MS-AFF: A Novel Semantic Segmentation Approach for Buried Object Based on Multi-scale Attentional Feature Fusion

Chao Lu (caius@shu.edu.cn)
Shanghai University https://orcid.org/0000-0001-7348-8658
Fansheng Chen
Chinese Academy of Sciences
Xiaofeng Su
Chinese Academy of Sciences
Dan Zeng
Shanghai University

Research Article

Keywords: Infrared Image, Deep Learning, Attention Mechanism, Semantic Segmentation

Posted Date: February 19th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-193757/v1

MS-AFF: A Novel Semantic Segmentation Approach for Buried Object Based on Multi-scale Attentional Feature Fusion

Chao Lu[1], Fansheng Chen[2], Xiaofeng Su[2], Dan Zeng[1*]

[1]Key Laboratory of Specialty Fiber Optics and Optical Access Networks, Joint International Research Laboratory of Specialty Fiber Optics and Advanced Communication, Shanghai Institute of Advanced Communication and Data Science Shanghai, ShangDa road 99 – China; caius@shu.edu.cn, dzeng@shu.edu.cn
[2]Key Laboratory of Intelligent Infrared Perception, Chinese Academy of Sciences. Shanghai, 200083, YuTian road 500 - China; cfs@mail.sitp.ac.cn

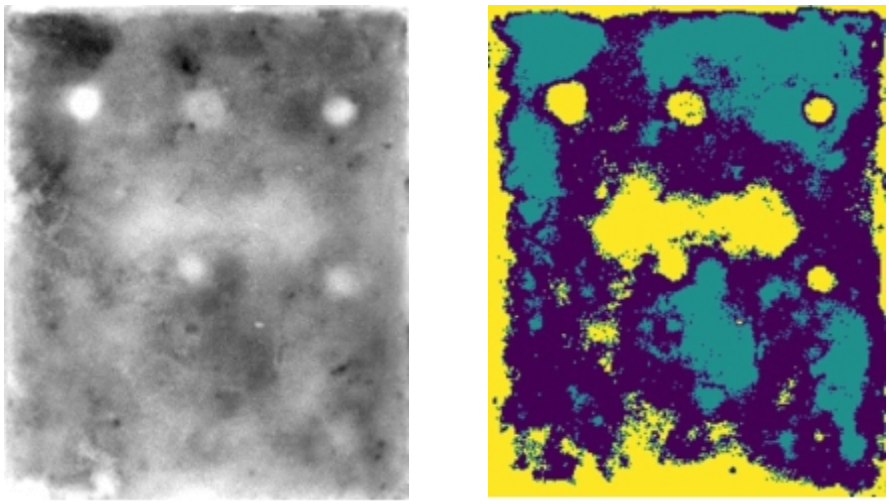
Abstract: Infrared technology is a widely used in precision guidance and mine detection since it can capture the heat radiated outward from the target object. We use infrared (IR) thermography to get the infrared image of the buried obje cts. Compared to the visible images, infrared images present poor resolution, low contrast, and fuzzy visual effect, which make it difficult to segment the target object, specifically in the complex backgrounds. In this condition, traditional segmentation methods cannot perform well in infrared images since they are easily disturbed by the noise and non-target objects in the images. With the advance of deep convolutional neural network (CNN), the deep learning-based methods have made significant improvements in semantic segmentation task. However, few of them research Infrared image semantic segmentation, which is a more challenging scenario compared to visible images. Moreover, the lack of an Infrared image dataset is also a problem for current methods based on deep learning. We raise a multi-scale attentional feature fusion (MS-AFF) module for infrared image semantic segmentation to solve this problem. Precisely, we integrate a series of feature maps from different levels by an atrous spatial pyramid structure. In this way, the model can obtain rich representation ability on the infrared images. Besides, a global spatial information attention module is employed to let the model focus on the target region and reduce disturbance in infrared images' background. In addition, we propose an infrared segmentation dataset based on the infrared thermal imaging system. Extensive experiments conducted in the infrared image segmentation dataset show the superiority of our method.

Keywords: Infrared Image; Deep Learning, Attention Mechanism;Semantic Segmentation.

1. Introduction

Infrared technology is one of the most powerful technologies used in precision guidance and mine detection. Image segmentation is the process of detecting objects or exciting areas from the input image. It is an essential step in object detection and recognition, tracking, and other related technologies. Its primary function is, which is mainly used to classify the object information in the image from the background. Conversely, due to the external environmental influence, such as temperature, airflow, radiation, and other factors, infrared images tend to have low resolution and high noise. This makes it a challenge to extract target objects from infrared images with complex backgrounds.

Some traditional image segmentation methods, such as threshold, which are based on the histogram of the bundle [1], the area expansion of [2], while the K-means method of [3] is adopted to image segmentation, as shown in Figure 1. As infrared images are generally obtained by measuring the object to the outside radiation heat gain, infrared images have poor resolution, low contrast, fuzzy visual effect, and there is no linear relationship between gray distribution and target reflection. Therefore, the K-means method can not achieve an excellent result. D. H. AlSaeed [4] has improved Otsu's method. However, these traditional algorithms only consider the gray level information between image pixels. When the gray level information, which is affected by noise, has no apparent linear relationship with the object, the segmentation effect is not noticeable.



(**a**) (**b**)

Figure 1. The image is segmented by the clustering method: (a) original image; (b) result of the clustering method.

Deep convolutional neural networks [28, 29] have achieved a collection of advancements in computer vision tasks. The CNN is gradually applied to extract image features instead of artificial design features, and the convolutional neural network is adopted to solve this problem. Using a CNN to solve traditional image problems has become a trend, and the field of image segmentation is no exception. FCN[18] is a sign of image segmentation. They employ a full CNN to extract image features, realizing end-to-end training of images of different sizes. Ignoring high-resolution feature maps will cause a loss of edge information. Moreover, Chen et al. [19] has proposed a combination of cnn and CRF to overcome the relatively low localization of deep convolutional neural networks. Combining the last layer of the neural network with the fully connected CRF can obtain more accurate boundary information. Another commonly used segmentation model based on deep learning is the codec structure. [20] belongs to the earlier semantic segmentation network using deconvolution. This model contains two parts: One is the encoder composed of VGG16. The other uses a deconvolution network, takes the encoder's output as input, and finally generates a pixel-level prediction probability map. The core of SegNet [21] is composed of an encoder network and a corresponding decoder structure, and the final pixel-level classification layer. Its main contribution is that the decoder performs nonlinear sampling on the input features of the resolution. Up-sampling maps and filters are trained

together to produce dense feature maps. SegNet cannot obtain global semantic information well, and misjudgments often occur. Unet proposed by Ronneberger et al. [22] is as well as a very classic codec network for medical image segmentation. Zhao et al. [17] proposed the Pyramid Scene Analysis Network (PSPN), It is based on various regions to aggregate context information and then find the relative information. Dilated convolution (also called dilated convolution) can spread the receptive field and reduce the computational cost without losing spatial resolution.

When setting different expansion rates, multi-scale contextual information can be captured. Deeplabv1[19] and deeplabv2[11] are the most popular segmentation methods. The latter uses dilated convolution to solve resolution reduction and uses atrous spatial pyramid pooling (ASPP) to catch objects at multiple scales with context information. Chen indicated deeplabv3[23] and deeplabv3+[24], which are parallel modules that use expanded convolution, and improved the ASPP structure, adding 1×1 convolution kernel and batch normalization to fuse features. [25] employed the semantic segmentation method to deal with the aluminum electrolyte image. Furthermore, ResNet is supposed to solve the problem of network degradation caused by network deepening. Although the above techniques are useful for public datasets, some limitations is existed in infrared images. Infrared images are obtained by measuring the heat radiated from the object and have poor resolution, low contrast, fuzzy visual effects, and there is no linear relationship between gray distribution and target reflection. The semantic information in the infrared image is scarce, so the above methods can not segment the infrared image well.

So as to solve these problems, we have created improvements based on PSPNet. PSPNet uses a pre-trained ReSnet [8] model with dilated convolution to extract image features and then uses [26] pyramid pooling module to obtain semantic information. Eventually, these features were merged, and a convolutional layer is used to generate the prediction result. After the pyramid pooling module, directly up-sampling will lose part of the spatial information. A feature fusion structure was designed based on ASPP. The Global Attention Up Sample (GAU) module [12] can better fuse the shallow and deep information and avoid information loss caused by violent up-sampling.

Furthermore, we also proposed a new type of attention module. This module integrates space and channel information, making the network focus on the target, thereby suppressing background and obtaining more accurate segmentation information. We also proposed an infrared segmentation dataset based on an infrared thermal imaging system. Our main contributions of this paper are as below:

(1) A multi-scale attentional feature fusion (MS-AFF) method for infrared images semantic segmentation is proposed, integrating a series of feature maps from different levels by an atrous spatial pyramid structure. Our model can obtain rich representation ability on the infrared images.

(2) We also propose a global spatial information attention module to let the model focus on the target region and reduce disturbance in the infrared background.

(3) We build a buried infrared segmentation dataset based on the infrared thermal imaging system. Extensive experiments in the infrared segmentation dataset show the advantages of our methods.

The remainder of the paper is constructed as follows. Sec. 2 show the methods of architecture and gives its analysis. Sec. 3 evaluate our network's performance in the infrared segmentation dataset, and the implementation details are given. Last, we provide the conclusion in Sec. 4.

2. Methods and Materials

In this section, based on the pyramid scene analysis network, we used the atrous convolutional resnet network to extract the image features to extract more features without losing the image resolution. In the later pyramid pooling, we used an improved aspp structure to replace the pyramid pooling module in the pyramid analysis network. We selected a GAU module to fuse multi-scale information in aspp. This structure could aggregate semantic information from multiple scales and better integrated shallow and in-depth semantic details, thereby providing more accurate positioning. We also designed a global spatial information attention module to focus on the target and ignore the background. The segmentation was more accurate, thereby improving the network's recognition ability in complex scenes. The main structure of the network, as shown in figure 2, it mainly contains three sections, a feature extraction part, a multi-scale attentional feature fusion (MS-AFF), and the last is a Global spatial information attention module.

