in this paper exhibits superiority over both traditional and deep learning-based techniques.

The segmentation performance of the model, as demonstrated on multiple datasets, is noteworthy. A more efficient approach to interactive image segmentation is presented, reducing the time and complexity of user interaction while ensuring accurate segmentation and practical value. However, the determination method of extreme points was not optimized in the study and could potentially impact segmentation accuracy. Thus, further enhancement of the Canny algorithm is necessary.

REFERENCES

[1] T. Ge and O. Darcy, "Study on the design of interactive distance multimedia teaching system based on VR technology," *Int. J. Continuing Eng. Educ. Life-Long Learn.*, vol. 32, no. 1, pp. 65–77, Mar. 2022, doi: 10.1504/ijceell.2022.121221.

[2] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3523–3542, Jul. 2022, doi: 10.1109/TPAMI.2021.3059968.

[3] S. Feng, H. Zhao, F. Shi, X. Cheng, M. Wang, Y. Ma, D. Xiang, W. Zhu, and X. Chen, "CPFNet: Context pyramid fusion network for medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 10, pp. 3008–3018, Oct. 2020, doi: 10.1109/TMI.2020.2983721.

[4] X. Luo, J. Chen, T. Song, and G. Wang, "Semi-supervised medical image segmentation through dual-task consistency," in *Proc. AAAI Conf. Artif. Intell.*, Feb. 2021, vol. 35, no. 10, pp. 8801–8809.

[5] R. Z. Ye, C. Noll, G. Richard, M. Lepage, É. E. Turcotte, and A. C. Carpentier, "DeepImageTranslator: A free, user-friendly graphical interface for image translation using deep-learning and its applications in 3D CT image analysis," *SLAS Technol.*, vol. 27, no. 1, pp. 76–84, Feb. 2022, doi: 10.1016/j.slast.2021.10.014.

[6] J. Miao, K. Sun, X. Liao, L. Leng, and J. Chu, "Human segmentation based on compressed deep convolutional neural network," *IEEE Access*, vol. 8, pp. 167585–167595, 2020, doi: 10.1109/ACCESS.2020.3023746.

[7] J. Chu, Y. Zhang, S. Li, L. Leng, and J. Miao, "Syncretic-NMS: A merging non-maximum suppression algorithm for instance segmentation," *IEEE Access*, vol. 8, pp. 114705–114714, 2020, doi: 10.1109/ACCESS.2020.3003917.

[8] Y. Zhang, J. Chu, L. Leng, and J. Miao, "Mask-refined R-CNN: A network for refining object details in instance segmentation," *Sensors*, vol. 20, no. 4, p. 1010, Feb. 2020, doi: 10.3390/s20041010.

[9] M. Wei, "A novel face recognition in uncontrolled environment based on block 2D-CS-LBP features and deep residual network," *Int. J. Intell. Comput. Cybern.*, vol. 13, no. 2, pp. 207–221, May 2020, doi: 10.1108/ijicc-02-2020-0017.

[10] M. F. Haque and D.-S. Kang, "Deep adversarial residual convolutional neural network for image generation and classification," *J. Adv. Inf. Technol. Converg.*, vol. 10, no. 1, pp. 111–120, Jul. 2020, doi: 10.14801/jaitc.2020.10.1.111.

[11] J. Cornelio, S. Mohd Razak, Y. Cho, H.-H. Liu, R. Vaidya, and B. Jafarpour, "Residual learning to integrate neural network and physics-based models for improved production prediction in unconventional reservoirs," *SPE J.*, vol. 27, no. 06, pp. 3328–3350, Dec. 2022, doi: 10.2118/210559-pa.

[12] H. Deng, Y. Zhang, R. Li, C. Hu, Z. Feng, and H. Li, "Combining residual attention mechanisms and generative adversarial networks for hippocampus segmentation," *Tsinghua Sci. Technol.*, vol. 27, no. 1, pp. 68–78, Feb. 2022, doi: 10.26599/TST.2020.9010056.

[13] B. Wu, D. Meng, L. Wang, N. Liu, and Y. Wang, "Seismic impedance inversion using fully convolutional residual network and transfer learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 12, pp. 2140–2144, Dec. 2020, doi: 10.1109/LGRS.2019.2963106.

[14] S. Seo, S. Ki, and M. Kim, "A novel just-noticeable-difference-based saliency-channel attention residual network for full-reference image quality predictions," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 7, pp. 2602–2616, Jul. 2021, doi: 10.1109/TCSVT.2020.3030895.

[15] X. Feng, X. Li, and J. Li, "Multi-scale fractal residual network for image super-resolution," *Int. J. Speech Technol.*, vol. 51, no. 4, pp. 1845–1856, Apr. 2021, doi: 10.1007/s10489-020-01909-8.

[16] A. Zaeemzadeh, N. Rahnavard, and M. Shah, "Norm-preservation: Why residual networks can become extremely deep?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 3980–3990, Nov. 2021, doi: 10.1109/TPAMI.2020.2990339.

