Cao YJ, Wu S, Liu C *et al.* Seg-CapNet: A capsule-based neural network for the segmentation of left ventricle from cardiac magnetic resonance imaging. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 36(2): 323–333 Mar. 2021. DOI 10.1007/s11390-021-0782-5

Seg-CapNet: A Capsule-Based Neural Network for the Segmentation of Left Ventricle from Cardiac Magnetic Resonance Imaging

Yang-Jie Cao[1], *Member, CCF*, Shuang Wu[1], Chang Liu[1], Nan Lin[1], Yuan Wang[2], Cong Yang[1,*], *Member, CCF* and Jie Li[1,3], *Senior Member, IEEE*

[1] *School of Software, Zhengzhou University, Zhengzhou 450000, China*
[2] *Center of Modern Analysis and Gene Sequencing, Zhengzhou University, Zhengzhou 450000, China*
[3] *Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200000, China*

E-mail: caoyj@zzu.edu.cn; ws_vivi@163.com; liuchangliu@126.com; {linnan, wyyc, wangyuanyc}@zzu.edu.cn
lijiecs@sjtu.edu.cn

Received July 5, 2020; accepted March 9, 2021.

Abstract Deep neural networks (DNNs) have been extensively studied in medical image segmentation. However, existing DNNs often need to train shape models for each object to be segmented, which may yield results that violate cardiac anatomical structure when segmenting cardiac magnetic resonance imaging (MRI). In this paper, we propose a capsule-based neural network, named Seg-CapNet, to model multiple regions simultaneously within a single training process. The Seg-CapNet model consists of the encoder and the decoder. The encoder transforms the input image into feature vectors that represent objects to be segmented by convolutional layers, capsule layers, and fully-connected layers. And the decoder transforms the feature vectors into segmentation masks by up-sampling. Feature maps of each down-sampling layer in the encoder are connected to the corresponding up-sampling layers, which are conducive to the backpropagation of the model. The output vectors of Seg-CapNet contain low-level image features such as grayscale and texture, as well as semantic features including the position and size of the objects, which is beneficial for improving the segmentation accuracy. The proposed model is validated on the open dataset of the Automated Cardiac Diagnosis Challenge 2017 (ACDC 2017) and the Sunnybrook Cardiac Magnetic Resonance Imaging (MRI) segmentation challenge. Experimental results show that the mean Dice coefficient of Seg-CapNet is increased by 4.7% and the average Hausdorff distance is reduced by 22%. The proposed model also reduces the model parameters and improves the training speed while obtaining the accurate segmentation of multiple regions.

Keywords capsule neural network, image segmentation, left ventricle segmentation, cardiac magnetic resonance imaging

1 Introduction

The task of image segmentation is to categorize each pixel into nonoverlapping regions according to the grayscale, color, texture, shape, and other image features. Image segmentation is a fundamental task of computer vision and image processing, laying the foundation for high-level computer vision tasks, such as object tracking and computer-aided diagnosis. In medical imaging, accurate segmentation of tissues enables the quantitative measurements of pathological indices such as histomorphology parameters of lesions, which provides a reliable basis for clinical diagnosis, treatment, and pathology research.

Segmentation of medical images is still an open issue due to the low contrast between tissues and the background, considerable noises, and blurred object boundaries. In the last decades, shape-based image segmentation methods including active shape model (ASM) [1, 2] and active appearance model (AAM) [3, 4]

Regular Paper

Special Section on AI and Big Data Analytics in Biology and Medicine

This work was supported by the Collaborative Innovation Major Project of Zhengzhou under Grant No. 20XTZX06013, and the National Natural Science Foundation of China under Grant No. 61932014.

[*] Corresponding Author

have attracted much attention. ASM has shown its potential in feature extraction and object detection. However, it only uses texture information of the object to select features, resulting in their sensitivity to the initial state, image noise, etc. Based on ASM, AAM makes full use of global texture information to establish a global grayscale model which reflects texture changes. Though AAM takes advantage of building a global representation of the object shape, its performance of capturing details of complex objects is still needed to be improved. Segmentation methods via multi-atlas [5] transform the image segmentation into image registration [6] by incorporating prior information. The segmentation is obtained by performing image registration of several manually delineated images on the target image to search object shapes. Because Atlas-based algorithms only use a limited number of labeled images, considerable deviations from actual shapes may occur when dealing with complex scenarios due to inadequate representation ability. At present, the deep neural network (DNN) has received extensive attention in the segmentation of medical images [7] due to its superior autonomous feature extraction and feature representation [8]. Moreover, compared with other medical image segmentation methods, DNNs deal with noise and unevenness in image segmentation in an intrinsic manner. However, existing DNNs train shape models for the multiple regions separately, which may lead to erroneous segmentation. For example, the predicted endocardium of the cardiac left ventricle may intersect with the predicted epicardium of the left ventricle.

To address this, we propose a capsule-based neural network, named Seg-CapNet, to model the endocardium and epicardium of the left ventricle within a single training process. We use the capsule network as the encoder to produce two vectors representing the endocardium and the epicardium so as to simultaneously train models for these two objects. In this way, it enables us to impose constraints to maintain the spatial relationship between them. Therefore, Sep-CapNet can model multiple regions in a parallel fashion. To maintain the spatial relationship between the endocardium and the epicardium, we propose a loss function named coverage ratio in addition to Dice-based loss to train the model parameters through backpropagation. The main contributions of this paper are summarized as follows.

Firstly, a capsule-based segmentation neural network is proposed. In contrast to fully convolutional network (FCN) based models, such as U-Net and SegNet, our model can extract more information about the object. Moreover, our model integrates the encoder and the decoder in one network instead of two separated networks used in existing capsule-based segmentation networks.

Secondly, we propose a new segmentation framework for the left ventricle of the heart, which can be easily extended to the segmentation of multiple regions, such as multi-organ segmentation from images of computed tomography (CT) or magnetic resonance imaging (MRI).

Finally, we propose a new loss function to maintain the spatial relationship between the endocardium and the epicardium, which tries to keep the segmentation results coherent with the cardiac anatomical structure.

The rest of this paper is organized as follows. Related work is reviewed in Section 2. The proposed model is introduced detailedly in Section 3, followed by experiments, and comparisons and analysis in Sections 4 and 5 respectively. The proposed model is concluded in Section 6.

2 Related Work

Cardiac MRI provides a qualitative estimation of cardiac functions and has important clinical significance for the early diagnosis of heart diseases. However, the considerable intensity inhomogeneity and high anatomical variability make the segmentation of cardiac MRI images an open issue. In contrast to existing medical image segmentation methods, DNN provides an end-to-end manner to extract objects from image data. Meanwhile, the great number of neuron connections and nonlinear transformation enable it to handle noise and nonuniformity.