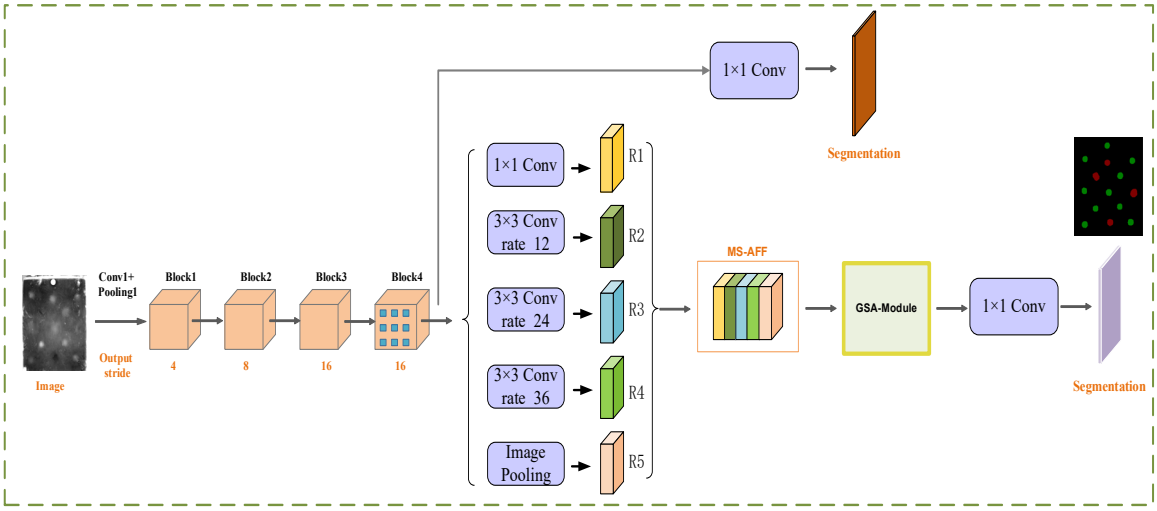


Figure 2. The main pipeline of infrared image segmentation. The infrared image is fed to the network to extract the feature. Then, MS-AFF integrates a series of feature maps. The GSA-Module to let the model focus on the target region. Finally, the network generates the segmentation result.

2.1 Atrous Convolution for Infrared Image Feature Extraction

The convolutional neural network's main job is to learn the infrared image features, and the extracted features have a decisive effect on the subsequent segmentation. Xia also suggested in [5] that feature extraction's accuracy directly affects the final classification accuracy. Feature extraction networks mainly include AlexNet[6], VGG[7], etc. The more layers of the network, the richer and more abstract the features that can be learned, but there is also a problem with it, which is the degradation of the network's learning ability. For this reason, the residual structure in the ResNet proposed by He Kaiming[8] figured

out this problem. We will use resnet101[8] as a backbone network. The direct downsampling of resnet will lose part of the spatial information, and the hole convolution can solve this problem. Atrous convolution can arbitrarily spread the receptive field and reduce the computational cost without losing spatial resolution, so it is generally used in object detection and segmentation. Wu [19] applied atrous convolution to capture more massive regional information, and Wang [20] selected atrous convolution extract features in resnet. In the last two stages of the backbone network, We cancel the down-sampling operation, use atrous convolution as the alternative to make up for the lost receptive field, and extract semantic information from different levels to enrich the information. Finally, We fulfill the [R1, R2, R3, R4, R5] feature map.

2.2 A multi-scale attentional feature fusion (MS-AFF)

The ASPP structure was first raised in DeepLabv2[11], composed of 4 hole convolutions with different atrous rates, and utilized hole convolution characteristics to extract multi-scale context information in parallel. As shown in figure 2, we use 1×1 convolution to replace the previous 24 and use the gap[13] to obtain global feature information to reduce information loss. Of course, the atrous convolution also has some problems. Because the result of the hole convolution of the current layer comes from the upper layer's independent combination, there is no mutual dependence, which will lose the local information. The information obtained from a long-distance does not correlate. Besides, a novel feature fusion structure was designed to solve this problem, according to ASPP, applying GAU[12] to fuse information between contexts, as indicated in Figure 3.

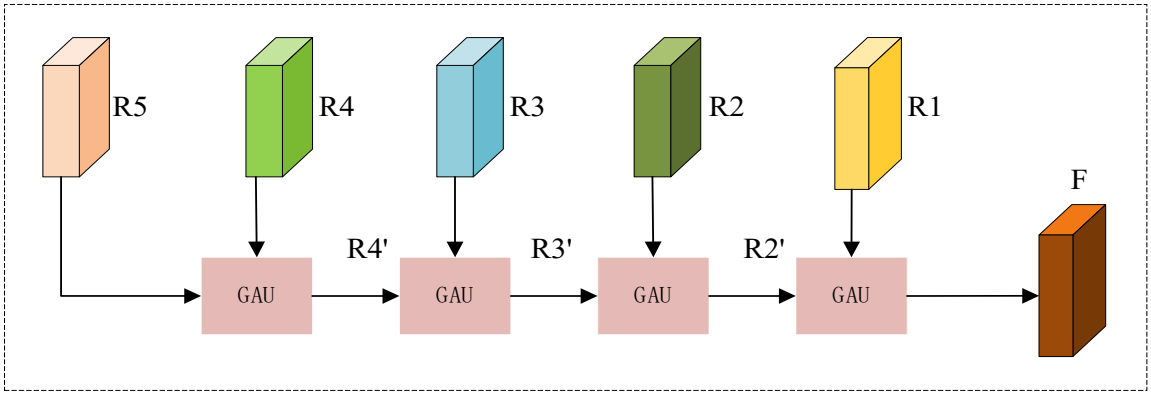


Figure 3. A multi-scale attentional feature fusion. The features of the adjacent feature maps are fused by the GAU.

GAU is a brand new type of pyramid attention model. In order to reduce the number of channels in the CNN feature map, it is necessary to perform a 3×3 convolution operation on low-level features. The global context information which is generated from high-level features undergoes 1×1 convolution firstly, batch normalization, and nonlinear transformation operations in turn and subsequently is multiplied by low-level features. Ultimately, it is important to accompany the high-level features with the weighted low-level features, and a gradual up-sampling process is performed. The GAU structure is listed in Figure 4.

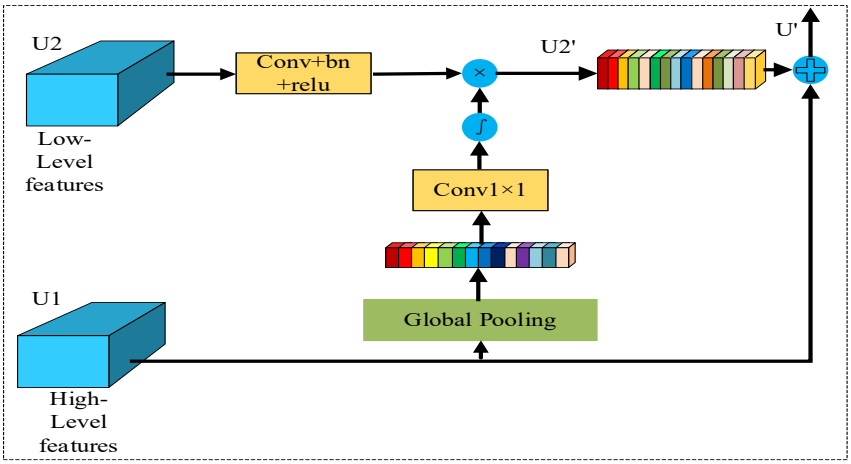


Figure 4. The network structure of the GAU.

According to Figure 3, it can be obtained that the shallow feature map $U_1 = [R_1, R_2, R_3, R_4] \epsilon R^{c \times h \times w}$, Through the GAU module, it is clear that high-level feature maps are $U_2 = [R_2', R_3', R_4', R_5'] \epsilon R^{c \times h \times w}$, the h and w mean the feature maps' height and width, respectively. R_4' can be achieved by R_4 and R_5 through GAU as follows:

$$R_4' = GAU(R_4, R_5) \tag{1}$$

Ultimately, R_2' and R_1 fed into GAU to get the final feature F. It combines all the information and has richer semantic features. For infrared images, the resolution is low, which can better mine the information of the images.

2.3 Global Spatial information Attention Module

Unlike RGB images, infrared images have stunted semantics and channel information. We design a global spatial information attention module to focus on the target area and ignore the background area. The feature map's essential features can be enhanced through the network, while the useless features can be suppressed. SENet[14] won the title of ImageNet 2017 image classification task. It is an attention on the channel dimension. The main work of CBAM [15] is based on the combination of channel attention and spatial attention. It is in[16] that proposed non-local modules to capture long-range dependencies. As the response of corresponding position, it is necessary to calculated the weighted sum of features on all positions. Inspired by [16], we put forward the global spatial information attention module to learn the discriminative spatial features. As shown in the figure 5, it is the structure of GSA network.

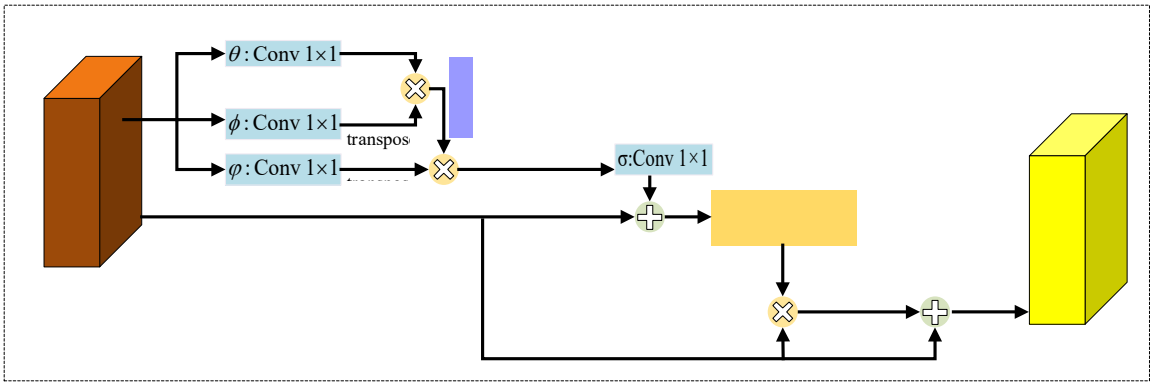


Figure 5. Network structure of the global spatial information attention module.

