IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 November 2022, accepted 20 November 2022, date of publication 28 November 2022, date of current version 2 February 2023.

Digital Object Identifier 10.1109/ACCESS.2022.3225101

RESEARCH ARTICLE

Medical Ultrasound Image Segmentation With Deep Learning Models

CHUANTAO WANG, JINHUA ZHANG, AND SIYU LIU
Beijing Engineering Research Center of Monitoring for Construction Safety, School of Mechanical-Electronic and Vehicle Engineering, Beijing University of Civil Engineering and Architecture, Beijing 102616, China

Corresponding author: Chuantao Wang (wangchuantao@bucea.edu.cn)

ABSTRACT This work aimed to adopt a transformer model combined with deep learning neural network to discuss the segmentation of medical ultrasound images. A network combining a transformer model with a deep neural network model (ConvTrans-Net) is proposed. The image content is preselected based on a multilayer perceptron and attention mechanism, different feature vectors are concatenated and fed into the multilayer perceptron, and the results of multiple attentions are mapped to a larger dimensional space using a feed-forward network. The lesion areas segmented by ultrasonic scan were analysed, and an attention mechanism and multilayer perceptron were combined to preselect image content. The performance and convergence of the model were analysed, and the Jaccard similarity coefficient precision and recall of the model were measured. In the experiment, two different iterative step sizes were selected, the convergence trend of the model increased with the increase in the number of iterative steps, and the model gradually stabilized. The Jaccard of ConvTrans-Net was 85.21%. The precision (85.17%) and recall (89.65%) were significantly higher than those of EfficientNet and DeepViT-L, and the differences were significant ($P < 0.05$). The experimental results show that the proposed model is stable, and the combination of Transformer model and deep learning neural network has a good effect on ultrasound image segmentation, which has some practical application value.

INDEX TERMS Deep learning, medical images, ultrasound, neural networks, image features.

I. INTRODUCTION

At present, medical imaging develops continuously and rapidly. Anatomic imaging has developed into functional and molecular imaging and has been integrated into intelligent algorithms [1]. To a great extent, not only is the definition of images improved, but the diagnostic accuracy of imaging is also significantly improved. Many medical imaging methods have emerged, such as X-ray, computed tomography (CT), ultrasonic imaging, and magnetic resonance imaging (MRI). All the above methods are widely applied in medical field [2], [3]. During the diagnosis of medical images, many manpower and material resources are consumed to observe lesion areas and potential lesion areas with the naked eye. In addition, it is time-consuming. The efficient detection and segmentation of medical images and diagnosis of lesion tissues by some automated approaches has become a matter of concern [4]. The use of traditional machines in the segmentation of medical images is usually affected by grayscale inhomogeneity, volume effect, and insignificant differences in grayscale between different soft tissues. Finally, medical image segmentation is nonideal [5].

Deep learning technology shows explosive growth and is widely applied in medical image segmentation. The development of medical technology changes from day to day. Medical diagnostic imaging equipment is computerized. Ultrasonic equipment can efficiently process medical images, carry out corresponding measurements, and intelligently export numbers [6]. Multiple image analysis and measurement methods can meet medical needs, such as the detection of cases. Data images can be used for data compression and image fusion to reduce technicians' burden and improve imaging quality. Data images can be networked. In addition, data image processing and computed-aided diagnosis can also meet medical needs [7], [8]. The iterative development of information intelligence, intelligent equipment in medical ultrasonic systems, intelligent machine translation, and intelligent human-machine interaction also creates a

The associate editor coordinating the review of this manuscript and approving it for publication was Yongjie Li.

good environment for the intelligent segmentation of medical images [9]. It can be assured that there is no image segmentation method suitable for all images at the current stage. Image segmentation is derived from medical equipment. In many systems, some images need to be segmented by users to achieve the desired effect because of the treatment plan [10]. With the iterative development of information intelligence, convolutional neural networks (CNNs) have been widely applied to medical images. Heidari et al. [11] used a deep learning model to identify COVID-19 victims, and excellent effects were achieved. Artificial intelligence (AI) has made great progress in computer-aided diagnostic systems. In medical ultrasonic systems, intelligent equipment, intelligent machine translation, and intelligent human-machine interactions also create good conditions for the intelligent segmentation of medical images. Great progress has been made in the extraction of image features, image segmentation, noise reduction, and artifact removal. In addition, two-dimensional images have been changed to three-dimensional images. CNNs have been used for the automatic segmentation of lesion sites by many scholars. In addition, training samples were input into the network for feature extraction rather than manual extraction. The network automatically adjusted parameters to learn image features, which improved classification precision [12].

Medical images usually have the characteristics of blurred boundary area, complex tissue texture, and low contrast. Early medical segmentation systems cannot perform segmentation well, and the segmentation algorithm for a specific task has become a hot spot in image segmentation. The design of manual features needs to rely on the prior knowledge of physicians, and the generalization ability is relatively weak and cannot be well connected with new tasks [13]. Therefore, traditional image segmentation techniques cannot achieve satisfactory results, and ultrasound images have the advantages of high sensitivity, noninvasiveness, convenient operation, and low cost. Real-time three-dimensional (four-dimensional) ultrasound imaging has become an effective diagnostic tool in imaging systems in the 21st century [14].

When ultrasound imaging detects patient lesions, the smoothness of the organ surface and the location of the lesions can be checked, and real-time image inspection can be performed [15]. Deep learning technology has achieved success in the construction of large-scale image datasets for natural image recognition [16]. The bilinear convolutional neural network (BCNN) is able to classify images very well. DL technology used by Heidari et al. [17] is divided into seven main categories, including long short-term memory (LSTM), self-organizing maps, conventional neural networks (CNNs), generative adversarial networks (GANs), recursive neural networks (RNNs), autocoders, and hybrid system methods. Chest CT/chest X-ray images of COVID-19 patients can be effectively identified by the above methods. CNN-based deep learning algorithms have made breakthroughs in various fields of image processing, such as semantic segmentation and image classification. Research in the field of computer vision includes U-net [18], full convolutional networks (FCNs) [19], and residual neural networks (ResNets) [20]. It has been proposed one after another and has received extensive attention from researchers at home and abroad. The use of deep learning technology to segment image images can make the semantic expression of imaging images clearer and increase the diagnostic efficiency. Yosinski et al. [21] showed that with increasing depth, the CNN network can obtain distinct features when learning the features of the underlying CNN, and the closer to the input, the more specific information the features show. Transformer model is a deep modelling framework mainly based on attention mechanisms. It can better capture long-range semantic features and support the development of operations, deep learning, and natural language production techniques for image description, text translation, dialogue generation, and automatic summarization [22]. Sowdaboina et al. [23] used machine learning to summarize and classify selections from time-series data. Huang et al. [24] proposed MISSFormer for multi-organ and heart segmentation tasks; Li et al. [25] proposed a TFCN network for capturing and propagating semantic features and filtering non-semantic features for the purpose of improving segmentation accuracy. Other scholars combined Transformer model with deep learning to accomplish the image segmentation task. Sun et al. [26] proposed HybridCTrm, a hybrid multimodal segmentation network based on Transformer and CNN, which solves the deficiency of long-range dependency and improves the generalization ability of the model.

In this work, the (Vision Transformer) ViT network model is improved, the self-attentive model is used for image segmentation, the ConvTrans-Net based on the Transformer model combined with a deep neural network is proposed, and the segmentation effect of the model on medical ultrasound images is discussed, which provides a reference value for practical work.