Convolutional neural network (CNN) based segmentation models are one type of the most widely used neural network architectures [9]. The neural network models proposed by Badrinarayanan *et al.* [10, 11] have achieved promising segmentation performance in many fields. On the basis of CNN, the above segmentation model replaces the last fully-connected layer with the convolutional layer. Then, the feature maps are restored to the original size by deconvolution operation [12] to predict the classification of each image pixel. And finally, the segmentation is transformed into a classification problem. Generally, this type of neural networks is the FCN model [13].

The majority of existing DNN algorithms in image segmentation originate from FCN. For example, a typical "encoder-decoder" structure in the segmentation field called U-Net is based on FCN and is mainly applied to medical image segmentation [14]. SegNet is based on the semantic segmentation task of FCN. The symmetrical structure of the encoder and the decoder on it is built to achieve pixel-level image segmentation. To maintain the spatial information during the down-sampling process, a context-coding network named CE-Net was proposed to capture semantic information and retain spatial information for 2D medical image segmentation [15]. FCN-based models have been widely used in medical segmentation tasks; however, they need to train shape models for the endocardium and the epicardium separately.

The loss function plays an important role in the training process. The commonly-used loss functions in FCN-based segmentation networks are Dice-based loss, cross-entropy loss, and their variants. These functions evaluate each pixel separately, judging whether a pixel is correctly predicted by the network. To obtain a better accuracy and robustness, many loss functions have been proposed, such as noise-robust loss [16], topology-preserving loss [17], Hausdorff-based loss [18], contour Dice coefficient loss [19]. These loss functions work well in the segmentation of a single object, but they cannot maintain the spatial relationship among multiple objects.

The concept of "capsules" in artificial neural networks was firstly introduced by Hinton *et al.* [20] Afterwards, Sabour *et al.* [21] introduced the capsule network which extracts both low- and high-level information through capsules updated by the dynamic routing algorithm. Some recent work applies the capsule network to segmentation tasks by transforming the segmentation into the classification problem. A capsule network model named SegCaps [22] was proposed by LaLonde for binary segmentation. Kromm and Rohr proposed an inception-based capsule network for the segmentation of vessel images [23]. He *et al.* combined the Fourier transform for LV region localization and the capsule network for the left ventricle segmentation [24]. Though these models can obtain segmentation via capsule network, only one target region can be modeled in each training process. Moreover, the encoder and the decoder of these models are built in two separated networks.

In this paper, we propose a capsule-based neural network and a spatial information aware loss function to simultaneously model the endocardium and the epicardium of the left ventricle.

3 Seg-CapNet

This section will mainly focus on the topology and loss function of the Seg-CapNet model. Fig.1 shows the overview of Seg-CapNet. The model consists of four main parts: the convolutional layers, capsule layers, fully-connected layers, and up-sampling layers. Seg-CapNet encodes multiple objects from the image as vectors containing grayscale, texture, location, orientation,

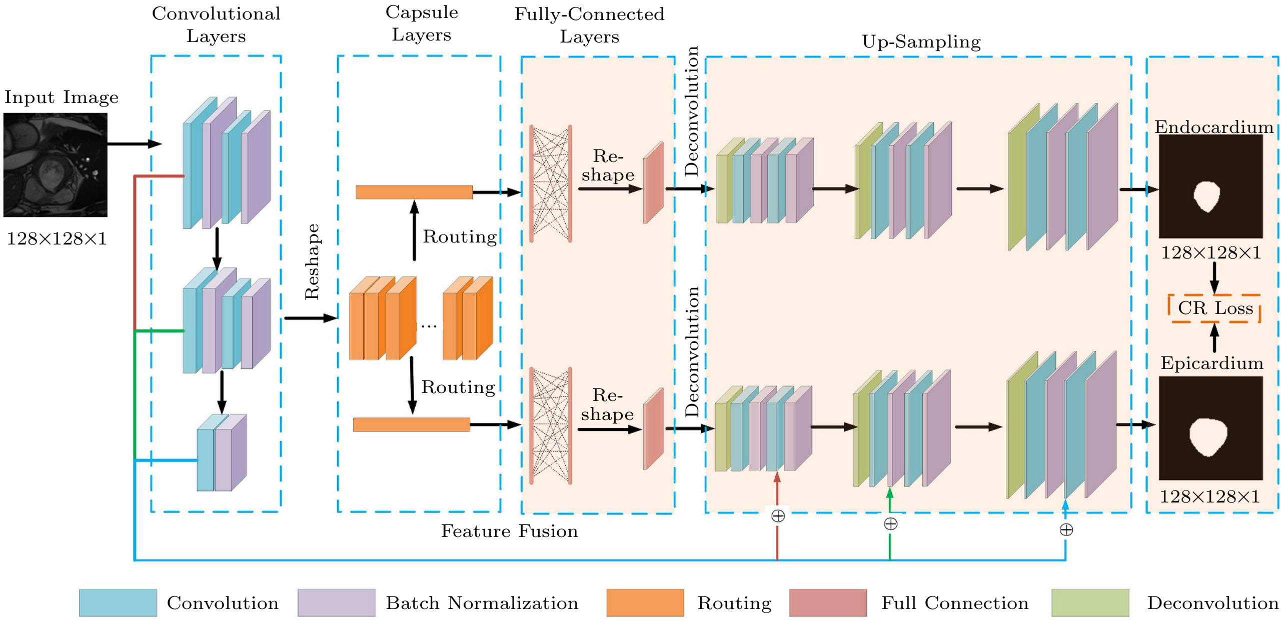


Fig.1. Overview of Seg-CapNet. The dashed boxes with colored backgrounds are the main contribution of our model. CR loss means coverage ratio loss which imposes constraints on the spatial relationship between predicted regions.

spatial, and other information of the object. Then, the segmentation is obtained by a decoder that consists of two parallel up-sampling processes.

The convolutional layers are composed of two down-sampling processes. To maintain the spatial information, Cap-SegNet implements down-sampling with strides instead of max pooling or average pooling. The capsule layers consist of the primary and the digital capsule layer. Feature maps obtained from the convolutional layers are reorganized as vectors that are connected with those in the first digital capsule layer. Two vectors are generated to represent the endocardium and the epicardium respectively. The number of vectors can be easily added or reduced according to the number of segmented objects. To reconstruct the spatial information and improve the quality of up-sampling layers, the fully-connected layers are added to map the vectors to a latent space. The fully-connected layers and up-sampling layers are divided into two parts, and each of them extracts one object. Each up-sampling process is composed of three deconvolutional blocks. In order to accelerate the convergence and improve segmentation accuracy, concatenation operation is performed to the last two deconvolutional layers. In order to maintain the spatial relationship between the endocardium and the epicardium, a novel loss function named coverage ratio (CR) loss is combined with Dice-based loss to train the model parameters.

