IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 May 2023, accepted 27 May 2023, date of publication 31 May 2023, date of current version 8 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3281468

RESEARCH ARTICLE

RUnT: A Network Combining Residual U-Net and Transformer for Vertebral Edge Feature Fusion Constrained Spine CT Image Segmentation

HAO XU[1], XINXIN CUI[1], CHAOFAN LI[2], ZHENYU TIAN[1], JING LIU[1], AND JIANLAN YANG[1,3]

[1]School of Information Engineering, Gansu University of Traditional Chinese Medicine, Lanzhou 730000, China
[2]Yancheng School of Clinical Medicine, Nanjing Medical University, Nanjing, Jiangsu 224008, China
[3]Orthopedic Traumatology Hospital, Quanzhou, Fujian 362019, China

Corresponding author: Jianlan Yang (FJYJL@gszy.edu.cn)

This work was supported in part by the School of Information Engineering, Gansu University of Traditional Chinese Medicine, China; and in part by the Quanzhou Orthopedic Hospital, Fujian Province, China.

ABSTRACT Scoliosis, spinal deformity and vertebral spondylolisthesis are spinal disorders with high incidence, which seriously affect people's lives and health. CT is an important medical tool for the detection and diagnosis of spinal disorders and provides a large amount of pathologically valid information in various clinical practices such as spine pathology assessment and computer-assisted surgical interventions. As the spine presents long span, complex shape of biological curve and high multi-stage similarity in the sagittal plane of CT images. Therefore, fast and accurate spine segmentation technology has become an important research direction for computer-aided diagnosis. We proposed an RUnT network based on the combination of residual U-Net feature extraction network and Vision Transformer structure for fast and efficient automatic segmentation of multiple vertebrae of the spine. The deep vertebral features are first extracted using the residual U-Net network to prevent gradient diffusion while improving the accuracy of vertebral contour segmentation. Then the multi-scale feature maps extracted by the residual structure containing rich vertebral superficial information are input to the edge segmentation module. We designed the vertebral contour feature extraction structure to refine the segmentation boundaries and ensure the segmentation consistency of each vertebra by combining the operations of deconvolution and convolution for three different scales of deep features.Finally, the global information extraction module based on Transformer structure is combined with the local feature extraction module to achieve the blending of global location information of vertebrae with local features through the self-attentive feature map of multi-scale volume. By mixing edge features with semantic features, the semantic confusion arising from the high similarity between vertebrae when the decoder extracts vertebral features is reduced. The model proposed in this paper is experimented on the CTSpine1K and VerSe 20 public datasets. The results show that the model proposed in this paper obtains the state-of-the-art segmentation performance with the average DSC scores of 88.4% and 81.5% on CTSpine1K and VerSe 20, respectively, while reducing the average distance of HD95 from 4.86 to 3.88.

INDEX TERMS Spinal vertebral segmentation, vision transformer, residual U-net, vertebral edge segmentation.

I. INTRODUCTION

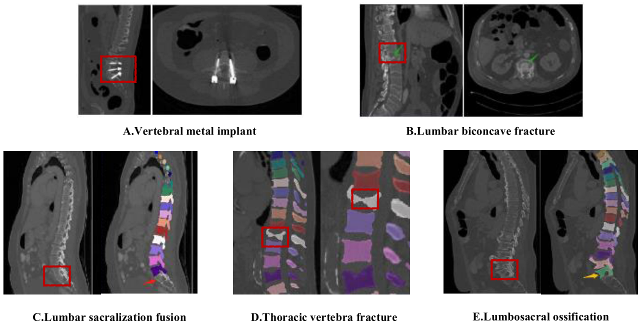
The spine is an extremely important skeletal structure in the human body that carries and conducts the combined load of the body's mechanical movements [1]. As the central load-bearing structure of the torso, the spine is responsible for the posterior support of the torso while protecting the central spinal cord pathways from sudden impact injuries, and is formed from top to bottom by the cervical (C1-C7), thoracic (T1-T12), lumbar (L1-L6), and sacral vertebrae through the

The associate editor coordinating the review of this manuscript and approving it for publication was M. Sabarimalai Manikandan.

bone joints. The cervical and thoracic vertebrae are subject to long-term mechanical stress and local injury resulting in disabling deformities due to scoliosis, which in the late clinical-stage can continue to squeeze internal organs and lead to death from pulmonary heart disease [2], while continuous compression of the spinal nerves in the lumbar spine can lead to severe pain with continuous spasms, resulting in paraplegia and deafness and blindness. Although the prevalence of spinal disorders is extremely high, early diagnosis cannot detect significant focal areas [3], so non-invasive detection of the spine using a computer-aided diagnostic system is of great clinical diagnostic value.

In recent years, clinical research on medical images of the spine has become increasingly important. In order to detect pathology in a timely manner for prevention and clinical treatment, surgeons mainly use digital radiography(DR) technology, computed tomography(CT) technology, and magnetic resonance(MR) technology to screen and make preliminary judgments about the spine [4]. The use of computer-aided diagnosis(CAD) systems based on medical imaging technology to extract 3D anatomical structures from patients' spine images can provide effective diagnostic information for clinical decisions such as vertebral fracture detection and identification of spinal deformities, allowing clinicians to precisely localize the patient's pathology and make timely surgical treatment plans, combining basic medicine with clinical medicine to provide individualized treatment strategies [5]. It also reduces the time of manual segmentation of the vertebrae and reduces the rate of misdiagnosis and missed diagnosis by radiologists. MR technology focuses on soft tissue imaging in the spinal canal, such as nerves and spinal cord, for the detection of spine-related diseases, and CT technology is preferred for the observation of high density skeletal lesions. The full-length CT image of the spine containing 24 vertebrae is reconstructed in 3D in the imaging workstation by multiple scans of the cervical, thoracic, and lumbar spine, and automatic vertebral segmentation from the CT image of the spine is of great clinical value in the diagnosis of vertebral diseases.

The spine presents a long span, complex shape of a biological curve, and high multi-stage similarity in the sagittal plane of CT images. The number of vertebrae in each site is large, and the vertebrae within the site are similar in height, while the vertebrae between sites are relatively different. Thus the following problems exist in the study of automatic segmentation of the spine: high-precision segmentation is difficult, and there is a high degree of similarity between vertebral instances in the same part, which makes the segmentation network semantically confused, over-reliance on the cross-sectional area of vertebrae to infer contextual information, but the human sacral vertebrae have a small cross-sectional area and the segmentation network cannot make judgements based on references. The robustness of the segmentation network is poor. The current network is good for healthy or slightly deformed vertebrae segmentation, but it cannot perform high-precision boundary segmentation for samples with metal implants or severe spinal deformities [6]. As shown in Figure.1, there are many abnormal features in CT imaging of human vertebrae that affect the segmentation results.


A.Vertebral metal implant

B.Lumbar biconcave fracture

C.Lumbar sacralization fusion

D.Thoracic vertebra fracture

E.Lumbosacral ossification

FIGURE 1. Bone tissue density is large relative to other organs, and CT is an important way to detect spinal diseases. Due to the obvious differences between A (metal implant) and B (bilateral fracture of vertebrae) relative to the characteristics of healthy vertebrae, the deep learning model is challenged, while C (lumbar sacralization) and E (sacral lumbarization) produce semantic segmentation confusion and cannot accurately segment the last vertebrae of the lumbar spine, and D (multiple fractures) appearance changes have a great impact on vertebral segmentation accuracy.

A. RELATED RESEARCH WORK

At present, the research methods of spine and vertebrae segmentation are divided into three categories. The first category is based on digital image processing methods to propose a vertebral segmentation network, the second category is based on machine learning theory to segment the vertebrae, and the third category is mainly based on convolutional neural networks (CNN) and U-net network to segment the vertebrae.

1) DIGITAL IMAGE PROCESSING VERTEBRAE SEGMENTATION

Researchers Klinder et al. first proposed a fully automated framework based on statistical shape model to localize the relative position of vertebrae and biomorphic curve information acquisition from spine images, using spine volume for curve reorganization and then generalized Hough transform to achieve spine detection function, using multi-level processing to identify and segment vertebrae, but this framework requires a large number of high-quality spine data set training and relatively high computational complexity is difficult to apply to clinical practice [7]. Korez et al. proposed an automatic spine localization and vertebrae segmentation model based on statistical model, using interpolation technique for maximum filling of missing vertebrae, using morphological operations for an initial segmentation of vertebrae, achieving individual detection of vertebrae by locating spine geometric positions, and finally using shape statistical model for high precision segmentation, which can process high resolution spine CT images to obtain full spine segmentation results, but the use of interpolation technique for filling leads to partial noise and false positive regions in the processed images affecting the segmentation results [8]. Štern et al. proposed to simulate the three-dimensional curve parameters of the spine through a geometric model, using 25 clinical parameters with 6 geometric parameters to approximate the vertebral shape; this method uses specific parameters for spherical model to fit the vertebral shape and has the problem of not fitting for deformed vertebral geometry [9]. Ibragimov et al. proposed a new lumbar spine segmentation framework based on transporting theory and strategic advantages reducing the computational complexity from 2D to 3D segmentation [10]. Castro-Mateos et al. similarly proposed an active contour model based on the biological curve of the vertebrae, which used the statistical interspace model(SIM) to model the intervertebral discs between adjacent vertebrae and then realized the segmentation of the vertebrae by calculating the relative position information of the discs combined with the vertebral biological curve information, but it relied too much on the manual selection of the central region features of the discs and manual initialization of the vertebral contour to provide clinical information in for spinal deformity and surgical planning [11]. Kadoury et al. used the markov random field(MRF) to divide the spine into intervertebral disc sets, by measuring the geometric characteristics between adjacent vertebrae, importing the intervertebral disc sets to achieve the consistency of the regional curve fitting of the spine, and improving the local curve of the vertebrae by constrained mesh relaxation technology [12]. In 2013, Kadoury et al. proposed spinal joint segmentation using stream embedding and higher-order MRF for CT and MR imaging on the original study, mapping foreground region pixels into low-dimensional stream space using higher-order MRF to localize and segment the underlying vertebrae, which requires a large number of higher-order MRF models with high computational complexity and lacks robustness to abnormal sensitivity to data noise [13].