II. EXPERIMENT AND INNOVATIONS

A. INNOVATIONS

1) VISION TRANSFORMER

Transformer model is an attention realization mechanism with the ability of parallel training. Transformer structure shows good results in natural language processing, such as text classification and machine classification. Compared with the traditional neural network transformer structure, the feature extraction ability is stronger, plays a pivotal position in the field of language processing, and has replaced the traditional CNN and cyclic neural network structure. Self-attention architecture-based Transformer is extensible and efficient. It is the preferred model for natural language processing. Inspired by the success of Transformer in the field of natural language processing, Vision Transformer combining natural language processing and computer vision is proposed and applied in images. Images were segmented into small pieces, and the linear insertion sequence of small pieces was used as the input of Transformer model to classify the

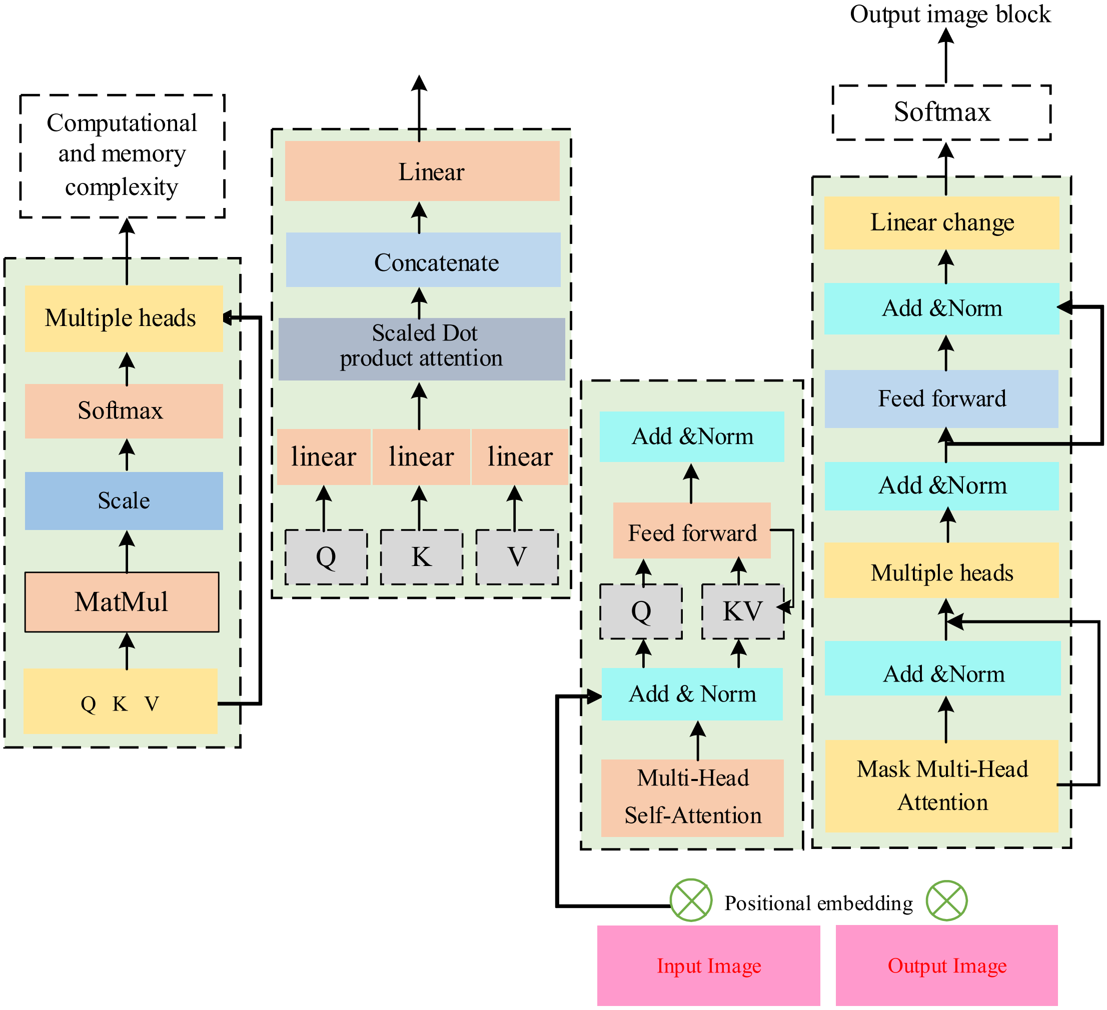


FIGURE 1. Structure of the transformer model.

supervised images. In Figure 1, the overall macro structure includes a decoder structure and an encoder structure. The left side of the figure is the encoder layer (encoder) of the Transformer, which is formed by stacking multiple encoders. The encoder structure is exactly the same, but the parameters are not shared. The encoding layer consists of 6 decoders, and the encoder includes a feedforward network layer and multiple self-attention layers. First, the input of the encoder needs to go through the self-attention layer. When encoding, this layer can see the input word, the encoding layer then considers the context vector, and the in-machine feedforward network layer is passed up layer by layer. The decoder has one more mask multi-head self-attention layer than the encoder. When the text sentence prediction is read, the information of the future moment can't be obtained. At this time, the existence of a special attention layer is needed. The decoder contains the same 6 layers as the encoder layer and one more mask multi-head self-attention layer.

2) FULL-ATTENTION LAYER CODING STRUCTURE

Full-attention and CNN-based coding were divided into two steps. Firstly, Transformer was used to extract the global feature mapping of the input images. Second, CNN was used to extract the local feature mapping of input images and then connected with the global feature mapping extracted by Transformer to effectively capture the global and local feature information about the input images.

Each time in the self-attention layer, the value vector (V), key vector (K), and query vector (Q) are calculated. Under the action of different weights, the embedded vector x obtained MV, MQ, MK, three for matrices of the same size, and the result obtained by QK's dot product was larger. To ensure the stability of the gradient and limit the obtained result, divide it by the dimension of the QK vector, and h_k represents the dimension of the QK vector. After normalization, the weighted score of each input vector was obtained. Equation (1) was expressed as follows:

$$\text{Attention}(Q, K, V) = soft\max(\frac{QK^R}{\sqrt{h_K}})v \tag{1}$$

It was assumed that an image $X \in R^{H \times M \times C}$ with the size of H × M and the number of channels being N was input. During feature extraction by Transformer, the input image should be adjusted to a regular square $X' \in R^{B \times B \times C}$ with a size of B × B. Then, the input image was segmented into multiple square sections to form a two-dimensional section sequence $X_g = \left[\mathrm{x}_g^1, \mathrm{x}_g^2, \mathrm{x}_g^3 \ldots, \mathrm{x}_g^N\right]$,$\mathrm{x}_g^i \in R^{N\times(c\times \mathrm{g}^2)}$. N refers to the number of sections, and x_g^i represents each section with a size of G × G. The linear projection Y mapped Xg into a D-dimensional section insertion sequence and fed $\mathrm{x}_\mathrm{g}Y$ into Transfer for coding. Y $\in R^{N\times(c\times \mathrm{g}^2)\times D}$ denotes the section insertion projection.