3.1 Convolutional Layers

Existing CNN models mainly adopt the pooling operation for the feature dimension reduction, data compression, and parameter quantity, so as to alleviate overfitting. However, this operation may result in the loss of spatial information. Seg-CapNet controls the size of feature maps through convolution layers with strides to maintain spatial information in the feature maps.

As shown in Fig.1, the size of the input image is 128×128. All convolution layers extract features by 9×9 kernels. There are 32 kernels in the first two convolutional layers, 64 kernels in the third and the fourth layers respectively, and 128 kernels in the last layer. During the down-sampling, convolutions with strides of 2 are used. The output of convolutional layers is activated by the Relu function [25] defined as (1).

$$f_{\mathrm{Relu}}(x) = \max(x, 0). \tag{1}$$

Compared with other activation functions, Relu converges faster and produces less gradient disappearance.

Batch normalization (BN) [26] is added in the convolution layers to accelerate the training, alleviate overfitting, and improve the generalization of SegCaps-Net.

3.2 Capsule Layers

Capsule layers consist of the primary capsule layer and the digital capsule layer. The primary capsule receives feature maps obtained from the convolutional layers and generates feature combinations. This layer firstly performs convolution operation on the output of the fifth convolutional layer. There are 128 original capsule convolutional kernels with the size of 9×9 and strides of 2. Output of the primary capsule layer is a tensor of size $8\,192 \times 8$. The digital capsule produces two vectors which represent features of the endocardium and the epicardium respectively. The size of each feature vector is 32. Parameters that connect the primary capsule layer to the digital capsule layer are updated through dynamic routing.

3.3 Fully-Connected Layers

The spatial relationships among pixels in the feature map are changed during the dynamic routing. Thus, two fully-connected layers are added to recover spatial information before up-sampling. In Seg-CapNet, the two vectors produced by the capsule layers are mapped to two 1 024-dimensional vectors. The vectors are then reshaped into tensors of size $4 \times 16 \times 16$. The output of the fully-connected layers is activated by the Relu function.

3.4 Up-Sampling Layers

The up-sampling layers are composed of three up-sampling blocks. Each block contains one deconvolutional layer and two convolution layers. The activation function in these layers is chosen as the Relu function. The output of the fully-connected layer is reshaped to tensors of size $4 \times 16 \times 16$, which is inputted to the first deconvolutional layer. In order to recover more details during the up-sampling process, feature fusion is performed by the concatenation between deconvolutional layers and the corresponding convolutional layers during the down-sampling, as shown in Fig.1. The size of kernels in each deconvolutional layer is chosen as 9×9. And the size of kernels in each convolution layer is chosen as 3×3. The number of kernels in the three blocks are 128, 64, and 64 respectively. The output of the third up-sampling process is a mask which has the same size with the input image. The segmentation result is a binary image, which is activated by the sigmoid function in the last layer. The sigmoid function is defined as (2).

$$f_{\text{sigmoid}}(x) = \frac{1}{1+\mathrm{e}^{-x}}. \tag{2}$$

3.5 Loss Function

In medical image segmentation, the Dice coefficient is usually used to evaluate the similarity between the segmentation results and the ground truth. Dice coefficient measures the overlap between the two compared regions. Let $\varOmega : \mathrm{R}^2 \to \mathbb{R}$ be the image domain, w and h be the width and the height of the image respectively. The Dice coefficient is defined as (3).

$$D(y_g, y_p) = \frac{2\,|y_g \cap y_p|}{|y_g| + |y_p|}, \tag{3}$$

where $y_g : \varOmega \mapsto \{0,\ 1\}^{w\times h}$ is the ground truth, 0 and 1 represent the background and the foreground pixels respectively, and $y_p : \varOmega \mapsto \{0,\ 1\}^{w\times h}$ is the predicted mask. $|\cdot|$ means the number of the elements of a collection.

The Dice coefficient varies from 0 to 1, and the closer it approaches to 1, the better the segmentation result is. To minimize the loss function, we use 1 − Dice coefficient as part of the loss function. The Dice-based loss function is defined as (4). D_i represents the Dice coefficient of the endocardium and D_o is the Dice coefficient of the epicardium.

$$L_D(y_{\mathrm{g}}, y_p) = (1 - D_i(y_g, y_p)) + (1 - D_o(y_g, y_p)). \tag{4}$$

Using the Dice-based loss function only to train the model parameters may yield undesirable results, as shown by the examples in Fig.2(b) and Fig.2(c). According to the cardiac anatomies, the predicted endocardium is totally enclosed by the predicted epicardium, as shown by the example in Fig.2(a). In Fig.2(b) and Fig.2(c), the shaded areas enclosed by the predicted endocardium stay outside of the outer contour, which violates the cardiac anatomical structure. To figure it out, we propose a new loss function, namely coverage ratio, as defined by (5) and (6).

$$R(y_{pi}, y_{po}) = \frac{|y_{pi} \times (1 - y_{po})|}{|y_{pi}|}, \tag{5}$$

where $y_{pi} : \varOmega \mapsto \{0,\ 1\}^{w\times h}$ and $y_{po} : \varOmega \mapsto \{0,\ 1\}^{w\times h}$ represent the segmentation of the endocardium and the epicardium respectively. In (5), the numerator computes the number of pixels simultaneously belonging to the region enclosed by the predicted endocardium and the region outside the predicted epicardium. And the denominator calculates the total number of pixels of the region surrounded by the endocardium. Thus, (5) calculates the ratio of the endocardium pixels that locate outside the epicardium. Based on (5), the proposed coverage ratio loss function is defined as (6).

$$L_{\mathrm{R}} = \mathrm{e}^{R(y_{pi}, y_{po})}. \tag{6}$$

In conclusion, the total loss function of Seg-CapNet is defined as shown in (7)

$$L = L_{\mathrm{D}} + L_{\mathrm{R}}, \tag{7}$$

where L_{D} and L_{R} are defined as (4) and (6) respectively.