By utilizing the correlation function in [11], we use dot product similarity as below:

$$f\left(F_i, F_j\right) = \theta\left(F_i\right)^T \phi\left(F_j\right) \tag{2}$$

i and j represent the index of position. We apply $\mathrm{Conv}\ 1 \times 1$ layer to get θ and ϕ which ϕ are feature embeddings in our attention module. The Attention map $\mathrm{N(X)}$ for each position can be computed by Softmax function, i.e.

$$N(F) = \frac{\exp\left(\theta\left(F_i\right)^T \phi\left(F_j\right)\right)}{\sum_{\forall j} \exp\left(\theta\left(F_i\right)^T \phi\left(F_j\right)\right)} \tag{3}$$

Then we used all the features of the relevant positions as weights and calculated the following:

$$X_i = \sum_{\forall j} \mathrm{N}(F)\varphi\left(F_j\right) \tag{4}$$

We can obtain the feature map X_i, which has the same size as F. The function σ, which consisted of $\mathrm{Conv}\ 1 \times 1$ one layer to computes the input's representation, then we can attain the output M.

$$M_i = W_\sigma X_i + F_i \tag{5}$$

Where represents the cross-channel transform. Then we send feature M into GATE_CONV. GATE_CONV consists of two 1×1 convolution layers and a sigmoid layer:

$$M' = \mathrm{Sigmoid}(\mathrm{conv}\ 1\times1(\mathrm{conv}\ 1\times1(M))) \tag{6}$$

Where denote Spatial information correlated with the shape, h and w is the shape of F.

$$Y=(M'+1)*F \tag{7}$$

Finally, we attained the final output y, focusing on the image's target and ignoring the background.

2.4 Loss Function

The loss function is also commonly accustomed to evaluate the degree of inconsistency between the model's predicted and the real. In the process, the smaller the loss value, the closer the model's predicted, and the better the robustness of the model, we use softmax loss to train the model as the output as follows.

$$L = \frac{1}{N}\sum_{i} -\log\left(\frac{f_{y_i}}{\sum_{j} e^{f_j}}\right) \tag{8}$$

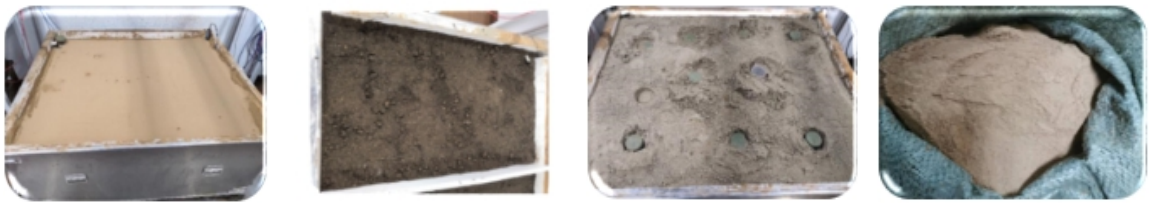
As shown in formula 8, N represents the pixel numbers, like PSPNet [17], we added an auxiliary loss function to the fourth stage of resnet101. A good auxiliary loss setting can help the network learn better performance. In our experiments, we set β 0.25 can get the perfect result.

$$L_{sum} = L_{main} + \beta L_{au} \tag{9}$$

3. EXPERIMENTS

3.1 Description of our dataset

Our dataset can simulate the process of temperature rising and cooling in an outdoor environment and detect buried or scattered targets by using the difference of soil surface temperature caused by different thermal conductivity characteristics of objects, or the difference between the target to be detected and the soil temperature. The target to be detected here is named Mellon. As shown in figure 6, experiments were carried out in the following four soils: soil, organic soil, fine sand, and yellow soil. Different pictures were collected by changing soil, voltage, heating/cooling time, humidity, and buried depth.



(a) **(b)** **(c)** **(d)**

Figure 6. Four experimental soil environments: (a) mineral soil; (b) organic soil;(c) fine sand soil; (d) yellow soil.

In the infrared imaging experiment, the format of infrared data followed the Chinese Academy of Sciences. Data was collected through the infrared camera made by the Chinese Academy of Sciences. The blackbody calibration data collected at the beginning of the experiment is converted—Dat data to infrared data. Raw format by image processing and its size is 320x256. Because of the contrast of raw format image is shallow. The pixel distribution is very concentrated. The display is usually pure white or pure black, so we should use numerical processing to map the maximum and minimum values. Raw data value in the range of 0 ~ 255, a 14 bit. The raw image is converted into an 8-bit high contrast. BMP image. After a series of preprocessing, the infrared image in BMP format is shown in Fig. 7.

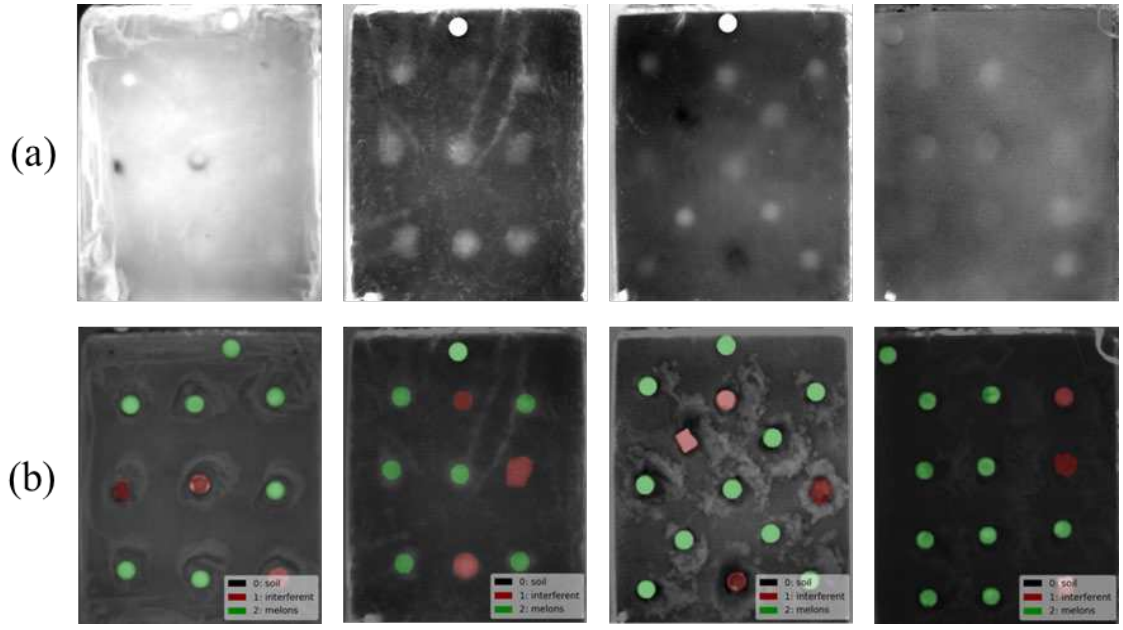


Figure 7. Infrared image after processing: (a)original image; (b) ground truth.

There are 66 batches of experiments in our dataset, among which 54 collections are divided as the training dataset, and 12 collections of experiments are used as the test dataset. There are 4379 BMP format images of infrared collected data in the training dataset, 742 images in the test dataset. As depicted in Figure 1, the real label is an original BMP format image. The dataset label is 320×256, where the label is a grayscale image. The training dataset and test dataset used in independent, thus ensuring the experimental results' validity.

3.2 Data augmentation

The deep model requires many data for training to have a good effect. Therefore, this article adopts MSDA, data augment in all experiments. We use random mirroring and random resizing between 0.5-2 for all training datasets, vertical flipping (50%), rotation -10◦ to 10◦, and so on. These methods were fine-tuned in the infrared image, such as rotating the image, change the angle of the image, and randomly crop 0.5-2 times so that the infrared image has more conditions to provide training. Due to the experiments' limitations, these data enhancements help improve the network's robustness and the model's data diversity.

3.3 Implementation details

We do our experiments at a computer equipped with Intel®Xeon(R) W-2123 CPU @ 3.60GHz×8, 32G memory, and two NVIDIA GTX2080TI. The operating system used is Ubuntu 16.04. We used a deep learning framework based on pytorch (version: 1.40) Paszke [27] suggested in 2017. Meanwhile, we applied the method with ResNet-101 [22] pre-trained model on ImageNet [23]. We train our model by using stochastic gradient descent(SGD). The batch size of each gpu card was set to 11 in the experiment details. The "poly" learning rate is which we adopted strategy. The learning rate is seen as follows:

$$bs\_lr \times (1 - \frac{iter}{\max\_iter})^{pow} \tag{10}$$

We set the bs_lr is 0.002, the max_iter is 120000, the weight decay is 0.0001, and the pow is 0.9.

3.4 Evaluation metrics

To quantitatively evaluate the performance of this method in infrared image segmentation tasks, we selected recall(R), precision(P), overall accuracy(OA), and F1score to estimate the segmentation result of our proposed method. The leading four evaluation indicators are only applicable to two categories. We regarded melons and non-melons as two categories, and interference and non-interference as the other two categories for accuracy evaluation.

$$OA = (TP + TN) / (TP + FP + FN + TN) \tag{11}$$

$$P = (TP) / (TP + FP) \tag{12}$$

$$R = (TP) / (TP + FN) \tag{13}$$

$$F_1 = 2 \times \frac{P \times R}{P + R} \tag{14}$$

True positive (TP) shows the correct prediction of the number of melon pixels, false positive (FP) represents the incorrect prediction of the number of interference pixels. True negative (TN) shows the correct number of melon pixels classified, and false negatives (FN) indicate interference pixel classification error. Besides, it is a segmentation task, we chose Pixel Accuracy (PA), Average Pixel Accuracy (MPA) to measure the segmentation effect of melons and interference objects.