The equation of section insertion is shown below (equation (2)).

$$Q_0 = X_g Y = \left[\mathrm{x}_g^1 Y, \mathrm{x}_g^2 Y, \mathrm{x}_g^3 Y \ldots, \mathrm{x}_g^N\right] \tag{2}$$

Transformer encoder contained multiple-layer perception modules and L-layer transportation multi-head self-attention modules. Q_L refers to the output of layer I (equations (3) and (4)).

$$Q_1' = MSA\left(LN\left(Q_{l-1}\right)\right) + Q_{l-1} \tag{3}$$

$$Q_1 = MLP\left(LN\left(Q_l'\right)\right) + Q_l' \tag{4}$$

I, LN(·), and Q_l' represent the number of layers, layer normalization, and the output of MSA in the I$^{\mathrm{th}}$ layer, respectively. Finally, $\mathrm{Q}_\mathrm{L} \in R^{N\times G\times G}$ was output through Transformer encoder.

Transformer was good at capturing global information, and the CNN could capture local information. The combination of Transformer and CNN can obtain more comprehensive image features. Multi-head attention, as the name suggested, referred to the segmentation of the model into multiple heads to form multiple subspaces. Transformer was endowed with a powerful structure by multi-head attention. The independent head could focus on global and local information to enhance more comprehensive image features. In addition, it is combined with input information to form a multichannel mechanism similar to CNN for feature extraction.

3) FEEDFORWARD NETWORKS AND POSITIONAL CODING

The function of the feed-forward network was to map the results of multi-head attention into a larger dimensional space. The feed-forward network included a linear activation function and a Relu activation function. Equation (5) was as follows:

$$\mathrm{FNN(G)} = \max(0, GM_1 + p_1)M_2 + p_2 \tag{5}$$

Words are an important part of the grammatical structure of sentences, and their arrangement order and importance in sentences play a crucial role in semantic expression.

The position encoding mentioned in the transformer model was the relative position encoding. The reference point of the encoding was the position of the input sequence, and the encoding was the distance to this position. The equation expressions (6) and (7) are as follows:

$$\mathrm{PE}_{(\mathrm{pos},2\mathrm{i})} = \sin(\mathrm{pos}/1000^{2i/d \bmod el}). \tag{6}$$

$$\mathrm{PE}_{(\mathrm{pos},2\mathrm{i}+1)} = \cos(\mathrm{pos}/1000^{2i/d \bmod el}) \tag{7}$$

pos, *d*, *2i*, and *2i+1* represent the position of the feature, the dimension of the position code, the dimension of the even integer, and the dimension of the odd integer, respectively ($2i \leq d, 2i \leq d$).

The correlation between sequence data was recorded by position code. Transformer directly inputs data and stores the positional relationship between data to greatly increase the computational speed and reduce the storage space.

The work flow of Transformer was as follows.

Step 1: The input matrix was input into the vector by word insertion algorithm. After that, position code was employed to obtain the position vector of words. The vector was summed with the position vector to obtain the model input.

Step 2: The vector matrix in Step 1 was transmitted into the encoder. It entered the feedforward neural network through multi-head attention layer. Then, the output was transmitted upwards to the next encoder.

Step 3: After passing through 6 encoders, an information matrix was formed and transmitted to 6 decoders. The matrix passes through multi-head attention layer, multi-head = attention layer, and feedforward connection layer in each encoder.

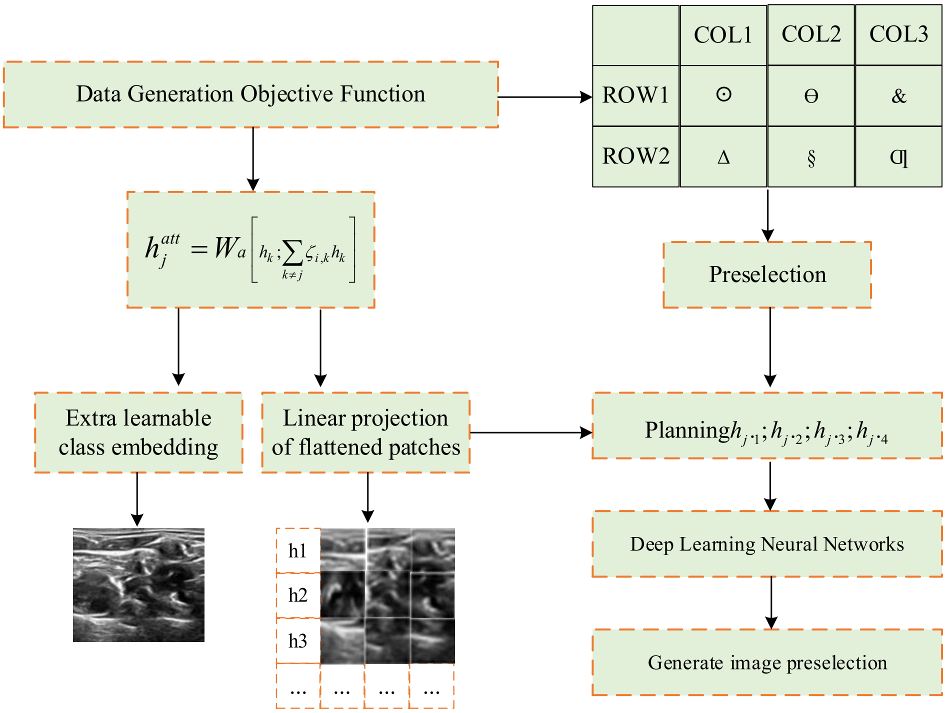
Step 4: The output of the decoder passes through the linear layer and softmax layer. After that, it was transformed into probability as the final output.

4) INFORMATION EXTRACTION METHOD

The flow chart of the information extraction method combined with the Transformer model and the deep learning network is shown in Figure 2. It mainly included the preselection of feature content and then through the deep learning network to realize the extraction of image information. In the figure, it was found that the key "words" can be extracted from the whole sentence text, and a similar network model was used in medical images. In the same way, the lesion location was extracted from the entire ultrasound image, and the region of interest was finally obtained after sorting.

ViT turned an H × W × C image into flattened 2D patches, which can be viewed as a sequence of flattened 2D blocks of size N ×(P^2 C). The size of the original image was (H, W), the channel was C, and the number of image blocks after the change was N (N = HW/P^2). The size of each image block was (P,P) channel or C, and the image blocks used by ViT were not pixels. Because the local information of the image block was more abundant, regardless of whether the image block obtained by using the CNN network or other network structures exceeded the information of a single pixel block, the Transformer was more complex to process long sequences. In the case of the original size of the image in CV, using the image patch was much smaller than using the pixel N.

Class was also a feature that needed to be used in the final classification layer. Different from the original Transformer, the position encoding Epos was not a fixed vector but a learnable vector. The forward process of the encoder in ViT was the same as that of the encoder in the Transformer.



	COL1	COL2	COL3
ROW1	⊙	Θ	&
ROW2	Δ	§	ɑ|

Data Generation Objective Function
$h_j^{att} = W_a \left[h_k ; \sum_{k \neq j} \zeta_{i,k} h_k \right]$
Preselection
Extra learnable class embedding
Linear projection of flattened patches
Planning$h_{j\cdot 1}; h_{j\cdot 2}; h_{j\cdot 3}; h_{j\cdot 4}$
Deep Learning Neural Networks
Generate image preselection
h1
h2
h3
…
…
…
…

FIGURE 2. Information extracted by transformer and CNN models.

When training ViT on big data, fine-tuning to downstream task datasets. It was better to use a larger image resolution than the pre-trained image during external fine-tuning. When the image resolution was larger, the total length of the sequence was longer for the same image block size. ViT can handle sequences of arbitrary length and can perform 2D interpolation on the trained positional codes.