2) MACHINE LEARNING VERTEBRAL SEGMENTATION

The second type of research is based on the theory of machine learning for vertebral segmentation. With the in-depth research of machine learning in image processing, segmentation models combining digital image technology and machine learning have been continuously proposed. In 2015, Suzani et al. first proposed a multi-layer perceptron-based lumbar segmentation region, statistically analyzed the voxel intensity and generated the original lumbar model, and gradually approached the real lumbar parameters through continuous iteration of local thresholds [14]. Chu et al. used random forest regression to locate and detect vertebrae, and used a hidden Markov model to generate a voxel distribution probability map to eliminate the segmentation ambiguity caused by the highly similar shape between vertebral bodies [15].

3) CONVOLUTIONAL NEURAL NETWORK VERTEBRAL SEGMENTATION

With the introduction of deep learning models, the third type of model is mainly based on CNN and U-net networks for vertebral segmentation. In 2016, Korez et al. used a convolutional neural network CNN for automatic vertebral segmentation of MR spine images, using multiple 3D convolutional layers with pooling layers to extract vertebral features, and mapping vertebral features to the mask of the spine through two fully connected layers to improve the robustness and accuracy of individual vertebral segmentation [16]. Sekuboyina et al. proposed a two-stage network for multi-label labeling of the lumbar spine. In the first stage, they used multilayer perceptron (MLP) to achieve local lumbar region localization, while in the second stage, they used the U-net classification model to classify the vertebrae. However, due to the relatively small training and validation datasets, there might be potential overfitting or underfitting problems with this model, thus requiring verification of experimental performance under different combinations of hyperparameters [17]. Janssens et al. proposed a cascaded fully convolutional networks (FCN) vertebral segmentation framework based on positioning FCN and segmentation FCN. In the first stage, training regression 3D FCN to realize the localization of the lumbar region, and in the second stage, 3D U-net was used to achieve multi-class segmentation for the lumbar region [18]. Lessmann et al. improved the two-stage vertebral segmentation network, using the low-resolution U-net network to identify each vertebra, and then using the CNN network to learn the refined low-resolution labels [19]. In 2019, Lessmann et al. proposed an automatic vertebral bone recognition and segmentation network based on an iterative fully convolutional neural network, which gradually refines the segmentation results by concatenating multiple FCNs, while introducing depth supervision and spatial upper and lower layer information to improve the segmentation performance, iterative instances to segment vertebrae and taking the maximum likelihood method to refine the boundaries of the segmented individual vertebrae, but the model requires multiple FCN training models in series, with significant improvement in training time and the number of parameters, and high computational complexity and model inference time [20]. Vania et al. used a 3D U-net network that fuses CNN and FCN to achieve fully automatic segmentation of the spine, and used category redundancy as a constraint to improve the accuracy of vertebral boundary segmentation [21]. Payer et al. proposed a segmentation method based on FCN in 2020, through the three stages of spine recognition, vertebral positioning, and vertebral segmentation to achieve sequential segmentation from spine to vertebrae, and used U-net to perform high-resolution segmentation on the identified vertebrae [22]. Nazir et al. proposed an embedded clustering and slicing U-net network, namely the ECSU-Net network, which consists of three modules: segmentation, intervertebral disc extraction, and image fusion. The segmentation module uses the embedded clustering method to perform rough segmentation on the spine, the intervertebral disc extraction module performs spine classification on rough segmentation and captures the spatial information between different vertebrae, and the fusion module stacks the segmented 2D images into 3D images [23]. Huang et al. proposed to use the Ortho2D model

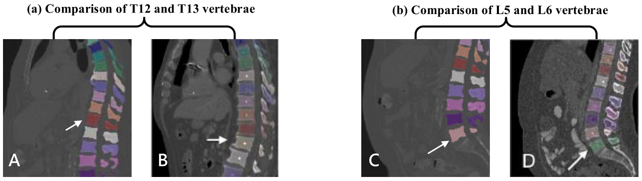

(a) Comparison of T12 and T13 vertebrae
(b) Comparison of L5 and L6 vertebrae
A
B
C
D

FIGURE 2. A: normal T1-12 vertebrae in the segmentation of the thoracic spine, B indicates that less than 5% of the population has the thirteenth thoracic vertebrae, i.e., T13, C indicates normal lumbar vertebrae L1-5, and D: the sixth block of lumbar vertebrae L6 that affects the generalizability of the spine segmentation model.

composed of two independent fast R-CNN networks to detect vertebrae and classify the sagittal and coronal planes of vertebrae in 2021 [24]. Pang et al. used the DGMSNet network on the MR spine image to generate the prediction path of the spine segmentation prediction and the detection path of the key point heat map, and weighted the segmentation loss and detection loss as a mixed supervision loss function to train the model to generate high-precision segmentation prediction [25]. Wu et al. proposed a 3D lumbar spine localization and segmentation network based on a 2D hybrid visual projection image fusion envelope (LVLS-HVPFE), which uses X-ray and CT scans to obtain 2D visual projection images and achieves 3D lumbar spine localization and segmentation by region growing algorithm with multi-distance weighted averaging strategy [26]. Zhao et al. proposed a Residual-atrous attention network (RA^2-Net) lumbar spine segmentation network model, using atrous encoder to learn multi-scale contextual information in MR images to improve lumbosacral from segmentation performance, and fusing deep features of the encoder with shallow features of the decoder to enhance local features of vertebrae through residual jump connection [27]. Meng et al. used graph optimization and statistics before localizing, segmenting and identifying the spine in CT images. The authors used a 3D U-net model to initialize the location information of the spine, and then applied a graph-cut algorithm to encode the processing and improve the segmentation accuracy through anatomical consistency cycles. However, the model requires large computational resources and multiple iterative cycles for encoding, and the inference time is too long, which makes it difficult to be applied in clinical settings [28].

The traditional vertebral segmentation model mainly locates the vertebral sites through the field-of-view (FOV) of the spine, and then stacks the obtained 2D segmentation labels into 3D vertebral shapes, ignoring the spatial location information of the vertebrae relative to the overall organ. As shown in Figure.2, the conventional segmentation model that does not fully consider the global position information of the vertebrae cannot accurately locate the 25th and 26th vertebrae by relying on the regions of interest(ROI) area obtained from the relative positions of the vertebrae, which reduces the generalization performance of the spinal vertebral segmentation model.

4) TRANSFORMER-BASED VERTEBRAL SEGMENTATION MODEL

In 2020, as the ViT [29] model first proposed to use the Transformer structure to process the image matrix, the combination of the Transformer structure and the CNN network has become a new research direction. The proposal of the Swin Transformer [30] further consolidated the potential of Transformer structure in image processing, and the image segmentation model based on the combination of Transformer structure and U-shaped network was further developed in the field of medical images. Syed Furqan Qadri et al. proposed a patch-based deep learning spine CT automatic segmentation method that divides CT data into small fast and obtains vertebral discriminative feature information from unsupervised data using stacked sparse autoencoder (SSAE) and used CNN for classification prediction and post-processing of image blocks on VerSe, CSI-Seg and lumbar spine datasets. The model segmentation performance was verified on VerSe, CSI-Seg and lumbar spine datasets. However, this method only considers the local information of vertebrae, ignoring the effect of global position information on the high similarity between vertebrae [31]. In contrast, our proposed segmentation model based on residual U-net and Transformer utilizes the Global Transformer structure of local volume-based multi-head self-attention (LV-MSA) and shift local volume-based multi-head self-attention(SLV-MAS) tandem in the Global Transformer structure can reduce the complexity of the computer while achieving the acquisition of global location information of vertebrae through the self-attentive feature maps at multiple volume scales, reducing the semantic confusion generated by the network when segmenting highly similar vertebrae instances. Tao et al. proposed a Transformer-based vertebral CT image automatic detection and positioning model Spine-Transformers, using the ResNet-50 network to process the input spine CT image, and using the lightweight Transformer to obtain different vertebral local feature maps for the extracted shallow features to form Multi-scale feature pyramids. However, the Spine-Transformers network only realizes partial detection and positioning of the vertebrae of the spine, and cannot realize the 3D segmentation of 24 vertebrae [32]. We propose a multiscale boundary fusion module to inverse fuse the multiscale feature maps sampled by the encoder, up-sample the local features containing multiscale shallow information to the same size and then use the convolutional layer to extract vertebral boundary information to constrain the vertebral deep features extracted by the decoder to ensure the consistency of each vertebral contour edge segmentation. You et al. proposed a 3D EG-Trans3DUNet vertebral segmentation model based on the TransUNet network, using the 3D U-Net network to obtain vertebral depth features and fusing the global position information of vertebrae captured by the Transformer model to improve vertebral segmentation accuracy [33]. However, the authors used the relatively small VerSe 20 spine dataset for training and evaluation without detailed comparison of the generalization ability of the model. We used the largest publicly available spine CT dataset, CTSpine1K [34], to input into the RUnT model for training, and added a residual structure to the encoder to increase the shallow vertebral information contained in the extracted depth features, and input the depth features containing multi-scale rich shallow information into the multi-scale edge segmentation module to extract vertebral contour features, which improved the vertebral contour feature segmentation accuracy.