Pixel Accuracy (PA)：The ratio of the correct number of classified pixels to the number of all pixels, the formula is as follows:

$$PA = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k} p_{ij}} \tag{15}$$

Mean Pixel Accuracy (MPA)：Calculate the ratio number of correct pixels in each class to the number of all pixels in that class and then make an average.

$$MPA = \frac{1}{k}\sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}} \tag{16}$$

Mean Intersection over Union (MIoU)：Calculate the IoU of each category and average. One type of IoU is calculated as follows; for example, $i = 1, P_{11}$ means true positives, that is, it belongs to category 1. The forecast is also category 1, $\sum_{j=0}^{k} P_{j1}$ indicates the number of pixels that belong to different types is expected to be class 1 (containsP_{11}), specific calculation formula as below.

$$MIoU = \frac{1}{k+1}\sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \tag{17}$$

3.5 Results and Analysis

In this part, We verified that our proposed segmentation method. Our method in this paper is capable to offer accurate segmentation of melon and Interference in complex infrared scenes, and all experiments data is collected from our experiments, the dataset called melons. To prove the advantage of our algorithm, we compared it with some state-of-the-art segmentation methods proposed in FCN-32s[18], SegNet[21], UNet[22], and deeplabv3[23]. Which of them were all introduced in part2. All experiments were performed under the same environment and parameters. Apparently, our methods can achieve excellent segmentation results.

Table1 shows various methods in melons dataset results. Among them, Mean IOU was the most common evaluation index in segmentation. PSPNet[17] behaved the worst performance among all networks. Deeplabv3[23] and SegNet[22] perform performed better than other networks. But our methods applied in which received the best result in all networks, proved the validity of our approach.

Table 1. Mean IOU on the infrared image segmentation test set.

Method	Mean IOU(%)
FCN-32s[18]	87.9
SegNet[21]	90.3
Unet[22]	89.1
PSPNet[17]	87.2
DeepLabv3[23]	89.5
Our	**90.54**

Table 2 shows the competitor of the different methods in same experimental setting. Our proposed method of precision and F1 of melons is significantly higher than other deep learning models. Our accuracy of melon detection reached 89.65%, F1 score of Melon detection is nearly 92.39%, and other evaluating indicators also achieve better results than most others.

Table 2. Other evaluate metrics on melons test set.

Method	Melon			Interference		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
FCN-32s[18]	86.1	95.2	90.1	87.36	92.9	90.1
SegNet[21]	88.1	95.7	91.8	90.2	95.5	92.8
Unet[22]	87.56	95.7	91.47	88.90	93.3	91.04
DeepLabv3[23]	88.59	95.05	91.71	88.98	94.13	91.48
Our	**89.65**	95.30	**92.39**	90.19	95.24	92.65

Table 3 shows the competitor of the three primary segmentation metrics under different methods. It can be found from Table 3 that the FCN-32s model is the worst in terms of the performance of these metrics among all networks. Secondly, the Unet model has a slight improvement over the FCN-32s model. Our model get the PA of Interference detection is nearly 99.84%, which of the best performance. We can also find our method was better than most of the models on the other evaluation metrics.

Table 3. Interference detection result in different methods.

Method	PA (%)	MPA (%)	MIOU (%)
FCN-32s[18]	99.79	93.64	90.87
SegNet[21]	99.84	95.09	93.20
Unet[22]	99.81	94.41	91.68
DeepLabv3[23]	99.82	94.46	92.06
Our	**99.84**	95.07	93.07

Table 4 also shows the competitor of the three primary segmentation metrics under different methods. That FCN-32s model also has the worst performance of these metrics among all networks. Secondly, SegNet and DeepLabv3 models have improved than other methods. Our proposed method can get the perfect effect at all three primary segmentation metrics. Our PA of melon detection is nearly 99.51%, our MPA is almost 94.75%, and the MIOU of our methods also achieved 92.67%, which is higher 0.54% than the second models. Table 3 and Table 4 show that the above methods can realize exact segmentation in the infrared segmentation dataset.

Table 4. melon detection result in different methods.

Method	PA (%)	MPA (%)	MIOU (%)
FCN-32s[18]	99.38	92.95	90.93
SegNet[21]	99.46	93.99	92.13
Unet[22]	99.44	93.71	91.85
DeepLabv3[23]	99.46	94.22	92.06
Our	**99.51**	**94.75**	**92.67**

Figure 8 also shows the segmentation result of of different methods on infrared segmentation dataset. The Green color is melons, and the red color is Interference. We can find that (a) is the test image,(g) is the ground truth. From the result, we can see that FCN-32s, DeepLabv3, and Segnet consist of false detection and misdetection. U-net exists some of the melons not detect. Our method's result can see (f), which detects all the melons and Interference. As shown in Table 1, our method achieved MIOU was 90.54%, which has the best accuracy than other models in our experiments.

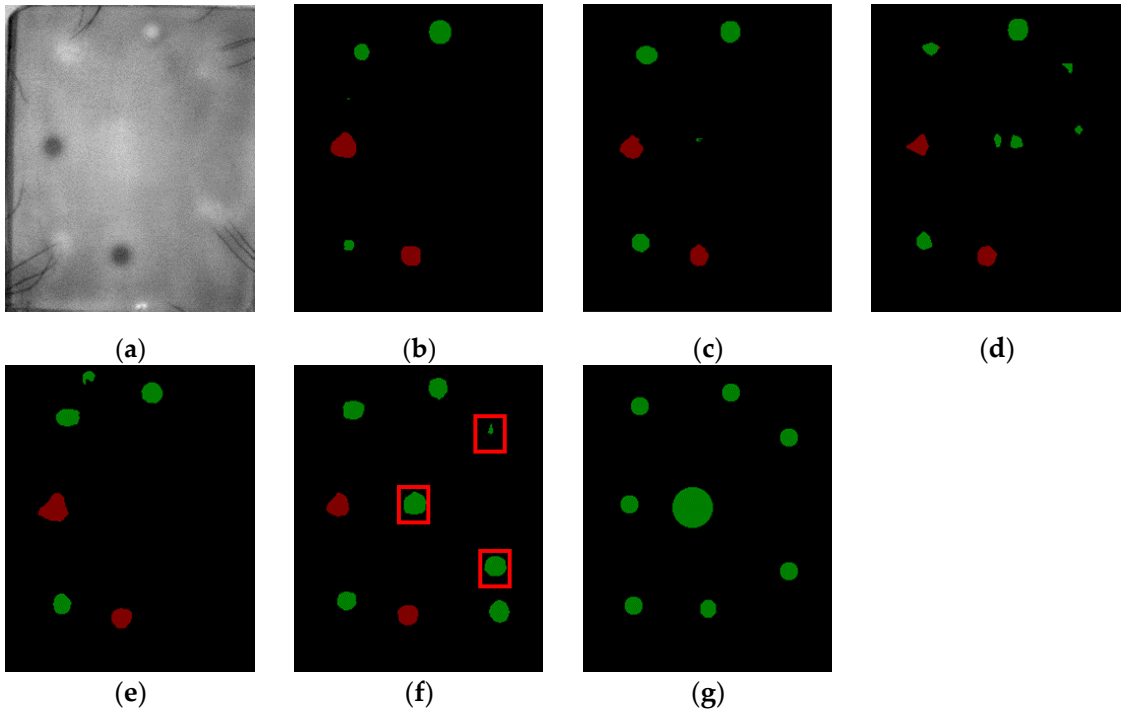


Figure 8. The comparison result of dierent methods on melons dataset. (a)Test image;(b)Segmentation performance on FCN-32s;(c)Segmentation performance on Unet (d) Segmentation performance on DeepLabv3; (e)Segmentation performance on segnet;(f)Segmentation performance on ours. (g) ground truth.

Figure 9 also shows the two test images segmentation performance on different methods. The image of the top row directly indicates that our work is more precise than others. They are more or less missed or falsely detected. The below image result also shows our method has better performance.

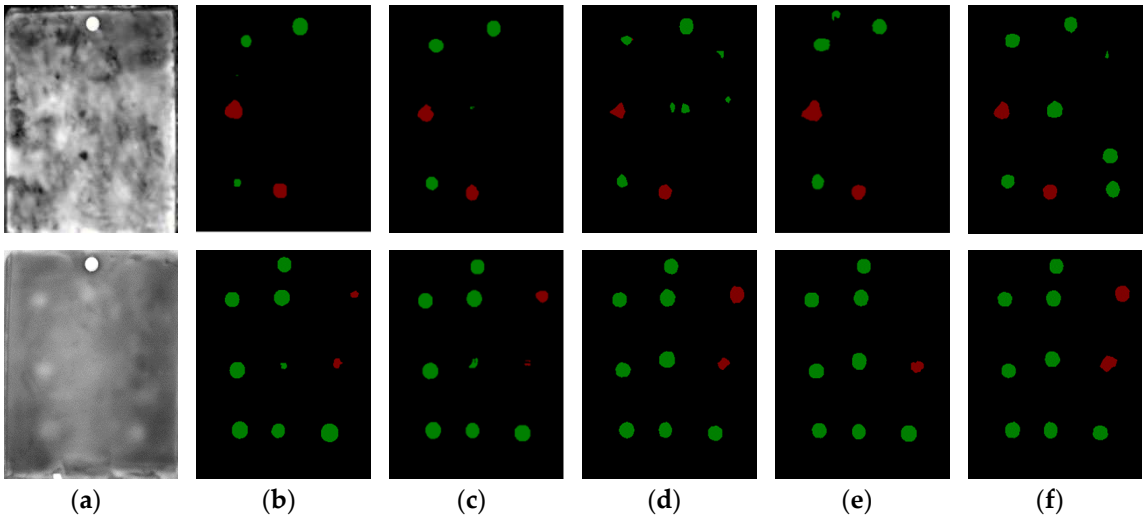
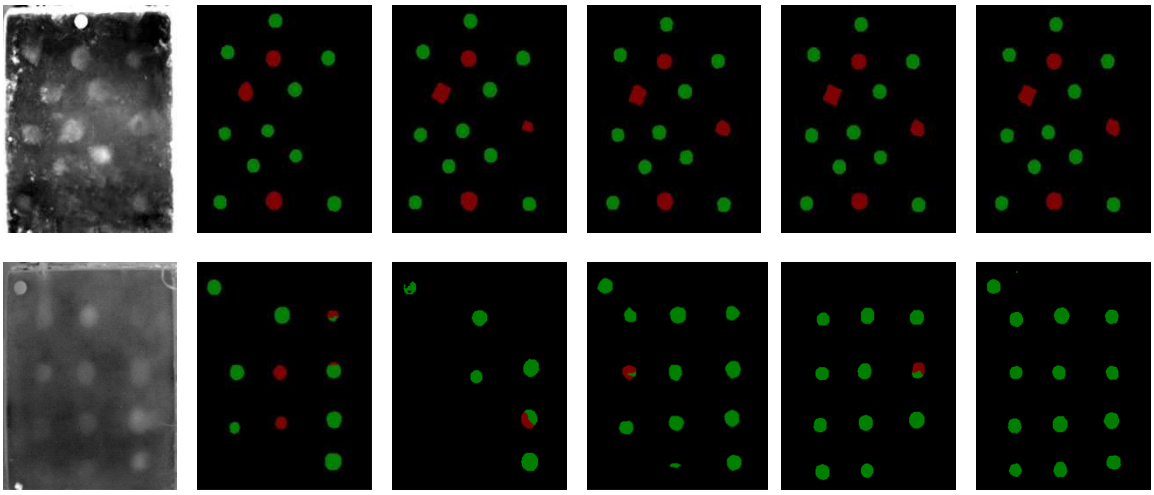