5) IMAGE CONTENT PRE-SELECTION

For the content in the image, the combination of the attention mechanism and multilayer perceptron was used in advance. The attention mechanism has been widely used in the field of visual images. Different feature vectors were spliced and input into the multilayer perceptron, and the feature vector formula of each record was expressed as shown in equation (8).

$$H_{\mathrm{j}} = f_{\mathrm{h}} \left(W_h \left[h_{j\cdot 1}; h_{j\cdot 2}; h_{j\cdot 3}; h_{j\cdot 4} \right] \right) + b_h \tag{8}$$

H_{j} was the implementation of the record encoding function, W_h, b_h were the parameter matrix, [;] represented the splicing vector, $h_{j\cdot 1}$ represented 4 different vector features, and f_{h} represented the linear rectification function. The activation function can output nonlinear results after linear changes. The association information between different records can record important information to determine the arrangement order of different information.

The attention score of each input was $\zeta_{\mathrm{j.k}}$, and an attention vector h_j^{att} was generated away from this score. Equations (9) and (10) were expressed as follows:

$$\zeta_{\mathrm{j.k}} \propto \exp \left(h_j^T W_a h_k \right) \tag{9}$$

$$h_j^{att} = W_a \left[h_k ; \sum_{k \neq j} \zeta_{j.k} h_k \right] \tag{10}$$

The sigmoid activation function was undertaken as the activation function of the neuron, and each new feature vector h_j^{cp} can be obtained as follows in equation (11):

$$h_j^{cp} = f_{si}(h) \bullet h_j \tag{11}$$

f_{si} is the activation function of the sigmoid neuron; $\bullet$ is the product of the elements in the corresponding feature vector, sigmoid(h_j^{cp}) $\in$[0,1] n, and Sigmoid neuron activation was achieved by equation (12):

$$f_{si}\,(\mathrm{x}) = \frac{1}{1 + \mathrm{e}^{-\mathrm{x}}} \tag{12}$$

The loss function in Transformer measures the inconsistency between the monthly collateral of the model and the true value. Usually, $L(Y, f(x))$ is used to indicate that the smaller the loss function is, the better the robustness of the model (equation (13)).

$$\omega * = arg \min_{\omega} \frac{1}{N} \sum_{i=1}^{N} L(yi, f(xi; \omega)) + \lambda \phi(\omega) \tag{13}$$

The former term in the equation is the loss term, and the latter term is the regular term, which prevents overfitting.

The gradient is not smooth at the point, and it is easy to skip the minimum point (equation (14)).

$$MSE = \frac{1}{\mathrm{n}} \sum_{1}^{n} (yi - \dot{y}i)^2 \tag{14}$$

Y is a definite value, taking +1 or −1, and $f(x)$ is a categorical value.

When $f(x)$ is greater than 1 or less than 1, the classification result is determined by the classifier, and the loss function loss is 0. When $f(x) \in(-1,1)$, the classifier is uncertain about the classification result, and the loss is not 0.

B. EXPERIMENT

The environment construction uses the deep learning Pytorch framework. The Pytorch network has high flexibility and supports dynamic graphs. python scripting language, easy to extend. 3D convolution is supported. If a network has many layers and each layer has a similar structure, it can write code to generate these layers in a loop, which is more concise. The experimental environment of the model implementation was given as follows. Ubuntu 16.04 was undertaken as the operating system, Pytorch = 1.4 for the deep learning framework, Python 3.7 was for the language, 128 G for the memory, E5-2680 v4 for the CPU, and Tesla K80 for the GPU model. The core frequency was 1784 Hz, and the memory frequency was 8058 MHz.

In terms of the experimental setting, the number of attention layers, the number of parallel layers of Transformer multi-layer attention, the output dimension d_{model}, the vector dimension of Q and K (d_k), the model optimization training epoch, the initial learning rate, and the batch were set as 64, 10, 344, 45, 25, 0.15, and 10, respectively.

1) IMAGE SEGMENTATION PROCESS

In the constructed transformer model, the Kaggle public dataset was used for network training. Kaggle Open Dataset was a popular data science competition platform that trains built network models with real data. Variables in the dataset included tumor size, tumor grade, and degree of tumor impact on lymph nodes. The ultrasound images of patients were selected in the dataset, a gold standard image obtained by segmentation was manually selected from the selected sample images, and the selected image formats were all nii files. The two-layer ultrasound image was obtained after the 3D image layered extraction. A total of 11,145 2D ultrasound images were extracted from the selected patient ultrasound images, the lesion images were screened out, and 9,574 images without lungs were obtained. After the final screening, a total of 7636 images in the training set were obtained, and then 5636 images were selected as the validation set, with a size of 512 * 512. From Figure 3, it can be observed that the tissue of the lesion was uneven, the grayscale density was low, and there was no obvious grayscale difference. After denoising the original data, the image was clearly visible, and the segmentation result was obtained. After the model was constructed and processed after segmentation, the edge features of the lesion were obtained.

A reasonable evaluation of the performance indicators can effectively evaluate the performance of the algorithm, and the ultrasound image cross-section was evaluated as a binary classification. The model prediction category and the real category were divided into true negative (TN), false positive (FP), true positive (TP), and false negative (FN).

The equation for calculating the accuracy rate was shown in equation (15). The higher the classification accuracy was, the better the performance of the algorithm. Precision refers to the proportion of true positive samples among all predicted positive samples, and recall refers to the proportion of positive samples predicted as positive samples for reagents. When the recall rate and precision rate (equations (16) and (17)) were both high, the harmonic average was higher. If one of them was low, the low and average will be pulled down, and its value will be close to the low number. The Jaccard index was shown in Equation (18).

$$Accuracy = \frac{TP + FN}{TP + FP + TN + FN} \tag{15}$$

$$Precision = \frac{TP}{TP + FP} \tag{16}$$

$$Recall = \frac{TP}{TP + FN} \tag{17}$$

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{18}$$

III. RESULTS

A. INNOVATIONS

There were 3,716 training images selected in the experiment. The graph softmax converged in 20 iteration steps. T1, T2, T3, T4, T5, and T6 represent the step sizes of different iterations, and the values were T1 = 50, T2 = 100, T3 = 150, T4 = 200, T5 = 250, and T6 = 300. The vertical axis represents the convergence gap $||x_{t+1}\text{-}x_t||$. As the iteration steps increased, the model gradually tended to be stable. The model convergence test results were illustrated in Figure 4 below.

B. MODEL EVALUATION

The validation set was employed to evaluate the perplexity of the model under the action of the conditional replication mechanism. The model proposed in this work was compared with the deep neural network (ConvTrans-Net) model compared with the NCP model proposed by Puduppully [27] and the WS-2017 model proposed by Wiseman [28]. The comparison performance results were shown in Figure 5. In the validation set, the precision of the ConvTrans-Net, NCP, and WS-2017 models in the validation set was 33.6%, 33.8%, and 28.1%, respectively, and the recall of the ConvTrans-Net, NCP, and WS-2017 models in the validation set was 51.2%, 52.9%, and 35.9%, respectively. The Precision and Recall of the ConvTrans-Net model were significantly higher than those of the other two models.