Therefore, we propose a spine vertebrae CT image segmentation model based on the combination of Transformer structure and residual U-net network to address the current problems, using the residual Conv structure to obtain a multi-scale feature pyramid of vertebrae, inputting shallow features into the parallel vertebrae boundary region, using the CNN structure to obtain the spine fuzzy boundary region features, and fusing the vertebrae edge features containing rich shallow information vertebral edge features are fused with the deep local features of vertebrae to further constrain the inconsistency of vertebral segmentation. The Global Transformer encoder is used to fuse with the Local Transformer encoder, and the global position information of 25 vertebrae captured by the Global Transformer is superimposed with the local features of vertebrae extracted by the Local Transformer to reduce the semantic confusion of the vertebral segmentation network for the highly similar local features of multiple vertebrae.

In summary, our main contributions can be summarized in the following four points:

(i) The combination of residual structure and encoder superimposes the deep features of vertebrae extracted from the convolutional layer with a layer of convolutionally filtered shallow information, which prevents gradient diffusion while inputting the edge segmentation module containing rich shallow vertebrae information to improve the accuracy of vertebrae contour segmentation.

(ii) The LV-MSA and SLV-MAS are connected in tandem in the Global Transformer module, which can reduce the computational complexity of high-resolution vertebral CT images while achieving the acquisition of global position information of vertebrae through self-attentive feature maps at multi-volume scales and reducing the semantic confusion caused by the high similarity between vertebrae when the decoder extracts vertebral features.

(iii) Using Global Transformer structure and Local Transformer structure channel information superposition, deep local features of vertebrae are superimposed with global position information to improve the accuracy of vertebral smooth region segmentation, and the vertebral edge features extracted by the vertebral edge module are spliced with the deep smooth region features of vertebrae extracted by the decoder to ensure the segmentation consistency of each vertebra.

(iv) To address the non-convex loss due to small datasets, we used the CTSpine1K large vertebral dataset with rich individual samples with the VerSe 20 dataset input model for training to improve the Query, Key, and Value matrix fitting speed in the multi-head self-attention(MSA) mechanism. It is demonstrated experimentally that our vertebral segmentation model outperforms other models in Dice metrics [6] and HD95(in mm) metrics [6]. The accuracy of our model in multi-label segmentation of spinal vertebrae with model robustness is demonstrated.

II. MATERIALS AND METHODS

The network we proposed is shown in Figure 3, and the overall structure is mainly composed of five parts: (i) The ResNet encoder structure for extracting vertebral depth features; (ii) The Local Transformer structure that captures the depth feature information of vertebrae and the Global Transformer structure that extracts the position information of 25 vertebrae; (iii) Multi-scale feature fusion vertebral Edge Segmentation module; (iv) A decoder structure for deconvolving deep features; (v) A predictive segmentation module that fuses vertebral contour information and depth features to ensure the consistency of multiple vertebral segmentations.

A. RESIDUAL U-NET STRUCTURE

The CTSpine1K dataset is preprocessed into the vertebral local dataset $D^L = I_i^S \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C}(i \in \mathbb{Z}^L)$ and the vertebral global dataset $D^G = \left\{I_i^S \in \mathbb{R}^{H\times W\times C}\right\}\left(i \in \mathbb{Z}^G\right)$, D^L extracts the local depth features of the vertebrae through the residual U-net encoder, and inputs the visual converter Local Transformer to capture the dependent information of the vertebral depth features, D^G is preprocessed by clipping and relative position encoding, and then input into the Global Transformer structure to extract the position information of 25 vertebrae, after the Global and Local Transformer output vertebral position information and local features are added, it is input to the decoder module for deconvolution operation. The residual U-net structure is divided into four stages $Stagei\ (i = 1, 2, 3, 4)$, as shown in Figure.3.

ResNet_Block extracts vertebral multiscale features $F_{l*}^{S_i}\ (i = 1, 2, 3, 4)$ mainly by 3D convolution kernel, l^* represents the number of layers of convolution in ResNet_Block, using ReLu function activation with GroupNorm for Batch dimension normalisation to obtain the output. Our previous research found that as the number of convolution layers deepens, the convolution kernel is used as a high-pass filter to continuously extract the boundary contours of vertebrae, which is likely to cause loss of contour curve information and enhanced filtering of features. The local depth features of vertebrae increase with the feature channel dimension, but the loss of local vertebral information is serious. Through the residual structure, the shallow input features containing rich information and the extracted depth features are added to the matrix to obtain $F_{l*}^{S_{i+1}} = F_{l*}^{S_i} + \sum_{i=2}^{S_{i-1}} \mathcal{F}\left(F_{l*}^{S_i}, W_{l*}^{S_i}\right)(i = 1, 2, 3, 4)$, $W_{l*}^{S_i}$ represents the training weight of the l^* layer convolution kernel of $ResNet\_Block\_i(i = 1, 2, 3)$, which adds the feature information of the hidden layer feature map.

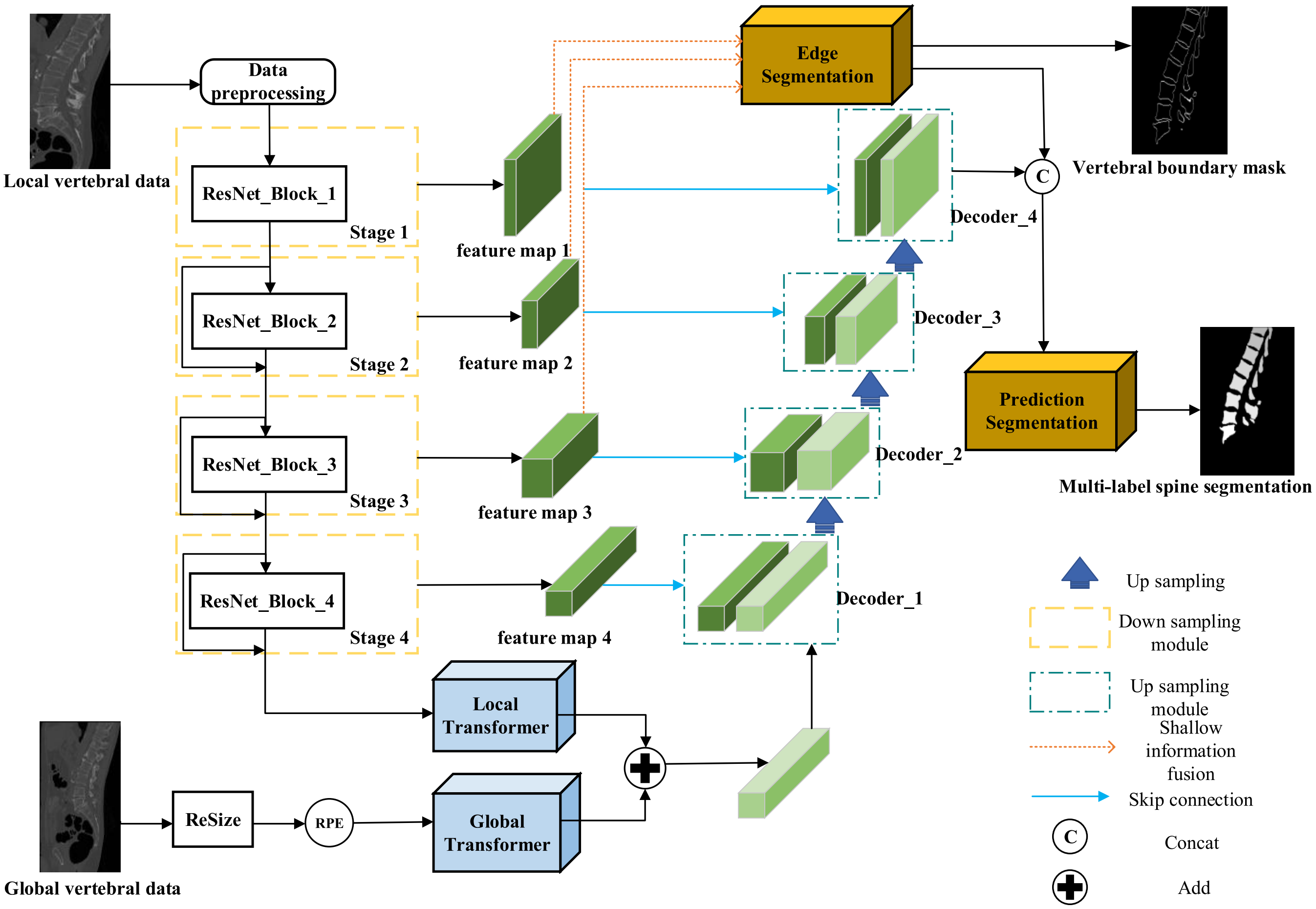


FIGURE 3. Overview of the architecture of the vertebral segmentation network based on Transformer and residual U-net networks. The RUnT network consists of five main components: (i) Encoder, which is formed by adding residual structure between each layer based on the U-Net encoder. (ii) Edge Segmentation module unifies the dimension and size of shallow features of different sizes using the deconvolution operation, and segment the vertebral edge curve by convolutional layers. (iii) Global Transformer module and Local Transformer module capture global position information and focus on local features for deep features of different scales. (iv) Decoder based on U-Net structure recovers the deep features extracted by the transformer module and constrains the vertebral bone boundary information extracted by the Edge Segmentation model for consistency. (v) The predictive segmentation structure classifies 25 classes of vertebrae for prediction.