Figure 9. The comparison result of different methods on melons dataset. (a)Test image;(b)Segmentation performance on FCN-32s;(c)Segmentation performance on Unet (d) Segmentation performance on DeepLabv3; (e)Segmentation performance on segnet;(f)Segmentation performance on our.

We can find that besides our result, the other methods predict melons as Interference from Figure 10. They make the pixel classification not correctly, but (f) shows the infrared

image segmentation result was superior. Our model provided the best results on melons, it's PA, MPA, and MIoU were 99.51%, 94.75%, and 92.67% respectively. These indicators are a good measure of the effect of segmentation.



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**)

Figure 10. The comparison result of different methods on melons dataset. (a)Test image;(b)Segmentation performance on FCN-32s;(c)Segmentation performance on Unet (d) Segmentation performance on DeepLabv3; (e)Segmentation performance onsegnet;(f)Segmentation performance on ours.

3.6 Ablation study

In this part, we explore the performance of each unit. The mean intersection over union (Mean IoU) is used to estimate the model performance. Table 5 shows the different parts of the models. We make the PSPnet as our baseline. It can show that the MS-AFF model has achieved nearly 2.62% improvement than baseline, the non-local Module has get 0.25% than the baseline, and the GSA Module has improved by 1.57%. Our proposed method improves the baseline from 87.21% to 90.54%. Table 5 shows that the proposed model contains all components (i.e., MS-AFF and GSA) that get the best performance.

Table 5. Detailed Mean IOU comparison of proposed our method.

Method	**Mean IOU(%)**
Baseline(PSPNet)	87.21
Baseline+Non-local	87.46(**0.25**)
Baseline+ MS-AFF	89.83(**2.62**)
Baseline+GSA	88.68(**1.57**)
Baseline+ MS-AFF +GSA(Ours)	**90.54**(**3.33**)

4. Conclusions

This paper focuses on buried object's infrared image segmentation. A multi-scale attentional feature fusion (MS-AFF) method is proposed for infrared image semantic segmentation. We integrate a series of feature maps from different levels by an atrous spatial pyramid structure to obtain rich representation ability on the infrared images. We also proposed a global spatial information attention module to let the model focus on the target region and reduce disturbance. We also present an infrared segmentation dataset based on an infrared thermal imaging system. In the experiment, compared with other methods in the infrared segmentation dataset, compared to other different methods, segmentation accuracy of this method is higher. It can be seen that the segmentation method which is proposed performed good effect on the buried object's infrared image dataset. The segmentation accuracy of this method is higher than other different methods. We can see that the proposed segmentation method has good performance on the buried object's infrared image dataset.

Acknowledgments:

This work was supported by the Natural Science Foundation of China (Project No. 61420106011, 61572307).

References

[1] Otsu, N.: A Threshold Selection Method from Gray-Level Histograms. IEEE Transactions on Systems Man & Cybernetics 9(1), 62-66 (2007).

[2] Nock, R., Nielsen, F.: Statistical region merging. IEEE Transactions on Pattern Analysis & Machine Intelligence 26(11), 1452 (2004).

[3] Dhanachandra, N., Manglem, K., Chanu, Y.J.: Image Segmentation Using K -means Clustering Algorithm and Subtractive Clustering Algorithm. Procedia Computer Science 54, 764-771 (2015).

[4] AlSaeed, D.H., Bouridane, A., ElZaart, A., Sammouda, R.: Two modified Otsu image segmentation methods based on Lognormal and Gamma distribution models. In: 2012 International Conference on Information Technology and e-Services 2012, pp. 1-5. IEEE .

[5] Xia, Liu, Shi, Weng, Liu: Cloud/snow recognition for multispectral satellite imagery based on a multidimensional deep residual network. International Journal of Remote Sensing (2019).

[6] Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Communications of the ACM 60(6), 84-90 (2017).

[7] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014).

[8] He, K., Zhang, X., Ren, S., Jian, S.: Deep Residual Learning for Image Recognition. In: IEEE Conference on Computer Vision & Pattern Recognition 2016.

[9] Wu, Z., Shen, C., Hengel, A.v.d.: Bridging category-level and instance-level semantic image segmentation. arXiv preprint arXiv:1605.06885 (2016).

[10] Wang, P., Chen, P., Yuan, Y., Liu, D., Huang, Z., Hou, X., Cottrell, G.: Understanding convolution for semantic segmentation. In: 2018 IEEE winter conference on applications of computer vision (WACV) 2018a, pp. 1451-1460. IEEE.

[11] Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Transactions on Pattern Analysis & Machine Intelligence 40(4), 834-848 (2018a).

[12] Li, H., Xiong, P., An, J., Wang, L.: Pyramid attention network for semantic segmentation. arXiv preprint arXiv:1805.10180 (2018).

[13] Lin, M., Chen, Q., Yan, S.: Network in network. arXiv preprint arXiv:1312.4400 (2013).

[14] Jie, H., Li, S., Gang, S., Albanie, S.: Squeeze-and-Excitation Networks. IEEE Transactions on Pattern Analysis & Machine Intelligence PP(99) (2017).

[15] Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: CBAM: Convolutional Block Attention Module. (2018).

[16] Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition 2018b, pp. 7794-7803.

[17] Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition 2017, pp. 2881-2890.

[18] Long, J., Shelhamer, E., Darrell, T.: Fully Convolutional Networks for Semantic Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 39(4), 640-651 (2015).

[19] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062 (2014).

[20] Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. In: Proceedings of the IEEE international conference on computer vision 2015, pp. 1520-1528.

[21] Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence 39(12), 2481-2495 (2017).

[22] Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention 2015.

[23] Chen, L.-C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017).

[24] Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. (2018b).

[25] Xu, Z.G., Wang, J., Wang, L.Y.: Infrared Image Semantic Segmentation Based on Improved DeepLab and Residual Network. In: 2018 10th International Conference on Modelling, Identification and Control (ICMIC) 2018.

[26] He, K., Zhang, X., Ren, S., Sun, J.: Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. IEEE Transactions on Pattern Analysis & Machine Intelligence 37(9), 1904-1916 (2014).

[27] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L.: Andreas Kö pf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. Advances in Neural Information Processing Systems 32.

[28] LeCun, Boser, Denker, Henderson, Howard, Hubbard, Jackel: Backpropagation Applied to Handwritten Zip Code Recognition. Neural Computation (1989).

[29] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 25, 1097-1105 (2012).

