TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 02/16 pp628–638
DOI: 10.26599/TST.2022.9010023
Volume 28, Number 4, August 2023

SUNet++: A Deep Network with Channel Attention for Small-Scale Object Segmentation on 3D Medical Images

Lan Zhang, Kejia Zhang*, and Haiwei Pan

Abstract: As a deep learning network with an encoder-decoder architecture, UNet and its series of improved versions have been widely used in medical image segmentation with great applications. However, when used to segment targets in 3D medical images such as magnetic resonance imaging (MRI), computed tomography (CT), these models do not model the relevance of images in vertical space, resulting in poor accurate analysis of consecutive slices of the same patient. On the other hand, the large amount of detail lost during the encoding process makes these models incapable of segmenting small-scale tumor targets. Aiming at the scene of small-scale target segmentation in 3D medical images, a fully new neural network model SUNet++ is proposed on the basis of UNet and UNet++. SUNet++ improves the existing models mainly in three aspects: 1) the modeling strategy of slice superposition is used to thoroughly excavate the three dimensional information of the data; 2) by adding an attention mechanism during the decoding process, small scale targets in the picture are retained and amplified; 3) in the up-sampling process, the transposed convolution operation is used to further enhance the effect of the model. In order to verify the effect of the model, we collected and produced a dataset of hyperintensity MRI liver-stage images containing over 400 cases of liver nodules. Experimental results on both public and proprietary datasets demonstrate the superiority of SUNet++ in small-scale target segmentation of three-dimensional medical images.

Key words: 3D medical images; small-scale target; segmentation; attention mechanism

1 Introduction

Assisted medical diagnosis is an important direction in the development of artificial intelligence. The main application is to identify and segment organs or lesions in medical images. In recent years, with the rise of deep learning, various models based on convolutional neural networks (CNNs), such as FCN[1], UNet[2], Deeplab[3], PSPNet[4], and RCNN[5, 6], have been widely used in medical image segmentation. In particular, UNet has received the most widespread attention because of its excellent application and efficient efficiency. Traditional UNet is based on an encoder-decoder architecture. During the encoding process, convolution and pooling operations are used to continuously reduce the dimensionality of the image and extract its high-level semantic features. In the decoding process, an up-sampling operation is used to gradually restore the scale of the image and finally form a mask. Meanwhile, skip connections are used to combine the high-level semantic feature maps with the corresponding low-level detailed feature maps. On this basis, to enhance the accuracy of segmentation, a series of improved versions of UNet are proposed. Just like UNet++[7], this model introduces dense skip connections to UNet in order to further reduce the semantic gap between encoder and decoder. However, when analyzing three-dimensional medical images such as magnetic resonance imaging (MRI) and computed tomography (CT), specifically when segmenting small-scale targets in the image, the

• Lan Zhang, Kejia Zhang, and Haiwei Pan are with the College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China. E-mail: zhanglan2015@hrbeu.edu.cn; kejiazhang@hrbeu.edu.cn; panhaiwei@hrbeu.edu.cn.
* To whom correspondence should be addressed.
Manuscript received: 2022-05-17; accepted: 2022-06-21

effects of these models are not satisfactory.

The characteristic of MRI, CT, and other medical images is to form continuous multiple scan pictures of the patient in a certain direction (generally in the vertical direction), thereby obtaining information about the relevant parts in three dimensions. Because the distance between adjacent scans is tiny, there is a strong correlation between adjoining scanned pictures. Models such as UNet were initially designed to analyze a single picture, and lack the ability to model three-dimensional images. Ignoring the correlation between pictures leads to low accuracy in 3D volume analysis. On the other hand, these models continuously reduce the dimensionality of the pictures through operations such as pooling in the encoding process to extract their high-level semantic features. In this process, the loss of a large amount of detailed information makes the model's accuracy in small-scale target segmentation poor. Although models such as UNet++ use a large number of intensive jump connections to compensate for the information loss in the encoding process, this kind of untargeted data enhancement technology can not effectively enhance the model's effect in small target segmentation, in addition to increasing the complexity of the model.

Although there are some improved versions of UNet, such as V-Net[8] and 3DUNet[9], three-dimensional modeling of pictures has been tried. However, the insertion of a large amount of invalid data in the modeling process severely affected its convergence effect. The loss of information in the encoding process has not been effectively resolved, resulting in the poor performance of these 3D versions of UNet models when segmenting small-scale targets.

Aiming at the scene of small-scale target segmentation on 3D medical images on the basis of UNet and UNet++, this paper proposes a new neural network model SUNet++. SUNet++'s improvements compared to UNet and UNet++ are mainly reflected in three aspects: 1) the modeling strategy of slice superposition is used to fully excavate the three dimensional information of the data; 2) by adding an attention mechanism during the decoding process, small scale targets in the picture are retained and amplified; 3) in the up-sampling process, the transposed convolution operation is used to further enhance the effect of the model.

In order to verify the effect of the model, we collected and produced a liver hyperintensity MRI dataset. At the same time, in order to ensure great data security, similar to previous methods[10, 11], our dataset hides the users' private information. The dataset contains more than 400 cases of liver nodule lesions (small-scale targets), which are marked by doctors with rich experience in diagnosis. In the process, we learned that even for experienced doctors, it is a very tough task to distinguish small-scale lesions in these MRI images. The experimental results on our dataset and the public dataset have proved the superiority of SUNet++ for small-scale target segmentation on three-dimensional medical images. To this end, the contribution of this paper includes two main aspects: (1) in order to adapt the small-scale target segmentation on 3D medical images, a neural network model SUNet++ is proposed to overcome the shortcomings of existing research works; (2) a liver hyperintensity MRI dataset containing more than 400 cases of liver nodules was collected and produced, and the superiority of the proposed model was verified on the dataset.

2 Related Work

Some commonly used methods for traditional CT and MRI segmentation are threshold-based[12], region-based[13], deformation-based[14], fuzzy-based[15], and neural network based[16]. However, medical images are complex and diverse, which will case a lot of issues, e.g., blurred tissue edges and unclear edges of lesions. Also, the accuracy of segmentation is low. Thus Long et al.[1] and Korez et al.[16] proposed a full convolutional network (FCN) structures and optimized the 3D FCN model algorithm to further improve the accuracy of segmentation. After that, many researchers used MRF algorithm[13] or CRF algorithm[12] to improve the segmentation results output by FCN. These methods further optimize the segmentation results.