As shown in Figure 6, the precisions of the ConvTrans-Net, NCP and WS-2017 models in the test set were 34.2%, 34.8%, and 29.5% in the validation set, respectively. The recalls of the ConvTrans-Net, NCP and WS-2017 models in the validation set were 51.2%, 52.9%, and 36.2%, respectively. The Precision and Recall of the ConvTrans-Net model were significantly higher than those of the other two models. The

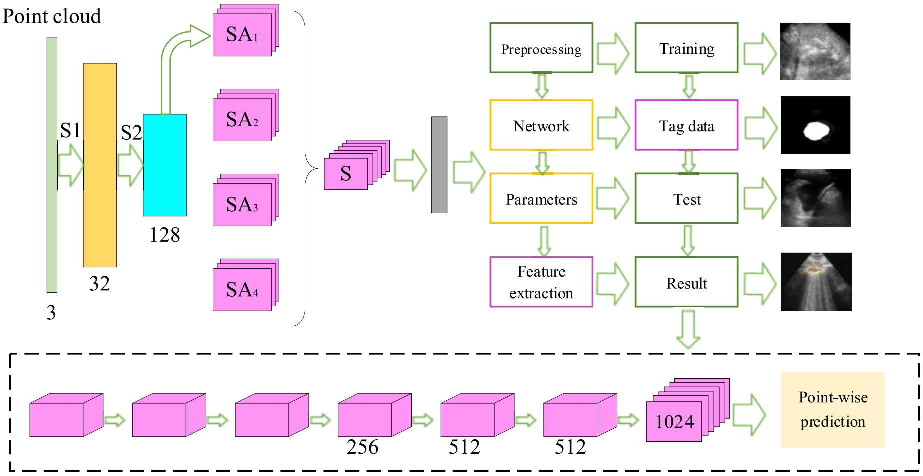


FIGURE 3. Experimental flowchart of image segmentation.

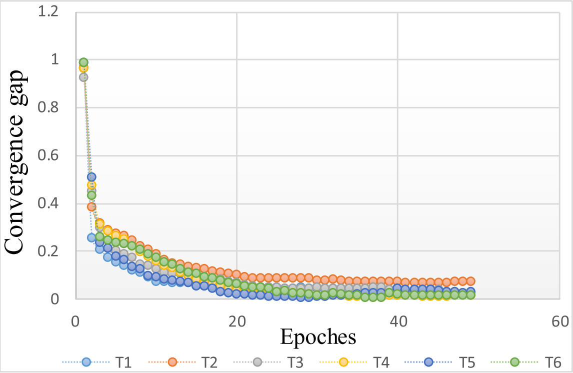


FIGURE 4. Model convergence test results.

intelligent algorithm based on deep learning had high performance and can provide a smoother speed for the segmentation of image features.

Considerable data training was required by image processing with the ViT model. The spatial information about images couldn't be modeled if the image processing method was poor. If the model depth was not sufficient, the number of layers couldn't be enlarged as the CNN did (1×10^6, 3×10^6, 10×10^6, and 100×10^6). ViT model training was analysed by different data sizes. Figure 7 shows that the BiT model had excellent performance in small datasets. With the growth of data size, ViT showed very high accuracy, which revealed a problem that ViT couldn't show excellent results both in large and small datasets.

C. IMAGE SEMENTATION RESULT

The size of the segmented image was 512 × 512 pixels, the red in the figure represents the result of the standard segmentation, and the yellow represents the result of ConvTrans-Net. It was not difficult to find from the figure that the results of the two were similar, indicating that the accuracy of the segmentation was relatively high. Figure 7 showed that the WS-2017 model had no obvious effect on image segmentation, and the segmentation effect of this paper was significantly better

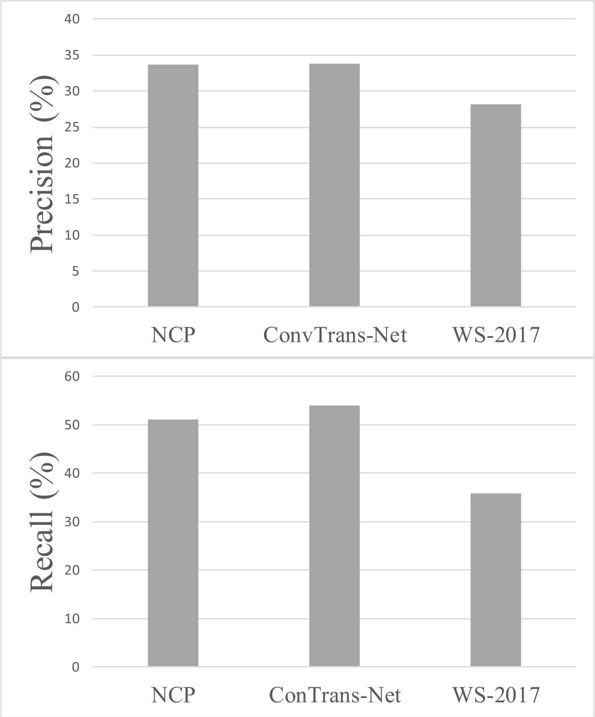


FIGURE 5. Comparison results of different models in the validation set. Figure A was the comparison result of the Precision of the three models; Figure B showed the comparison result of the Recall of the three models.

than that of NCP and WS-2017 models. Figure 8 showed the labelled images of the same target compared with the three models. It was obvious that the label images in this work were better than the NCP and WS-2017 models, and the image imaging was more obvious.

D. COMPARISON OF THE EFFECT OF DIFFERENT SEMENTATION METHODS

ResU-Net and EfficientNet algorithms were based on CNNs. A large amount of data needed to be trained, and a huge amount of data was required to achieve the best result. IGPT, Deep ViT-L, and T2T-ViT image classification algorithms

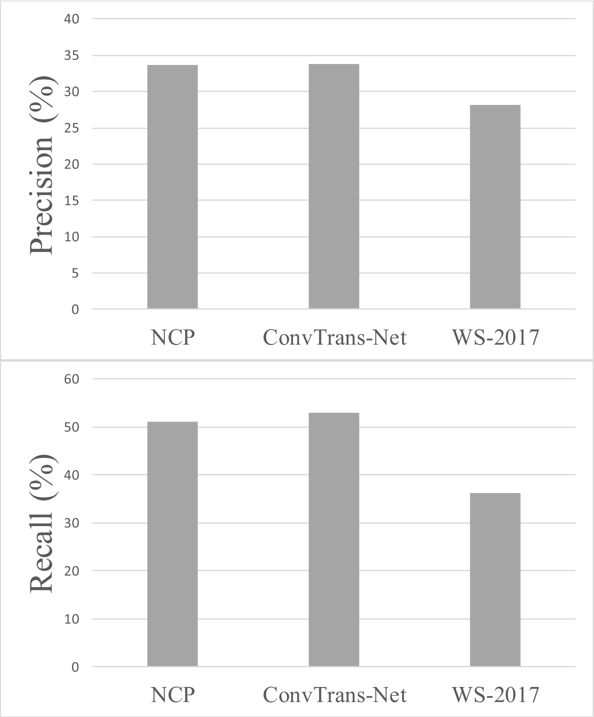


FIGURE 6. Comparison results of different models in the test set. Figure A was the comparison result of the Precision of the three models; Figure B showed the comparison result of the Recall of the three models.

