A Human Target Infrared Image Segmentation Approach Based on Convolution Neural Network

Chao Liu¹, Qingping Hu², Yuan Yao¹

¹Institute of Systems Engineering, Academy of Military Sciences, People’s Liberation Army, Beijing, China
²Department of Ordnance Engineering, Naval University of Engineering, Wuhan, China

E-mail:generaladolph@163.com

Abstract. In order to effectively segment the human target under complex background constraints, we present an infrared target segmentation method based on deep convolution neural network, and proposes the loss function based on the intersection-over-union for network optimization. Firstly, we design the network architecture which consists of a contracting path to capture the feature content and a symmetric expanding path that enables precise localization. And then rely on the powerful data amplification technology to effectively train the available sample data. The experimental results show that the network can make full use of the prior information of the data to study the characteristics of the human target, which can use less training data for end-to-end training in the human body target infrared image segmentation. And segmentation effect is superior to the traditional image segmentation algorithm. In addition, the network segmentation speed is very fast, 320 ×256 size image segmentation takes less than 0.2 seconds, to meet the human body target image segmentation of the effectiveness and real-time needs.

1. Introduction
Infrared imaging is widely used in the field of safety protection, automobile-assisted driving and other monitoring because of its all-weather work, detectable hidden heat targets, passive concealment, and long-distance operation. The resulting infrared human target segmentation technology has become one of the research hotspots in the development of machine vision[1].

At present, for infrared human targets, scholars have studied a large number of segmentation algorithms, generally can be divided into edge-based methods, regional growth methods, threshold methods and feature clustering algorithms, such as the largest class-based variance method, the maximum expectation method, the maximum entropy segmentation method and so on[2–5]. These traditional human target segmentation algorithms have a good
segmentation effect when the target and background grayscale distribution are different. However, these methods are often split or under-segment when applied to complex environment images where the target and background grayscale features are mixed, or the target occupies fewer pixels.

In recent years, with the improvement of hardware computing performance such as GPU, deep convolutional neural network has made a major breakthrough in many machine visual tasks, and the research on image segmentation using convolutional neural network has been widely concerned by scholars, but it is different from the traditional image recognition task. The task of image recognition is to accurately attribute the image to a certain class according to the target contained in the image, without marking the location of the target in the image, and the image segmentation needs to accurately indicate the location of different areas such as the target and background in the image, that is to say each pixel position of the image contains its own output label [6,7].

The difficulty of convolutional neural network for image segmentation problem is that the output image approximated by the original image calculated by the convolutional neural network is under-sampled, the location information is lost, and the probability that each pixel of the image belongs to a certain category. At present, there are two main ideas for using deep convolution network to deal with image segmentation: one is to transform the well-trained classification network for the full convolution network, and then the output of the middle layer is interpolated to obtain the original image size, to obtain the split result, the typical application of this idea is the semantic segmentation of the image[8,9]. The other is to randomly select several rectangular areas from the image, use these rectangular areas as training data to extract the characteristics of the target area, and thus train the classifier to detect the rectangular area containing the target, a typical application of this idea is the target detection in the image [10,11].

In 2015, researchers at the University of Freiburg in Germany proposed a U-Net convolution network for the division of medical ultrasound images, which has the advantage of producing more accurate segmentation with few training images [12]. Inspired by the U-Net network structure, for the study of human target segmentation, we designed an infrared target segmentation convolution neural network with encoder-decoder style type. The main idea is that the encoding layer extract and combine low-level features, and the decoding layer expands the encoding layer by replacing the pooling operation with the upper sampling operation. And connect the high-resolution features of the encoding layer and the upper sampling layer with a jump structure to ensure that subsequent convolution layers can learn more accurate output. The experimental results show that the trained network can correctly identify the human target in the infrared image, divide the image into a white foreground and black background containing the human object, and be able to perform better than the human person.

2. Data pre-processing
In the absence of an available infrared data set for training and testing, we used the infrared observation mode of the microlight/infrared observer developed by the project team to collect a large number of pedestrian-containing 640×480 resolution video data. From which we intercepted 180 images with complex backgrounds and including human targets of different scales as data sets. The data sets is then labeled using Photoshop CS6 software, the area
where the infrared human target is located as white (pixel value is 255), and the remaining areas are given black (pixel value is 0), as shown in figure 1, and then the data set is divided into training and test sets.