The UNet network structure[2] is based on FCN. UNet is suitable for medical image segmentation. The up-sampling stage and down-sampling stage of UNet use the same number of levels of convolution operations, and the skip connection structure is used to connect the down-sampling layer and the up-sampling layer, so that the features extracted by the down-sampling layer can be directly transferred to the up-sampling layer. This makes the pixel positioning of the UNet network more precise and the segmentation accuracy higher.

3DUNet network structure[9] realizes 3D volume segmentation by inputting a continuous 2D slice sequence of 3D volume. V-net[8], a 3D deformed structure, uses the 3D convolution kernel to perform

the image processing. Both long-hop and short-hop connection structures are used in UNet structure[17] to improve the segmentation accuracy. Kamnitsas et al.[18] and Ghafoorian et al.[19] used multi-scale convolution to extract global and local image information. Reference [20] used the UNet network and joined the jump connection structure, so that the network structure can still get a good segmentation result with less training data.

UNet++[7] redesigned the jump path, which makes it easier to optimize semantically similar feature maps. This structure integrates the advantages of long and short connections, thus can capture features of different levels. UNet3plus[21] redesigned the interconnection between encoder and decoder. Therefore, fine particle details can be captured from full scale. It is helpful for accurate segmentation, particularly for organs of different scales in the medical images volume.

Currently popular datasets, such as CHAOS[22], segment abdominal organs from CT and MRI data. The cardiac magnetic resonance imaging dataset[23] consists of a sequence of short-axis cardiac magnetic resonance images collected from 33 subjects, with a total of 7980 2D images. ACDC[24] consists of 150 patients and has three categories of heart. Reference [25] proposed the PROMISE12 challenge—MICCAI prostate magnetic resonance images segmentation. The challenge data includes patients with benign diseases (such as benign prostatic hyperplasia) and prostate cancer. LiTS[26] competition dataset contains 200 cases. Due to the low contrast with normal tissues and various gray levels, the dataset of 200 cases is limited, which is likely to cause over-fitting. It is not easy to observe tumors with CT scan images. These datasets are not good enough to distinguish the tumors. Most medical datasets involve more personal privacy, in order to protect private data in large datasets. Some methods have been proposed to prevent sensitive information from being attacked[27–29], and have excellent effect. A method of privacy data protection for multiple parties is important[30].

Most of the existing methods focused on the optimization of the single-layer slice information extraction process. Take V-Net and other 3D models as examples. They simply added the overall three-dimensional information. Large-scale labels can maintain the balance of positive and negative samples. However, for small-scale labels, a huge number of negative samples will be introduced. Also, for small-scale label datasets, there are currently no good results.

3 Methodology

This section introduces the SUNet++ model, which is based on UNet and UNet++. The structure is shown in Fig. 1.

3.1 Encoder-decoder structure

SUNet++ still uses the encoder-decoder structure. The pixel size of input training slices of the model is $5 \times 512 \times 512$. The input and output tensor size of each node is shown in Table 1.

3.1.1 Encoder

Five coding units are used (E_i in Fig. 1) to extract high-level features of training slices. Suppose the size of the input training slices X is $c_i \times w_i \times h_i$, where c_i is the number of channels, and w_i , h_i are the width and height of the feature training slices. $c_0 = 5$, $w_0 = h_0 = 5$. E_i includes a residual part and an SElayer part. Its structure is shown in Fig. 2a.

In the residual structure, the tensor X passes through a convolutional layer with the 1×1 kernel and a batch normal (BN) layer to obtain a tensor X_1 ($2c_i \times w_i \times h_i$).

$$X_1 = BN\left(Conv_{1\times1}\left(X\right)\right) \in \mathbb{R}^{2c_i \times w_i \times h_i} \quad (1)$$

Then, the tensor X_1 passes through a convolutional layer with the 3×3 kernel and a batch normal layer to obtain a tensor X_2 ($2c_i \times w_i \times h_i$).

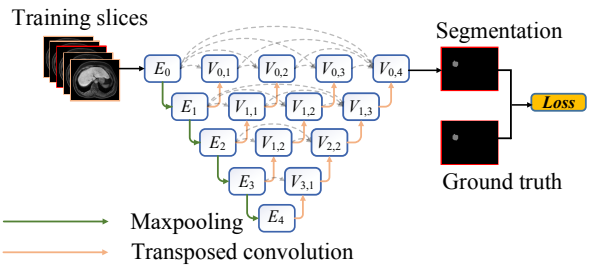
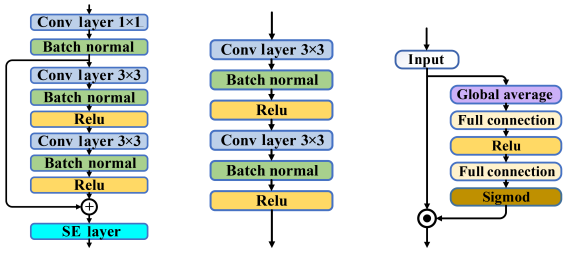


Fig. 1 The structure of the SUNet++ model.

Table 1 Pixel size of SUNet++ structure.

	Input	Output	Block Layer	Input	Output	Block Layer	Input	Output
E_0	5×512×512	32×512×512	$V_{0,1}$	96×512×512	32×512×512	$V_{1,2}$	256×256×256	64×256×256
E_1	32×256×256	64×256×256	$V_{1,1}$	192×256×256	64×256×256	$V_{2,2}$	512×128×128	128×128×128
E_2	64×128×128	128×128×128	$V_{2,1}$	384×128×128	128×128×128	$V_{0,3}$	160×512×512	32×512×512
E_3	128×64×64	256×64×64	$V_{3,1}$	76×64×64	256×64×64	$V_{1,3}$	320×256×256	64×256×256
E_4	256×32×32	512×32×32	$V_{0,2}$	128×512×512	32×512×512	$V_{0,4}$	192×512×512	32×512×512



(a) E_i (b) $V_{i,j}$ (c) SElayer

Fig. 2 Detailed structure of coding units in SUNet++.