Figures

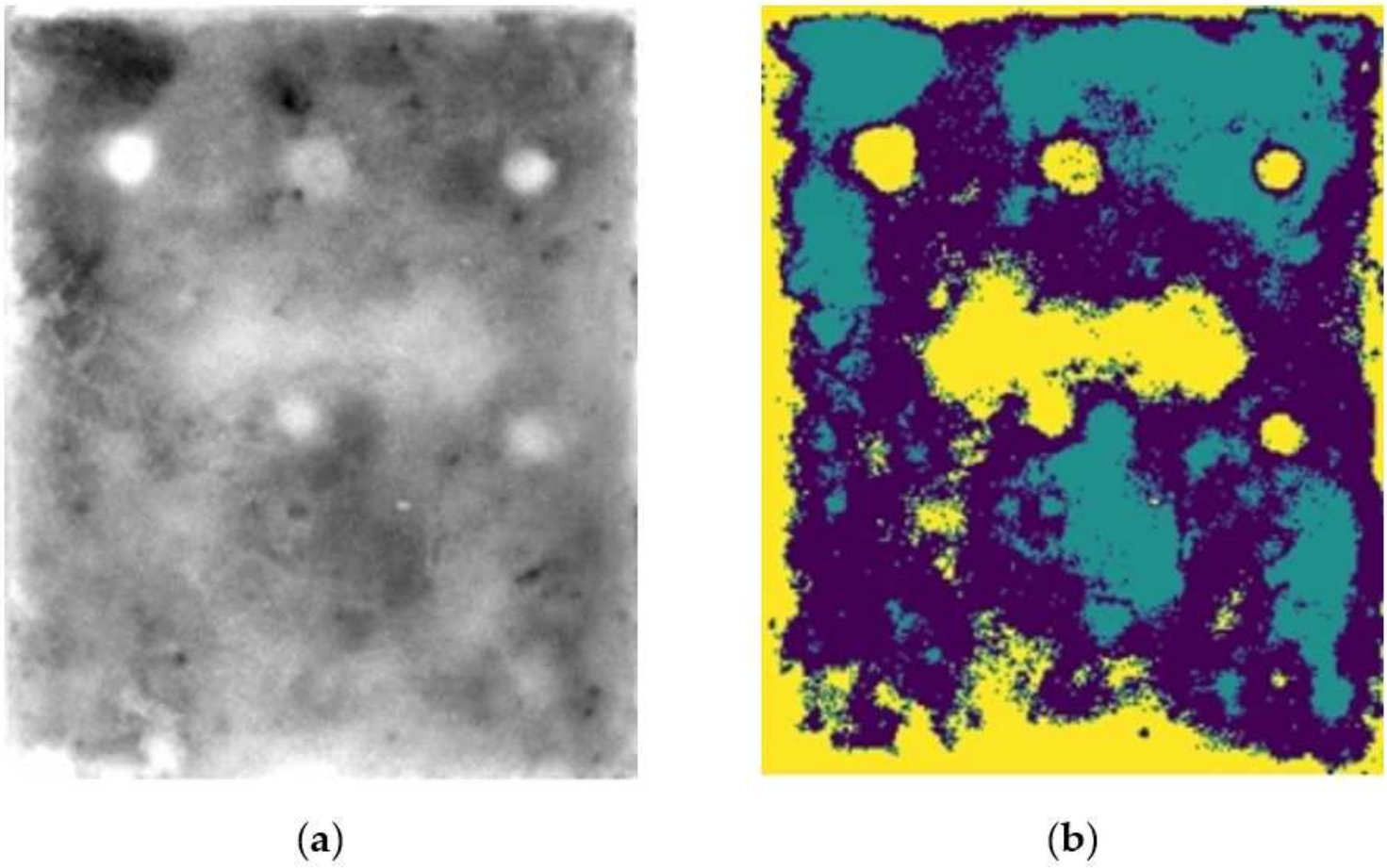


Figure 1

The image is segmented by the clustering method: (a) original image; (b) result of the clustering method.

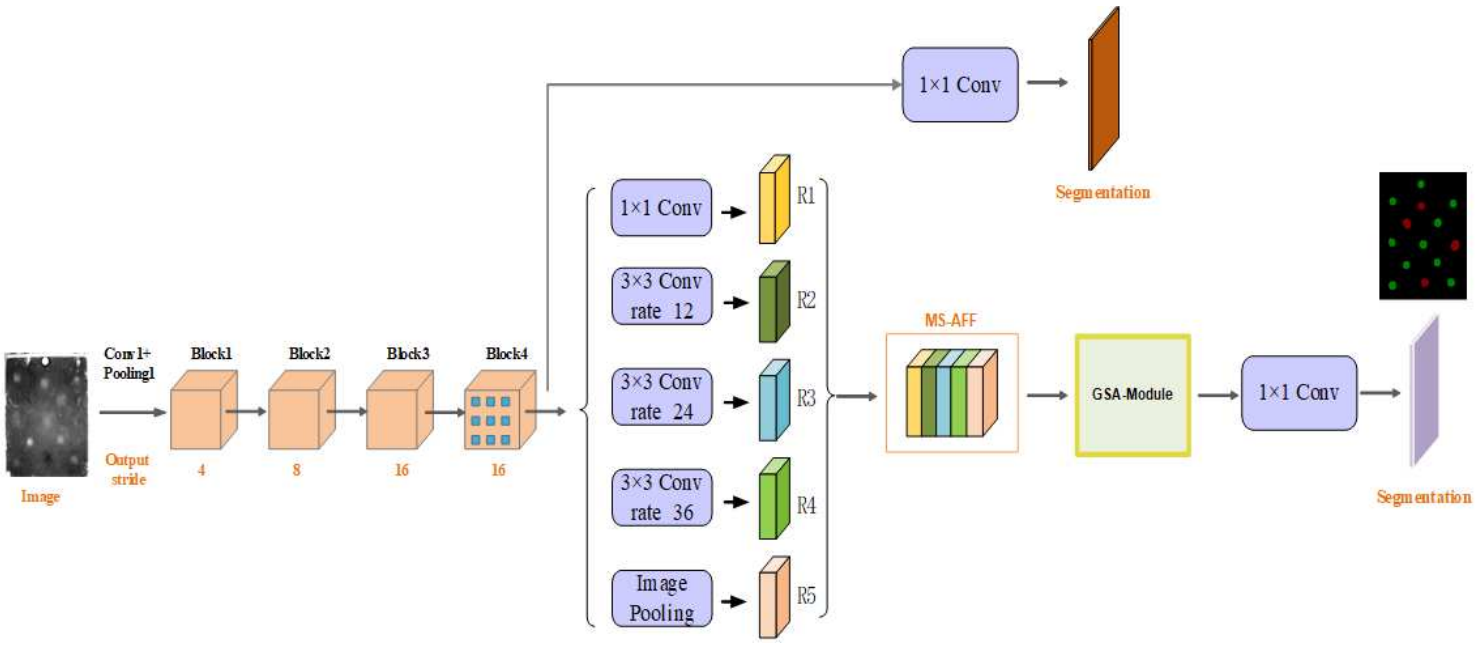


Figure 2

The main pipeline of infrared image segmentation. The infrared image is fed to the network to extract the feature. Then, MS-AFF integrates a series of feature maps. The GSA-Module to let the model focus on the target region. Finally, the network generates the segmentation result.

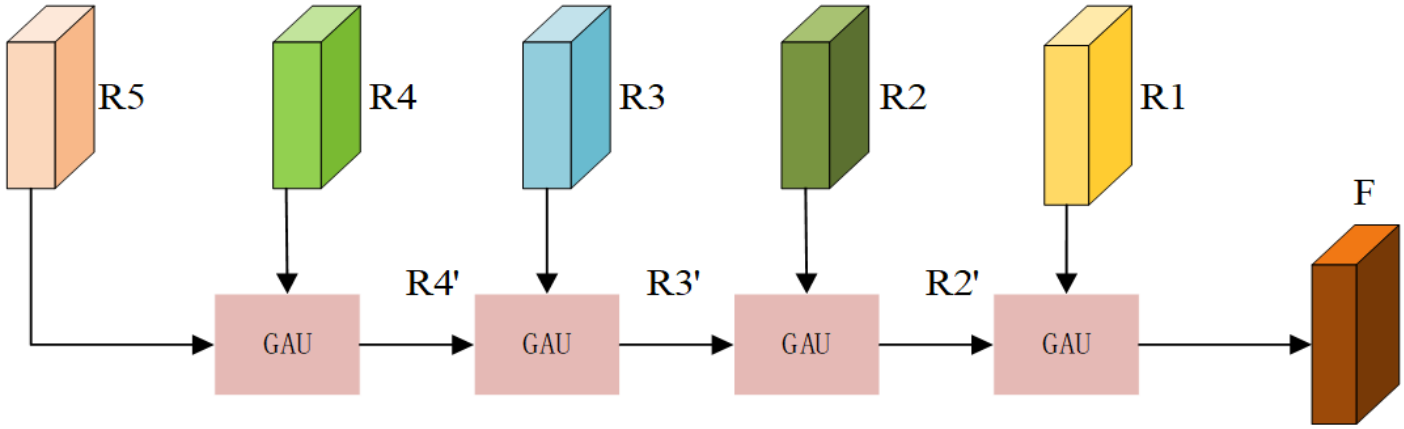


Figure 3

A multi-scale attentional feature fusion. The features of the adjacent feature maps are fused by the GAU.

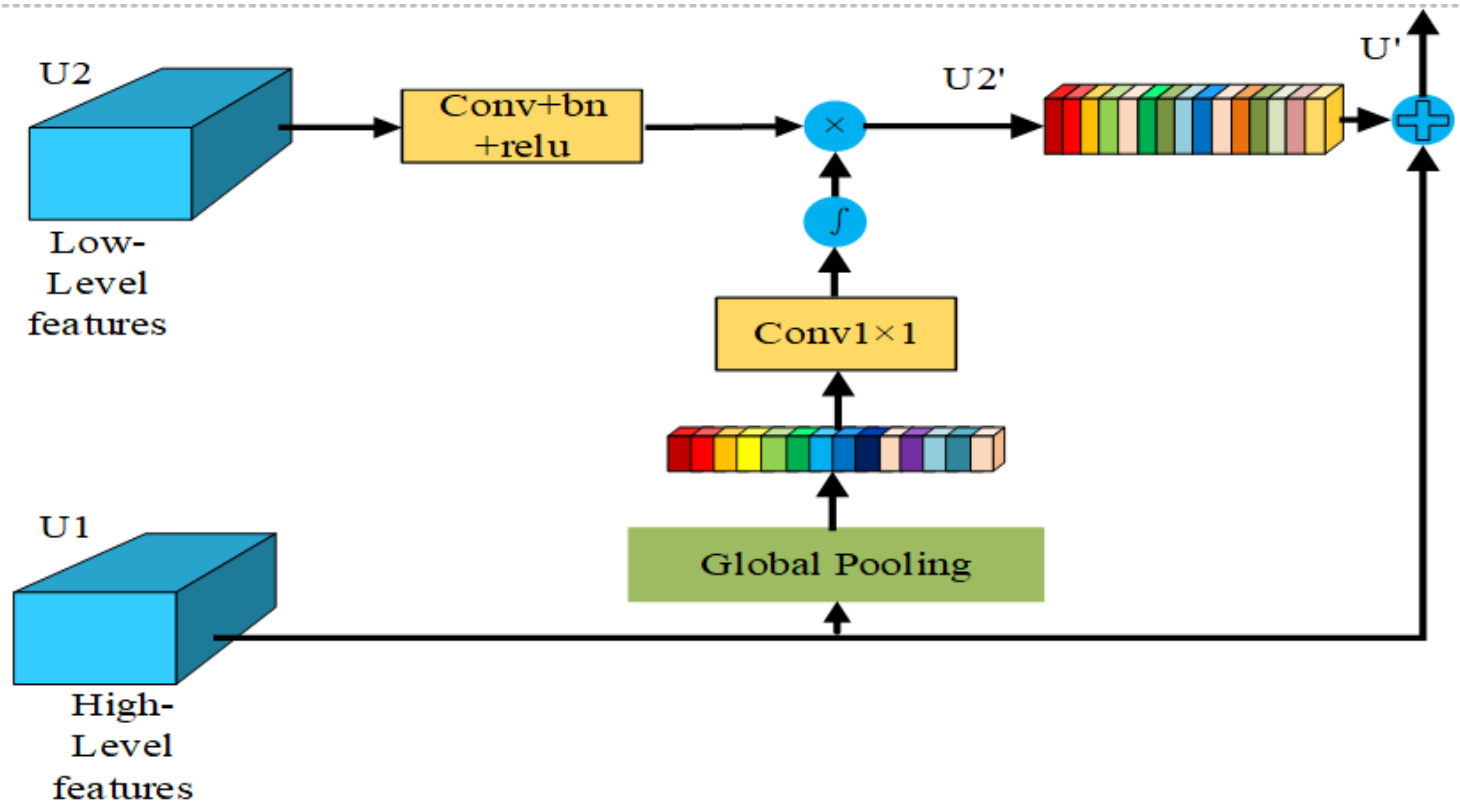


Figure 4

The network structure of the GAU.