![Image](figure1.png)

**Figure 1.** Typical image of the training set.

Infrared target segmentation convolution neural network is mainly divided by learning the characteristics of the human body, so the input block should be required to contain a complete human image, but due to the limitations of our laboratory hardware, the maximum image size is 128×128 that can be trained at present, and cannot directly carry out the 640×480 resolution image training. If the image block is extracted directly for training, the lack of image blocks containing the complete human target will cause the network to fail to take full advantage of the overall characteristic information of the target, so we do 2 times the subsampling of the image of the training set, and get the 320×240 resolution as the image set, so that most of the 128×128 image blocks extracted from it contain complete human target information. As shown in figure 2.

![Image](figure2.png)

**Figure 2.** The 128×128 image block.

For our task, with limited data, we amplify data by applying elastic transformations and noise additions. Amplify images (including images and labels) by rotating 45 degrees, scaling 15-25%, cutting operations, and vertical and horizontal flips, and adding Gaussian noise. Allow network learning to change this distortion and noise, increase the generalization ability. Finally, in order to overcome the gradient explosion in the training process, the training set is normalized.

### 3. Network Architecture

The infrared target segmentation convolution neural network designed in this paper consists of the contracting path of the left implementation feature extraction and the expanding path for precise positioning on the right side, while the high-resolution features from the compression layer and the upper sampling layer are combined using the jump structure to ensure that the subsequent convolution layer can learn more accurate output. The entire network consists of a convolution layer, an activation layer, a normalization layer, an upper sampling layer, and a Dropout layer, and its basic structure is shown in figure 3.
The contraction path on the left follows the typical architecture of the convolution network, consisting mainly of three modules, each consisting of two convolution modules and one subsampling module: each convolution module first applies a $3 \times 3$ convolution layer CONV to capture the input infrared image feature, and the convolution layer uses a zeroing operation to ensure that the output shape does not change. Then with an activation layer to provide a nonlinear source for the network, the essence is an exponential activation function ELU, and finally with a batch normalization layer BN used to normalize the network parameters, improve the convergence speed of the network. The subsampling module uses the MAXPOOL, a maximum pooled operation of $2 \times 2$ with a stride of size 2, and in each down sampling step, we double the number of feature channels to ensure that the network parameters are equal.

The right-hand extension path is symmetrically distributed relative to the compression path and also consists of three units, each consisting of two convolution modules and one up sampling module: each up-sampling module UPSAMPLE first increases the spatial resolution of the feature map through bidirectional interpolation and halves the number of feature channels. It is then connected to the feature map of the corresponding dimension from the shrink path. Then follow the convolution module of the extension path, first apply the DROPOUT layer to modify the network structure during training, inhibit the network overfitting, then apply a $3 \times 3$ convolution layer CONV for feature positioning, and finally, exponential function activation ELU provides nonlinearity for the network. Because we wanted to predict the probability that pixels were human targets, we used the Sigmoid activation function SIGMOID on the last layer of convolution.

4. Loss Function
As a method of data expression, deep neural network can theoretically approach any function, the essence of its training is to find the right weight and bias through a learning algorithm,
ensure that the output of the network can fit all the training inputs, generally through the loss function to quantify this goal. At present, the gradient descent learning technique is mainly used to solve the minimization of loss function, so it is also required that the loss function has a guidable nature.

In the study of two-value image segmentation, the network model can make the probability that the point belongs to the target class for each pixel of the input image, and generally uses binary cross entropy as the loss function. The training target is that the binary cross entropy sum of all pixels is during training, and its expression is shown in equation (1):

$$C = -\frac{1}{N} \sum_{i=1}^{N} \left[ g_i \log(p_i) + (1 - g_i) \log(1 - p_i) \right] = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{2} g_{ij} \log(p_{ij})$$

(1)

Among them, $j=1, 2$ is the class label (representing the foreground human target class, background class), $N$ is the number of pixels contained in the image block, $g_i$ is the true segmentation pixel (i is the sample index), $p_{ij} \in (0,1)$($\sum_{j=1}^{2} p_{ij} = 1 \forall i, j$) is the probability of the predicted binary split pixel belong to $j=1, 2$ class.