$$X_2 = Relu\left(BN\left(Conv_{3\times3}\left(X_1\right)\right)\right) \in \mathbb{R}^{2c_i\times w_i\times h_i} \quad (2)$$

Then, X_2 passes through a convolutional layer with a 3×3 convolution kernel, a batch normal layer, and a Relu layer to obtain X_3 $(2c_i \times w_i \times h_i)$.

$$X_3 = Relu\left(BN\left(Conv_{3\times3}\left(X_2\right)\right)\right) \in \mathbb{R}^{2c_i\times w_i\times h_i} \quad (3)$$

After that, skip connecting X_1 to X_3, the output of the residual structure is X' $(2c_i \times w_i \times h_i)$:

$$X' = F_{Resdul}(X) = X_1 + X_3 \in \mathbb{R}^{2c_i\times w_i\times h_i} \quad (4)$$

Specially, if $i = 0, X_1, X_2, X_3 \in \mathbb{R}^{32\times w_i\times h_i}$. By using the residual structure, the model overcomes the degradation problem of the multilayer downsampling network.

After the residual part, X' is passed to the channel attention mechanism, SElayer. And, the output is X'' $(2c_i \times w_i \times h_i)$. SElayer majors in retaining and amplifying the characteristics of small-scale targets. Its specific structure is described in following part. Finally, X'' is passed through a 2×2 maxpooling layer. The layer changes its size to $2c_i \times w_i \times h_i$ and passes it to E_{i+1}. In addition, X'' is passed to the upper decoding unit $V_{i-1,1}$ with a size of $2c_i \times 2w_i \times 2h_i$, after being subjected to 4×4 transposed convolution.

3.1.2 Decoder

In the decoding process, multiple cascaded decoding units are used ($V_{i,j}, i = 0,1,\ldots,3\,; j = 1,2,\ldots,4$ in Fig. 1). The input of $V_{i,j}$ consists of two parts: 1) the feature map obtained by the lower decoding unit $V_{i+1,j-1}$, after transposed convolution (the size is $4c_i \times w_i \times h_i$); 2) the feature maps output by the same layer decoding unit $V_{i,0}, V_{i,1}, \ldots, V_{i,j-1}$. The size of each feature map is $2c_i \times w_i \times h_i$. $V_{i,0} = E_i$. Because the same size of the feature map, $Y_{i,j}$ of $V_{i,j}$ can be obtained by superimposing on the channel. Its size is $(4+2j)\,c_i \times w_i \times h_i$. $V_{i,j}$ contains two basic structures of VGG[31]. Each structure contains a 3×3 convolutional layer, a batch normal, and an activation layer. The structure is shown in Fig. 2b.

First, $Y_{i,j}$ passes through the first VGG structure to get the feature map $Y^1_{i,j}$ $(2c_i \times w_i \times h_i)$.

$$Y^1_{i,j} = Relu\left(BN\left(Conv_{3\times3}\left(Y_{i,j}\right)\right)\right) \in \mathbb{R}^{2c_i\times w_i\times h_i} \quad (5)$$

Second, $Y^1_{i,j}$ passes through the next VGG structure to get the feature map $Y^2_{i,j}$ $(2c_i \times w_i \times h_i)$.

$$Y'_{i,j} = Relu\left(BN\left(Conv_{3\times3}\left(Y^1_{i,j}\right)\right)\right) \in \mathbb{R}^{2c_i\times w_i\times h_i} \quad (6)$$

$Y'_{i,j}$, the output of the decoding block $V_{i,j}$, continues to two directions: 1) though a 4×4 transposed convolution, whose step size is 2, its size is changed into $2c_i \times 2w_i \times 2h_i$. After that, it is passed to the upper layer decoding unit $V_{i-1,j+1}$. 2) passed to the same layer decoding block $V_{i,j+1}, V_{i,j+2}, \ldots, V_{i,4-i}$ by skip connection. Finally, the output $Y'_{0,4}$ of the decoding units $V_{0,4}$ is subjected to a 1×1 convolution operation. And the models's output is Y' $(q\times w_0\times h_0)$, the segmentation. Especially, q is the number of tumor types.

3.1.3 SElayer

The SUNet++ model uses the SElayer in each encoding unit. This function is to add an attention mechanism to different channels of the feature map, so as to retain and amplify the features of small-scale targets. In UNet and UNet++, after multiple feature extraction (encoding), the target part has a tendency to be captured by the feature map of a certain channel. Due to the undifferentiated treatment of all channels, the edge information will be ablated in multiple down-sampling operations (maxpooling). Therefore, this makes it easy for these models to miss some small-scale targets. In the SElayer module, we use global pooling and fully connected layer (FC layer) to continue exploring the feature map. Then, feed the result of feature extraction to different channels through the attention mechanism, so as to give higher weight to the feature map of the captured lesion. This process will retain the most useful information. The structure of the SElayer module is shown in Fig. 2c.

Assume that the input of SElayer is a feature map X $(c\times w\times h)$. Firstly, the module uses global average pooling to obtain the global features in each channel feature map.

$$z[k] = F_{GP}(X[k]) = \frac{1}{wh}\sum_{i=1}^{w}\sum_{j=1}^{h} X[k](i,j) \quad (7)$$

$X[k]$ $(k = 1,2,\ldots,c)$ is the feature map of the k-th channel. $X[k](i,j)$ is the value of point (i,j) in $X[k]$. And $z[k]$ is the calculation result of the k-th channel. After global average pooling, the calculation result of

the input feature map X is

$$z = F_{\mathrm{GP}}(X) \quad \in \mathbb{R}^c \tag{8}$$

Next, pass z to two fully connected layers with activation functions. The first fully connected layer uses the parameter matrix $W_1 \in \mathbb{R}^{\frac{c}{u} \times c}$ to reduce the dimension of z. Then the ReLU function is used to activate. The second fully connected layer uses the parameter matrix $W_2 \in \mathbb{R}^{c \times \frac{c}{u}}$ to restore the intermediate results to the original dimensions, and uses the sigmoid function to activate. Then, each channel's attention coefficient is obtained. The dimensionality reduction coefficient u is a hyperparameter. The calculation process of the attention coefficient can be expressed as

$$s = F_{\mathrm{Att}}(X) = \sigma\left(W_2 \times ReLU\left(W_1 \times F_{\mathrm{GP}}(X)\right)\right) \in \mathbb{R}^c \tag{9}$$

Finally, multiply the attention coefficient by the feature map on the corresponding channel to enlarge or reduce the information on different channels.

$$X' = F_{\mathrm{SElayer}}(X) = s \odot X \in \mathbb{R}^{c \times w \times h} \tag{10}$$

where $\odot$ is the Hadamard product.