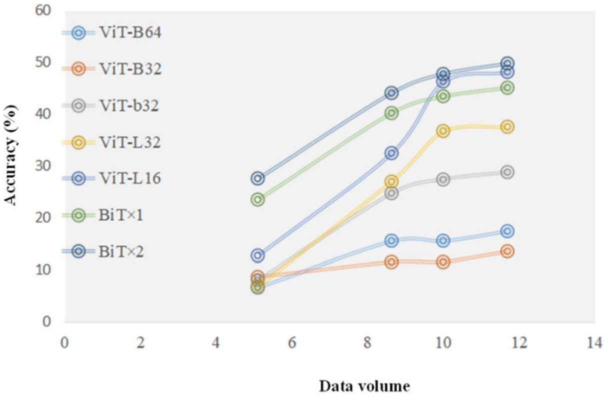


FIGURE 7. Influences of data size on models.

were based on Transformer. As a result, the computational resources needed during network training were dramatically reduced. As presented in Figure 9, the accuracy of Adaboost [29], Random Forests [30], ResU-Net [31], EfficientNet [32], IGPT [33], Deep ViT-L [34], T2T-ViT [35], and ConvTrans-Net amounted to 77.34%, 80.99%, 76.20%, 83.6%, 69%, 82.00%, 80.70%, and 85.17%, respectively (Figure 10). The proposed ultrasonic image segmentation method had very high accuracy.

IV. DISCUSSION

In the field of computer vision, Transformer is applied increasingly frequently. The self-attention mechanism was combined with a common CNN architecture in many vision tasks [36]. However, the shortcomings of Transformer were very apparent. Its inductive bias capability was poor without the transitional invariance and local sensitivity of CNN. In addition, it couldn't generalize massive data. Second, it was unable to process high-resolution feature images and often lost the small targets in images. Third, it usually lost position information during data input. The local image features could hardly be described with a single scale, and local features couldn't be effectively captured by Transformer. In the research, the neighborhood coding insertion method was improved in Transformer model. CNN and Transformer drew on each other's merits to offset their own weakness and integrated with each other to obtain global or local information. Multi-scale neighborhood search was adopted to construct the multiple neighborhoods of the input features of attention coders to describe the features with different spatial scales. The classification network constructed by multiple stacked coders not only focused on a point but also took local and global network features into account. During the down-sampling of neighborhood coding, the input resolution of coders could be effectively reduced [37]. Wang et al. [38] proposed a novel joint deep learning architecture, which consisted of two main parts. A transformer encoder that uses scaled dot product attention to extract dependencies across distances in time series. Combined with convolutional neural networks, built to repair the insensitivity of self-attention mechanisms to local features, new opportunities arise from the addition of supervised data. Che et al. [39] combined a CNN and transformer network to extract temporal information in ECG signals and was able to perform arrhythmia classification with acceptable accuracy. The model can help cardiologists assist in the diagnosis of heart disease and improve the efficiency of medical services. This work adopted a transformer combined with a deep learning network to explore its segmentation of ultrasound images of lung diseases, showing good results. Vision Transformer (VIT) can improve the image feature encoding module and the location feature encoding module. Dai et al. [40] applied it to the analysis task of medical images and obtained a good multi-modal classification task. Dosovitskiy et al. [41] used the VIT model to introduce a self-attention model in the field of image classification, and the picture can be divided into 3 × 3 small pictures. After the image is encoded, it is combined with the position encoding to be input into the self-attention model, which can achieve a good classification effect. Ahmad et al. [42] utilized light-weight deep learning model to segment livers and applied random Gaussian distribution to weight initialization. The method had good performance on each benchmark data. Furqan et al. [43] showed that a patch-based deep belief network model could automatically select features from image blocks and obtain regions of interest (ROIs) from CT images. In addition, they applied unsupervised feature reduction contrast divergence algorithm to weight initialization and optimized weights layer by layer during supervised fine-tuning. It possessed a very high accuracy in vertebral body segmentation.

The Jaccard, precision, and recall of ConvTrans-Net were 85.21%, 85.17%, and 89.65%, respectively, which

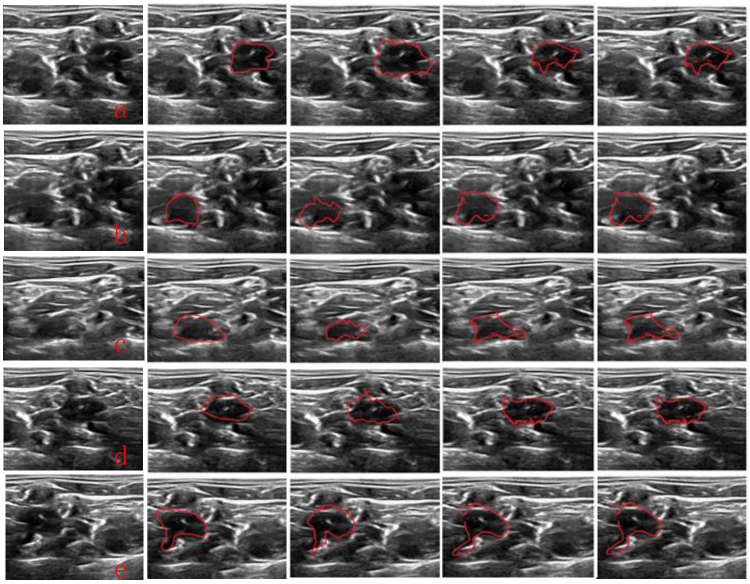


FIGURE 8. Image segmentation results. (a) original image; (b) manual segmented image; (c) image segmented by NCP model, (d) image segmented by WS-2017 model, (e) image segmented by ConvTrans-Net.

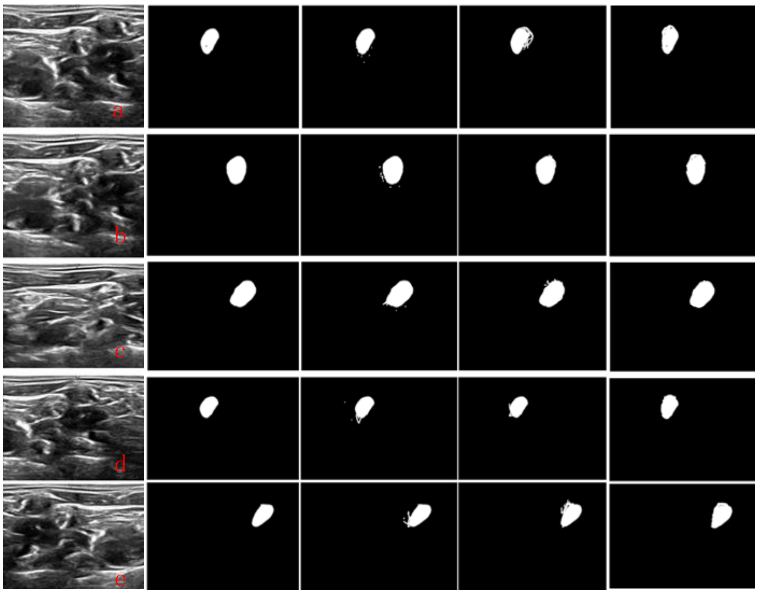


FIGURE 9. Example of target sample segmentation. (a) Original image, (b)Standard labelled image; (c) WS-2017 model; (d) NCP model; (e) Label image of ConvTrans-Net.

were significantly higher than those of EfficientNet and DeepViT-L, and the differences were significant ($P < 0.05$). The convergence performance shows that the built model is stable. Medical image itself was complex, and there were differences in size, location, and shape among different images at different positions. The ultrasonic image features were obtained by combining Transformer model with deep learning. In addition, more space information was preserved. There was no significant difference between the proposed network and U-net. It was possible that the difference in the effect on simple tasks between them was not obvious with a deep network. The segmentation of medical images by Transformer still faces many challenges. Medical image datasets were small, and a certain sample size was needed to fully exploit its advantage in capturing long distances. The attention mechanism of Transformer focused only on image blocks. After image serialization, attention weights were calculated only between image blocks, while the internal information of image blocks was ignored. During the segmentation, identification, or detection of small targets and tasks with blurred boundaries, the key information between pixels affected model precision [44]. Image serialization was necessary for image processing with Transformer. Medical images have high resolution and many pixels. The