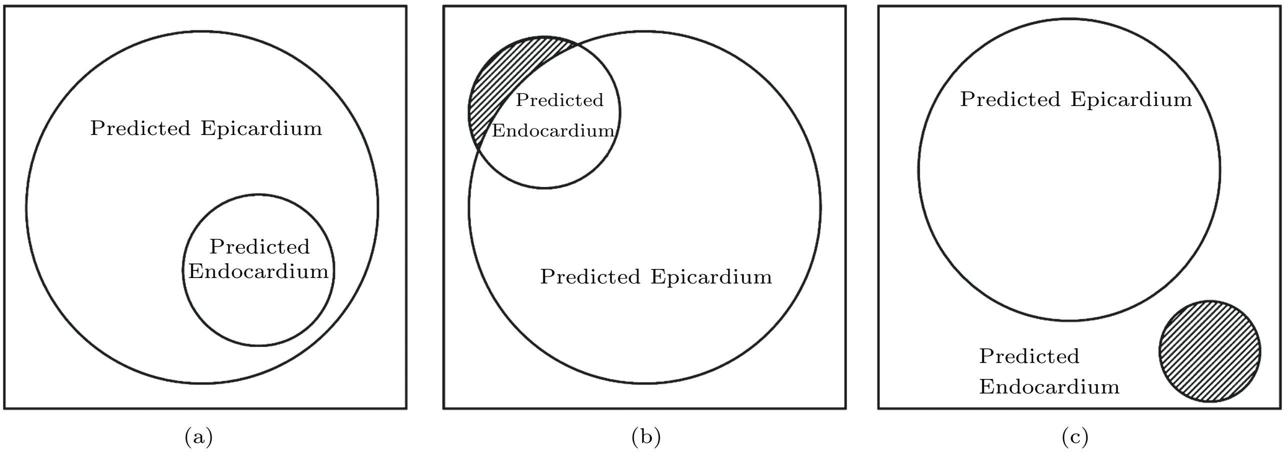


Fig. 2. Three typical spatial relationship between the predicted endocardium and epicardium. (a) The predicted endocardium is completely surrounded by the predicted epicardium. (b) The predicted endocardium intersects with the predicted epicardium. (c) The endocardium lies outside of the predicted epicardium.

4 Experiments

The proposed model is implemented with Python 3.6 and TensorFlow on Nvidia Tesla K80 GPU with 10 G video memory, Intel e5-2650 CPU, and 10 G main memory. The learning rate is set to 0.001.

4.1 Data Processing

To train and to validate the proposed model, ACDC 2017 [27] and SunnyBrook[①] datasets are processed. In cardiac MRI images, the heart generally locates near the image center. Therefore, a region is cropped near the image center. The size and the number of training and testing images are shown in Table 1.

Table 1. Partition of Two Datasets

Dataset	Size	Number of Training Images	Number of Testing Images
ACDC 2017	128 × 128	1 512	390
SunnyBrook	128 × 128	135	147

4.2 Evaluation of Segmentation Results

In order to evaluate the performance of the model, the average of the Dice coefficient and Hausdorff distance (HD) obtained from our model is calculated for evaluation. The results of the two datasets are shown in Table 2. The second column indicates the mean of Dice and HD of the endocardium of the left ventricle. The third column depicts the mean of Dice and HD of the epicardium of the left ventricle.

Table 2. Results of Our Model in Terms of Dice and HD

Dataset	Endocardium		Epicardium	
	Dice	HD	Dice	HD
ACDC 2017	0.927 5	3.231 6	0.943 9	2.909 8
SunnyBrook	0.897 5	8.941 9	0.907 0	7.369 9

4.3 Visual Segmentation Results

Fig.3 shows several cases of Seg-CapNet on the ACDC 2017 and the SunnyBrook datasets. It is worth noting that the testing images are randomly selected from the two datasets.

5 Comparisons and Analysis

The Seg-CapNet model is compared with some other FCN-based models including SegNet [11], U-Net [14], CE-Net [16], Deeplabv3 [28], and U-Net++ [29] in terms of model parameters, segmentation accuracy, robustness, and consistency to evaluate the performance of Seg-CapNet. U-Net and SegNet are widely used in image segmentation. Deeplabv3, CE-Net, and U-Net++ are newly proposed FCN-based image segmentation models and have shown their potential in medical image segmentation.

5.1 Comparison on Model Parameters

The number of model parameters decides the speed of generating the segmentation result. Thanks to its superior feature extraction, Seg-CapNet has fewer parameters compared with other models, as shown in Table 3.

From the topology of Seg-CapNet, it can be observed that the main computation is consumed by the dynamic routing algorithm which requires 4 194 304 parameters for training. One can reduce the feature maps in the convolution layers such that fewer primary capsules will be generated. In this way, the complexity could be reduced.

To compare the real speed of obtaining a segmentation, Seg-CapNet, U-Net, CE-Net, and SegNet are tested in terms of the segmentation time. There are 390 and 147 testing images collected from ACDC and SunnyBrook datasets respectively. The results are shown in Fig.4. The average segmentation time of Seg-CapNet is significantly reduced compared with the other models. It is worth noting that Seg-CapNet obtains the predicted endocardium and epicardium within a single test, while the other compared models require to test twice.

5.2 Comparison on Segmentation Accuracy, Robustness, and Consistency

Firstly, we compare Seg-CapNet with U-Net, Deeplabv3, and SegNet on several images from SunnyBrook and ACDC 2017 datasets, as shown in Fig.5. It can be observed that Seg-CapNet has better segmentation results on the whole.

Secondly, we use objective evaluation metrics, such as Dice coefficient, Jaccard similarity coefficient (JSC), HD, and average perpendicular distance (APD), to compare Seg-CapNet with U-Net, Deeplabv3, CE-Net, SegNet, and U-Net++. Table 4 and Table 5 depict

[①] Radau P, Lu Y, Connelly K *et al*. Evaluation framework for algorithms segmenting short-axis cardiac MRI. http://hdl.handle.net/10380/3070, Feb. 2021.