3.2 3D superposition and transposed convolution

Observing liver slices, experts generally distinguish blood vessels from tumors through three-dimensional information provided by multiple serial slices. Compared with the previous dataset, large-scale labels can ignore three-dimensional information, and better results can be acquired from two-dimensional slices alone. However, the small-scale tags in this dataset need to fully consider the three-dimensional information of the tags. There are some models that consider the three-dimensional characteristics of medical images, such as VNnet[8], 3DUNet[9], etc. These methods put the training slices into the model. However, the small label dataset contains a large amount of unlabeled data. The existence of negative samples causes uneven sample distribution, which affects the convergence effect of the algorithm. Therefore, a method is needed to accurately retain the three-dimensional information of the images. Also, the number of positive and negative samples needs to be balanced to ensure a better convergence effect.

Simulating the human observation method, SUNet++ adopts a slice superposition strategy for the input to retain the three-dimensional information of the label. At the same time, it can flexibly remove negative samples. After preprocessing, training slices of a patient is arranged vertically into a series of slices $S_1, S_2, \ldots, S_m$, and each slice is adjusted to a single-channel slice with a size of 512×512. SUNet++ generates masks for each slice in turn to segment small target tumors. In order to generate the mask of slice S_i, the input of the model is $S_{i-l}, S_{i-l+1}, \ldots, S_i, \ldots, S_{i+l-1}, S_{i+l}$. The feature map X_i ($(2l+1) \times 512 \times 512$) is formed by superimposing in turn.

$$X_i = [S_{i-l}, S_{i-l+1}, \ldots, S_i, \ldots, S_{i+l-1}, S_{i+l}] \tag{11}$$

The number of superimposed channels $p = 2l + 1$ is one of the hyperparameters of the model. For edge slicing, a mirror image extension method is used. If $i - k < 1$, $S_{i-k} = S_{i+k}$; If $i + k > m$, $S_{i+k} = S_{i-k}$.

In the previous semantic segmentation methods, bilinear interpolation is usually used as the resolution method for restoring the feature map. The parameter value is sufficient, so no more detailed results can be obtained, and the performance is poor in the upsampling of small-scale labels. We use transposed convolution to achieve up-sampling because this process is learnable, and high-resolution information can be fully recovered during the parameter adjustment process, avoiding errors introduced in the bilinear interpolation process, and retaining the clear boundaries.

3.3 Loss

For input X, suppose its label is Y, and the model output result is Y'. Among them, the size of Y and Y' are both $q \times w_0 \times h_0$. q represents the number of tag categories. One of the loss functions used by the model is the dice coefficient, which is an ensemble similarity measurement function. It is usually used to calculate the similarity value range of two samples as the value range [0, 1], which is defined as:

$$L_{\mathrm{dice}} = \sum_{k=1}^{q} \left(1 - \frac{Y[k] \times Y'[k]}{Y[k] + Y'[k]}\right) \tag{12}$$

$Y[k]$ is the real mask of the k-th label, $Y'[k]$ is the mask generated by the model for the k-th label. L_{dice} represents the division of the tumor area, which is more suitable for extremely uneven samples. In general, using L_{dice} will adversely affect backpropagation and easily make training unstable. Therefore, we add the classification loss L_{BCE} that can be compared to the label.

$$L_{\mathrm{BCE}} = -\frac{1}{q}\sum_{k=1}^{q}\left(Y[k] \times \ln Y'[k]\right) - \frac{1}{q}\sum_{k=1}^{q}\left((1-Y[k]) \times \ln\left(1-Y'[k]\right)\right) \tag{13}$$

Combining binary cross entropy (BCE) loss and dice loss to calculate the loss from the two aspects of tumor

area and classification. The overall loss is as follows.

$$L_{\text{total}} = a \times L_{\text{BCE}} + L_{\text{dice}} \tag{14}$$

where a is the combination coefficient, $a = 0.5$.

4 Dataset

4.1 Dataset introduction

We collected Gd-EOB-DTPA MR enhanced hepatobiliary images from a tumor hospital from January 2020 to December 2020. At the same time, four medical experts used professional software to label them. We draw on some relevant methods and theories[32–35] to ensure that the private information of data is not leaked, and the security of information is protected. Thus, a small-scale enhanced hepatobiliary phase images (SEHPI) dataset is formed. Moreover, the dataset is a liver-specific contrast agent dynamically enhanced the hepatobiliary MR image. Different from MR images of non-liver-specific contrast agents, specific contrast media hepatobiliary stage images are helpful for the screening and diagnosis of liver nodules, and it is easier to delineate the boundaries of the lesions. During the labeling process, the tumors are distinguished and classified. They are liver cancer, metastases, cysts, and hemangioma (Fig. 3).

In the previous literature, most of them major in CT images of liver cancer nodules, and less work on MR images. Most of the literature is segmentation and classification of liver cancer lesions, and there is no specific classification and segmentation of common liver nodules. Our dataset performs segmentation and classification of common liver lesions on hepatobiliary MRI stage images, which can assist imaging doctors to diagnose liver diseases. Medical images of digital imaging and communications in medicine (DICOM) formate were shown. All of the medical slices are single-channel grayscale images, and the range of each pixel value is $[0, 5000]$. The dataset contains a total of 400 labeled cases.

