CIT. Journal of Computing and Information Technology, Vol. 31, No. 3, September 2023, 167–183
doi: 10.20532/cit.2023.1005740

An Innovative Deep Learning Approach for Image Semantic and Instance Segmentation

Chuangchuang Chen[1], Guang Gao[1], Linlin Liu[1] and Yangyang Qiao[2]

[1]School of Network Engineering, Zhoukou Normal University, Zhoukou, Henan, China
[2]School of Information Engineering, Zhengzhou Technology and Business University, Zhengzhou, China

In this study, we propose a segmentation model based on convolutional neural networks (CNNs) to address image segmentation challenges in computer vision. Prior to designing the model, the activation function and other modules of the convolutional neural network were optimized to meet specific requirements. The segmentation task was transformed into binary classification problem to simplify network calculations and improve efficiency. Additionally, the model utilized a mask map obtained from the semantic segmentation model to aid in instance segmentation. Class activation technology was introduced to extract feature mapping maps. The corresponding thermal maps were obtained to achieve target instance segmentation. To further validate the effectiveness of the segmentation model, simulation experiments were conducted on semantic segmentation and instance segmentation respectively. The results show that the accuracy of the basic semantic segmentation model reached 87.58%, while the average accuracy of the entire class of the optimized instance segmentation model reached 97.9%. Therefore, the research and design of image segmentation models demonstrate high accuracy and good robustness.

ACM CCS (2012) Classification: Computing methodologies → Artificial intelligence → Computer vision

Keywords: CNN, Full Supervision, Image Segmentation, Thermal Diagram, Global Pooling

1. Introduction

Deep learning (DL has become essential in human-computer interaction, and has been extensively studied in the academic community, particularly in fields like computer vision, including image segmentation (IS) and object detection. In recent years, DL has made great progress in the field of image processing. Image processing technology is often applied in various aspects such as traffic management, robotics, and intelligent driving, truly integrating into people's daily lives and work [1].

However, DL confronts challenges, notably the need to improve the accuracy of the model through a comprehensive large-scale dataset. In the context of big data, it undoubtedly becomes an obstacle to DL networks. The significant growth of data leads to the continuous extension of model training time, ultimately leading to an increase in network computing costs. Therefore, improving network training is a crucial point in current image processing technology, especially for tasks like semantic segmentation and IS, which are vital for intelligent human-machine interaction [2].

Semantic segmentation involves classifying images at the pixel level. Compared with traditional feature extraction methods, it demonstrates significant improvements in algorithm accuracy and timeliness, and finds extensive applications in transportation, remote sensing, medicine, and other fields. On the other hand, instance segmentation entails predicting target pixel region, which requires more accurate region labels and class perception mask information, thus presenting greater operational complexity.

Currently, most semantic segmentation relies heavily on labeled datasets, requiring large-scale datasets for training support and high-per-

formance hardware. Instance segmentation models are typically built upon this foundation [3]. However, traditional IS methods have some obvious limitations when dealing with the real image of complex scenes. They often rely on low-level image features such as color, texture, edges, and so on. These features are often not robust enough for dealing with complex or uneven scenes, resulting in unsatisfactory segmentation results. Additionally, they may lack the ability to generalize, and their performance tends to drop dramatically when image conditions, such as lighting, angle, and background, change. Furthermore, the performance of these methods is often negatively impacted by noise and artifacts due to the lack of effective noise suppression or feature extraction mechanisms.

To address the training difficulty of fully supervised IS models, a Fully Convolutional Neural Network (FCN)-based IS model is proposed and optimized using class activation techniques. The optimizers used to train CNN are discussed in depth, which is helpful for the selection of different optimizers for the subsequent research and increases the discussion of practical properties. At the same time, a new Fully Convolutional Neural Network (FCN) model is proposed for semantic segmentation of vehicles.

In addition, the paper tackles the inadequacy of vehicle semantic segmentation datasets by augmenting and expanding existing datasets. To improve the accuracy and efficiency of vehicle instance segmentation, researchers combine the mask information generated by a semantic segmentation network with the heat map generated by a class activation graph algorithm. This approach offers a novel perspective and technical pathway for future research. It has important practical significance and theoretical value for the research and application of vehicle detection, autonomous driving vision systems, and other related fields.

The rest of this study is organized as follows. Section 2 reports on the current research status of DL networks and IS techniques. In Section 3, the design of semantic segmentation and instance segmentation models is presented. The fourth section presents the results of the evaluation of the IS network. The fifth section summarizes the experimental results and concludes the paper.

2. Related Work

Image processing technology is currently focal point in computer vision research, with DL standing out as the most widely used and essential technology in this domain. P. Mukasa *et al.* [4] applied image processing technology to the field of agricultural planting. Due to the ploidy of watermelon seeds, this technology can have a significant impact on their yield. Therefore, seed categories can be classified through multivariate and DL models. Meanwhile, it introduced deep labv 3+and Resnet 18 DL networks for further optimization. The experimental results showed a classification accuracy of 95.5%, which was 26% and 11.2% higher than the Data Driven - Soft Independent Modeling of Class Analogy (DD-SIMCA) and Support Vector Machine (SVM) models, respectively.

C. Bowd *et al.* [5] applied DL to medical image classification and trained it on the feature-based optical coherence tomography angiography optic nerve head dataset using Virtual Games Global (VGG) 16 CNN. The experimental results showed that the area accuracy under the recall curve reaches 0.97, indicating a significant improvement effect.

P. Munoz-Benavent *et al.* [6] addressed the challenge of automatic measurement of fish size, crucial for perceiving changes in the body and environment of fish. Since traditional segmentation network models were computationally overly demanding, a CNN-based IS algorithm was proposed, which mitigated the complex process of parameter adjustment. The experimental results indicated that the number of measurements has increased by 2.45 times.

G. Yuan *et al.* [7] applied DL to image reconstruction and proposed a fast bilateral network that can be applied to grayscale and color image reconstruction. This study introduced bilinear interpolation and CNN to achieve image compression, and the overall reconstruction process was divided into two parts: texture and contour. The experimental results showed that the image reconstruction technology has excellent timeliness, accuracy, and robustness.

It is evident that there is a strong connection between DL technology and image processing, with IS emerging as a critical yet challenging aspect that urgently requires improvement.

J. Niedballa *et al.* [8] highlighted the prevalent use of CNNs in IS. However, the lack of a comprehensive toolbox is one of the main reasons why the research community has not made progress. Based on this, they introduced a R package "images", which implemented DL segmentation workflow by constructing U-Net and U-Net++, including data preprocessing, model training, and testing. The experimental results indicated that the Dice score of the model is 0.91.

B. Ji [9] applied neutral C-means clustering to color IS, initially requiring an optimized linear clustering algorithm to obtain adaptive local spatial neighborhoods. Afterwards, an objective function based on neutral C-means clustering was introduced, incorporating local neighbor data to classify membership levels based on certainty and uncertainty. The experimental results showed that the method has good noise resistance and performance.

X. Yan *et al.* [10] constructed local pre-fitting image functions based on regional features and optimized edge indication functions based on edge features to achieve IS. Additionally, regularization functions were introduced to improve the overall model's timeliness, while addressing the sensitive parameter impact of traditional penalty terms.

Numerous studies have shown that DL is widely used in IS techniques, but it usually requires a large number of labeled datasets to train the model, which increases the computational burden of the model and weakens its segmentation speed. Hence, this study proposes a semantic segmentation and instance segmentation model based on an improved FCN. This model applies the mask graph obtained from semantic segmentation to instance segmentation, and optimizes it using class activation techniques, greatly improving the timeliness of the model.

3. Research Model

CNNs are a commonly used algorithm for image semantic segmentation, utilizing end-to-end training networks to achieve pixel-level prediction. The optimized FCN is studied to improve the accuracy of semantic segmentation. Through certain improvements, the key boundary segmentation mask map output by the model is further applied to image instance segmentation.