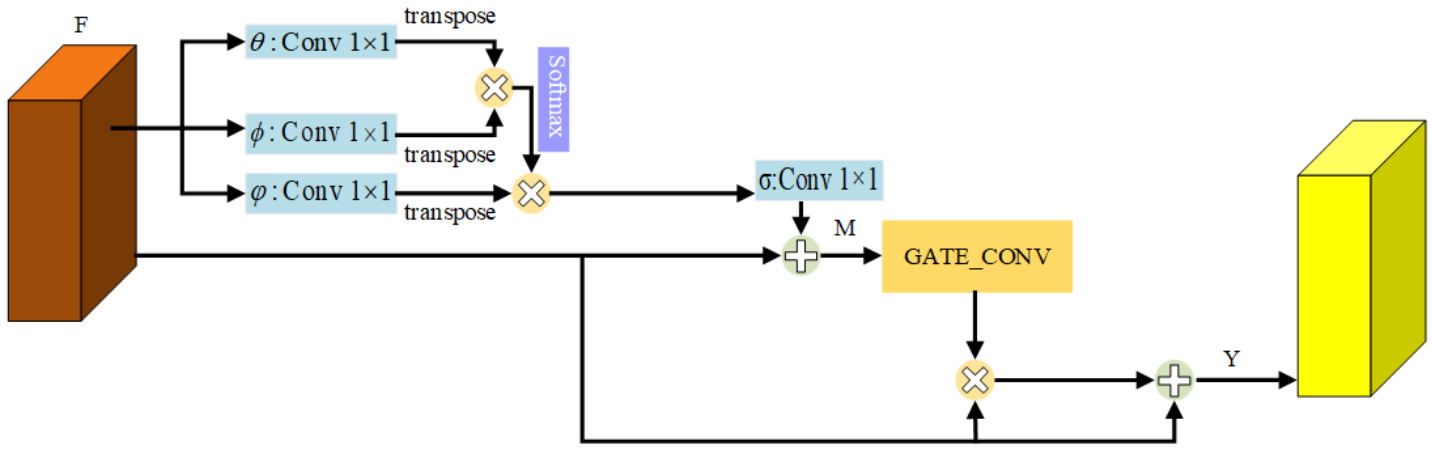


Figure 5

Network structure of the global spatial information attention module.

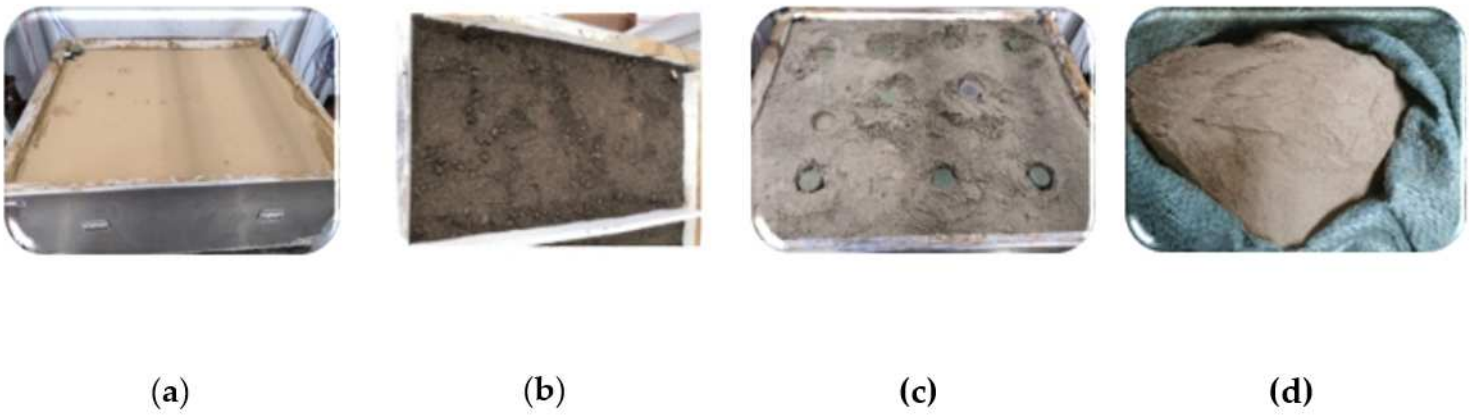


Figure 6

Four experimental soil environments: (a) mineral soil; (b) organic soil;(c) fine sand soil; (d) yellow soil.

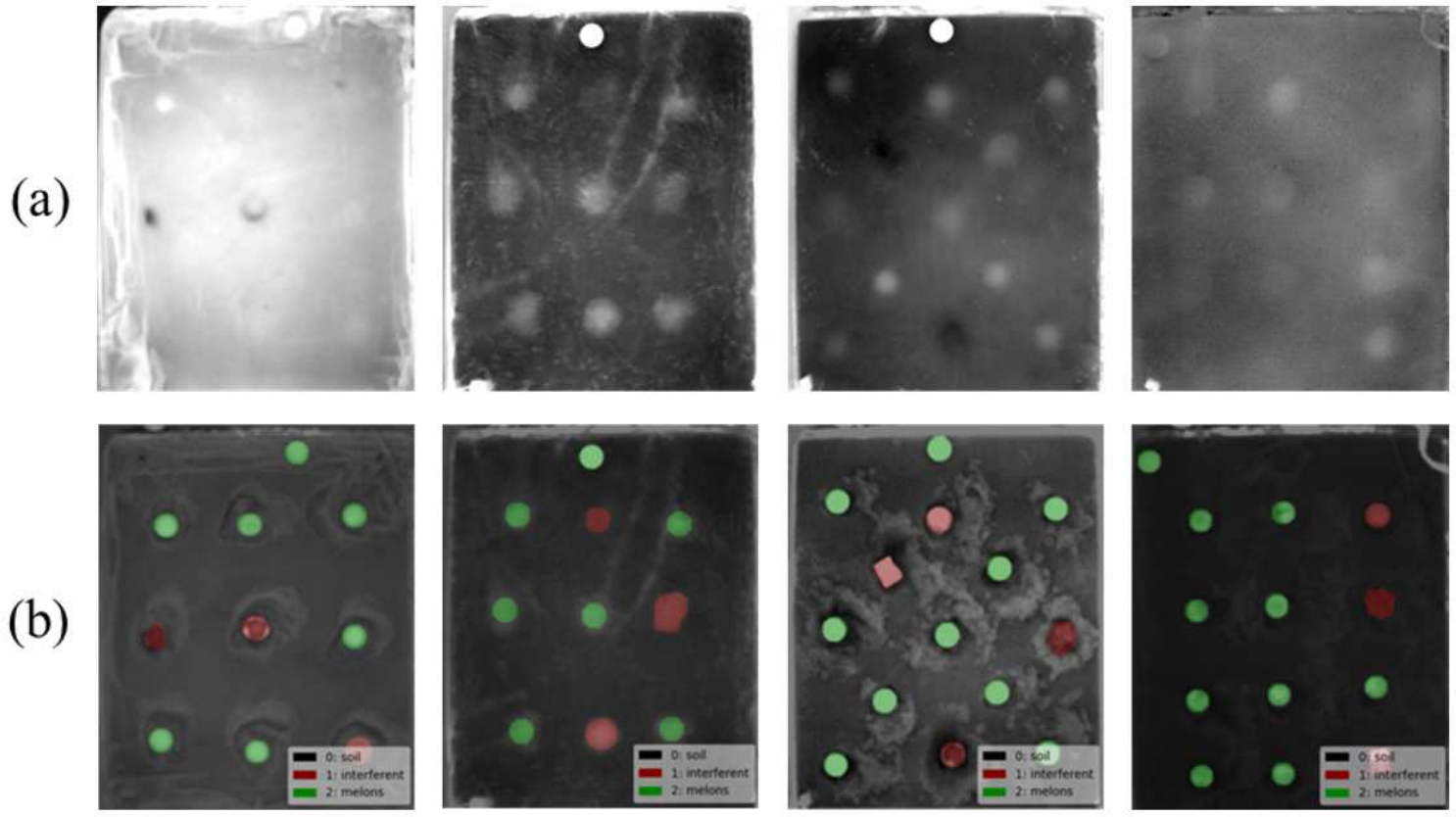


Figure 7

Infrared image after processing: (a)original image; (b) ground truth.