Our dataset contains medical data acquired on different devices and uses different acquisition protocols. The data in this dataset is representative of the clinical variability and challenges encountered in the clinical environment. In order to measure the size of the segmented parts, we compared the constructed SEHPI dataset with two other public datasets, LiTS[26] and BraTS[36]. The LiTS dataset contains 200 cases, and the images are CT scan images. The dataset segmented the liver and liver tumors. The image dimension of each case is $(90, 512, 512)$. The BraTS dataset is a brain tumor segmentation competition dataset. It contains 285 cases. Each case is four sets of multi-modal CT images $(155, 240, 240)$. Table 2 shows the results of the comparison. The rate is the percentage of the number of marked pixels in the total number of pixels in the picture. Since the LiTS dataset needs to segment the enlarged organ—the liver, its label accounts for the largest proportion. The target of the BraTS dataset that needs to be segmented is brain tumors, which are greater than SEHPI but considerably smaller than LiTS. The target to be segmented in the SEHPI dataset is liver nodules. Compared with organ targets and common tumor targets, SEHPI targets are much smaller. Similarly, in the LiTS dataset with individual tumor labels (LiTS Tumor), the proportion of labels is very small, only 0.63%, which is a small-label dataset. Due to the particularity of the target, the traditional medical image segmentation method is easy to lose the target in the process of downsampling, resulting in its low accuracy on SEHPI.

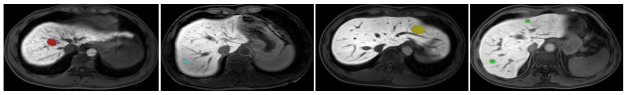


Fig. 3 SEHPI. Label 1 in red represents liver cancer; Label 2 in blue represents metastatic tumor; Label 3 in yellow represents cyst; Label 4 in green represents hemangioma.

4.2 Dataset preprocessing

We used the SimpleITK package to preprocess operations on the original images in order to unify the data format. First, we unify the image dimensions. The size of the images is not consistent, due to the use of different equipment for data collection. Most of the image sizes are less than or equal to 512 pixel×512 pixel. So, we adjust the size of all images to 512 pixel × 512 pixel by using a bilinear interpolation sampling strategy. Meanwhile, most cases consist of 90 slices. We use the z-axis linear interpolation strategy, in order to change the slice spacing. We change each case slices to 90. In all sampling processes, the adjacent interpolation method is adopted for the label position to ensure the consistency of the label position as much as possible. As a consequence, the image size of each case are unified as $(90, 512, 512)$.

Next, in order to shield the noise in the image, we

Table 2 The proportion of labels in the dataset.

(%)

	SEHPI	LiTS Tumor	LiTS	BraTS
Rate	0.5290	0.6327	11.1641	2.1462

used the quantiles of the two pixel values of 0.99 and 0.01 as the upper and lower bounds to correct the image.

$$X'(i,j)=\begin{cases}\mathcal{P}_{0.01}(X), & X(i,j)\in(-\infty,\mathcal{P}_{0.01}(X))\\ X(i,j), & X(i,j)\in[\mathcal{P}_{0.01}(X),\mathcal{P}_{0.99}(X)]\\ \mathcal{P}_{0.99}(X), & X(i,j)\in(\mathcal{P}_{0.99}(X),+\infty)\end{cases} \tag{15}$$

X' is the corrected image, $\mathcal{P}_{0.01}(X)$ and $\mathcal{P}_{0.99}(X)$ are the quantiles of pixels on the original image X that are 0.01 and 0.99. Since the intensity of each pixel ranges between 0 and 5000, the minimum value in the normalized slice corresponds to the 0 intensity in the unnormalized slice. In order to highlight the 0 intensity in the background, the background is replaced with –9 to track the background with 0 intensity pixels when sampling random slices. Such correction can get the optimal dataset of the article with better accuracy.

5 Experiment

5.1 Evaluation index

For input X, suppose its corresponding label is Y, and the model output Y'. The number of channels of Y and Y' are both q (the number of tag categories). We use the following indicators to evaluate the effect of the model.

5.1.1 Dice

For the evaluation criteria in the segmentation process, the dice similariy coefficient (DSC) is mainly used. It is a measure of ensemble similarity, which is usually used to calculate the similarity of two samples.

$$Dice=\sum_{k=1}^{q}\frac{2Y[k]\times Y'[k]+\delta}{Y[k]+Y'[k]+\delta} \tag{16}$$

where $\delta=1\times10^{-5}$.

5.1.2 Positive predicted value

Positive predicted value (PPV) reflects the ability of the classifier or model to correctly predict the accuracy of positive samples, that is, how many of the predicted positive samples are real positive samples. The higher value, the better performance.

$$PPV=\sum_{k=1}^{q}\frac{2Y[k]\times Y'[k]+\delta}{Y'[k]+\delta} \tag{17}$$

5.1.3 Sensitivity

Sensitivity reflects the ability of the classifier or model to correctly predict the full degree of positive samples, and increases the prediction of positive samples as positive samples. It relects the proportion of positive samples predicted as positive samples in the total positive samples. The higher value, the better performance.

$$Sensitivity=\sum_{k=1}^{q}\frac{2Y[k]\times Y'[k]+\delta}{Y[k]+\delta} \tag{18}$$

5.1.4 Hausdorff 95

The DSC is more sensitive to the filling inside the mask, and the Hausdorff distance is more sensitive to the boundary of the segmentation. B and B' are the masks of ground truth and the model's output. $\|\cdot\|_2$ represents the L_2 distance between b and b'.

$$dist(B,B')=\max_{b\in B}\min_{b'\in B'}\|b-b'\|_2 \tag{19}$$

$$d_{\mathrm{H}}(B,B')=\max\{dist(B,B'),dist(B',B)\} \tag{20}$$

We use Hausdorff 95 to eliminate the influence of a tiny subset of outliers.

$$d_{95\%\mathrm{HD}}=d_{\mathrm{H}}\times95\% \tag{21}$$

5.2 SElayer ablation experiment

In order to better verify that SElayer can really pay attention to a specific channel during model training, we visualize the feature maps generated by different slice SElayer after the first decoder and the corresponding weight values, as shown in Fig. 4. The two slices on the left, with a smaller weight, do not contain the tumor, and the three slices on the right contain the visualization bar below the tumor (in the red box). The weights of these five slices are respectively 0.12, 0.23, 0.92, 0.74, 0.69. It can be found that the weight of the slice containing the tumor is greatly higher than that of the slice not containing the tumor. Also, the weight of the middle slice is the highest.