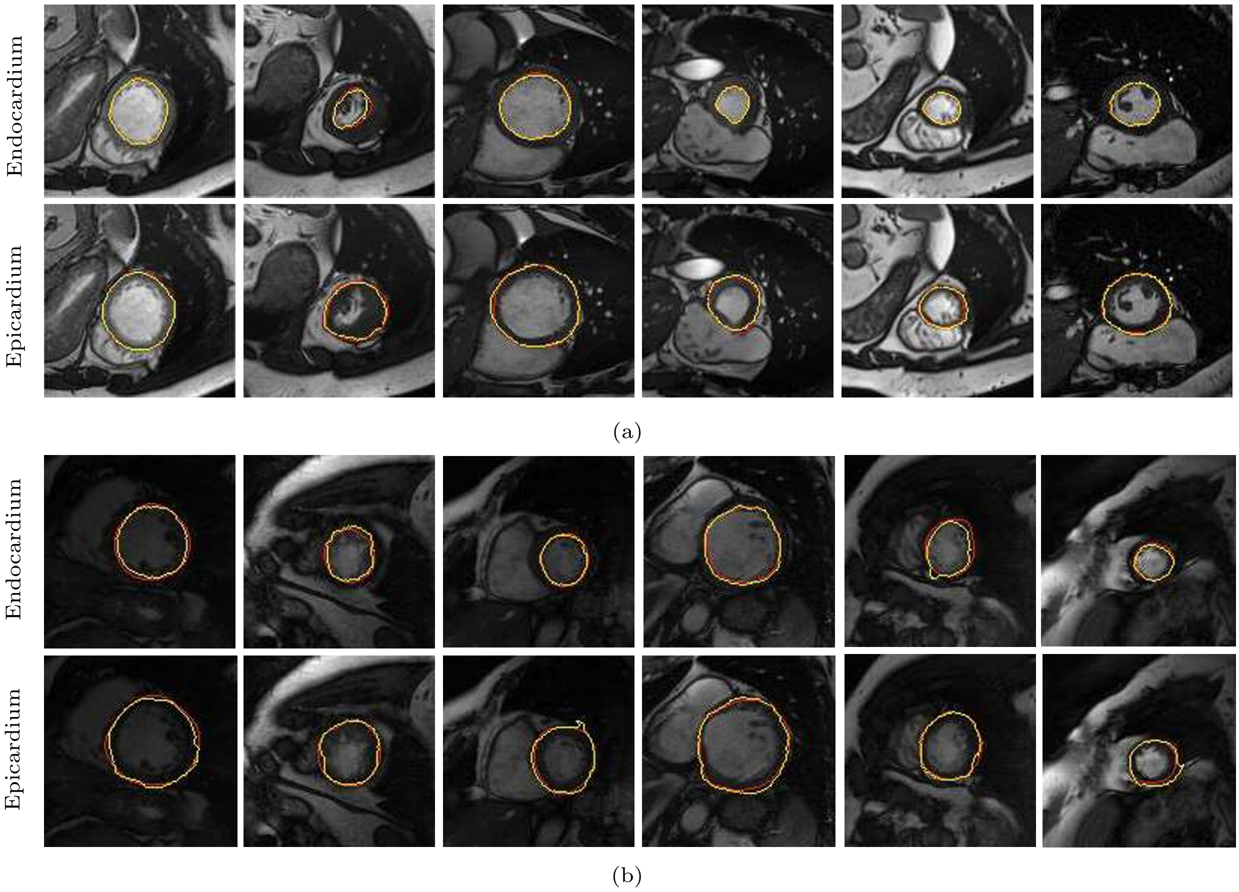


Fig.3. Visual results of Seg-CapNet. (a) Results on images from the ACDC 2017 dataset. (b) Results on images from the SunnyBrook dataset. The ground truth (in red) and predicted contours (in yellow) are simultaneously plotted on the test images.

the performance of the compared models on the above-mentioned metrics. It can be observed that Seg-CapNet outperforms the other models on these metrics. We also use the box-plot of Dice coefficient to compare the accuracy and robustness of the compared models, as shown in Fig.6. It can be seen that Seg-CapNet obtains a bigger average Dice value, indicating the accuracy of our model. Moreover, the interval between the first quartile and the third quartile is much smaller, which depicts the robustness of Seg-CapNet.

Table 3. Comparison on Model Parameter Size

Model	Parameter Size
Seg-CapNet	16 704 195
SegNet	29 440 901
U-Net	21 796 613
CE-Net	31 603 821
Deeplabv3	20 140 929
U-Net++	27 842 179

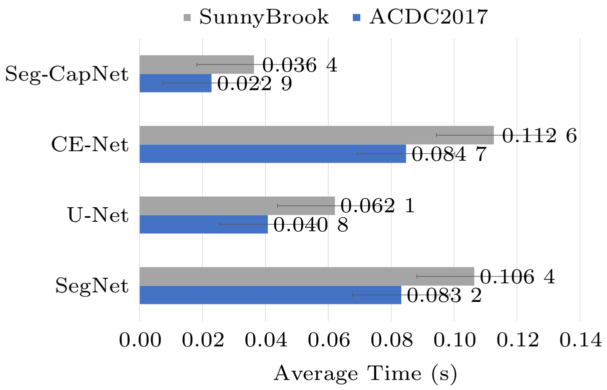


Fig.4. Comparison on time consumption.

Finally, we calculate end-systolic volume (ESV) which is a clinical index of the left ventricle to further illustrate the superiority of Seg-CapNet. Then we compare the computed ESV with the golden standard in terms of the Bland-Altman plot, as shown in Fig.7. It can be concluded that Seg-CapNet obtains smaller difference and deviation values, which indicates that our model has a better consistency.