3.1. DL-Based Classic CNN Model

The classical CNN is composed of convolutional layers, pooling layers, fully connected layers, and activation functions. The sliding window calculation value of the convolution layer needs to be biased to obtain the final corresponding pixel output value, as shown in equation (1) [11].

$$x_n = f\left(\sum_{i=1}^{I} conv(x_i, K_{ni}) + b_n\right) \tag{1}$$

In equation (1), x_n/b_n respectively represent the n-th output value and its corresponding bias size. I is the input feature level. x_i is the i-th input value. K_{ni} represents the component size of the convolutional kernel in the n-th value of i output. The study selects maximum pooling as the pooling layer method, as shown in equation (2).

$$x_{i,j} = \max\left(\left\{x_{i\times j+k,\ j\times s+k}\right\},\ k = 0,1,\ ...,K\right) \tag{2}$$

In equation (2), $x_{i,j}$ is the downsampling output value of coordinate point (i, j), k represents the fixed edge length for downsampling, and s represents the sliding step size of a fixed area. The two-dimensional feature map obtained after convolutional pooling contains all spatial position data of pixels, as shown in Figure 1 [12].

The fully connected layer is used to transform feature maps into one-dimensional feature vectors, and neurons with pre and post-layer relationships are connected through weights [13]. The Softmax layer is responsible for classifying input features, and after the fully connected layer, it is suitable for multi-segmented types of images, as shown in equation (3).

$$p(y^i = k \mid x^i;\ \omega) = \frac{\exp((\omega_k)^T x^i)}{\sum_{j=1}^{k} \exp((\omega_j)^T x^i)} \tag{3}$$

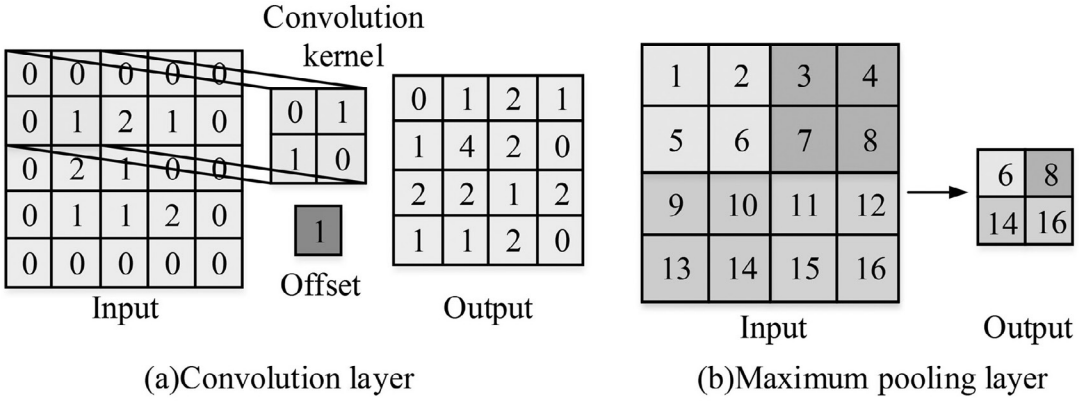


Figure 1. Schematic diagram of the convolution – pooling process.

In equation (3), $p(y^i = k \mid x^i; \omega)$ represents the probability that the i-th sample feature belongs to the k-th class. T represents the length of the sample vector. $\omega = [\omega_1, \omega_2, ..., \omega_k]$ represents the weight value. The activation function improves the representation ability and universal applicability of the original model through its nonlinear transformation form. Common activation functions include the Sigmoid function, Tanh function, and ReLU function. The corresponding visualization diagram is shown in Figure 2.

The Tanh function is essentially an optimization function of the Sigmoid function, which expands the mapping range and feature representation ability of the initial convolution. However, both types of functions exhibit gradient vanishing phenomena, where the gradients approach zero during the backpropagation. Therefore, some input features cannot be trained, and their corresponding expressions are shown in equation (4) [14].

$$\begin{cases} \delta(x) = \dfrac{1}{1+e^{-x}} \\ \mathrm{Tanh}(x) = 2\delta(x) - 1 = \dfrac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \end{cases} \tag{4}$$

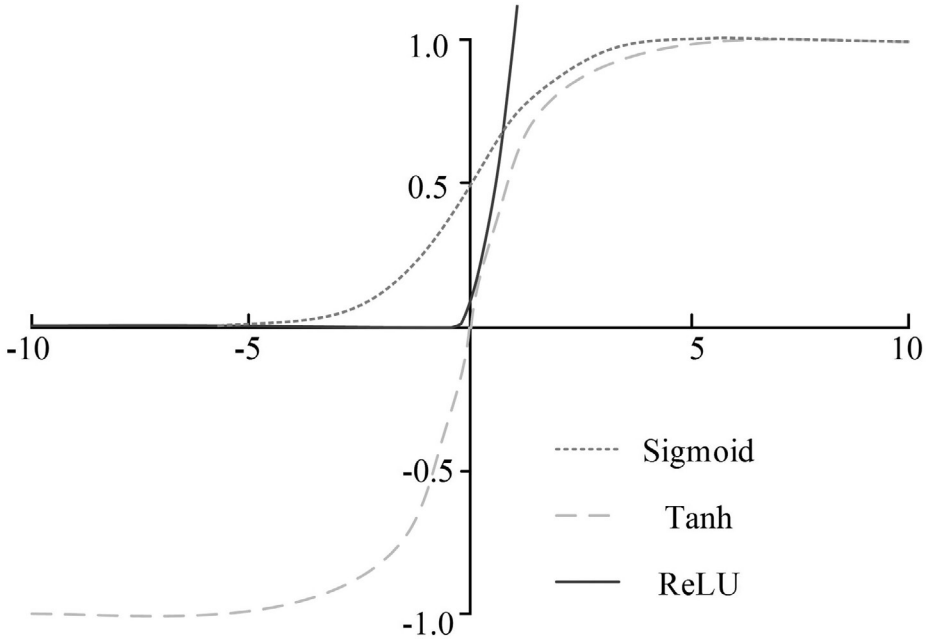


Figure 2. Visual analysis of different activation functions.

In equation (4), $\delta(x)$/Tanh(x) represents the Sigmoid function and the Tanh function, respectively. x represents the input value. Relatively speaking, the ReLU function achieves higher training speed and also solves the defect of vanishing gradient. But its negative half-axis output is all zero, which will cause data sparsity and the problem of parameter values not being updated during the backpropagation, as shown in equation (5).

$$\mathrm{Re}\,LU(x) = \max(0, x) \tag{5}$$

After careful consideration, the ReLU activation function was selected for this study. Additionally, the dropout layer was incorporated to enhance generalization through regularization. This layer, placed after the fully connected layer, temporarily deactivates certain neurons during the training process [15], mitigating overfitting of the model and enhancing feature sparsity.

Due to the increasing complexity of network models, the choice of optimization function becomes an essential part. In this study, the Adam function was selected for network training. It is similar to momentum-based methods, with low gradient fluctuations and rapid convergence [16].

CNNs can divide pixels in an image into different categories or regions to achieve semantic segmentation and instance segmentation. In semantic segmentation, the key role of CNN is to understand and label the content of each part of the image, for example, to distinguish between roads, pedestrians, vehicles, buildings, *etc.* However, instance segmentation is not only necessary to identify the object categories in the image but also to distinguish different object instances. For example, in a picture of multiple cars, instance segmentation not only identifies the vehicle but also the corresponding model. FCN is an important CNN architecture used for IS. Unlike traditional CNN, FCN converts the fully connected layer into a convolutional layer, allowing the network to accept input images of any size and retain spatial information, making it very suitable for IS tasks. However, the current classification and labeling methods require iterative optimization.

3.2. Improved FCNs for Image Semantic Segmentation

Image semantic segmentation is a form of target segmentation that divides pixels based on semantic classification. CNNs are commonly used in image semantic segmentation, are trained on manually annotated image data to predict pixel-level classification [17]. This study used an FCN based on the VGG16Net network to optimize the initial CNN.

The VGG16Net network uses a unified 3x3 convolution kernel. This small-sized kernel increases the network's depth, improves its learning ability, and maintains its non-linear characteristics while keeping a small receptive field. The overall network consists of 16 layers and is able to capture more complex and abstract features in the image. Because of its simple structure and effective training method, the VGG16Net network structure has good generalization ability and is relatively easy to understand and implement. Its basic structure is similar to CNN networks, where the convolutional layer utilizes weight sharing to achieve parameter dimensionality reduction, and the fully connected layer is also replaced to ensure the preservation of two-dimensional spatial data in the model. The feature map obtained by deconvolution is as large as the original feature map [18]. The dimensionality reduction of the above convolution layer and pooling layer can solve the overfitting problem well.