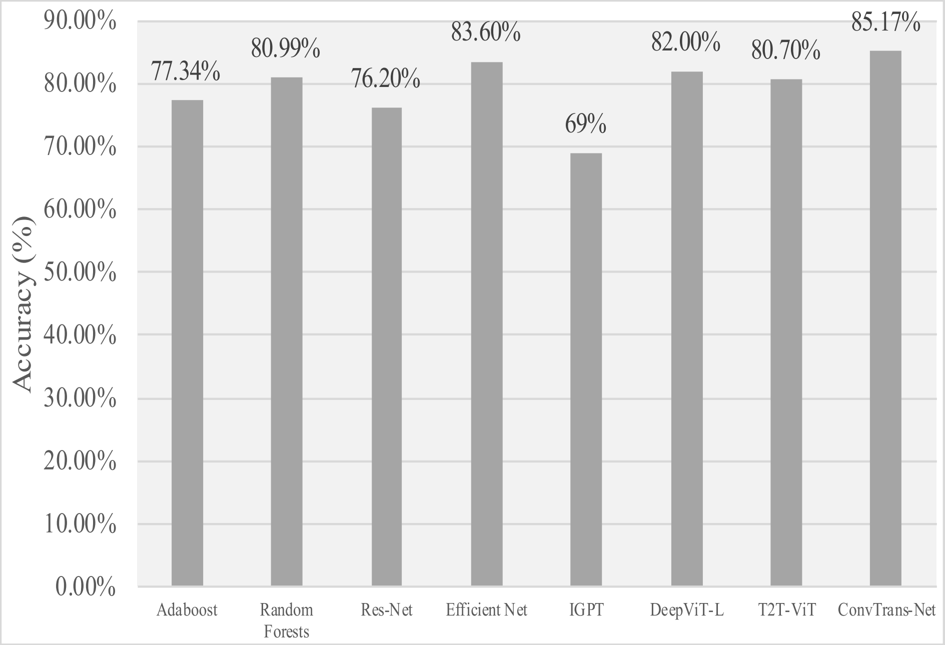


FIGURE 10. Comparison of image segmentation methods.

segmentation of high-resolution medical images still leads to an increased amount of calculation.

V. CONCLUSION

In this work, Transformer model was combined with deep neural network to segment medical ultrasonic images. Validation set and test set were verified. The constructed model had excellent performance, so it could be directly applied in the extraction of ultrasonic image features to generate coherent image information. The limitations of this research lie in the small number of image training sets, which usually cause overfitting, poor generalization ability, and time-consuming network training. The simplification of network structure is also a research trend in the future. It was found that Transformer model had great potential and was still a research focus in the field of computer vision. In the future, there will be a good developmental trend.

REFERENCES

[1] J. C. Gore, "Artificial intelligence in medical imaging," *Magn. Reson. Imag.*, vol. 68, pp. A1–A4, May 2020, doi: 10.1016/j.mri.2019.12.006.

[2] J. H. Scatliff and P. J. Morris, "From Röntgen to magnetic resonance imaging," *North Carolina Med. J.*, vol. 75, no. 2, pp. 111–113, Mar. 2014, doi: 10.18043/ncm.75.2.111.

[3] M. Eichelberg, K. Kleber, and M. Kämmerer, "Cybersecurity in PACS and medical imaging: An overview," *J. Digit. Imag.*, vol. 33, no. 6, pp. 1527–1542, Dec. 2020, doi: 10.1007/s10278-020-00393-3.

[4] H. K. Huang, "Medical imaging, PACS, and imaging informatics: Retrospective," *Radiol. Phys. Technol.*, vol. 7, no. 1, pp. 5–24, Jan. 2014, doi: 10.1007/s12194-013-0245-y.

[5] S. A. Miozzi, "Narcolepsy and advanced medical imaging," *Radiol. Technol.*, vol. 93, no. 1, pp. 106–109, Sep. 2021.

[6] X. Zhou, Y. Li, and W. Liang, "CNN-RNN based intelligent recommendation for online medical pre-diagnosis support," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 3, pp. 912–921, May 2021, doi: 10.1109/TCBB.2020.2994780.

[7] F. Nensa, A. Demircioglu, and C. Rischpler, "Artificial intelligence in nuclear medicine," *J. Nucl. Med.*, vol. 60, no. 2, pp. 29S–37S, Sep. 2019, doi: 10.2967/jnumed.118.220590.

[8] F. Lombardi and S. Marinai, "Deep learning for historical document analysis and recognition—A survey," *J. Imag.*, vol. 6, no. 10, p. 110, Oct. 2020, doi: 10.3390/jimaging6100110.

[9] A. Alexander, A. Jiang, C. Ferreira, and D. Zurkiya, "An intelligent future for medical imaging: A market outlook on artificial intelligence for medical imaging," *J. Amer. College Radiol.*, vol. 17, no. 1, pp. 165–170, Jan. 2020, doi: 10.1016/j.jacr.2019.07.019.

[10] P. H. S. Kalmet, S. Sanduleanu, S. Primakov, G. Wu, A. Jochems, T. Refaee, A. Ibrahim, L. V. Hulst, P. Lambin, and M. Poeze, "Deep learning in fracture detection: A narrative review," *Acta Orthopaedica*, vol. 91, no. 2, pp. 215–220, Mar. 2020, doi: 10.1080/17453674.2019.1711323.

[11] A. Heidari, S. Toumaj, N. J. Navimipour, and M. Unal, "A privacy-aware method for COVID-19 detection in chest CT images using lightweight deep conventional neural network and blockchain," *Comput. Biol. Med.*, vol. 145, Jun. 2022, Art. no. 105461, doi: 10.1016/j.compbiomed.2022.105461.

[12] Z. Akkus, J. Cai, A. Boonrod, A. Zeinoddini, A. D. Weston, K. A. Philbrick, and B. J. Erickson, "A survey of deep-learning applications in ultrasound: Artificial intelligence–powered ultrasound for improving clinical workflow," *J. Amer. College Radiol.*, vol. 16, no. 9, pp. 1318–1328, Sep. 2019, doi: 10.1016/j.jacr.2019.06.004.

[13] M. Avanzo, L. Wei, J. Stancanello, M. Vallières, A. Rao, O. Morin, S. A. Mattonen, and I. E. Naqa, "Machine and deep learning methods for radiomics," *Med. Phys.*, vol. 47, no. 5, May 2020, doi: 10.1002/mp.13678.

[14] B. Lafci, E. Mercep, S. Morscher, X. L. Dean-Ben, and D. Razansky, "Deep learning for automatic segmentation of hybrid optoacoustic ultrasound (OPUS) images," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 68, no. 3, pp. 688–696, Mar. 2021, doi: 10.1109/TUFFC.2020.3022324.

[15] F. Marzola, N. van Alfen, J. Doorduin, and K. M. Meiburger, "Deep learning segmentation of transverse musculoskeletal ultrasound images for neuromuscular disease assessment," *Comput. Biol. Med.*, vol. 135, Aug. 2021, Art. no. 104623, doi: 10.1016/j.compbiomed.2021.104623.