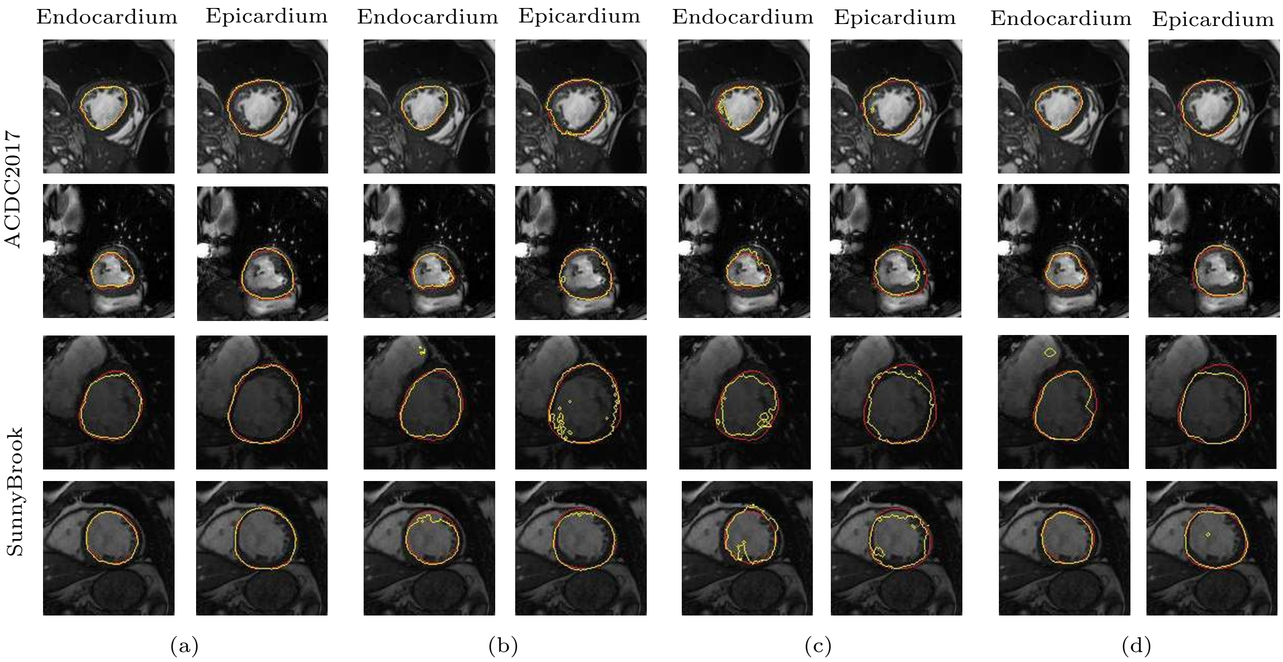


Fig.5. Visual comparison among the compared models. (a) Results of Seg-CapNet. (b) Results of U-Net. (c) Results of SegNet. (d) Results of Deeplab v3. It is worth noting that red contours represent the ground truth and yellow contours are the predicted results.

Table 4. Model Comparison on the ACDC 2017 Dataset

Model	Endocardium				Epicardium			
	Dice	JSC	HD	APD	Dice	JSC	HD	APD
Seg-CapNet	**0.927 5**	**0.124 9**	**3.231 6**	**0.899 3**	**0.943 9**	**0.093 7**	**2.909 8**	**0.980 7**
U-Net	0.914 1	0.144 9	4.592 6	1.375 8	0.925 5	0.115 3	8.879 6	1.604 5
SegNet	0.844 2	0.232 1	9.290 8	2.568 7	0.886 7	0.171 9	9.877 8	2.378 6
Deeplabv3	0.890 5	0.177 9	5.764 6	1.492 9	0.918 3	0.131 7	6.391 5	1.749 5
CE-Net	0.917 4	0.187 7	7.051 6	1.669 3	0.920 6	0.138 2	6.815 7	1.954 0
U-Net++	0.906 2	0.156 8	5.833 6	1.296 1	0.942 5	0.188 2	7.663 0	1.692 7

Note: The best result in each column is highlighted in bold.

Table 5. Model Comparison on the SunnyBrook Dataset

Model	Endocardium				Epicardium			
	Dice	JSC	HD	APD	Dice	JSC	HD	APD
Seg-CapNet	**0.897 5**	**0.158 2**	**8.941 9**	**2.094 8**	0.907 0	**0.146 9**	7.369 9	**2.061 5**
U-Net	0.885 8	0.174 2	14.104 3	2.468 8	**0.907 5**	0.159 0	**7.340 7**	2.989 9
SegNet	0.798 8	0.312 2	13.281 7	4.251 4	0.794 2	0.316 7	12.915 8	5.247 5
Deeplabv3	0.849 9	0.239 7	17.172 1	3.746 6	0.873 0	0.209 2	13.017 3	2.546 1
CE-Net	0.878 2	0.250 8	16.969 2	4.011 6	0.894 7	0.216 6	14.714 1	3.210 3
U-Net++	0.864 3	0.207 3	14.882 1	2.993 8	0.895 4	0.184 3	9.632 5	2.740 7

Note: The best result in each column is highlighted in bold.

6 Conclusions

In this paper, we proposed a capsule-based network, namely Seg-CapNet, to simultaneously segment multiple regions, more specifically, the endocardium and the epicardium of the left ventricle from cardiac MRI images. Seg-CapNet transforms the endocardium and the epicardium into two vectors representing the object entity information so as to achieve the parallel segmentation through up-sampling. We proposed a new loss function that imposes a constraint on the predicted masks to follow cardiac morphological knowledge. Compared with state-of-the-art, Seg-CapNet could not only extract the endocardium and the epi-

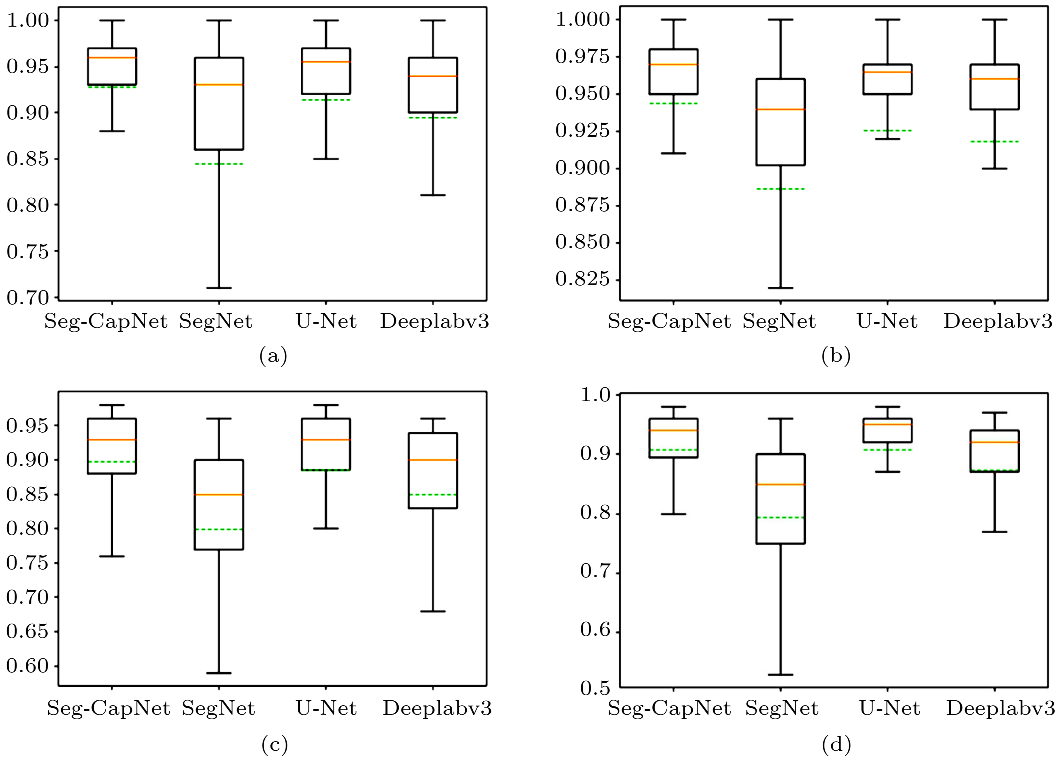


Fig.6. Dice coefficient obtained by the compared models. (a) Endocardium on ACDC 2017. (b) Epicardium on ACDC 2017. (c) Endocardium on Sunnybrook. (d) Epicardium on Sunnybrook.