Traditional CNN models use fully connected layers to convert feature maps to feature vectors and achieve final classification through the Softmax layer. However, the weight matrix in the fully connected layer remains unchanged, and its input feature map size is fixed. By replacing it with convolutional layers, the upsampling requirements for semantic segmentation can be met, ensuring the maintenance of spatial features [19]. The size of the output feature map of the classification network will decrease when it passes through the pooling layer, so upsampling is necessary to restore the original feature map size and ensure that the pixels between the feature map and the input map correspond one-to-one. Bilinear interpolation is used to achieve upsampling, which involves inserting new pixels between pixels through an algorithm for im-

age magnification. The downsampling process of feature images is shown in Figure 3.

The difference between FCN and CNN is that the output feature map of the former only changes in the number of features, while the size of the image remains unchanged. However, the accuracy of the feature map obtained using only upsampling technology does not meet the requirements. Therefore, it is still necessary to introduce a network fusion structure to restore the feature maps to the previous pooling convolutional layer, implement reverse upsampling operations, and loop until the size matches the original image. The fusion of network layers has greatly enriched detailed data and improved the accuracy of semantic segmentation [20]. The Softmax layer at the end of the network is the key to classification, and it is necessary to use hypothesis functions to calculate the probability of different categories, as shown in equation (6).

$$h_\theta\left(x^{(i)}\right)=\begin{bmatrix} p\left(y^{(i)}=1\middle|x^{(i)};\theta\right) \\ p\left(y^{(i)}=1\middle|x^{(i)};\theta\right) \\ \cdots \\ p\left(y^{(i)}=k\middle|x^{(i)};\theta\right) \end{bmatrix} = \frac{1}{\sum_{j=1}^{k} e^{\theta_j^T x(i)}}\begin{bmatrix} e^{\theta_1^T x(i)} \\ e^{\theta_2^T x(i)} \\ \cdots \\ e^{\theta_k^T x(i)} \end{bmatrix} \tag{6}$$

In equation (6), $\{(x^{(i)}, y^{(i)}), ..., (x^m, y^m)\}$ represents the training set. $y^i \in \{1, 2, ..., k\}$ represents output. $\theta_1, \theta_2, ..., \theta_k \in {}^{n+1}$ represents parameters. The sum of the probability values of all output vectors is 1. The cost function, also known as the loss function, is the description of the loss value of the hypothesis function for various parameter values. It is the learning criterion and optimization problem for the model establishment, determining the prediction accuracy of the model, as shown in equation (7).

$$L(\theta)=-\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{k}1\left\{y^{(i)}=j\right\}\log\frac{e^{\theta_j^T x(i)}}{\sum_{j=1}^{k}e^{\theta_j^T x(i)}}\right] + \frac{\lambda}{2}\sum_{i=1}^{k}\sum_{j=0}^{n}\theta_{ij}^2 \tag{7}$$

In equation (7), 1 represents the performance function, and if the equation in parentheses is true, 1 is taken, while if it is opposite, 0 is taken. $\frac{\lambda}{2}\sum_{i=1}^{k}\sum_{j=0}^{n}\theta_{ij}^2(\lambda>0)$ represents the weight attenuation term. The function of the cost function is to reduce the gradient and ultimately converge to the global optimal value. The overall FCN semantic segmentation model is shown in Figure 4.

The foundation of the FCN model is VGG-16Net, which consists of 13 convolutional layers, 5 pooling layers, and two fully connected layers. Among them, the fully connected layer is replaced with a deep convolutional layer for deconvolution to achieve IS. However, its retention effect on detail features is poor, so a

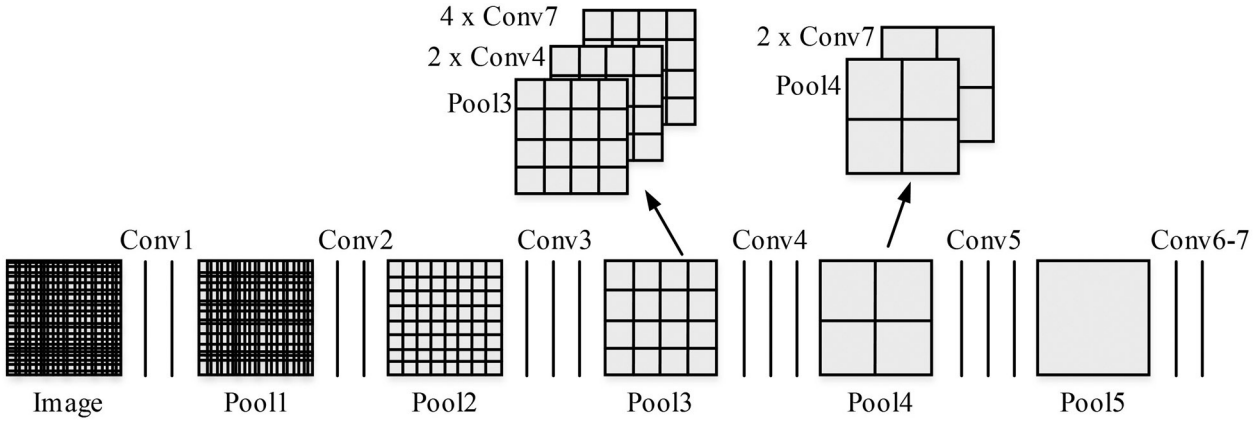


Figure 3. Downsampling process of feature image in FCN model.

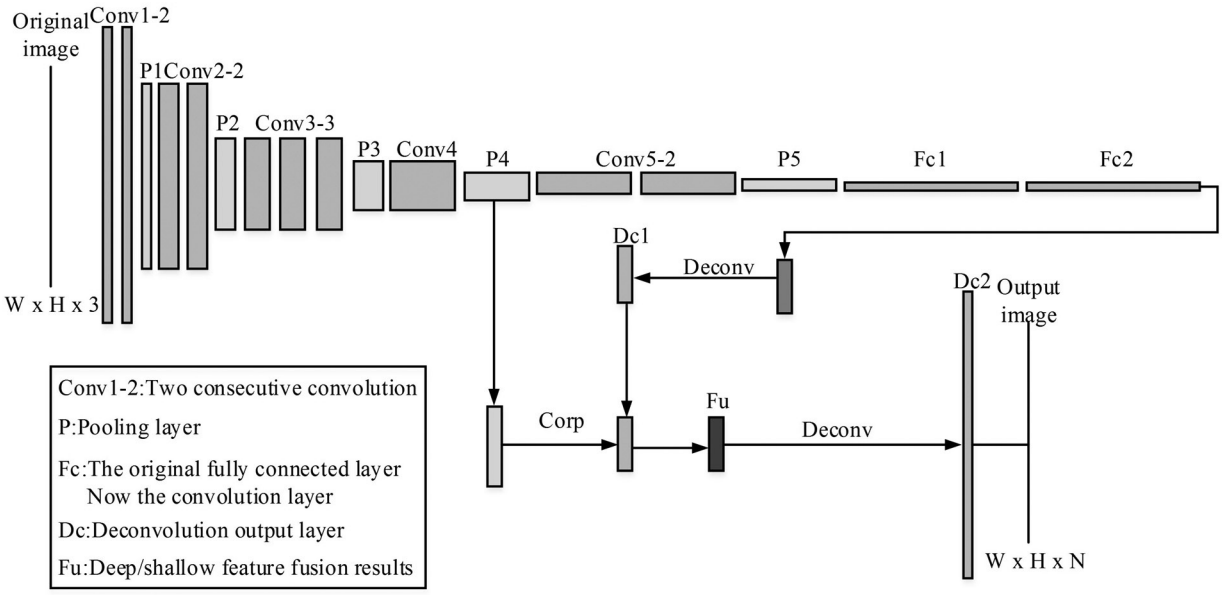


Figure 4. Semantic segmentation model based on full CNN.

shallow network is introduced to combine and achieve high precision IS output. Semantic segmentation is the categorization of each pixel, while instance segmentation is the subdivision of the category [21]. Therefore, the instance segmentation faces pixel-level classification and specific object classification of the same category, as shown in Figure 5.