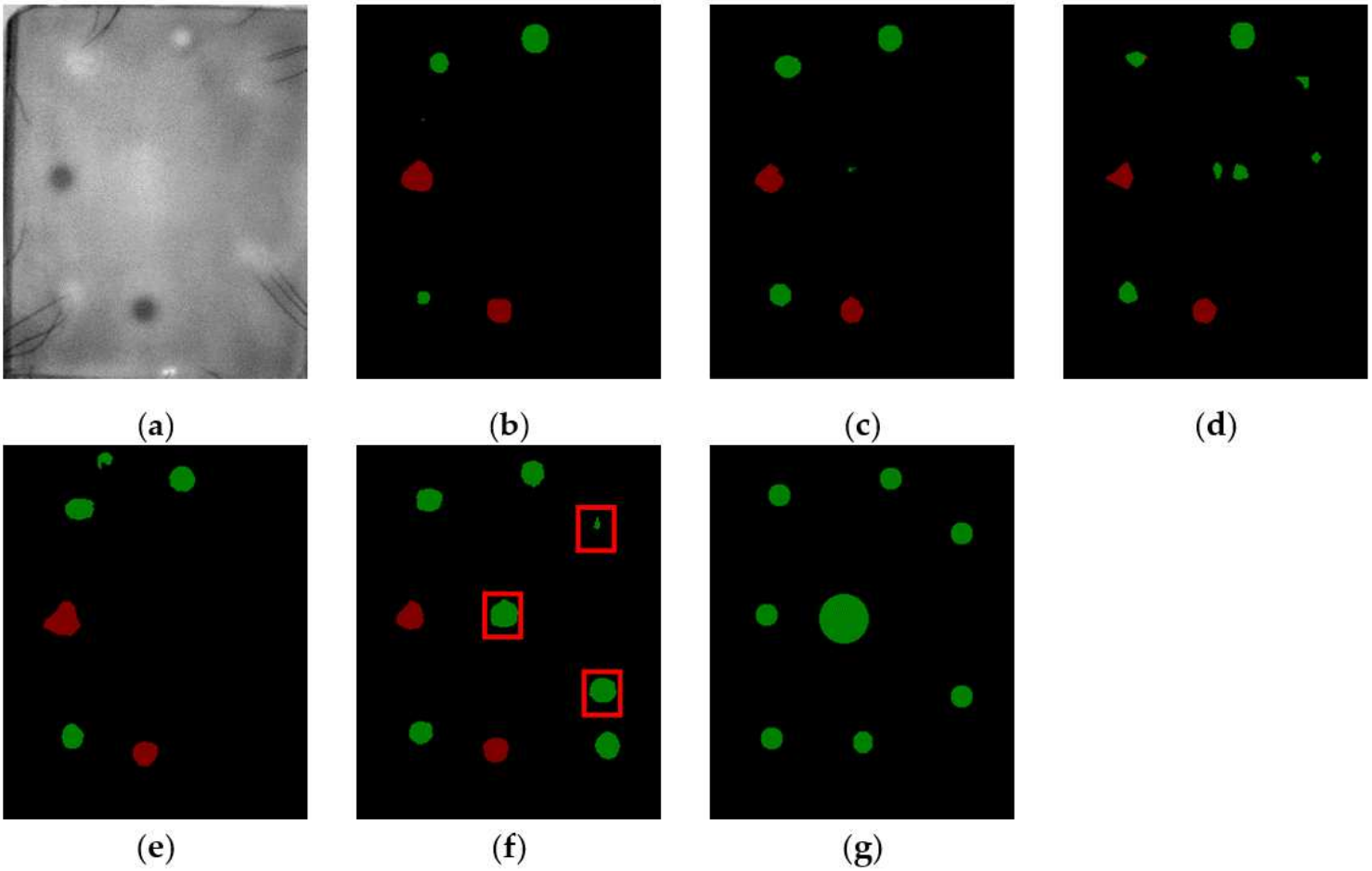


Figure 8

The comparison result of dierent methods on melons dataset. (a)Test image;(b)Segmentation performance on FCN-32s;(c)Segmentation performance on Unet (d) Segmentation performance on DeepLabv3; (e)Segmentation performance on segnet;(f)Segmentation performance on ours. (g) ground truth.

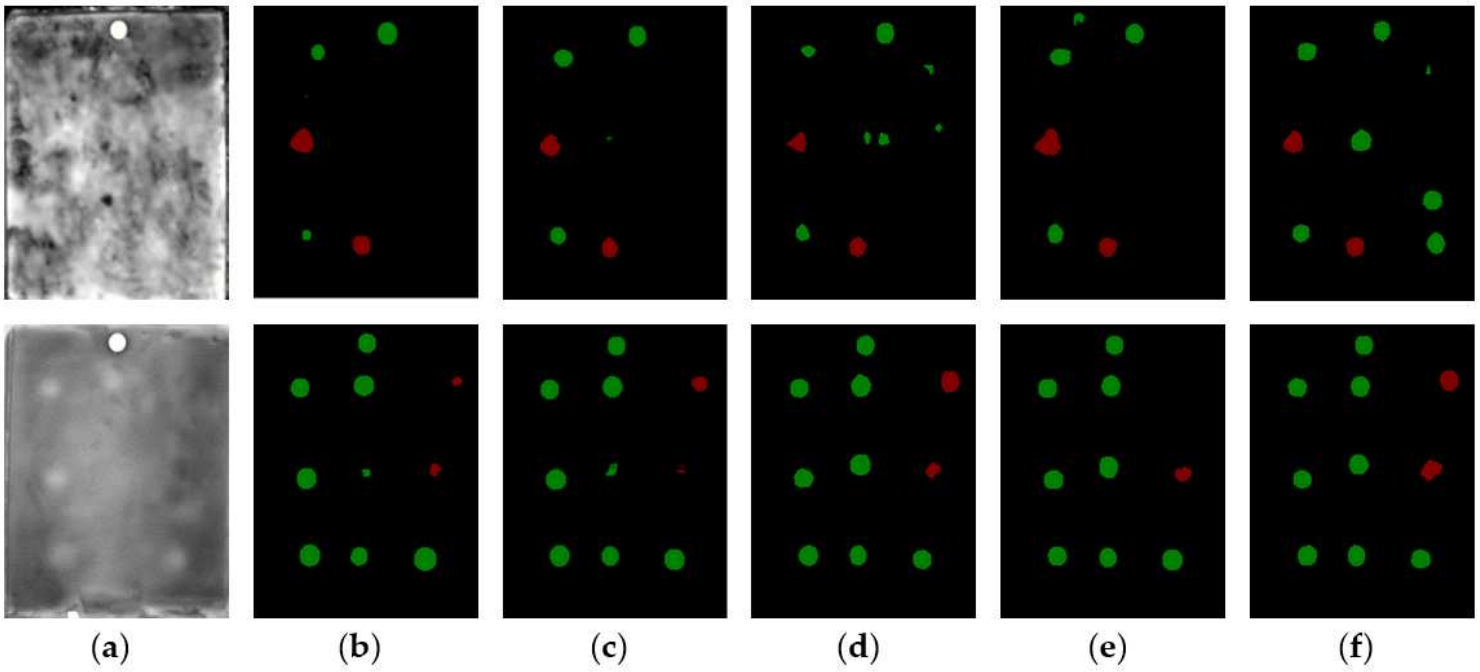


Figure 9

The comparison result of different methods on melons dataset. (a)Test image;(b)Segmentation performance on FCN-32s;(c)Segmentation performance on Unet (d) Segmentation performance on DeepLabv3; (e)Segmentation performance on segnet;(f)Segmentation performance on our.

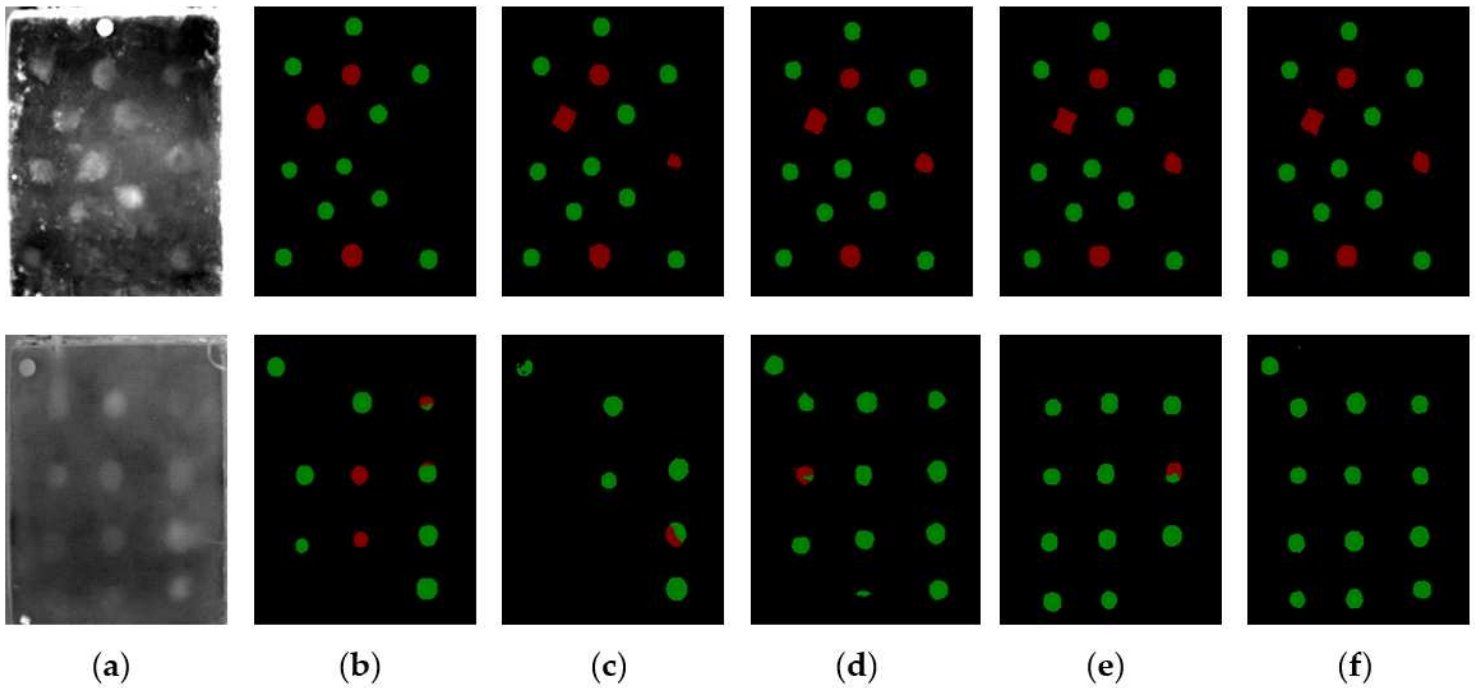


Figure 10

The comparison result of different methods on melons dataset. (a)Test image;(b)Segmentation performance on FCN-32s;(c)Segmentation performance on Unet (d) Segmentation performance on DeepLabv3; (e)Segmentation performance onsegnet;(f)Segmentation performance on ours.