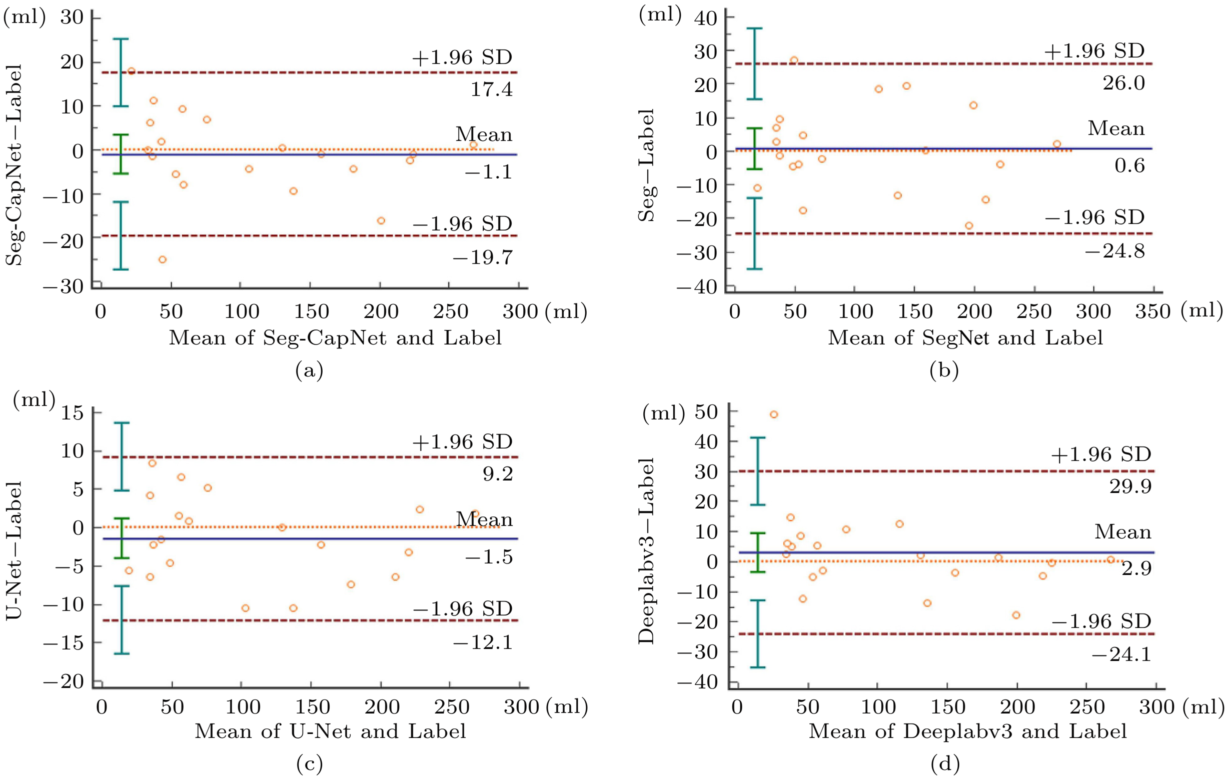


Fig.7. ESV obtained by the compared models on the ACDC 2017 dataset. (a) ESV of Seg-CapNet. (b) ESV of SegNet. (c) ESV of U-Net. (d) ESV of Deeplabv3.

cardium of the left ventricle simultaneously but also perform better in terms of Dice and HD on ACDC 2017 and Sunnybrook datasets. Besides, Seg-CapNet only requires about half the quantity of parameters to be trained, which can also save computation cost during testing. The proposed segmentation framework can be easily extended to segment more regions by adding capsule vectors and deconvolutional layers. Future work will focus on the improvement of dynamic routing which consumes a large portion of computation in the capsule neural network.

References

[1] Cootes T F, Taylor C J, Cooper D H *et al.* Active shape models—Their training and application. *Computer Vision and Image Understanding*, 1995, 61(1): 38-59. DOI: 10.1006/cviu.1995.1004.

[2] Soliman A, Khalifa F, Elnakib A *et al.* Accurate lungs segmentation on CT chest images by adaptive appearance-guided shape modeling. *IEEE Transactions on Medical Imaging*, 2016, 36(1): 263-276. DOI: 10.1109/TMI.2016.2606370.

[3] Cootes T F, Edwards G J, Taylor C J. Active appearance models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2001, 23(6): 681-685. DOI: 10.1109/34.927467.

[4] Matthews l, Baker S. Active appearance models revisited. *International Journal of Computer Vision*, 2004, 60: 135-164. DOI: 10.1023/B:VISI.0000029666.37597.d3

[5] Wachinger C, Fritscher K, Sharp G *et al.* Contour-driven atlas-based segmentation. *IEEE Transactions on Medical Imaging*, 2015, 34(12): 2492-2505. DOI: 10.1109/TMI.2015.2442753.

[6] Maintz J B, Viergever M A. A survey of medical image registration. *Medical Image Analysis*, 1998, 2(1): 1-36. DOI: 10.1016/S1361-8415(01)80026-8.

[7] Litjens G, Kooi T, Bejnordi B E *et al.* A survey on deep learning in medical image analysis. *Medical Image Analysis*, 2017, 42: 60-88. DOI: 10.1016/j.media.2017.07.005.

[8] LeCun Y, Bengio Y, Hinton G E. Deep learning. *Nature*, 2015, 521(7553): 436-444. DOI: 10.1038/nature14539.

[9] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In *Proc. the 26th Int. Conference on Neural Information Processing Systems*, December 2012, pp.1097-1105. DOI: 10.5555/2999134.2999257.

[10] Badrinarayanan V, Handa V, Cipolla R. SegNet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. arXiv:1505.07293, 2015. https://arxiv.org/pdf/1505.07293.pdf, March, 2020.

[11] Badrinarayanan V, Kendall A, Cipolla R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 39(12): 2481-2495. DOI: 10.1109/TPAMI.2016.2644615.

[12] Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation. In *Proc. the 2015 IEEE International Conference on Computer Vision*, December 2015, pp.1520-1528. DOI: 10.1109/ICCV.2015.178.

[13] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In *Proc. the 2015 IEEE International Conference on Computer Vision and Pattern Recognition*, June 2015, pp.3431-3440. DOI: 10.1109/CVPR.2015.7298965.