There are two traditional methods for instance segmentation image detection. The first involves embedding a Region Proposal Network (RPN) structure within the network. The second employs border extraction algorithms such as Selective Search outside the network. Both of these methods are based on candidate region extraction, and an increase in the number of candidate boxes often slows down the training speed of the model, resulting in poor IS performance. The IS effect of the fully supervised network model is better than that of the unsupervised model [22]. Therefore, this study continues to use the fully supervised semantic segmentation model mentioned above, aiming to use the output key boundary segmentation mask map to achieve instance segmentation of the target. Its form is different from the fully supervised instance segmentation model, which first requires pixel-level semantic segmentation through masked image labels. It is necessary to introduce a class activation function to generate a thermal map to achieve instance perception of the target. Therefore, this network model does not require further training on the instance dataset [23].

Thermal maps, also known as Class Activation Mapping (CAM), are suitable for locating discriminative regions. The fully connected layer in CNN networks hinders the implementation of object detection, and Global Average Pooling (GAP) can not only replace the fully connected layer but also maintain the network's target localization function, greatly reducing its number of parameters. CAM technology can visualize the extracted features to facilitate the target localization process, and this module is usually located behind the deepest convolutional layer. The feature mapping map output from this directly enters the GAP module to calculate the weights of different categories. It corresponds to the feature mapping map one by one to calculate the product sum, as shown in Figure 6.

In Figure 6, the calculation of GAP is shown in equation (8).

$$F_k = \sum_{x,y} f_k(x,y) \tag{8}$$

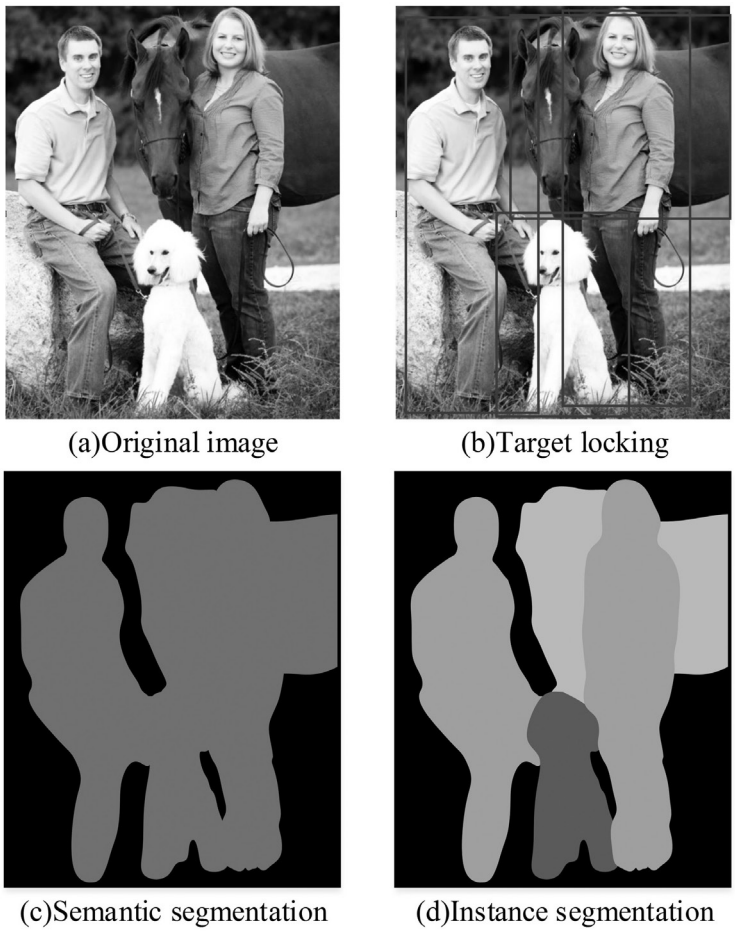


Figure 5. Differences between image semantic segmentation and instance segmentation.

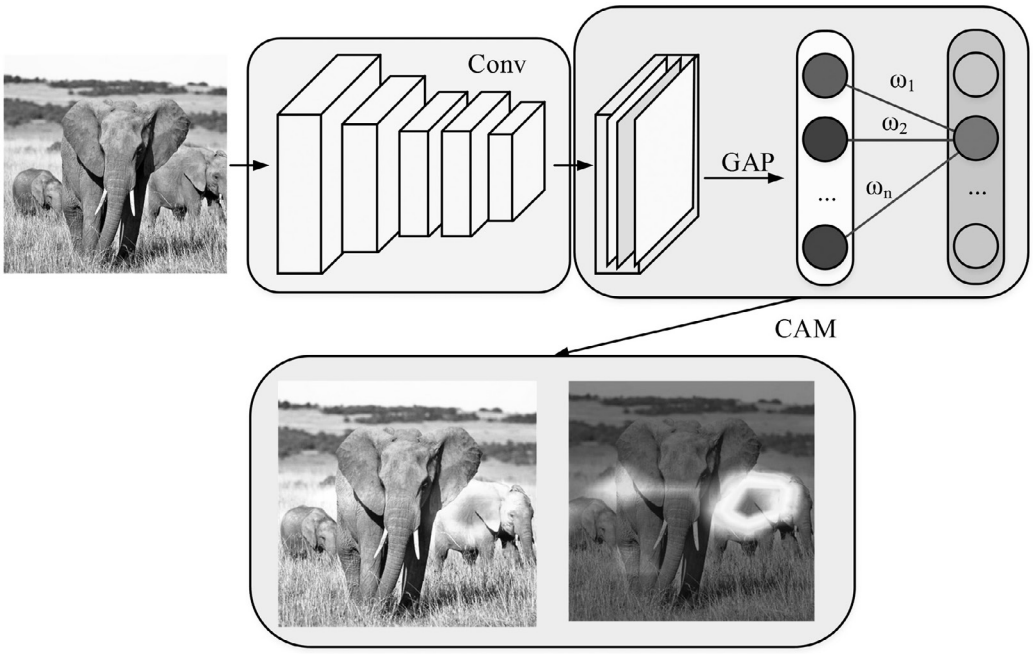


Figure 6. Class activation diagram flow.

In equation (8), F_k represents the calculated GAP value for each unit. (x, y) represents the coordinates of a unit in the final convolutional layer. $f_k(x, y)$ represents the activation value of the k unit in the final convolutional layer at coordinate (x, y). The input values for each category in the softmax layer are shown in equation (9).

$$S_c = \sum_k \omega_k^c F_k \tag{9}$$

In equation (9), ω_k^c represents the weight of k units in the k category. The output values of different category attributes in the softmax layer are shown in equation (10).

$$P_c = \frac{\exp(S_c)}{\sum_c \exp(S_c)} \tag{10}$$

Due to the negligible impact of bias on classification visualization, it is set to 0. The spatial element values of the class activation function mapping M_c of the category are shown in equation (11).

$$M_c(x, y) = \sum_k \omega_k^c f_k(x, y) \tag{11}$$

From equations (8) to (11), the final input values of various types in the softmax layer can be obtained, as shown in equation (12).

$$S_c = \sum_{x,y} \sum_k \omega_k^c f_k(x, y) = \sum_{x,y} M_c(x, y) \tag{12}$$

The evaluation indicators for image instance segmentation include pixel accuracy, class average accuracy, and average region overlap. The pixel accuracy is shown in equation (13) [24].

$$P_{\text{Pixel}} = \frac{\sum_i n_{ii}}{\sum_i t_i} \tag{13}$$

In equation (13), n_{ii} represents the number of pixels for the predicted class i and the actual class i. t_i represents the total number of pixels in the class. The average accuracy of the class is shown in equation (14).

$$P_{\text{Class average}} = \frac{1}{n_{cl}} \sum_i \frac{n_{ii}}{t_i} \tag{14}$$

In equation (14), n_{cl} represents the total number of categories in the dataset. The average area overlap, also known as the average IU (Intersection over Union, Mean lU), is shown in equation (15).

$$\text{Average IU} = \frac{1}{n_{cl}} \sum_i \frac{n_{ii}}{(t_i + \sum_j n_{ji} - n_{ii})} \tag{15}$$

The average IU represents the accuracy and completeness of the segmented region and is the final criterion for evaluating image instance segmentation.

4. Result and Discussion

To assess the performance of the optimized FCN for instance segmentation, simulation experiments were conducted on both semantic segmentation and instance segmentation of the model. Before conducting the experiments, the research improved the existing dataset by utilizing traffic monitoring in a certain city to increase the information content of the model data to ensure that the model training yielded optimal results.