Because the essence of deep neural network training is to find the weight and bias that can minimize the loss function, different forms of loss function can guide the network to learn in different directions, which has a great influence on the final image segmentation effect. The Jaccard index and the Dice index are used as evaluation index of image segmentation effect, which are used to evaluate the degree of match between the model's prediction results and the actual situation, so the learning degree of the network can be quantified, and the binary vector form of the two indices can be differential, so it can be used as a loss function of the two-value image segmentation network. Let's analyze the motivations of these two indices as loss functions and calculate their derivatives.

The Jaccard index, also known as intersection over union, is collated as the overlap between the predicted and real target marker values. Which is the intersection of prediction results and true values divided by their union, as shown in equation (2):

$$J = \frac{TP}{TP + FP + FN} = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

(2)

Where A indicates the prediction results, B indicates the actual situation, TP is a true positive region, FP is a false positive region, FN is a false negative region, the visual effect is shown in figure 4.

Although the equation (2) can be very intuitive to explain the physical significance of the Jaccard index, but because of the inclusion of intersections and other operations, so can not be
differential, cannot be used as a loss function of the neural network. We use the binary vector $P$ and $G$ to rewrite the equation (2) based on vector operation, as shown in equation (3):

$$J = \frac{|P \cdot G|}{|P|^2 + |G|^2 - |P \cdot G|} = \frac{\sum_{i}^{N} p_i g_i}{\sum_{i}^{N} p_i^2 + \sum_{i}^{N} g_i^2 - \sum_{i}^{N} p_i g_i}$$  \hspace{1cm} (3)$$

Where $N$ is the number of pixels contained in the image block, $P$ is the predicted binary segmentation pixel, and $G$ is the true segmentation pixel, The derivative of each activation for Jaccard is:

$$\frac{\partial J}{\partial p_i} = \frac{g_i \left( \sum_{i}^{N} p_i^2 + \sum_{i}^{N} g_i^2 - \sum_{i}^{N} p_i g_i \right) - (2p_i - g_i) \sum_{i}^{N} p_i g_i}{\left( \sum_{i}^{N} p_i^2 + \sum_{i}^{N} g_i^2 - \sum_{i}^{N} p_i g_i \right)^2}$$  \hspace{1cm} (4)$$

However, the direct use of The Jaccard function as a loss function there are certain limitations, first of all, when $A$, $B$ coincide, the prediction results and the actual results are the same, the prediction is the best, but the Jaccard index is the maximum 1, our aim is to minimize the loss function, so $1 - \text{Jaccard}$ should be used as a loss function, Second, in cases where real human objectives are not included in the real image block (as is the case when the data set is extracted in the image block), the loss is maximized even if the output segmentation plot contains very few false positive areas, so it must be taken into account that for samples containing very few or no human instances, it is not ideal as a loss function. To overcome this problem, we adding an constant item $S$ to the molecule sand and denominator of the Jaccard index, so our loss function based on the Jaccard index can be defined as:

$$J_{\text{loss}} = 1 - \frac{|P \cdot G| + S}{|P|^2 + |G|^2 - |P \cdot G| + S} = 1 - \frac{\sum_{i}^{N} p_i g_i + S}{\sum_{i}^{N} p_i^2 + \sum_{i}^{N} g_i^2 - \sum_{i}^{N} p_i g_i + S}$$  \hspace{1cm} (5)$$

The Dice index, also known as the F1 score or Dice similar coefficient, calculates only the secondary positive region in the numerator and denominator compared to the Jaccard index, which is not significantly different from the form of the Jaccard index. However, since it does not meet the triangulation, it can be considered a semi-measured version of the Jaccard index, as shown in equation (6):

$$D = \frac{2TP}{2TP + FP + FN} = \frac{2 |A \cap B|}{|A| + |B|}$$  \hspace{1cm} (6)$$

Similarly, we can override the equation (6) using the vector operations of binary vectors $P$ and $G$, as shown in the equation (7):

$$D = \frac{2|P \cdot G|}{|P|^2 + |G|^2} = \frac{2 \sum_{i}^{N} p_i g_i}{\sum_{i}^{N} p_i^2 + \sum_{i}^{N} g_i^2}$$  \hspace{1cm} (7)$$
Where $N$ is the number of pixels contained in the image block, $p_i \in P$ is the predicted binary segmentation pixel, and $g_i \in G$ is the true segmentation pixel, the Derivative for each active dice index is:

$$\frac{\partial D}{\partial p_j} = 2 g_i \left( \sum_{i}^N p_i^2 + \sum_{i}^N g_i^2 \right) - 2 p_j \sum_{i}^N p_i g_i \left( \sum_{i}^N p_i^2 + \sum_{i}^N g_i^2 \right)^2$$