[14] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In *Proc. the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention*, October 2015, pp.234-241. DOI: 10.1007/978-3-319-24574-4_28.

[15] Gu Z W, Cheng J, Fu H Z *et al.* CE-Net: Context encoder network for 2D medical image segmentation. *IEEE Transactions on Medical Imaging*, 2019, 38(10): 2281-2292. DOI: 10.1109/TMI.2019.2903562.

[16] Wang G, Liu X, Li C *et al.* A noise-robust framework for automatic segmentation of COVID-19 pneumonia lesions from CT images. *IEEE Transactions on Medical Imaging*, 2020, 39(8): 2653-2663. DOI: 10.1109/TMI.2020.3000314.

[17] Hu X, Li F, Samaras D *et al.* Topology-preserving deep image segmentation. In *Proc. the 33rd Annual Conference of Neural Information Processing Systems,* December 2019, pp.5658-5669.

[18] Karimi D, Salcudean S E. Reducing the Hausdorff Distance in medical image segmentation with convolutional neural networks. *IEEE Transactions on Medical Imaging*, 2020, 39(2): 499-513. DOI: 10.1109/TMI.2019.2930068.

[19] Moltz J H, Hänsch A, Lassen-Schmidt B *et al.* Learning a loss function for segmentation: A feasibility study. In *Proc. the 17th IEEE Int. Biomedical Imaging Symp.*, April 2020, pp.357-360. DOI: 10.1109/ISBI45749.2020.9098557.

[20] Hinton G E, Alex K, Wang S D. Transforming auto-encoders. In *Proc. the 21st Int. Conference on Artificial Neural Networks*, June 2011, pp.44-51. DOI: 10.1007/978-3-642-21735-7_6.

[21] Sabour S, Frosst N, Hinton G E. Dynamic routing between capsules. In *Proc. the 31 st Int. Conference on Neural Information Processing Systems*, December 2017, pp.3856-3866.

[22] LaLonde R, Bagci U. Capsules for object segmentation. arXiv:1804.04241, 2018. https://arxiv.org/pdf/1804.04241v1.pdf, March, 2020.

[23] Kromm C, Rohr K. Inception capsule network for retinal blood vessel segmentation and centerline extraction. In *Proc. the 17th IEEE Int. Biomedical Imaging Symp.*, April 2020, pp.1223-1226. DOI: 10.1109/ISBI45749.2020.9098538.

[24] He Y, Qin W, Wu Y *et al.* Automatic left ventricle segmentation from cardiac magnetic resonance images using a capsule network. *Journal of X-Ray Science and Technology*, 2020, 28(3):541-553. DOI: 10.3233/XST-190621.

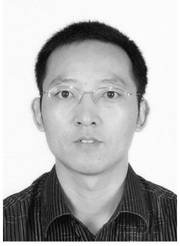
[25] Hara K, Saito D, Shouno H. Analysis of function of rectified linear unit used in deep learning. In *Proc. the 2015 International Joint Conference on Neural Networks*, July 2015. DOI: 10.1109/IJCNN.2015.7280578.

[26] Loffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv:1502.03167, 2015. https://arxiv.org/pdf/1502.03167.pdf, March 2020.

[27] Bernard O, Lalande A, Zotti C *et al.* Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging*, 2018, 37(11): 2514-2525. DOI: 10.1109/TMI.2018.2837502.

[28] Chen L, Papandreou G, Schroff F *et al.* Rethinking atrous convolution for semantic image segmentation. arXiv: 1706.05587, 2017. https://arxiv.org/abs/1706.05587, June 2020.

[29] Zhou Z W, Siddiquee M, Tajbakhsh N *et al.* UNet++: A nested U-Net architecture for medical image segmentation. In *Proc. the 4th International Workshop on Deep Learning in Medical Image Analysis*, September 2018, pp.3-11. DOI: 10.1007/978-3-030- 00889-5_1.



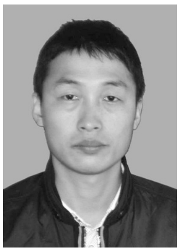
Yang-Jie Cao is currently an associate professor of the School of Software, Zhengzhou University, Zhengzhou. He received his Ph.D. degree in computer science from Xi'an Jiaotong University, Xi'an, in 2012, and his M.S. degree in computer science from Zhengzhou University, Zhengzhou, in 2006. His current research interests include computer vision and intelligent computing, artificial intelligence, and high-performance computing.

Shuang Wu is currently a Master student majoring in computer science at Zhengzhou University, Zhengzhou. Her current research interests include deep neural networks, computer vision, Internet of Things, and medical image processing.

Chang Liu is currently a software engineer of the Bank of Zhengzhou, Zhengzhou. He received his M.S. degree in computer science from Zhengzhou University, Zhengzhou, in 2020. His current research interests include artificial intelligence and medical image processing.

Nan Lin is currently an associate professor of the School of Software, Zhengzhou University, Zhengzhou. She received her M.S. degree in computer science from Huazhong University of Science and Technology, Wuhan, in 2003. Her current research interests include intelligent systems and artificial intelligence.

Yuan Wang is currently an experimentalist of the Center of Modern Analysis and Gene Sequencing, Zhengzhou University, Zhengzhou. She received her Ph.D. degree in material science from Northwestern Polytechnical University, Xi'an, in 2018, and her B.S. degree in material science from Xi'an University of Science and Technology, Xi'an, in 2009. Her research interests include computer vision and its application in the analysis of scanning electron microscopes.



Cong Yang is currently a lecturer at the School of Software, Zhengzhou University, Zhengzhou. He received his Ph.D. degree in computer science from Xi'an Jiaotong University, Xi'an, in 2017, and his B.S. degree in information security from Chongqing University, Chongqing, in 2010. His current research interests include deep neural networks, computer vision, and medical image processing.

Jie Li is a professor in the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, and an adjunct professor in Zhengzhou University, Zhengzhou. He received his Dr. Eng. degree from the University of Electro-Communications, Tokyo, in 1993, his M.E. degree in electronic engineering and communication systems from China Academy of Posts and Telecommunications, Beijing, in 1985, and his B.E. degree in computer science from Zhejiang University, Hangzhou, in 1982. His current research interests are in big data and AI, blockchain, edge computing, networking and security, OS, modeling, and performance evaluation of information systems. He is the co-chair of IEEE Technical Community on Big Data, the founding chair of IEEE ComSoc Technical Committee on Big Data, and the co-chair of IEEE Big Data Community. He serves as an associated editor for many IEEE journals and transactions. He has also served on the program committees for several international conferences.