4.1. Performance Verification of FCN IS Model

The experiment was conducted on the Ubuntu 14.04 operating system platform, utilizing the DL framework TensorFlow built on an NVIDIA GTX 970 GPU. A fully supervised DL model requires training through a large number of datasets. The PASCAL VOC2012 dataset is a classic dataset suitable for semantic segmentation validation and also a standard dataset for evaluating CNN performance. There are a total of 21 semantic categories in the dataset, with only 1 background category and the rest being foreground categories. Among them, the vehicle dataset has 700 pieces, which is insufficient to support model training.

To address this limitation,, this study utilizes the PASCAL VOC2012 dataset format for semantic segmentation datasets and supplements it with data collected traffic monitoring videos from a specific city. This dataset includes 1000

vehicle images captured under various lighting conditions and angles, ensuring diverse representation of vehicle operating scenarios, as illustrated in Figure 7.

This study compared the image semantic segmentation performance between the binary classification FCN model and the 21-class FCN model using different datasets. The experimental results are shown in Figure 8.

From Figure 8 (a), it is evident that the model trained using only the VOC dataset exhibits significantly lower accuracy compared to the others, while the performance of the 2-class network is better than that of the 21-class model. The pixel accuracy of the 2-class network in the self-made vehicle dataset and VOC dataset is 0.93 ± 0.03, with a median of 93%. Compared with other semantic segmentation models, its

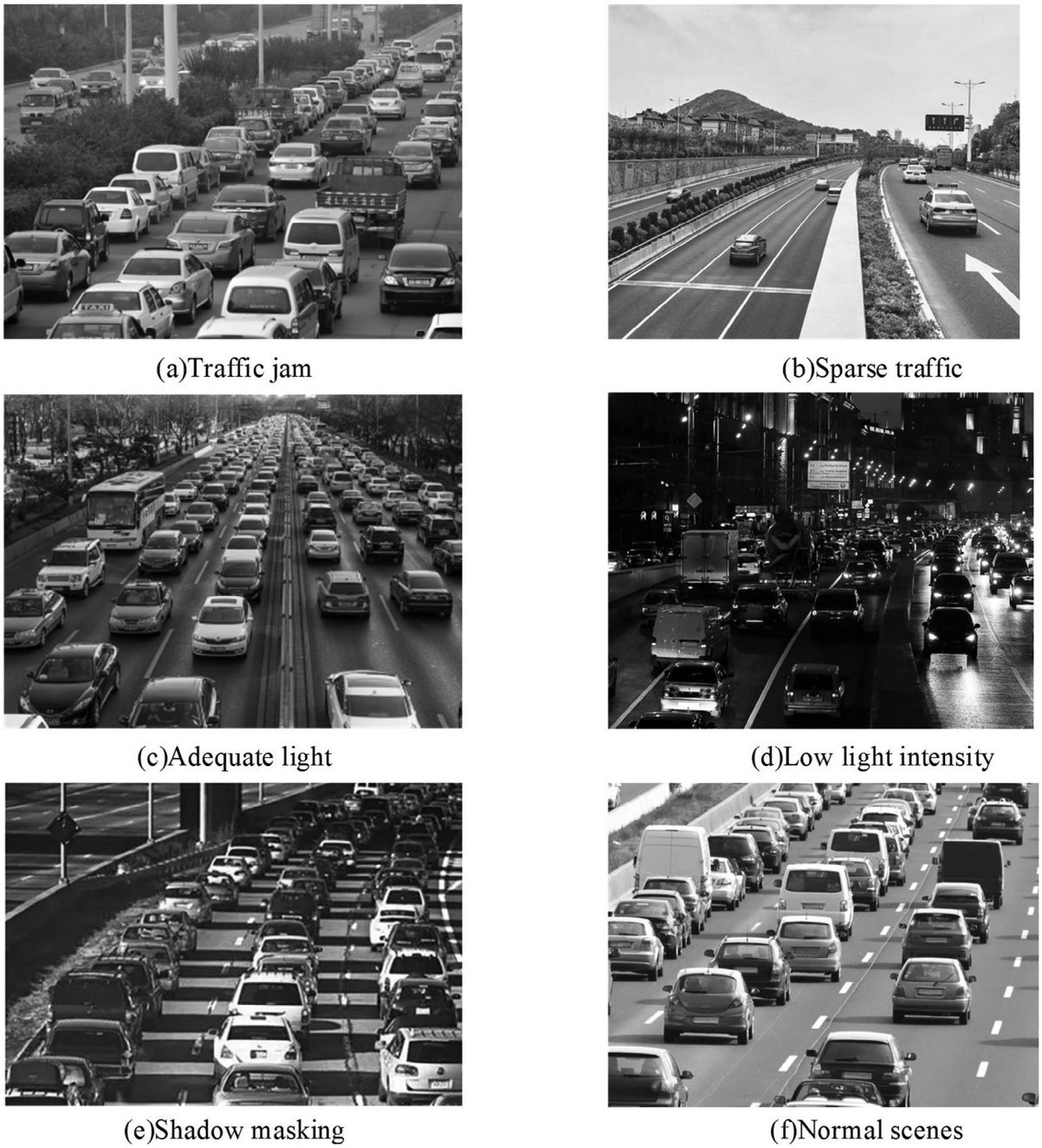
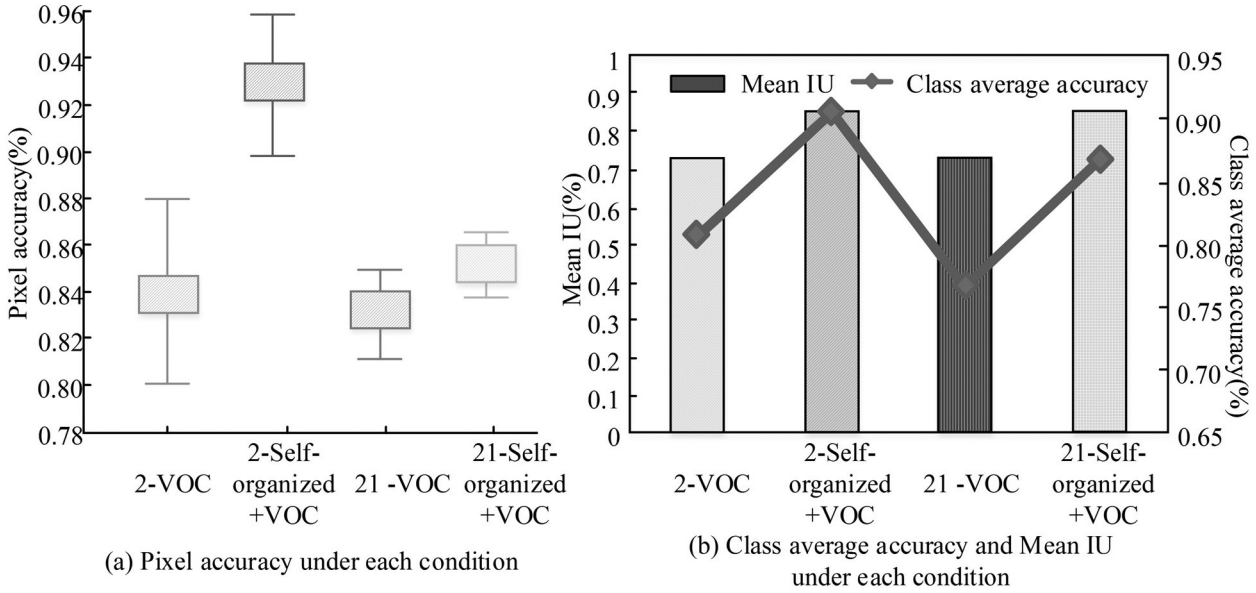


Figure 7. Vehicle traffic images under different conditions.