[16] A. Heidari, N. J. Navimipour, M. Unal, and S. Toumaj, "Machine learning applications for COVID-19 outbreak management," *Neural Comput. Appl.*, vol. 34, no. 18, pp. 15313–15348, Sep. 2022, doi: 10.1007/s00521-022-07424-w.

[17] A. Heidari, N. J. Navimipour, M. Unal, and S. Toumaj, "The COVID-19 epidemic analysis and diagnosis using deep learning: A systematic literature review and future directions," *Comput. Biol. Med.*, vol. 141, Feb. 2022, Art. no. 105141, doi: 10.1016/j.compbiomed.2021.105141.

[18] O. Ronneberger, P. Fischer, and T. Brox, *U-Net: Convolutional Networks for Biomedical Image Segmentation*. New York, NY, USA: Springer, 2015, pp. 234–241.

[19] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–10.

[20] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE CVPR*, Jul. 2017, pp. 4700–4708.

[21] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3320–3328.

[22] X. Huang, M. A. L. Bell, and K. Ding, "Deep learning for ultrasound beamforming in flexible array transducer," *IEEE Trans. Med. Imag.*, vol. 40, no. 11, pp. 3178–3189, Nov. 2021, doi: 10.1109/TMI.2021.3087450.

[23] P. K. V. Sowdaboina, S. Chakraborti, and S. Sripada, "Learning to summarize time series data," in *Computational Linguistics and Intelligent Text Processing*. Berlin, Germany: Springer, 2014, pp. 515–528.

[24] X. Huang, Z. Deng, D. Li, and X. Yuan, "MISSFormer: An effective medical image segmentation transformer," 2021, *arXiv:2109.07162*.

[25] Z. Li, D. Li, C. Xu, W. Wang, Q. Hong, Q. Li, and J. Tian, "TFCNs: A CNN-transformer hybrid network for medical image segmentation," 2022, *arXiv:2207.03450*.

[26] Q. Sun, N. Fang, Z. Liu, L. Zhao, Y. Wen, and H. Lin, "HybridCTrm: Bridging CNN and transformer for multimodal brain image segmentation," *J. Healthcare Eng.*, vol. 2021, pp. 1–10, Oct. 2021.

[27] R. Puduppully, L. Dong, and M. Lapata, "Data-to-text generation with content selection and planning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 6908–6915.

[28] S. Wiseman, S. Shieber, and A. Rush, "Challenges in data-to-document generation," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Stroudsburg, PA, USA: Association for Computational Linguistics, 2017, pp. 2253–2263.

[29] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, Aug. 1997.

[30] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[32] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 1–10.

[33] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever, "Generative pretraining from pixels," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1691–1703.

[34] D. Zhou, B. Kang, X. Jin, L. Yang, X. Lian, Z. Jiang, Q. Hou, and J. Feng, "DeepViT: Towards deeper vision transformer," 2021, *arXiv:2103.11886*.

[35] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F. E. H. Tay, J. Feng, and S. Yan, "Tokens-to-token VIT: Training vision transformers from scratch on ImageNet," 2021, *arXiv:21001.11986*.

[36] X.-X. Yin, L. Sun, Y. Fu, R. Lu, and Y. Zhang, "U-Net-based medical image segmentation," *J. Healthcare Eng.*, vol. 2022, Apr. 2022, Art. no. 4189781, doi: 10.1155/2022/4189781.

[37] X. Wei, M. Gao, R. Yu, Z. Liu, Q. Gu, X. Liu, Z. Zheng, X. Zheng, J. Zhu, and S. Zhang, "Ensemble deep learning model for multicenter classification of thyroid nodules on ultrasound images," *Med. Sci. Monitor, Int. Medical J. Exp. Clin. Res.*, vol. 26, Jun. 2020, Art. no. e926096, doi: 10.12659/MSM.926096.

[38] H.-K. Wang, Y. Cheng, and K. Song, "Remaining useful life estimation of aircraft engines using a joint deep learning model based on TCNN and transformer," *Comput. Intell. Neurosci.*, vol. 2021, pp. 1–14, Nov. 2021, doi: 10.1155/2021/5185938.

[39] C. Che, P. Zhang, M. Zhu, Y. Qu, and B. Jin, "Constrained transformer network for ECG signal processing and arrhythmia classification," *BMC Med. Informat. Decis. Making*, vol. 21, no. 1, Dec. 2021, doi: 10.1186/s12911-021-01546-2.

[40] Y. Dai, Y. Gao, and F. Liu, "TransMed: Transformers advance multimodal medical image classification," *Diagnostics*, vol. 11, no. 8, p. 1384, Jul. 2021, doi: 10.3390/diagnostics11081384.

[41] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[42] M. Ahmad, S. F. Qadri, S. Qadri, I. A. Saeed, S. S. Zareen, Z. Iqbal, A. Alabrah, H. M. Alaghbari, and S. M. Mizanur Rahman, "A lightweight convolutional neural network model for liver segmentation in medical diagnosis," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–16, Mar. 2022, doi: 10.1155/2022/7954333.

[43] S. F. Qadri, D. Ai, G. Hu, M. Ahmad, Y. Huang, Y. Wang, and J. Yang, "Automatic deep feature learning via patch-based deep belief network for vertebrae segmentation in CT images," *Appl. Sci.*, vol. 9, no. 1, p. 69, Dec. 2018.

[44] U. Jung and H. Choi, "Active echo signals and image optimization techniques via software filter correction of ultrasound system," *Appl. Acoust.*, vol. 188, Jan. 2022, Art. no. 108519.

CHUANTAO WANG was born in Hubei, China, in 1981. He received the B.S. degree in mathematics and applied mathematics from Qiqihaer University, in 2004, the M.S. degree in computational mathematics from Chongqing University, in 2007, and the Ph.D. degree in system engineering from Beijing Jiaotong University, in 2011. From 2011 to 2016, he was an Assistant Professor at the Industrial Engineering Department, Beijing University of Civil Engineering and Architecture, Beijing, China. He is with the Beijing Engineering Research Center of Monitoring for Construction Safety, School of Mechanical-Electronic and Vehicle Engineering, Beijing University of Civil Engineering and Architecture. Since 2017, he has been an Associate Professor with the Industrial Engineering Department, Beijing University of Civil Engineering and Architecture. He has authored two books and more than 30 articles. His research interests include data mining, deep learning, and natural language processing.

JINHUA ZHANG was born in Hebei, China, in 1997. He received the bachelor's degree from the Beijing Institute of Graphic Communication, Beijing, China, in 2020. He is currently pursuing the master's degree in industrial engineering with the Beijing University of Civil Engineering and Architecture, Beijing.

He is with the Beijing Engineering Research Center of Monitoring for Construction Safety, School of Mechanical-Electronic and Vehicle Engineering, Beijing University of Civil Engineering and Architecture. His current research interests include data mining, deep learning, and natural language processing, and image processing.

SIYU LIU was born in Beijing, China, in 1998. She received the bachelor's degree from the Beijing University of Civil Engineering and Architecture, Beijing, in 2020, where she is currently pursuing the master's degree in industrial engineering and management.

She is with the Beijing Engineering Research Center of Monitoring for Construction Safety, School of Mechanical-Electronic and Vehicle Engineering, Beijing University of Civil Engineering and Architecture. Her current research interests include data mining, deep learning, natural language processing, and image processing.