Due to the high semantic similarity between the vertebrae in the vertebral segmentation process, so to constrain the segmentation consistency of each vertebrae, we extracted shallow feature maps containing rich vertebral information from $Stagei_{1,2,3}$ as multi-scale feature pyramids and input them into the Edge Segmentation module, the basic CNN structure is used to realize the information segmentation of the boundary contour of the vertebrae. The encoder based on the residual structure extracts the deep features containing the multi-scale information of the vertebrae and performs skip connection with the deep features diluted by the decoder, so as to realize the extraction of the deep abstract features of the vertebrae while containing sufficient semantic information. The output of the residual U-net structure is spliced with the contour features of the vertebral boundary segmentation to further integrate the local features of the vertebrae with the global information, and improve the segmentation effect of the contour information and structural features of the vertebrae.

B. LOCAL INFORMATION AND GLOBAL INFORMATION EXTRACTION MODULE

The ResNet structure in the encoder proposes vertebral depth features, but studies have shown that the ResNet network is not robust to high-frequency signal processing, and high-frequency signals are continuously amplified as the number of convolutional layers deepens. The Transformer module based on the MSA mechanism has good high-frequency signal filtering ability and gathers the deep features extracted by the ResNet structure. Therefore, we proposed to use a network combining ResNet and Transformer modules, and input the deep features extracted by the encoder into the Transformer module to smooth the spatial features and enhance the receptive field. Since the Local Transformer structure (Figure 4(a)) can only extract vertebral features and cannot capture the relative global position information of vertebral local features, we refer to the nnformer [35] network framework and proposed a Global Transformer structure as shown in Figure 4(b). It is used to extract the relative position

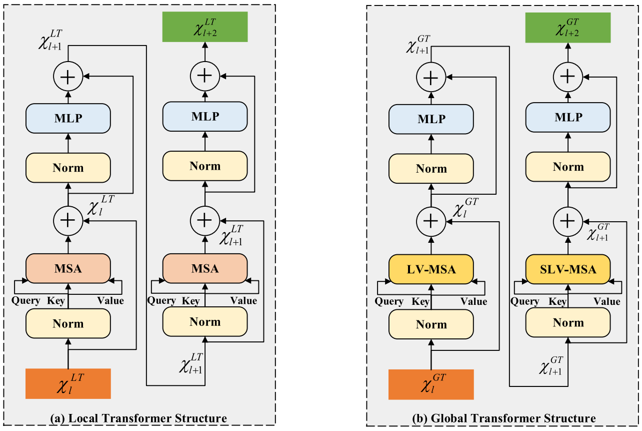


FIGURE 4. (a) The input in the Local Transformer structure is the deep vertebral feature set χ_l^{LT} (l represents the index of the layer in the Transformer structure) after ResNet downsampling, and the local feature attention of the spine is calculated using the MSA mechanism. Norm represents the layer regularization. MLP stands for multilayer perceptron composed of two layers of neural networks. (b) The input of the Global Transformer structure is the vertebral global feature χ_l^{GT}, which contains more vertebral superficial semantic information than the deep feature χ_l^{LT}, and the complexity of the traditional MSA calculation is too high, so we used LV-MSA and SLV-MSA tandem to extract the vertebral global position information while reducing the computational complexity.

information between the vertebra features and the whole. Using LV-MSA and SLV-MAS to calculate the different self-attention characteristic maps under the multi-volume scale in series. Since the input χ_l^{GT} (l represents the index of the layer in the Transformer structure) of the Global Transformer module is the original data of the 3D CT image, using the traditional multi-head attention mechanism for calculation will take up too many GPU resources. On a patch with a volume of $\{h \times w \times c\}$, the computational complexity based on LV-MSA (Equation 2) and SLV-MSA (Equation 3) is reduced by about 97% and 98% compared to the computational complexity of MSA (Equation (1)).

$$\Omega\,(MSA) = 4hwC^2 + 2(hw)^2C \tag{1}$$

$$\Omega\,(LV - MSA) = 4hwcD^2 + 2S_hS_wS_chwcD \tag{2}$$

$$\Omega\,(SLV - MSA) = 4hwcD^2 + 2V_hV_wV_chwcD \tag{3}$$

where D represents the length of the data sequence input into the Global Transformer structure, $\{S_h, S_w, S_c\}$ represents the local image volume size of the input LV-MSA-based Transformer structure, $\{V_h, V_w, V_c\}$ represents the local image volume size of the input SLV-MSA-based Transformer structure, $(V_h, V_w) = \alpha\,(S_h, S_w)\,V_c = \beta S_c$,we empirically set the hyperparameters $\alpha, \beta = 0.5$ to alternate the attention information under the two volume sizes.

The calculation process of the Local Transformer structure and the Global Transformer structure is as follows:

$$\begin{aligned} \chi_{l+1}^{LT} &= MSA(Norm\left(\chi_l^{LT}\right) + \chi_l^{LT}\, l = 0, 1, 2 \ldots\ldots L \\ \chi_{l+2}^{LT} &= MSA\left(Norm\left(\chi_{l+1}^{LT}\right)\right) + \chi_{l+1}^{LT} \end{aligned} \tag{4}$$

$$\begin{aligned} \chi_{l+1}^{GT} &= LV - MSA(Norm(\chi_l^{GT})) + \chi_l^{GT}\, l = 0, 1, 2 \ldots .L \\ \chi_{l+2}^{GT} &= SLV - MSA(Norm(\chi_{l+1}^{GT})) + \chi_{l+1}^{GT} \end{aligned} \tag{5}$$

In Equation (4)(5), χ_l^{LT} is the input of the Local Transformer module, and χ_l^{GT} is the input of the Global Transformer module, where l represents the number of layers of the Transformer structure, and $L = 11$ represents that the Local Transformer module and the Global Transformer module each have 12 floors.

C. MULTI-SCALE FEATURE EDGE SEGMENTION MODULE

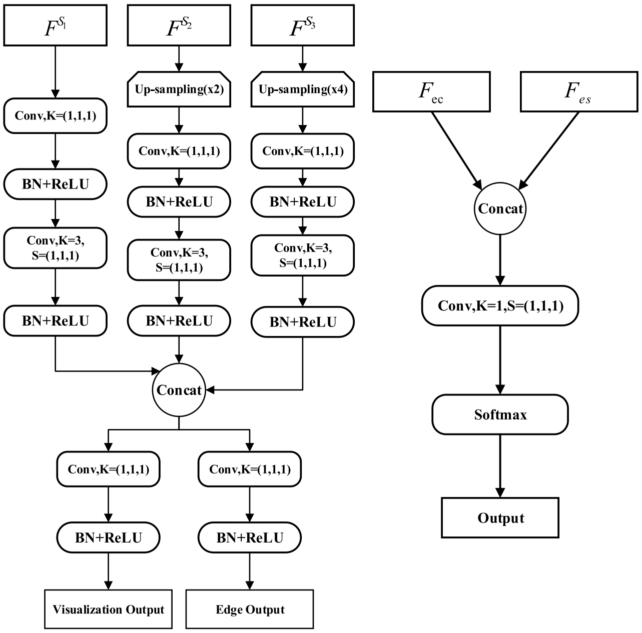
In order to accurately segment the 3D biological curve contours of the vertebrae, we proposed to use the hidden layer features extracted from the downsampled residual module in Figure 3 to segment the boundary contours of the vertebrae. As the number of downsampling layers deepens, the feature information extracted by the convolution operation contain more abstract vertebral features, and the loss of deep feature semantic information leads to the poor segmentation accuracy of the current convolution operation-based vertebral segmentation network for vertebral boundary contours. While the shallow features extracted by the residual convolution operation of the shallow layer contain the boundary contour detail information of the vertebrae, we used $ResNet\_Block\_i (i = 1, 2, 3)$ to propose 3 scales of shallow features to input into the Edge Segmentation module, the structure diagram of the boundary segmentation module is shown in Figure 5. F^{S_i} represents feature maps extracted from different depths, F^{S_2} and F^{S_3} due to the size difference of H_{S_i}, W_{S_i} in $\left\{B_{S_i}, C_{S_i}, H_{S_i}, W_{S_i}\right\}$ need to pass the upsampling module to restore the size, through the $3 \times 3 \times 3$ convolution kernel extracts the local details of the shallow vertebrae and then using the $1 \times 1 \times 1$ convolution kernel controls the number of feature channels, and through the Concat operation realizes the splicing of feature channels at the edge of vertebrae. The spliced multidimensional features are reduced in dimension to F^{S_1} by convolution kernel $1 \times 1 \times 1$, the output vertebral edge feature map is fused with the vertebral feature after upsampling by the decoder, through the boundary of a single vertebra constrains the scale range of vertebral segmentation, and the deep fusion of local information and overall features is realized to ensure the consistency of single vertebral segmentation.