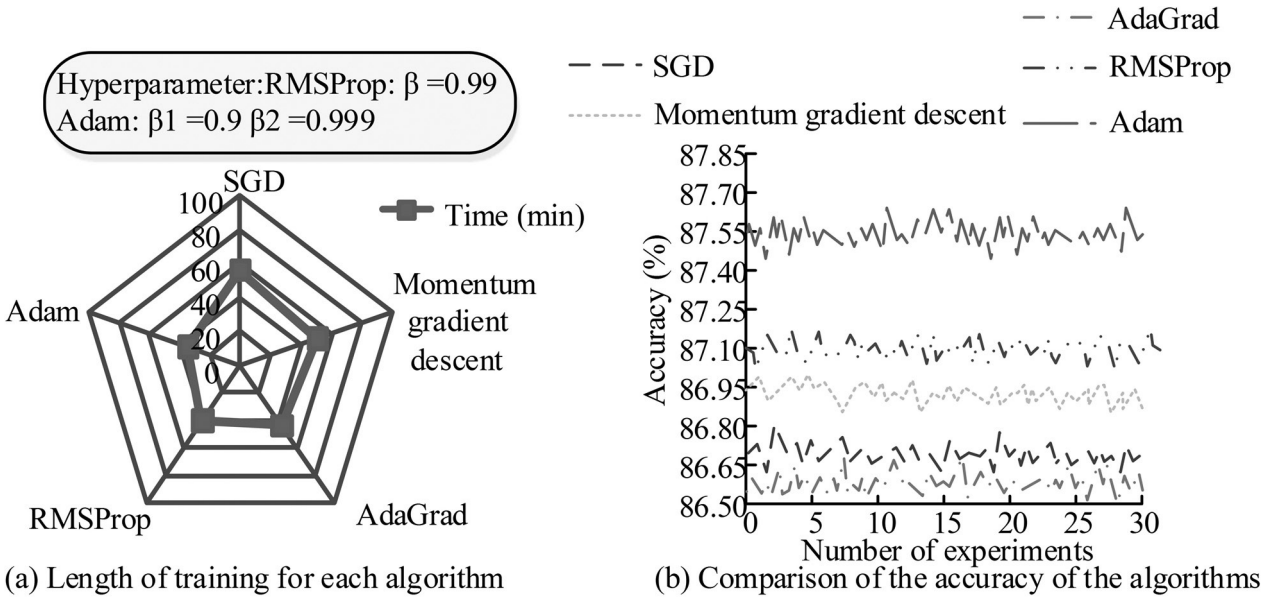
(a) Pixel accuracy under each condition

(b) Class average accuracy and Mean IU under each condition

Figure 8. Semantic segmentation performance of the model under different classification numbers and data sets.

pixel accuracy has improved by an average of 8.33%. In Figure 8 (b), the average class accuracy of the 2-class network in the self-made vehicle dataset and VOC dataset is 90.64%, while the other models are 80.77%, 86.80%, and 76.68%, respectively. The average class accuracy of the 2-class classification network in the mixed dataset has improved by 9.22% compared to the other models on average. Furthermore, its average IU increased by 14.38% on average.

The third layer of research conducted further comparative analysis of the optimization functions selected by the model. The study introduced Stochastic Gradient Descent (SGD), Momentum Gradient Descent (MGD), Adaptive Dynamic Learning Algorithm (AdaGrad), and Root Mean Square Propagation (RMSProp) to compare the accuracy and training time of the above function models. The results are shown in Figure 9.



(a) Length of training for each algorithm

(b) Comparison of the accuracy of the algorithms

Figure 9. Comparison of semantic segmentation performance of models under different optimizers.

Figure 9 shows that the semantic segmentation accuracy of the models under each optimization function does not differ significantly. The accuracy rates from the SGD model clockwise to the Adam model are 86.72%, 86.94%, 86.62%, 87.12%, and 87.58%, respectively. Therefore, the semantic segmentation accuracy of the Adam optimization function model is on average 0.73% higher than other models.

Although the differences in accuracy among different models are small, they have widened the gap in training time performance. The training time of both the SGD model and the momentum gradient descent model is over 50 minutes because the gradient descent function is trained by searching for the local minimum value of the loss function. This search is carried out in the opposite direction of the gradient, and the change length is fixed, resulting in a naturally longer training time. The training duration for AdaGrad and RMSProp models has decreased to 43.8 minutes and 40.5 minutes, respectively. Both belong to the learning rate adaptive gradient descent method, which sets a corresponding learning rate for each parameter. Therefore, during training, the parameters need to be adaptively operated on the learning rate. The training duration of the Adam model is only 34.4 minutes, which is 12.7 minutes less than the average training duration of other functions. Additionally, the Adam model has shown a 36.81% improvement in timeliness. This is because other optimization functions cannot cope with more complex network structures, and increasing the number of parameters will limit its convergence speed. However, the gradient fluctuation of the Adam function is small, making it naturally more adaptable to complex network environments.

4.2. Performance Verification and Comparison of FCN Instance Segmentation

The model was implemented in Python using Tensor Flow framework, and executed on a Linux system. Both the instance segmentation and semantic segmentation models share consistent deconvolution layer parameters. Figure 10 depicts the class activation diagram generated by the model under conditions of sparse traffic volume, congested traffic volume, and insufficient lighting.

The model achieves visualization of class activation maps by backpropagating feature maps to the original image. In Figure 10, the red part represents the area with higher median values in the feature map. Therefore, when the target is relatively large and complete, its corresponding thermal map color is darker. Conversely, when

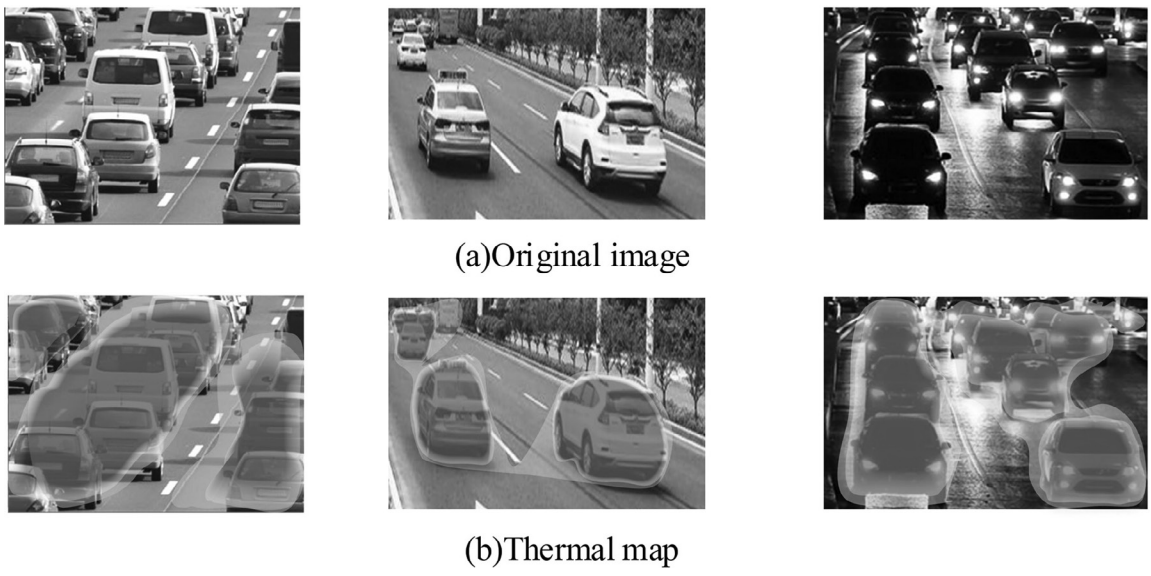


Figure 10. Comparison of original images and thermal maps under different conditions.

the target is obstructed or the light is weak, the thermal map will become lighter. This study further validated and analyzed the performance of the improved FCN instance segmentation model, as shown in Figure 11.

Figure 11 shows that the pixel accuracy of the instance segmentation model based on the improved FCN reached 98.4%, which is 5.4% higher compared to the model before optimization. The average class accuracy reached 97.32%, which is 6.68% higher compared to the model before optimization. The average regional overlap is 92.14%. The above data indicates that the optimized instance segmentation model has improved the accuracy and completeness of IS task.

This study selected Mean Average Precision (MAP) as the evaluation index, and introduced the instance segmentation model proposed in [21][23], and W. Zhou *et al.* for comparative analysis [24]. The experimental results are shown in Figure 12.