[17] P. Akhenia, K. Bhavsar, J. Panchal, and V. Vakharia, "Fault severity classification of ball bearing using SinGAN and deep convolutional neural network," *Proc. Inst. Mech. Eng., C, J. Mech. Eng. Sci.*, vol. 236, no. 7, pp. 3864–3877, Apr. 2022, doi: 10.1177/09544062211043132.

[18] X. Xu, K. Meng, X. Xing, and C. Chen, "Adaptive low-resolution palmprint image recognition based on channel attention mechanism and modified deep residual network," *KSII Trans. Internet Inform. Syst.*, vol. 16, no. 3, pp. 757–770, Mar. 2022, doi: 10.3837/tiis.2022.03.001.

[19] Z. Jiang, Z. Li, S. Yang, X. Fan, and R. Liu, "Target oriented perceptual adversarial fusion network for underwater image enhancement," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 6584–6598, Oct. 2022, doi: 10.1109/TCSVT.2022.3174817.

[20] G. Fichtinger, J. Troccaz, and T. Haidegger, "Image-guided interventional robotics: Lost in translation?" *Proc. IEEE*, vol. 110, no. 7, pp. 932–950, Jul. 2022, doi: 10.1109/JPROC.2022.3166253.

[21] Z.-L. Ni, G.-B. Bian, Z. Li, X.-H. Zhou, R.-Q. Li, and Z.-G. Hou, "Space squeeze reasoning and low-rank bilinear feature fusion for surgical image segmentation," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 7, pp. 3209–3217, Jul. 2022, doi: 10.1109/JBHI.2022.3154925.

[22] J. Zhang, K. Yang, A. Constantinescu, K. Peng, K. Müller, and R. Stiefelhagen, "Trans4Trans: Efficient transformer for transparent object and semantic scene segmentation in real-world navigation assistance," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 19173–19186, Oct. 2022, doi: 10.1109/TITS.2022.3161141.

[23] L. H. Shehab, O. M. Fahmy, S. M. Gasser, and M. S. El-Mahallawy, "An efficient brain tumor image segmentation based on deep residual networks (ResNets)," *J. King Saud Univ.-Eng. Sci.*, vol. 33, no. 6, pp. 404–412, Sep. 2021, doi: 10.1016/j.jksues.2020.06.001.

[24] R. Lan, L. Sun, Z. Liu, H. Lu, Z. Su, C. Pang, and X. Luo, "Cascading and enhanced residual networks for accurate single-image super-resolution," *IEEE Trans. Cybern.*, vol. 51, no. 1, pp. 115–125, Jan. 2021, doi: 10.1109/TCYB.2019.2952710.

[25] Y. He, L. Dai, and H. Zhang, "Multi-branch deep residual learning for clustering and beamforming in user-centric network," *IEEE Commun. Lett.*, vol. 24, no. 10, pp. 2221–2225, Oct. 2020, doi: 10.1109/LCOMM.2020.3005947.

[26] X. Xu, Z. Fang, J. Zhang, Q. He, D. Yu, L. Qi, and W. Dou, "Edge content caching with deep spatiotemporal residual network for IoV in smart city," *ACM Trans. Sensor Netw.*, vol. 17, no. 3, pp. 1–33, Aug. 2021, doi: 10.1145/3447032.

[27] S. Huang, R. Dai, J. Huang, Y. Yao, Y. Gao, F. Ning, and Z. Feng, "Automatic modulation classification using gated recurrent residual network," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 7795–7807, Aug. 2020, doi: 10.1109/JIOT.2020.2991052.

[28] G. Li, J. Su, and Y. Li, "An aircraft detection algorithm in SAR image based on improved faster R-CNN," *J. Beijing. Univ. Aeronaut. Astronaut.*, vol. 47, no. 1, pp. 159–168, May 2021, doi: 10.13700/j.bh.1001-5965.2020.0004.

[29] M. Bellver, C. Ventura, C. Silberer, I. Kazakos, J. Torres, and X. Giro-i-Nieto, "A closer look at referring expressions for video object segmentation," *Multimedia Tools Appl.*, vol. 82, no. 3, pp. 4419–4438, Jan. 2023, doi: 10.1007/s11042-022-13413-x.

[30] P. Naga Srinivasu and V. E. Balas, "Self-learning network-based segmentation for real-time brain M.R. images through HARIS," *PeerJ Comput. Sci.*, vol. 7, p. e654, Aug. 2021, doi: 10.7717/peerj-cs.654.

FENG YANG was born in Liaoning, China, in 1976. He received the bachelor's degree in computer software from Liaoning University, China, in 1999, and the M.Eng. degree in computer software from the Dalian University of Technology, in 2008. Since 1999, he has been a Lecturer with Anshan Normal University, Liaoning. His professional title is Associate Professor. He has authored four books and over ten articles. His current research interests include image recognition, cloud computing, and wireless sensor networks.

DAN GENG was born in Liaoning, China, in 1977. She received the bachelor's degree in computer science from Northeast University, China, in 1999, and the M.Eng. degree in computer software from the Dalian University of Technology, in 2008. Since 1999, she has been with Anshan Normal University, Liaoning. Her professional title is Associate Researcher and the Deputy Director of the Big Data Research Institute, Anshan Normal University. Her current research interests include image recognition, cloud computing, and wireless sensor networks.