D. LOSS FUNCTION

This paper proposed a vertebral segmentation model loss function $\mathcal{L}$ such as Equation (6), which is mainly composed of two parts of the loss function, the overall label loss function $\mathcal{L}_s$ and the vertebral edge segmentation loss function $\mathcal{L}_e$.

$$\mathcal{L} = \alpha\mathcal{L}_s + \beta L_e \tag{6}$$

Among them, we used the vertebral Edge Segmentation loss function as a supplement to the overall segmentation loss function, because of the weight parameters $\alpha = 0.7$, $\beta = 0.3$. Since vertebral segmentation is a multi-label task, the number of vertebrae labels contained in a single training sample varies greatly, so there is a serious label category imbalance. Therefore, we used a loss function combining cross entropy



(a) Edge Segmentation Module

(b) Prediction Segmentation Module

FIGURE 5. (a) Structure diagram of Edge Segmentation module, $F^{S_i}(i = 1, 2, 3)$ represents the hidden layer features output from Stage i in the encoder residual module, and the feature map downsampled from the upper layer contains more semantic information of the vertebrae, and the fine boundary contour segmentation of the vertebrae is performed by using multi-scale feature map stitching fusion. (b) represents the predictive segmentation structure diagram, F_{ec} represents the shallow features of the vertebrae obtained by upsampling on the encoder, F_{es} represents the vertebral edge information output by the Edge Segmentation module, and the feature channel is segmented by a filter with a convolution kernel of 1 × 1×1. The 25 class labels of the vertebrae are obtained using the softmax activation function.

and Dice loss function for training (Equation 7).

$$\begin{aligned}\mathcal{L}_s &= \gamma \mathcal{L}_{CE} + \delta \mathcal{L}_{dice}\\ \mathcal{L}_{CE} &= -\frac{1}{N}\sum_{i}\sum_{c=1}^{M} y_{ic} log(\hat{y}_{ic})\\ \mathcal{L}_{dice} &= 1 - \frac{1}{M}\sum_{c}^{M}\frac{2\sum_{i}^{N^*} y_{ic}\hat{y}_{ic}}{\sum_{i}^{N^*} y_{ic} + \sum_{i}^{N^*} \hat{y}_{ic}}\end{aligned} \tag{7}$$

Among them, $\mathcal{L}_{CE}$ is the multi-class cross entropy loss function, N is the total number of samples, M is the total number of label categories, y_{ic} represents the value of 1 when the real label of sample i is c, otherwise the value is 0, $\hat{y}_{ic}$ represents the probability value when the predicted label value of sample i is c. N^* in $\mathcal{L}_{dice}$ represents the total number of voxels in the sample, and the remaining parameters have the same meaning as $\mathcal{L}_{CE}$.

For the vertebral edge loss function, we used a deeply supervised approach to train the vertebral Edge Segmentation module, which segmented the semantic information of the shallow vertebral edge contours into binarized labels, so to reduce the voxel expansion of the vertebral edges and improve the boundary accuracy of the shallow segmentation of the vertebral edge contours, we used the binary loss function BCE for training, as shown in Equation (8).

$$\mathcal{L}_e = -\lambda \sum_{i} y_i log y_i - \mu \sum_{i} (1 - y_i)\, log(1 - y_i) \tag{8}$$

Since the proportion of voxels in the vertebral edge contours is much smaller than in the background, we set the foreground hyperparameter λ to 0.8 and the background weight parameter μ to 0.2 to reduce the $\mathcal{L}_e$ loss value.

E. EUALUATION METRICS

In this paper, the accuracy of vertebral segmentation is measured by using the DSC to calculate the segmentation effect of 25 vertebrae. The overall segmentation result of the spine is evaluated by the average DSC. The calculation of the DSC score of the vertebra is shown in Equation 9, T represents the expert segmentation result, P represents the model prediction result, and i represents the index value of the spinal vertebra.

$$Dice\,(P, T) = \frac{1}{N}\sum_{i=1}^{N}\frac{2\,|P_i \cap T_i|}{|P_i| + |T_i|} \tag{9}$$

At the same time, the HD95 evaluation index is used to calculate the spatial distance between the vertebrae prediction segmentation result set and the real vertebrae set. The calculation formulas of the prediction set and ground truth set are shown in Equation 10.

$$\begin{aligned}HD(P, T) = \frac{1}{N}\sum_{i=1}^{N} max\,\{&sup_{p\in P_i} inf_{\,t\in T_i} d\,(p, t)\,,\\ &sup_{t\in T_i} inf_{\,p\in P_i} d(p, t)\}\end{aligned} \tag{10}$$

Among them, P_i represents the surface distance set of the vertebral prediction segmentation mask whose index is i, T_i represents the vertebral surface distance set whose real label index is$i$, and $d(p, t)$ represents the Euclidean distance between point p and point t in the set of P_i surfaces and the set of T_i surfaces. Although HD95 produces large outliers in the calculation of the thoracic T13 vertebra and the lumbar L6 vertebra, the calculation results of the remaining vertebrae can be analyzed to obtain the average performance of the model segmentation performance results.

III. EXPERIMENTS

A. DATA PREPARATION

In order to solve the problem of slow fitting rate during the training of small data sets for the Q, K and V matrix parameters in the Local Transformer module of the vertebral segmentation model and the Global Transformer module of the MSA calculation Equation 11.

$$Attention\,(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d}} + R)V \tag{11}$$

In the formula, Q, K, and V represent Query, Key and Value, and R represents the relative position code.

We selected the large spine dataset CTSpine1K released in 2021 for training. Compared with the traditional small

TABLE 1. Overview of the status of each spinal vertebral CT dataset.

Dataset	CSI-Seg	Dataset-5	xVertSeg	VerSe 19	VerSe 20	CTSpine1K
Training	10	10	15	80	103	610
Validation	10	-	10	40	216	197
Test	-	-		40	-	198
Vertebrae	120	50	125	1725	4141	11172
(Cer/Tho/Lum)	(-/12/-)	(-/-/50)	(-/-/125)	(220/884/621)	(581/2255/1305)	(419/5942/4712)
Mean size	-	-	(962,969,212)	(512,512,640±197)	(512,512,640±197)	(512,512,504±155)
Patients(n)	18	-	25	141	300	1005
Source and Year	2014	2014	2016	2019	2020	2021

datasets CSI-Seg 2014, Dataset-5 2014 and xVertSeg 2016, the CTSpine1K dataset contains 1005 training samples and is stored in NIFTI format the training set contains 419 cervical vertebrae, 5942 thoracic vertebrae, and 4712 lumbar vertebrae. A comprehensive comparison between the CTSpine1K dataset and other datasets is shown in Table 1, compared with the previous three data sets, the amount of annotated training data has increased by 60, 60, and 39.66 times, respectively. Compared with the CTSpine1K data set, although the VerSe 19 and VerSe 20 data sets are currently the most popular datasets, the training samples of the VerSe series data sets are only 15.92% and 31.74%, and some samples of the VerSe se-ries dataset are regionally cropped to contain only CT images of the spine, lacking the information of the surrounding organs and tissues of the spine, leading the MSA-based Transformer structure has a relatively small receptive field when calculating global attention and cannot learn the global position information of the vertebra.

B. DATA PRE-PROCESSING

The CTSpine1K dataset consists of four sets of vertebral data sets (COLONOG, HNSCC-3DCT-RT, MSD T10, COVID-19), and the training set, validation set, and test set are divided in a ratio of 3:1:1. The input data is divided into two parts, the local vertebra dataset is trained through the residual U-net structure, and the global vertebra dataset is input into the Transformer structure for learning.

Our preprocessing step for the vertebral CT image dataset is mainly divided into two parts, the first step is the batch resampling of the dataset, and the second step is the normalization operation of the resampled dataset. In our experiments, we found that when resampling the data into $\boldsymbol{1mm} \times \boldsymbol{1mm} \times \boldsymbol{1mm}$ voxel space, we found that when the size of a single sample reached $\mathbf{256} \times \mathbf{256} \times \mathbf{180}$, the video memory size was insufficient when inputting into the GPU server for training, and for this reason, we proposed to perform resize operation to compress the vertebral images after the resampling operation, but after several experiments, we found that the combination of resampling and resize operation. However, after several experiments, we found that the combination of resampling and resize operation resulted in a serious loss of the original data information and the model could not learn the foreground region information.To reduce the loss of original data information, we improved the resize operation to a crop operation but the random crop led to a serious imbalance in the foreground data categories, and the model could only fit the local features of the random categories during training, resulting in some vertebrae with less category information could not be learned. Finally, we compressed the resampled data directly to a fixed size without changing the voxel spacing to allow the network model to learn all the category information in each training session.

The CT values of the vertebrae in the local vertebral dataset range from 100 to 3000, and background CT values of peripheral tissues are relatively small. Therefore, we used the threshold segmentation method to preserve the bone tissue in the original dataset, and cut out the irrelevant background, retaining a size of 160×160×96 vertebrae image. The resampling method of linear interpolation was used to unify the image resolutions of different scanning devices into a voxel space of $\boldsymbol{1mm} \times \boldsymbol{1mm} \times \boldsymbol{1mm}$, and the vertebral labels were resampled to a voxel space of $\boldsymbol{1mm} \times \boldsymbol{1mm} \times \boldsymbol{1mm}$ by nearest neighbor interpolation. The resampled image is denoised by bilateral filtering to ensure the detailed features of the vertebral edge contour, and the contrast of the vertebral image is enhanced by window adjustment. After many tests, we selected the window width of 1100Hu and the window level of 550Hu. Due to the slow training process of the Transformer structure, to improve the hyperparameter fitting rate, the z-score normalization method is used to normalize the input data to [0,1]. In the vertebral Edge Segmentation module, to obtain the real edge label of the vertebra, we used the Canny operator to perform contour segmentation on the ground-truth labels of vertebrae, used the non-maximum value to suppress the width of the edge voxel, and realized the detection and connection of the vertebra outline through the double threshold.