(8)

As mentioned above, the Dice index, as a different version of the Jaccard index, has similar functions, and there are similar limitations to the Dice index as a loss function, so draw on the Jaccard loss function design. We can define a loss function based on the Dice index as:

$$D_{loss} = -\frac{2|P\cap G| + S}{|P| + |G| + S} = -\frac{2 \sum_{i}^N p_i g_i + S}{\sum_{i}^N p_i^2 + \sum_{i}^N g_i^2 + S}$$

(9)

5. Training and tests

The size of the input image we define is 128×128 infrared grayscale image block, with the output is 128×128 single-channel prediction mask module. In terms of optimization strategy, the SGD optimizer is not effective, using Adam directly as the optimizer, learning rate is $lr = 10^{-5}$, and finally using the previously designed network structure on the GTX850 GPU training. Because the training of the network model designed in this paper is end-to-end training, it is possible to test images of any size. Prior to this, we analyze the three loss functions applied to image segmentation, first of all, we compare the effect of the three loss functions under the network built in this paper, as shown in figure 5, where figure 5(a) is original image, figure 5 (b) is segmentation result with dice loss function, figure 5 (c) is segmentation result with jaccard loss function , figure 5 (d) is segmentation result with cross entropy loss function.

![Figure 5. The segmentation effect of different loss functions.](image_url)

From the comparison in figure 5, it can be seen that all three loss functions can obtain a better effect of human target segmentation, but the comparison of the three can be seen: the segmentation of the target is relatively smooth with Dice loss function , and there is no adhesion between the targets; There is a certain fracture of the human target of the function of the Jaccard loss function, and there is some adhesion between the multiple targets; the outline...
of segmentation target is not smooth enough with the binary cross-entropy loss function. And the analysis of dice loss function has some advantages over the other two loss functions in the division of the infrared human object. Figure 6 is the result of the division of different scale targets under complex scenarios using the Dice loss function, where figure 6(a) is Large-scale targets, figure 6(b) is small-scale targets, figure 6(c) is complex scenarios.

![Figure 6](image-url)

**Figure 6.** The results of the segmentation with different scale targets in complex scenarios.

As can be seen from figure 6(a): The convolutional neural network designed in this paper has a good segmentation effect on the larger scale human target, and the division error is also small. As can be seen from figure 6(b), there are smaller human targets in the test original map, and the target and surrounding environment grayscale is relatively close, the human eye is difficult to distinguish, the network can correctly divide the target. Although the target is small, the environment is complex in Figure 6(c), the segmentation target is accurately predicted. The convolutional neural network designed in this paper has a good segmentation effect on the infrared human target in different scenes, which proves the universality of the network.

6. Conclusion
In this paper, a network structure containing the compression path for feature content extraction and the symmetrical extension path capable of precise positioning is proposed, and three different loss functions are defined and then the available sample data can be effectively trained by strong data amplification technology. This method of segmentation has the characteristics of distinguishing the detail texture of the image and applying it to the complex background, and can effectively extract the human target of different scales under the complex scene, and the segmentation effect is better than the traditional image segmentation algorithm.

References
[1] He Fuliang, Guo Yongcai, orgasm. Improved PCNN method for infrared image segmentation of human targets in complex environments. *Journal of Optics* 2017 (2) 183-92
[2] Xu Xinzheng, Ding Shifei, Shi Zhongzhi. New theories and new methods for image segmentation. *Journal of Electronics* 2010 (S1) 76-82
[3] He Jun, Ge Hong, Wang Yufeng. A review of the study of image segmentation algorithms.
Convolutional Encoder-Decoder Architecture for Image Segmentation[J]. arXiv:1511.00561 [cs], 2015.

BELL S, LAWRENCE ZITNICK C, BALA K, et al. Inside-Outside Net: Detecting Objects in Context With Skip Pooling and Recurrent Neural Networks[C]//2016: 2874–83

RENN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[G]//CORTES C, et al. Advances in Neural Information Processing Systems 28. Curran Associates, Inc., 2015 91–9

RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional Networks for Biomedical Image Segmentation[J]. arXiv:1505.04597 [cs], 2015