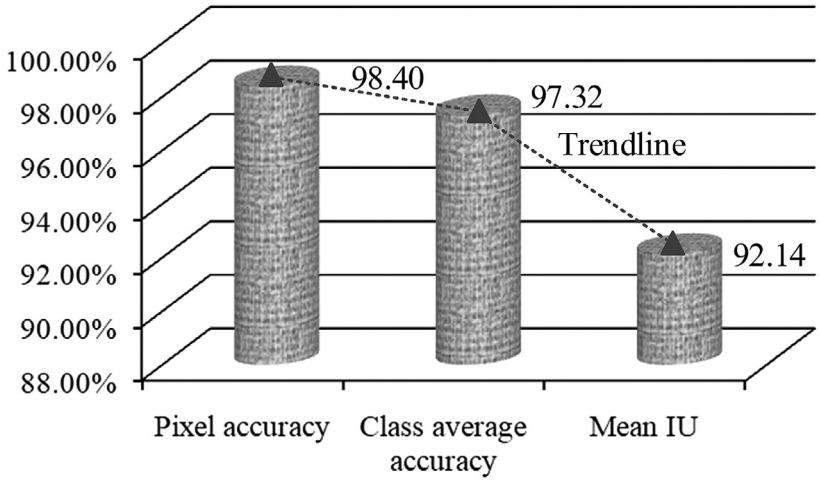


Figure 11. Performance analysis of the improved FCN instance segmentation model.

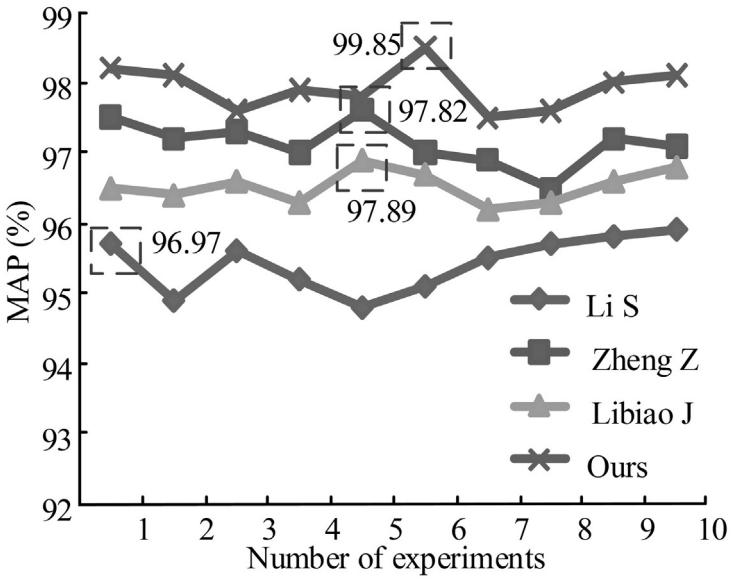


Figure 12. MAP performance analysis of segmentation model for different instances.

Figure 12 shows the changes in MAP indicators of different instance segmentation models in 10 simulation experiments, and each IS model is relatively stable. The instance segmentation model based on an optimized FCN proposed in this study has an average accuracy of 97.9% for all classes, which is the best among the comparison models. The average MAP of the instance segmentation model proposed by Li S *et al.* [21] is 95.7%. The average MAP of the instance segmentation model proposed by Zheng Z *et al.* [23] is 97.2%. The average MAP of the instance segmentation model proposed by W. Zhou *et al.* is 96.6% [24]. Therefore, the IS model proposed in this study is 2.2%, 0.7%, and 1.3% higher than the other three models, respectively.

To investigate the model's performance in comparison to recent research models on other datasets, the Coupled Deformation Model (CDM) proposed by A. Kumar *et al.* [25], and the U-Net optimization model proposed by J. Fan *et al.* [25–26] were introduced in this study. The experiment was conducted using public data sets Cityscapes and iSAID, and the experimental results are shown in Table 1.

In Table 1, SSIM represents the Mean Structural Similarity Index Measure, including the analysis of IS at three levels: brightness, contrast and structure. PSNR represents the Peak Signal-to-Noise Ratio of IS results. The models in the Cityscapes dataset exhibit minimal differences when the number of dimensions is 5, with an SSIM of approximately 0.745 and a PSNR of around 21.70.

With the increase of dimension, the performance of the model is significantly improved. For dimension 10, the SSIM value of the proposed model is 7.19% higher than that of the CDM model and 5.23% higher than that of the U-Net optimization model. Similarly, the PSNR value of the proposed model is 9.80% higher than that of the CDM model and 8.21% higher than that of the U-Net optimization model.

On the iSAID dataset with a dimension of 10, the SSIM and PSNR values of the proposed model are, on average, 5.83% and 8.25% higher than the other models, respectively. When the dimension reaches 15, the gap between the two data models is relatively larger. The SSIM val-

Table 1. Compares the performance of each model in different data sets.

Data sets	Dimensionality	Model		
		CDM	U-Net	Ours
Cityscapes	5	SSIM: 0.747	SSIM: 0.742	SSIM: 0.748
		PSNR: 21.83	PSNR: 21.66	PSNR: 21.95
	10	SSIM: 0.852	SSIM: 0.867	SSIM: 0.918
		PSNR: 25.49	SNR: 25.94	PSNR: 28.26
	15	SSIM: 0.916	SSIM: 0.925	SSIM: 0.948
		PSNR: 30.46	PSNR: 30.04	PSNR: 33.27
iSAID	5	SSIM: 0.773	SSIM: 0.785	SSIM: 0.816
		PSNR: 22.38	PSNR: 23.05	PSNR: 23.74
	10	SSIM: 0.901	SSIM: 0.893	SSIM: 0.948
		PSNR: 27.49	PSNR: 28.06	PSNR: 30.05
	15	SSIM: 0.952	SSIM: 0.923	SSIM: 0.970
		PSNR: 31.01	PSNR: 30.76	PSNR: 34.58

ue of the proposed model is on average 12.05% higher than that of the other models, and the PSNR value is on average 13.57% higher than that of the other models. These differences were statistically significant ($P<0.05$). In conclusion, the proposed model has most effective instance segmentation performance.

5. Conclusion

This study employed a Fully Convolutional Network (FCN) to develop a semantic segmentation model for images. First, the network's basic parameters were determined, and the original fully connected layer was replaced with a convolutional layer. An instance segmentation model was constructed utilizing the output mask graph, while class activation technology was introduced to improve instance recognition of the target.

Subsequently, simulation validation was conducted on semantic segmentation and instance segmentation models respectively. In the former, different classification forms and dataset selections were analyzed. The data showed an average increase of 8.33% in pixel accuracy, 9.22% in class average accuracy, and 14.38% in average Intersection over Union (IU). In comparison with SGD, AdaGrad, and RMSProp optimization functions, the model achieved an accuracy of 87.58%, surpassing other models by an average of 0.73%.

The runtime of the experiment was 34.4 minutes, and the timeliness was improved by an average of 36.81%. The instance segmentation experiment yielded good results under various conditions, including different light intensities and sparsity. The pixel accuracy was 98.4%, which was 5.4% higher than the semantic segmentation model. The average class accuracy reached 97.32%, with an improvement of 6.68%. The average regional overlap was 92.14%.

In comparison with the segmentation models proposed in other relevant studies, the average accuracy of the entire class was 97.9%, which is higher than the other three models by 2.2%, 0.7%, and 1.3%, respectively. This research primarily focused on the network structure did not target specific segmentation objectives. Therefore, designing a specialized loss function is imperative to address the issue of unbalanced label quantities among classes.