Due to the relatively high computational complexity of the Global Transformer structure, using the original spine dataset to input into the Global Transformer structure would lead to a continuous increase in model training time. Because we preserve the global position information of the original image by pre-experimentation, the original image and labels of the global vertebrae dataset are resampled to a uniform resolution of $\boldsymbol{1mm} \times \boldsymbol{1mm} \times \boldsymbol{1mm}$, voxel space, and then the superficial information of the vertebrae is extracted by two convolution operations (k=3) while changing the size of the image, which is processed by relative position encoding(RPE) and input into the Global Transformer structure to learn the superficial information of vertebrae.

TABLE 2. Training information for 8 vertebral segmentation networks.

Methods	Structure	VerSe 20 Input Size ($h \times w \times c$)	CTSpine1K Input Size ($h \times w \times c$)
Chen D[6]	3D U-net+3D ResNet50	128 × 160 × 96	160 × 160 × 96
Payer C[6]	3D U-net	128 × 160 × 96	160 × 160 × 96
Zhang A[6]	V-Net	128 × 160 × 96	160 × 160 × 96
nnU-net[36]	3D U-net	512 × 512 × 512	512 × 512 × 500
RUnT	ResUNet+Transformer	128 × 160 × 96	160 × 160 × 96
RUnT-Eg	ResUNet+Transformer+Edge	128 × 160 × 96	160 × 160 × 96
RUnT-GT	ResUNet+Transformer+GT	128 × 160 × 96	160 × 160 × 96
RUnT-Eg&GT	ResUNet+Transformer+Eg+GT	128 × 160 × 96	160 × 160 × 96

TABLE 3. RUnT-Eg & GT achieved the highest mean DSC (%) for a single spine on the CTSpine1K dataset.

Methods	Cervical (C1-C7)	Thoracic (T1-T12)	Lumbar (L1-L6)	Label 25	Average
Chen D[6]	89.7	90.6	79.2	70.9	87.8
Payer C[6]	84.5	87.6	72.2	71.2	83.0
Zhang A[6]	86.0	88.2	75.1	75.1	84.5
nnU-net[36]	87.2	89.9	**80.4**	74.5	86.9
RUnT	88.5	89.6	77.5	**85.5**	86.4
RUnT-Eg	87.7	90.4	75.6	79.8	87.5
RUnT-GT	82.1	83.7	72.8	75.2	80.6
RUnT-Eg&GT	**91.5**	**91.8**	78.0	81.5	**88.4**

TABLE 4. RUnT-Eg & GT obtained the minimum HD95 (mm) of a single vertebra in the CTSpine1K dataset.

Methods	Cervical (C1-C7)	Thoracic (T1-T12)	Lumbar (L1-L6)	Label 25	Average
Chen D[6]	**1.35**	3.18	9.78	20.75	4.25
Payer C 6	2.73	3.07	12.96	35.45	5.35
Zhang A[6]	2.40	3.92	10.32	25.47	5.03
nnU-net[36]	3.21	3.17	12.40	19.84	5.40
RUnT	1.58	3.29	**8.60**	10.78	4.08
RUnT-Eg	1.62	**2.58**	9.57	9.73	3.99
RUnT-GT	2.20	3.50	10.22	17.54	4.75
RUnT-Eg&GT	1.43	2.86	8.76	**8.97**	**3.88**

C. EXPERIMENTAL DETAILS

In order to verify the segmentation performance of the network on the vertebrae, using 8 kinds of segmentation networks were to train and test the segmentation performance of the CTSpine1K dataset and VerSe series dataset, including 4 state-of-the-art vertebral segmentation networks (namely Chen D [6], Payer C [6], Zhang A [6], nnU-net [36]), and our proposed Residual U-net combined with Transformer model (hereinafter referred to as RUnT) and its 3 variants. The information of the 8 segmentation is shown in Table 2.

IV. RESULTS AND ANALYSIS

A. SEGMENTATION RESULT ANALYSIS

1) SEGMENTATION RESULTS OF CTSPINE1K DATASET

To shorten the training and inference time of the RUnT model, we divide the CTSpine1K dataset into a training set, a test set and a validation set according to 3:1:1 by randomly selecting 300 cases as a small-scale pre-training dataset. We use the population-based training (PBT) training method to initialize the random hyperparameters, perform information interaction and strategy optimization during the parallel training of the model, replicate the weights according to the obtained optimal hyperparameter training combinations and add random noise for iterative training. The Verse 20 dataset is trained in parallel based on the training results of other segmentation models as the initialized hyperparameters.

We trained on the CTSpine1K training set and verify the vertebral segmentation performance on the test set, using an A100-PCIE-40GB device for training, the optimizer uses AdamW, the initial lr is 0.01, the default weight_decay is 0.001, and using the evaluation indicators DSC (%) and HD95 (in mm) described in Section II-E to evaluate the segmentation results. The segmentation results of the 8 networks are presented in Table 3 and Table 4. Among them, RUnT-Eg & GT obtained the highest average DSC score of 88.4 among all networks, Compared with the current highest segmentation model Chen D [6], it has increased by 0.6%, and has increased by 14.9% in the segmentation of special label 25, which improves the robustness of the model. In the HD95 (in mm) score, the RUnT-Eg & GT model shortened the mean HD95 distance by 0.37 and improved it by 6% over the current optimal model Chen D [6], achieving optimal performance in the mean DSC score in the cervical (C1-C7) and thoracic (T1-T12) regions. In contrast, in the lumbar spine (L1-L6) region due to the voxel fusion of the L6 vertebrae with the sacral spine, it is relatively difficult for the training of label 25 to achieve complete contour segmentation, resulting in a relatively low DSC for the RUnT-Eg & GT model in the lumbar region. The experimental results show that the variant model RUnT without the Global Transformer module and the edge detection module Eg outperforms all models in the lumbar spine test, and we will elaborate on the findings in Section IV-B of the ablation experiment. In Figure 6 we show six typical samples of the model RunT-Eg&GT with the highest mean DSC scores in the experiment, with the six samples mainly targeting the thoracic and lumbar spine sites. In general, compared with the four models of Chen D [6] Payer C [6] Zhang A [6] nnU-net [36], the DSC score and HD95(in mm) of our proposed network have been improved by an average of 3% and 21.5%, proving that our network outperforms other segmentation networks on the CTSpine1K dataset.

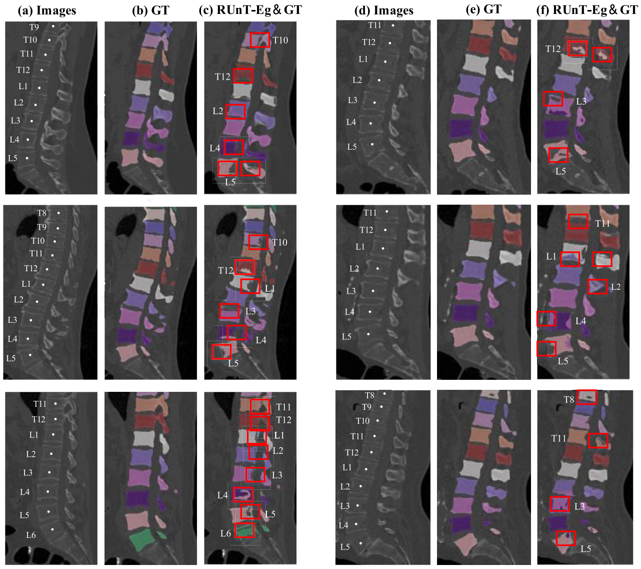


FIGURE 6. Six typical samples segmented by the RUnT-Eg & GT model in the CTSpine1K dataset, with red boxes representing regions different from GT (False-Positive regions or False-Negative regions). The box indexes represent the names of the vertebrae in the spine where the FN or FP areas are located.

TABLE 5. RUnT-Eg & GT achieved a DSC (%) on par with the current state-of-the-art segmenta-tion model in the VerSe20 dataset.

Methods	Cervical (C1-C7)	Thoracic (T1-T12)	Lumbar (L1-L6)	Label 25	Average
Chen D[6]	**80.7**	85.9	74.3	66.7	**81.7**
Payer C[6]	76.6	80.3	68.4	19.7	76.4
Zhang A[6]	73.1	79.6	69.2	15.7	75.3
nnU-net[36]	73.1	81.6	70.4	14.0	76.5
RUnT	75.5	82.2	71.7	52.2	77.8
RUnT-Eg	76.6	83.9	73.9	64.8	79.5
RUnT-GT	68.9	79.4	70.3	48.9	74.3
RUnT-Eg&GT	78.4	**86.6**	**74.7**	**64.9**	81.5

2) SEGMENTATION RESULTS OF VERSE 20 DATASET

In order to verify the robustness and universality of the network, we used the VerSe 20 dataset to verify the robustness of the model, and used the same equipment configured in Section I) to conduct experiments. The experimental segmentation results of the 8 networks are shown in Table 5 and Table 6. Due to the small amount of data in the VerSe 20 dataset and the foreground of some training samples being cropped, due to the improvement of 3D residual U-net structure and Global Transformer structure of the RUnT-Eg&GT model, leading there is currently no pretraining network parameters to learn. Therefore, on the test set of the VerSe 20 dataset, the RUnT-Eg&GT model did not achieve the optimal DSC score, which is close to the score of the current optimal segmentation model Chen D [6]. However, the RUnT-Eg&GT and its variant models achieved an average shortest distance of 4.86 on the HD95 distance, which verified that the RUnT-Eg&GT model has a good segmentation effect in small vertebral datasets, and tested the universality and robustness of our proposed model in vertebral segmentation.