At the same time, we visualize the feature maps of tumor slices in different models, as shown in Fig. 5. The above picture is the output of the original UNet++, and under layer is the output of SUNet++ at the same node. It can be clearly observed that the spot of the tumor cannot be seen in the first three pictures of UNet++; SUNet++

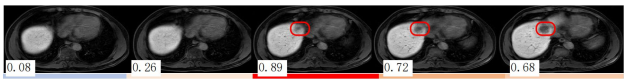


Fig. 4 Continuous slices weight visualization.

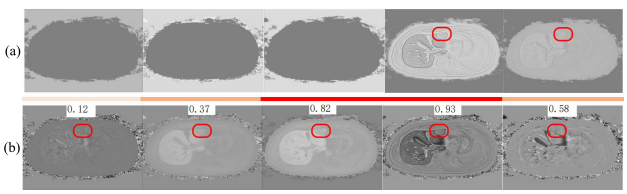


Fig. 5 Feature maps of the same node generated by (a) UNet++ and (b) SUNet++.

highlights the tumor and has better expressiveness. The boundary of the tumor is most clear. After SElayer, slices with higher weights retain more comprehensive information, which can retain more image information. Therefore, SUNet++ is conducive to the generation of small-scale label masks.

5.3 Hyperparameter selection

The reduction rate r introduced in SElayer is an important hyperparameter. It allows changing the capacity and calculation cost of the SELayer in the model. In order to find this relationship, we conducted a series of experiments with different values of r. In the experiment, the dice value is chosen to use the average of the four label dice values. The comparison in Table 3 shows that performance does not increase monotonically with the increase in capacity. This may be the result of SElayer being able to overfit the channel dependence of the training set. In particular, it was found that setting $r = 16$ achieved a good balance between accuracy, and used in code.

When choosing the number of slices to be superimposed p, we compare the influence of different layers. The dice value of the results in the experiment uses the average value of the four label dice, the experimental results are shown in Table 3. It can be found that the model training effect is not positively correlated with the number of superimposed layers, because the introduction of more useless slices. When the number of slices is equal to the number of general slices of the case, the input of the model degenerates to the input of 3DUNet. At the same time, the more superimposed layers, the more model parameters. After comparing the model training effects, we choose 5 layers.

5.4 Training

In SUNet++, the output of the last sigmod layer is composed of four foregrounds probability maps. Foreground voxels which have higher probability (> 0.5) than those belonging to the background, are considered to be part of the segmentation. In the training process, the method of bilateral optimization is adopted, 40 of the 400 examples are selected as the test set, and 40% of the training set is selected as the value set. Conditions for the model to stop training: when the best dice result in the val_set does not improve after 20 epochs, the model stops training. Because value does not participate in training, it can ensure that the training achieves the best results without overfitting. The batch size is 10. The momentum is 0.9. The initial learning rate is 0.0003, and the final model parameters are 69.39 MB.

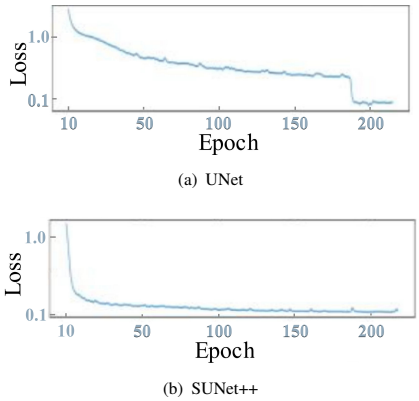
Table 3 Hyperparameter selection result.

Hyperparameter		Dice
r	4	0.8895
	8	0.8953
	16	**0.9251**
	32	0.9107
p	1	0.8218
	3	0.8832
	5	**0.9107**
	7	0.8983

As mentioned earlier, we suppress useless information and improve the convergence effect by deleting the unlabeled slices. Figure 6 compares the training process of the ordinary UNet model without removing the slices from the data and the training process of SUNet++. The dataset is LiTS Tumor. It can be observed that the convergence process of the SUNet++ model is faster and stable. The UNet model has the problem of gradient explosion after 150 rounds. This may be due to the fact that the small-scale label dataset has less information and more negative samples. At the same time, there are too many hidden layers in the ordinary UNet model when training small-scale tags, which leads to gradient explosion.

5.5 Result

For comparison, we used the original UNet[2], UNet++[7], as well as the nnUNet[37] and UNet3plus[21] architecture. These models can retain the details of each scale or capture fine-grained details from the full scale. All four models have the ability to retain small-scale labels. The experimental results are shown in Table 4. It can be seen that the four labels of SUNet++'s dice index increased by an average of 2.73%, the PPV index increased by an average of 1.71%, the sensitivity index increased by an average of 0.21%, and the Hausdorff 95 index increased



(a) UNet

(b) SUNet++

Fig. 6 Training on the LiTS Tumor dataset.

Table 4 SEHPI dataset experimental results.

Model	Dice				Sensitivity			
	Label 1	Label 2	Label 3	Label 4	Label 1	Label 2	Label 3	Label 4
UNet	0.8605	0.8761	0.9168	0.9207	**0.9783**	0.9821	0.9903	0.9671
UNet++	0.8684	0.8965	0.8576	0.9156	0.9506	**0.9834**	0.9791	0.9673
UNet3plus	0.8619	0.8965	0.9134	0.8423	0.9733	0.9776	0.9876	0.9627
nnUNet	0.8912	0.8931	0.8729	0.8933	0.9714	0.9812	0.9890	0.9693
SUNet++	**0.8987**	**0.9152**	**0.9422**	**0.9446**	0.9757	0.978	**0.9912**	**0.9707**
Model	PPV				Hausdorff 95			
	Label 1	Label 2	Label 3	Label 4	Label 1	Label 2	Label 3	Label 4
UNet	0.8756	0.8886	0.9228	0.9441	0.6218	0.4616	0.2589	**0.2081**
UNet++	0.9104	0.929	0.8754	0.9468	0.5916	**0.3113**	0.4470	0.3167
UNet3plus	0.8828	0.9154	0.9227	0.8614	0.5696	0.3645	0.2640	0.5364
nnUNet	0.8911	0.9233	0.9391	**0.9741**	**0.4236**	0.3321	0.2069	0.2091
SUNet++	**0.9181**	**0.9321**	**0.9492**	0.9732	0.4683	0.3323	**0.1618**	0.2919

by 42.39%. The resulting visualization is shown in Fig. 7. After comparing multiple and single tumor cases, we can find that SUNet++ can accurately locate multiple tumors and classify them more accurately than other models. At the same time, it can get a more exact boundary.