References

[1] K. J. Singh *et al.*, "Computer-Vision Based Object Detection and Recognition for Service Robot in Indoor Environment", *Computers, Materials and Continuum*, vol. 17, no. 7, pp. 197–213, 2022. http://dx.doi.org/10.32604/cmc.2022.022989

[2] T. Wu *et al.*, "Image Segmentation via Fischer-Burmeister Total Variation and Thresholding", *Advances IN Applied Mathematics AND Mechanics*, vol. 14, no. 4, pp. 960–988, 2022. http://dx.doi.org/10.4208/aamm.OA-2021-0126

[3] Y. Yang *et al.*, "A Convergent Iterative Support Shrinking Algorithm for Non-Lipschitz Multiphase Image Labeling Model", *Journal of Scientific Computing*, vol. 96, no. 2, pp. 47–47, 2023. http://dx.doi.org/10.1007/s10915-023-02268-5

[4] P. Mukasa *et al.*, "Nondestructive Discrimination of Seedless from Seeded Watermelon Seeds by Using Multivariate and Deep Learning Image Analysis", *Computers and Electronics in Agriculture*, vol. 194, pp. 106799–106809, 2022. http://dx.doi.org/10.1016/j.compag.2022.106799

[5] C. Bowd *et al.*, "Deep Learning Image Analysis of Optical Coherence Tomography Angiography Measured Vessel Density Improves Classification of Healthy and Glaucoma Eyes", *American Journal of Ophthalmology*, vol. 236, no. 1, pp. 298–308, 2022. http://dx.doi.org/10.1016/j.ajo.2021.11.008

[6] P. Munoz-Benavent *et al.*, "Impact Evaluation of Deep Learning on Image Segmentation for Automatic Bluefin Tuna Sizing", *Aquacultural Engineering*, vol. 99, pp. 102299–102309, 2022. http://dx.doi.org/10.1016/j.aquaeng.2022.102299

[7] G. Yuan *et al.*, "Fast Bilateral Complementary Network for Deep Learning Compressed Sensing Image Reconstruction", *IET Image Processing*, vol. 16, no. 13, pp. 3485–3498, 2022. http://dx.doi.org/10.1049/ipr2.12545

[8] J. Niedballa *et al.*, "Images: An R Package for Deep Learning-based Image Segmentation", *Methods in Ecology and Evolution*, vol. 13, no. 11, pp. 2363–2371, 2021. http://dx.doi.org/10.1111/2041-210X.13984

[9] B. Ji *et al.*, "An Effective Color Image Segmentation Approach Using Superpixel-neutrosophic C-means Clustering and Gradient-structural Similarity", *Optik*, vol. 260, no. 1, pp. 169039–169059, 2022. http://dx.doi.org/10.1016/j.ijleo.2022.169039

[10] X. Yan and G. Weng, "Hybrid Active Contour Model Driven by Optimized Local Pre-fitting Image Energy for Fast Image Segmentation", *Applied Mathematical Modelling*, vol. 101, no. 1, pp. 586–599, 2022. http://dx.doi.org/10.1016/j.apm.2021.09.002

[11] W. He *et al.*, "Differentiable Automatic Data Augmentation by Proximal Update for Medical Image Segmentation", *Acta Automatica Sinica: English*, vol. 9, no. 7, pp. 1315–1318, 2022. http://dx.doi.org/10.1016/j.apm.2021.09.002

[12] S. L. Lim *et al.*, "Comparing Machine Learning and Deep Learning Based Approaches to Detect Customer Sentiment from Product Reviews", *Journal of System and Management Sciences*, vol. 13, no. 2, pp. 101–110, 2023. http://dx.doi.org/10.33168/JSMS.2023.0207

[13] Z. Yang, "Image Segmentation of Cucumber Seedlings Based on Genetic Algorithm", *Sustainability*, vol. 15, no. 4, pp. 3089–3089, 2023. http://dx.doi.org/10.3390/su15043089

[14] P. Bajcsy *et al.*, "Approaches to Training Multiclass Semantic Image Segmentation of Damage in Concrete", *Journal of Microscopy*, vol. 279, no. 2, pp. 98–113, 2020. http://dx.doi.org/10.1111/jmi.12906

[15] R. Ke *et al.*, "Multi-task Deep Learning for Image Segmentation Using Recursive Approximation Tasks", *IEEE Transactions on Image Processing*, vol. 30, no. 99, pp. 2555–2567, 2020. http://dx.doi.org/10.1109/TIP.2021.3062726

[16] Y. Fang *et al.*, "ST-SIGMA: Spatio-temporal Semantics and Interaction Graph Aggregation for Multi-agent Perception and Trajectory Forecasting", *CAAI Transactions on Intelligence Technology*, vol. 7, no. 4, pp. 744–757, 2022. http://dx.doi.org/10.1049/cit2.12145

[17] Y. Wu *et al.*, "Deep Instance Segmentation and 6D Object Pose Estimation in Cluttered Scenes for Robotic Autonomous Grasping", *Industrial Robot*, vol. 47, no. 4, pp. 0259–0273, 2020. http://dx.doi.org/10.1108/IR-12-2019-0259

[18] Y. Yang and X. Song, "Research on Face Intelligent Perception Technology Integrating Deep Learning Under Different Illumination Intensities", *Journal of Computational and Cognitive Engineering*, vol. 1, no. 1, pp. 32–36, 2022. http://dx.doi.org/10.47852/bonviewJCCE19919

[19] L. Liu *et al.*, "A Survey on U-shaped Networks in Medical Image Segmentations", *Neuro Computing*, vol. 409, no. 10, pp. 244–258, 2020. http://dx.doi.org/10.1016/j.neucom.2020.05.070

[20] C. J. Xiong *et al.*, "Deepfakes Detection Using Computer Vision and Deep Learning Approaches", *Journal of System and Management Sciences*, vol. 12, no. 5, pp. 21–35, 2022. http://dx.doi.org/10.33168/JSMS.2022.0502

[21] S. Li *et al.*, "SPM-IS: An Auto-algorithm to Acquire a Mature Soybean Phenotype Based on Instance Segmentation", *The Crop Journal*, vol. 10, no. 5, pp. 1412–1423, 2022. http://dx.doi.org/10.1016/j.cj.2021.05.014

[22] M. Greguric *et al.*, "Towards the Spatial Analysis of Motorway Safety in the Connected Environment by Using Explainable Deep Learning", *Knowledge-based Systems*, vol. 269, no. 7, pp. 1.1–1.17, 2023. http://dx.doi.org/10.1016/j.knosys.2023.110523

[23] Z. Zheng *et al.*, "Enhancing Geometric Factors in Model Learning and Inference for Object Detection and Instance Segmentation", *IEEE*, vol. 52, no. 8, pp. 8574–8586, 2022. http://dx.doi.org/10.1109/TCYB.2021.3095305

[24] W. Zhou *et al.*, "Adaptive Sinh Transformation Gaussian Quadrature for 2D Potential Problems Using Deep Learning", *Engineering Analysis with Boundary Elements*, vol. 155, no. 12, pp. 197–211, 2023. http://dx.doi.org/10.1016/j.enganabound.2023.06.002

[25] A. Kumar *et al.*, "CDM: A Coupled Deformable Model for Image Segmentation with Speckle Noise and Severe Intensity Inhomogeneity", *Chaos, Solitons & Fractals*, vol. 173, no. 3, pp. 104385–104396, 2023. http://dx.doi.org/10.1016/j.chaos.2023.113551

[26] J. Fan *et al.*, "Macerals Particle Characteristics Analysis of Tar-rich Coal in Northern Shaanxi Based on Image Segmentation Models Via the U-Net Variants and Image Feature Extraction", *Fuel*, vol. 341, no. 1, pp. 127757–127757, 2023. http://dx.doi.org/10.1016/j.fuel.2023.127757

Received: October 2023
Revised: January 2024
Accepted: January 2024

Contact addresses:
Chuangchuang Chen*
School of Network Engineering
Zhoukou Normal University
Zhoukou
Henan
China
e-mail: zknuchenchuang@163.com
*Corresponding author

Guang Gao
School of Network Engineering
Zhoukou Normal University
Zhoukou
Henan
China
e-mail: zzgaoguang@sina.com

Linlin Liu
School of Network Engineering
Zhoukou Normal University
Zhoukou
Henan
China
e-mail: linlin8665@outlook.com

Yangyang Qiao
School of Information Engineering
Zhengzhou Technology and Business University
Zhengzhou
China
e-mail: qdychy@163.com

Chuangchuang Chen obtained the master's degree from the University of South China, in 2016. He is currently working as a researcher at the Zhoukou Normal University, China. His expertise lies in the fields of computer vision and artificial intelligence.

Guang Gao obtained a master's degree from the Southwest Jiaotong University, China, in 2011. Currently, he serves as the head of the Software Application Teaching and Research Department at Zhoukou Normal College. His research focuses on big data processing and intelligent image processing.

Linlin Liu obtained her master's degree from the Shanghai University, China, in 2013. Currently, she serves as the head of the Information Perception Research Office at Zhoukou Normal University. Her research focuses on deep learning.

Yangyang Qiao obtained his master's degree from the North China University of Technology, China, in 2015. Currently, he serves as the head of the Department of Computer Applications at Zhengzhou University of Commerce. His research focuses on data mining, big data, and artificial intelligence.