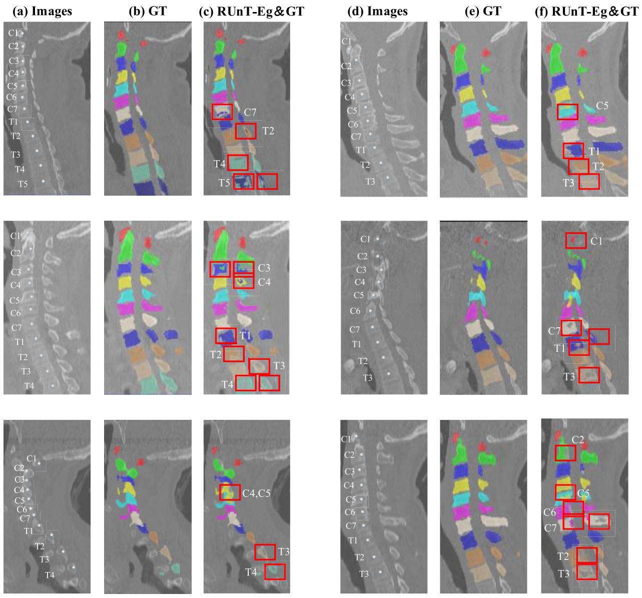


FIGURE 7. Six typical segmentation samples of the RUnT-Eg & GT model on the VerSe 20 dataset.

TABLE 6. RUnT-Eg & GT obtained the minimum HD95 (mm) of a single vertebra in the VerSe20 dataset.

Methods	Cervical (C1-C7)	Thoracic (T1-T12)	Lumbar (L1-L6)	Label 25	Average
Chen D[6]	**2.22**	4.15	10.55	48.74	5.14
Payer C[6]	2.83	**4.04**	14.30	40.78	6.16
Zhang A[6]	2.89	4.95	12.71	34.21	6.23
nnU-net[36]	2.41	4.76	13.06	33.36	6.10
RUnT	4.10	5.01	13.78	18.25	6.39
RUnT-Eg	3.71	4.60	11.61	13.71	5.62
RUnT-GT	6.33	5.34	12.55	25.65	6.33
RUnT-Eg&GT	3.28	4.17	**9.49**	**13.08**	**4.86**

The visual segmentation results of the VerSe 20 test set are fully displayed in Figure 7.

B. EXPERIMENTAL ABLATION STUDIES

In order to evaluate the effectiveness of the proposed module, we conducted ablation experimental research through 16 experiments, and the experimental results are clearly shown in Table 7. The backbone network uses the UnT model with a pure U-net structure combined with the Transformer structure and its three variants as a control experimental group.The experimental data of the residual U-net network and the three variant networks use the experimental data of Section I) and Section II).

In the CTSpine1K dataset test, the RUnT model is superior to the UnT model in terms of DSC score and HD95 distance, which can prove the effectiveness of the residual structure in extracting deep features of vertebrae. Compared with the RUnT model, RUnT-Eg increased the average DSC scores of the cervical spine (C1-C7) and thoracic spine (T1-T12) by 0.5% and 1.5%, respectively, and shortened the average HD95 dis-tance by 5.6% and 7.8%, respectively, with a 1.2%

TABLE 7. Analysis of experimental results of RUnT-Eg&GT ablation.

Model	CTSpine1K					VerSe 20				
	C1-C7	T1-T12	L1-L6	25	Average	C1-C7	T1-T12	L1-L6	25	Average
	DSC/HD95	DSC/HD95	DSC/HD95	DSC/HD95		DSC/HD95	DSC/HD95	DSC/HD95	DSC/HD95	
UnT	86.7/3.02	86.1/4.35	**81.7**/8.04	52.4/18.22	85.4/4.72	85.0/3.16	85.7/4.35	71.9/8.72	51.3/19.26	80.2/5.07
UnT-Eg	87.2/2.85	87.4/4.01	72.8/7.42	68.9/10.77	86.7/4.02	**86.9/2.99**	86.4/**3.01**	71.1/7.43	61.9/13.09	82.3/4.55
UnT-GT	84.3/3.19	85.9/3.87	76.8/6.44	65.8/12.74	83.2/5.76	83.3/3.84	86.9/4.17	72.8/6.47	64.7/13.08	80.9/4.63
UnT-Eg＆GT	86.3/2.59	88.9/3.61	75.4/**5.27**	72.7/10.87	86.9/4.01	86.1/3.30	**88.5**/3.70	74.1/**6.34**	73.8/**8.90**	**83.8/4.22**
RUnT	88.5/1.58	89.6/3.29	77.5/8.60	75.5/10.78	86.4/4.08	75.5/4.10	82.2/5.01	71.7/13.78	52.2/18.25	77.8/5.39
RUnT-Eg	87.7/1.62	90.4/**2.58**	75.6/9.57	75.8/9.73	87.5/3.99	76.6/3.71	83.9/4.60	73.9/11.61	64.8/13.71	79.5/5.62
RUnT-GT	82.1/2.20	83.7/3.50	72.8/10.22	79.2/17.54	80.6/4.75	68.9/6.33	79.4/5.34	70.3/12.55	48.9/25.65	74.3/5.33
RUnT-Eg＆GT	**91.5/1.43**	**91.8**/2.86	70.8/8.76	**81.5/8.97**	**88.4/3.88**	78.4/3.28	86.6/4.17	**74.7**/9.49	**64.9**/13.08	81.5/4.86

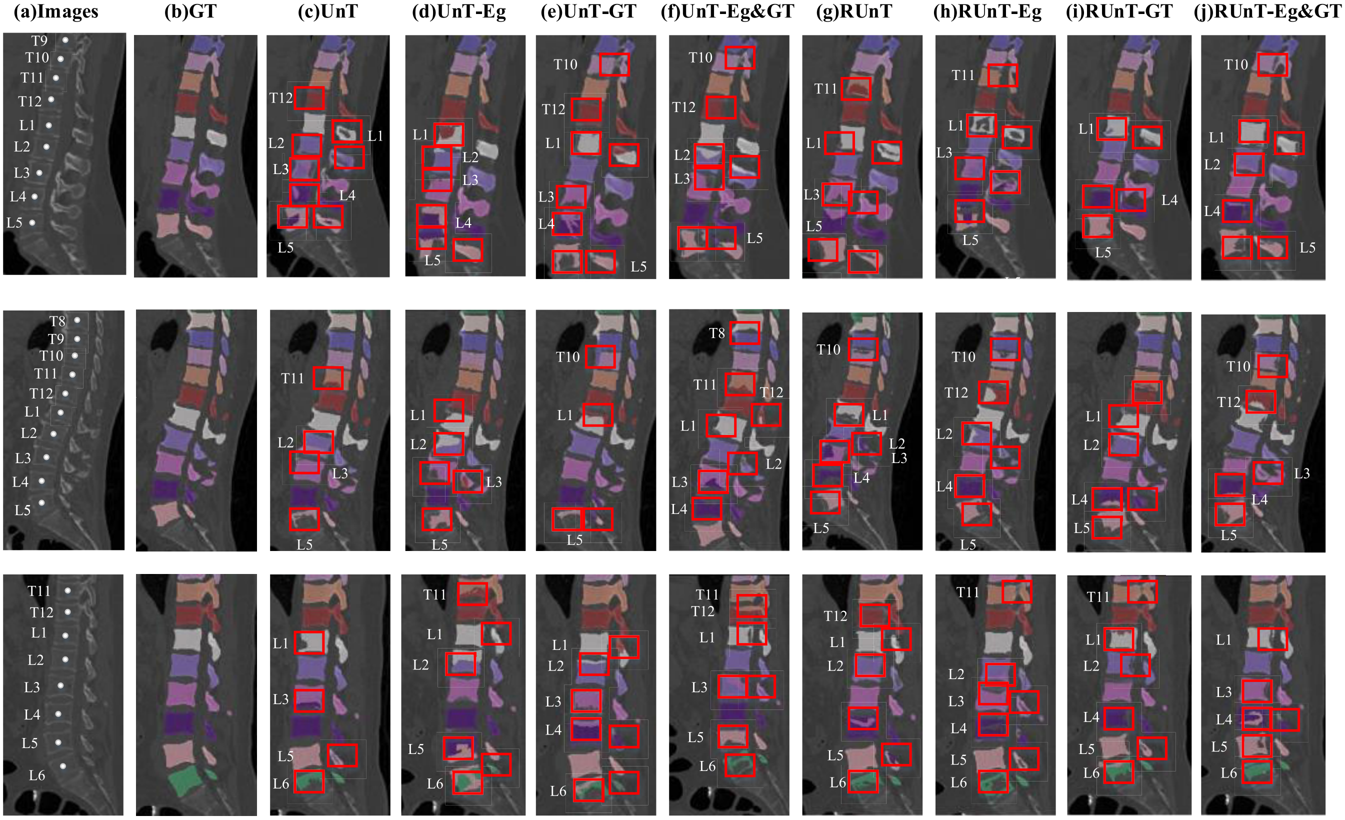


FIGURE 8. Three typical segmentation samples from the CTSpine1K test set were selected for the eight models, and the third sample contains the 25th label, with the red border indicating the FP and FN regions, and the visualization results show the improved performance of our proposed model RUnT-Eg&GT in vertebral segmentation.

increase in the overall segmentation average DSC score and a 2.2% reduction in the average HD95 distance, indicating that the Edge Segmentation model Eg extracted the shallow fusion information of the vertebra through multi-scale features is effective in vertebral segmentation. From the visualization results in Figure 8, (d)UnT-Eg segmented lumbar spine has a large number of false positive areas and false negative areas on the edge of the vertebrae compared to (c)UnT segmented results, and the edge of the vertebral contour is smoother, while the (g)RUnT model segmentation results compared with (h)RUnT-Eg model, there are more false negative areas at the edge of the outline of the vertebra, verifying that after adding the Edge Segmentation module, it is helpful to supervise and constrain the edge area of the vertebrae, ensuring the consistency of each vertebral segmentation.