Moreover, we compare the performance of these models on LiTS Tumor dataset. This dataset contains 200 cases. Also, in order to obtain the experimental results, we eliminate the pointless slices in the dataset to ensure the regular convergence of the model. In the model implementation, the batch size, momentum and the initial learning rate are equal to the last experiment. The same method of bilateral optimization is used to train the model. We compared the four models in the LiTS Tumor dataset (with the liver label removed). In Table 5 the results are as follows: compared with the second best performing model nnUNet, dice increased by 4.96%, PPV increased by 1.21%, sensitivity increased by 1.58%, and Hausdorff 95 decreased by 10.08%.

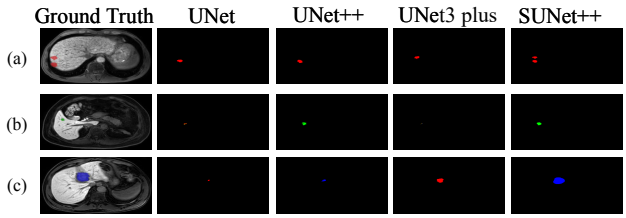


Fig. 7 Visualization of experimental results.

Table 5 LiTS Tumor dataset experimental results.

Model	Dice	PPV	Sensitivity	Hausdorff 95
UNet	0.3813	0.5984	0.6300	4.9384
UNet++	0.3992	0.5724	0.7233	3.9812
UNet3plus	0.4013	0.5432	0.8092	4.3324
nnUNet	0.4501	0.6012	0.8201	3.1742
SUNet++	**0.4724**	**0.6067**	**0.8214**	**2.8541**

6 Conclusion

In order to solve the problem of medical image segmentation with small-scale labels, we propose SUNet++. The architecture takes advantage of a redesigned encoder units. The redesigned encoder units combine the residual structure and the layer attention mechanism in order to focus on the importance of different slices. At the same time, a larger small-scale labeled liver segmentation dataset is proposed to provide better experimental objects for liver segmentation. Experiments show that SUNet++ has outstanding performance in small-scale label segmentation.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (No. 62072135), Natural Science Foundation of Ningxia Hui Autonomous Region (No. 2022AAC03346), and Fundamental Research Funds for the Central Universities (No. 3072020CF0602).

References

[1] J. Long, E. Shelhamer, and T. Darrell, Fully convolutional networks for semantic segmentation. in *Proc. 2015 IEEE Conf.on Computer Vision and Pattern Recognition*, Boston, MA, USA, 2015, pp. 3431–3440, 2015.

[2] O. Ronneberger, P. Fischer, and T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in *Proc. 18th Int. Conf. on Medical Image Computing and Computer-Assisted Intervention*, Munich, Germany, 2015, pp. 234–241.

[3] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, Semantic image segmentation with deep convolutional nets and fully connected CRFs, presented at the 3rd Int. Conf. on Learning Representations, San Diego, CA, USA, 2015.

[4] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, Pyramid scene

parsing network. in *Proc. 2017 IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 6230–6239.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in *Proc. 2014 IEEE Conf. on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 2014, pp. 580–587, 2014.

[6] S. Ren, K. He, R. Girshick, and J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017.

[7] Z. Zhou, M. R. Siddiquee, N. Tajbakhsh, and J. Liang, UNet++: A nested U-net architecture for medical image segmentation, in *Proc. 4^{th} Int. Workshop on Deep Learning in Medical Image Analysis*, Granada, Spain, 2018, pp. 3–11.

[8] F. Milletari, N. Navab, and S. A. Ahmadi, V-Net: Fully convolutional neural networks for volumetric medical image segmentation, in *Proc. 4^{th} Int. Conf. on 3D Vision (3DV)*, Stanford, CA, USA, 2016, pp. 565–571.

[9] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, 3D U-Net: Learning dense volumetric segmentation from sparse annotation, in *Proc. 19^{th} Int. Conf. on Medical Image Computing and Computer-Assisted Intervention*, Athens, Greece, 2016, pp. 424–432.

[10] Z. Cai, X. Zheng, and J. Yu, A differential-private framework for urban traffic flows estimation via taxi companies, *IEEE Trans. Industr. Inform.*, vol. 15, no. 12, pp. 6492–6499, 2019.

[11] Y. Liang, Z. Cai, J. Yu, Q. Han, and Y. Li, Deep learning based inference of private information using embedded sensors in smart devices, *IEEE Network*, vol. 32, no. 4, pp. 8–14, 2018.

[12] A. Alansary, K. Kamnitsas, A. Davidson, R. Khlebnikov, M. Rajchl, C. Malamateniou, M. Rutherford, J. V. Hajnal, B. Glocker, D. Rueckert, et al., Fast fully automatic segmentation of the human placenta from motion corrupted MRI, in *Proc. 19^{th} Int. Conf. on Medical Image Computing and Computer-Assisted Intervention*, Athens, Greece, 2016, pp. 589–597.

[13] M. Shakeri, S. Tsogkas, E. Ferrante, S. Lippe, S. Kadoury, N. Paragios, and I. Kokkinos, Sub-cortical brain structure segmentation using F-CNN'S, in *Proc. 13th Int. Sympos. on Biomedical Imaging (ISBI)*, Prague, Czech Republic, 2016, pp. 269–272.

[14] P. Moeskops, J. M. Wolterink, B. H. M. van der Velden, K. G. A. Gilhuijs, T. Leiner, M. A. Viergever, and I. Išgum, Deep learning for multi-task medical image segmentation in multiple modalities, in *Proc. 19^{th} Int. Conf. on Medical Image Computing and Computer-Assisted Intervention*, Athens, Greece, 2016, pp. 478–486.

[15] X. Zhou, T. Ito, R. Takayama, S. Wang, T. Hara, and H. Fujita, Three-dimensional CT image segmentation by combining 2D fully convolutional network with 3D majority voting, in *Proc. 1^{st} Int. Workshop on Deep Learning in Medical Image Analysis*, Athens, Greece, 2016, pp. 111–120.

[16] R. Korez, B. Likar, F. Pernuš, and T. Vrtovec, Model-based segmentation of vertebral bodies from MR images with 3D CNNs, in *Proc. 19^{th} Int. Conf. on Medical Image Computing and Computer-Assisted Intervention*, Athens, Greece, 2016, pp. 433–441.

[17] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, The importance of skip connections in biomedical image segmentation, in *Proc. 1^{st} Int. Workshop on Deep Learning in Medical Image Analysis*, Athens, Greece, 2016, pp. 179–187.

[18] K. Kamnitsas, C. Ledig, V. F. J. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation, *Med. Image Anal.*, vol. 36, pp. 61–78, 2017.

[19] M. Ghafoorian, N. Karssemeijer, T. Heskes, I. W. M. van Uder, F. E. de Leeuw, E. Marchiori, B. van Ginneken, and B. Platel, Non-uniform patch sampling with deep convolutional neural networks for white matter hyperintensity segmentation. in *Proc. 13^{th} Int. Sympos. on Biomedical Imaging*, Prague, Czech Republic, 2016, pp. 1414–1417.

[20] T. Brosch, L. Y. W. Tang, Y. Yoo, D. K. B. Li, A. Traboulsee, and R. Tam, Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation, *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1229–1239, 2016.

[21] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y. W. Chen, and J. Wu, UNet 3+: A full-scale connected UNet for medical image segmentation, in *Proc. 2020 IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Barcelona, Spain, 2020, pp. 1055–1059.

[22] A. E. Kavur, N. S. Gezer, M. Barış S. Aslan, P. H. Conze, V. Groza, D. D. Pham, S. Chatterjee, P. Ernst, S. Özkan, et al., CHAOS challenge-combined (CT-MR) healthy abdominal organ segmentation, *Med. Image Anal.*, vol. 69, p. 101950, 2021.

[23] A. Andreopoulos and J. K. Tsotsos, Efficient and generalizable statistical models of shape and appearance for analysis of cardiac MRI, *Med. Image Anal.*, vol. 12, no. 3, pp. 335–357, 2008.

[24] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P. A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester, et al., Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Trans. Med. Imaging*, vol. 37, no. 11, pp. 2514–2525, 2018.

[25] G. Litjens, R. Toth, W. van de Ven, C. Hoeks, S. Kerkstra, B. van Ginneken, G. Vincent, G. Guillard, N. Birbeck, J. Zhang, et al., Evaluation of prostate segmentation algorithms for MRI: The PROMISE12 challenge, *Med. Image Anal.*, vol. 18, no. 2, pp. 359–373, 2014.

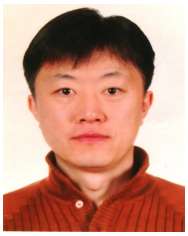
[26] P. Christ, LiTS-liver tumor segmentation challenge (LiTS17), https://academictorrents.com/details/27772adef6f563a1ecc0ae19a528b956e6c803ce, 2017.

[27] Z. Cai, Z. He, X. Guan, and Y. Li, Collective data-sanitization for preventing sensitive information inference attacks in social networks, *IEEE Trans. Dependable Secure Comput.*, vol. 15, no. 4, pp. 577–590, 2018.

[28] Z. Cai and X. Zheng, A private and efficient mechanism for data uploading in smart cyber-physical systems, *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 2, pp. 766–775, 2020.
[29] X. Zheng and Z. Cai, Privacy-preserved data sharing towards multiple parties in industrial IoTs. *IEEE J. Sel. Areas Commun.*, vol. 38, no. 5, pp. 968–979, 2020.
[30] Z. Cai, Z. Xiong, H. Xu, P. Wang, W. Li, and Y. Pan, Generative adversarial networks: A survey toward private and secure applications, *ACM Comput. Surv.*, vol. 54, no. 6, p. 132, 2022.
[31] K. Simanyan and A. Zisserman, Very deep convolutional network for large-scale image recognition, arXiv preprint arXiv: 1409.1556, 2015.
[32] J. Wang, Z. Cai, and J. Yu, Achieving personalized k-anonymity-based content privacy for autonomous vehicles in CPS, *IEEE Trans. Industr. Inform.*, vol. 16, no. 6, pp. 4242–4251, 2020.
[33] X. Zheng, Z. Cai, J. Li, and H. Gao, Location-privacy-aware review publication mechanism for local business service systems, in *Proc. 2017 IEEE Conf. on Computer Communications*, Atlanta, GA, USA, 2017, pp. 1–9.
[34] X. Zheng, Z. Cai, J. Yu, C. Wang, and Y. Li, Follow but no track: Privacy preserved profile publishing in cyber-physical social systems, *IEEE Internet Things J.*, vol. 4, no. 6, pp. 1868–1878, 2017.
[35] X. Zheng, Z. Cai, and Y. Li, Data linkage in smart internet of things systems: A consideration from a privacy perspective, *IEEE Commun. Mag.*, vol. 56, no. 9, pp. 55–61, 2018.
[36] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al., The multimodal brain tumor image segmentation benchmark (BRATS), *IEEE Trans. Med. Imaging*, vol. 34, no. 10, pp. 1993–2024, 2015.
[37] F. Isersee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S, Kohl, J. Wasserthal, G. Koehler, T. Norajrtra, S. Wirkert, et al., nnU-Net: Self-adapting framework for U-Net-based modical image segmantation, arXiv preprint arXiv: 1809.10486, 2018.

Lan Zhang received the BS degree from Harbin Engineering University in 2019, and is pursuing the PhD degree at Harbin Engineering University. Here research interests mainly include deep learning, computer vision, and medical and health big data.

Kejia Zhang received the BS degree from Beijing Normal University in 2004 and the PhD degree from Harbin Institute of Technology in 2012. He is currently an associate professor at Harbin University of Engineering. He has published a number of SCI-indexed papers as the first author, and has presided over one national research project. His research interests mainly include wireless sensor network, Internet of Things, privacy protection, deep learning, and computer vision.



Haiwei Pan is currently a professor in the Department of Computer Science and Technology at Harbin Engineering University. He received the PhD degree from Harbin University of Technology in 2006. He worked as a post-doctoral fellow at Harbin University of Engineering in 2009. In 2014, he was a visiting scholar at City University of Hong Kong. His research interests include big data analysis, artificial intelligence, and smart medical care.