For the UnT-GT model and the RUnT-GT model that include the Global Trans-former module, although they lag behind the UnT, UnT-Eg, RUnT, and RUnT-Eg mod-els in terms of DSC score and HD95 distance, but on the segmentation result of the special label 25(i.e.L6), the UnT-GT model and RUnT-GT perform well only second to the UnT-Eg&GT and RUnT-Eg&GT models. Therefore, it can be concluded that although the Global Transformer structure is not significantly helpful for deep feature extraction of vertebrae, the Global Transformer structure is conducive to extracting the global position information of vertebrae for the model to establish global dependencies, which can help determine the number of vertebrae and improve thesegmentation accuracy of special vertebrae. By combining the vertebral Edge Segmentation module and the Global Transformer model, the comprehensive model UnT-Eg&GT and RUnT-Eg&GT obtained the highest score of 86.9 and 88.4 respectively on the overall average DSC index of the vertebrae, and obtained the shortest HD95 distances of 4.01 and 3.88.

On the VerSe 20 dataset, due to the difference in data foreground clipping and data volume, the average DSC score of the RUnT model compared with the UnT model decreased by 2.9%, and the HD95 distance increased by 6.7%, indicating that the U-net structure has a stronger ability to extract deep features and a larger area of local receptive field than the residual U-net structure in the small dataset. However, the UnT-Eg and RUnT-Eg models including the Edge Segmentation module still have higher DSC scores and shorter HD95 distances than the UnT model and the RUnT model, indicating the effectiveness of the vertebral edge module in improving the accuracy of vertebral segmentation. The segmentation results of Label 25 demonstrate the effectiveness of the Global Transformer module in extracting location information. While the comprehensive model UnT-Eg&GT and RUnT-Eg&GT, compared with the CTSpine1K dataset segmentation test results, have decreased by 3.6% and 8.4%, but are still better than other models.

Overall, we believe that although the improved segmentation accuracy is limited, it still demonstrates the generalization and robustness of our model in vertebral seg-mentation.

V. CONCLUSION

Our proposed multi-label vertebral segmentation network based on Transformer and residual U-net structure was trained and tested on CTSpine1K and VerSe 20 datasets, and the experimental results demonstrated that our proposed segmentation network outperformed current publicly available segmentation networks in terms of accuracy of multi-label vertebral segmentation. The depth features of the vertebrae are extracted through the residual structure, using multi-scale shallow information fusion to extract the boundary contour information of vertebrae, using the clipped local spine dataset and the global spine dataset to input the Transformer structure respectively to fuse the local deep feature of the vertebra and the global shallow position information of the vertebra, after fusion and splicing with vertebral boundary contour features to achieve the consistency of vertebral segmentation.

In our work, we found that the training cost of the experimental network proposed in this paper is relatively high, due to the sample diversity of the CTSpine1K dataset leads to a relatively long inference time for the network in the training process, and the segmentation accuracy is not significantly improved compared to the VerSe series dataset. Due to the reason of the network architecture, the pretrained Transformer model cannot be used, so the hyperparameter fitting speed is too slow when calculating the attention distribution of the vertebrae compared to the pretraining network of ImageNet-1K, ImageNet-22K, and Synapse medical datasets, inference time is significantly increased.

Future work is mainly aimed at improving the MSA mechanism in Transformer, reducing the computational complexity and the number of parameters, and studying the impact of pretraining based on general datasets and medical datasets on the accuracy of vertebral segmentation.

ACKNOWLEDGMENT

(Hao Xu and Jianlan Yang contributed equally to this work.)

REFERENCES

[1] J. Yao, J. E. Burns, H. Munoz, and R. M. Summers, "Detection of vertebral body fractures based on cortical shell unwrapping," in *Proc. 15th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, Nice, France. Berlin, Germany: Springer, Oct. 2012, pp. 509–516.

[2] F. Altaf, A. Gibson, Z. Dannawi, and H. Noordeen, "Adolescent idiopathic scoliosis," *Bmj*, vol. 346, p. 2508, Apr. 2013.

[3] D. C. Howlett, K. J. Drinkwater, N. Mahmood, J. Illes, J. Griffin, and K. Javaid, "Radiology reporting of osteoporotic vertebral fragility fractures on computed tomography studies: Results of a U.K. National audit," *Eur. Radiol.*, vol. 30, no. 9, pp. 4713–4723, Sep. 2020.

[4] T. M. Emch and M. T. Modic, "Imaging of lumbar degenerative disk disease: History and current state," *Skeletal Radiol.*, vol. 40, no. 9, pp. 1175–1189, Sep. 2011.

[5] M. Mediouni, D. R. Schlatterer, H. Madry, M. Cucchiarini, and B. Rai, "A review of translational medicine. The future paradigm: How can we connect the orthopedic dots better?" *Current Med. Res. Opinion*, vol. 34, no. 7, pp. 1217–1229, Jul. 2018.

[6] A. Sekuboyina et al., "VerSe: A vertebrae labelling and segmentation benchmark for multi-detector CT images," *Med. Image Anal.*, vol. 73, Oct. 2021, Art. no. 102166.

[7] T. Klinder, J. Ostermann, M. Ehm, A. Franz, R. Kneser, and C. Lorenz, "Automated model-based vertebra detection, identification, and segmentation in CT images," *Med. Image Anal.*, vol. 13, no. 3, pp. 471–482, Jun. 2009.

[8] R. Korez, B. Ibragimov, B. Likar, F. Pernuš, and T. Vrtovec, "A framework for automated spine and vertebrae interpolation-based detection and model-based segmentation," *IEEE Trans. Med. Imag.*, vol. 34, no. 8, pp. 1649–1662, Aug. 2015.

[9] D. Štern, B. Likar, F. Pernuš, and T. Vrtovec, "Parametric modelling and segmentation of vertebral bodies in 3D CT and MR spine images," *Phys. Med. Biol.*, vol. 56, no. 23, pp. 7505–7522, Dec. 2011.

[10] B. Ibragimov, B. Likar, F. Pernuš, and T. Vrtovec, "Shape representation for efficient landmark-based segmentation in 3-D," *IEEE Trans. Med. Imag.*, vol. 33, no. 4, pp. 861–874, Apr. 2014.

[11] I. Castro-Mateos, J. M. Pozo, M. Pereañez, K. Lekadir, A. Lazary, and A. F. Frangi, "Statistical interspace models (SIMs): Application to robust 3D spine segmentation," *IEEE Trans. Med. Imag.*, vol. 34, no. 8, pp. 1663–1675, Aug. 2015.

[12] S. Kadoury, H. Labelle, and N. Paragios, "Automatic inference of articulated spine models in CT images using high-order Markov random fields," *Med. Image Anal.*, vol. 15, no. 4, pp. 426–437, Aug. 2011.

[13] S. Kadoury, H. Labelle, and N. Paragios, "Spine segmentation in medical images using manifold embeddings and higher-order MRFs," *IEEE Trans. Med. Imag.*, vol. 32, no. 7, pp. 1227–1238, Jul. 2013.

[14] A. Suzani, A. Rasoulian, A. Seitel, S. Fels, R. N. Rohling, and P. Abolmaesumi, "Deep learning for automatic localization, identification, and segmentation of vertebral bodies in volumetric MR images," *Proc. SPIE*, vol. 9415, Mar. 2015, Art. no. 941514.

[15] C. Chu, D. L. Belavý, G. Armbrecht, M. Bansmann, D. Felsenberg, and G. Zheng, "Fully automatic localization and segmentation of 3D vertebral bodies from CT/MR images via a learning-based method," *PLoS ONE*, vol. 10, no. 11, Nov. 2015, Art. no. e0143327.

[16] R. Korez, B. Likar, F. Pernuš, and T. Vrtovec, "Model-based segmentation of vertebral bodies from MR images with 3D CNNs," in *Proc. 19th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Athens, Greece. Cham, Switzerland: Springer, Oct. 2016, pp. 433–441.

[17] A. Sekuboyina, A. Valentinitsch, J. S. Kirschke, and B. H. Menze, “A localisation-segmentation approach for multi-label annotation of lumbar vertebrae using deep nets,” 2017, *arXiv:1703.04347*.

[18] R. Janssens, G. Zeng, and G. Zheng, “Fully automatic segmentation of lumbar vertebrae from CT images using cascaded 3D fully convolutional networks,” in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 893–897.

[19] N. Lessmann, B. van Ginneken, and I. Išgum, “Iterative convolutional neural networks for automatic vertebra identification and segmentation in CT images,” *Proc. SPIE*, vol. 10574, Mar. 2018, Art. no. 1057408.

[20] N. Lessmann, B. van Ginneken, P. A. de Jong, and I. Išgum, “Iterative fully convolutional neural networks for automatic vertebra segmentation and identification,” *Med. Image Anal.*, vol. 53, pp. 142–155, Apr. 2019.

[21] M. Vania, D. Mureja, and D. Lee, “Automatic spine segmentation from CT images using convolutional neural network via redundant generation of class labels,” *J. Comput. Des. Eng.*, vol. 6, no. 2, pp. 224–232, Apr. 2019.

[22] C. Payer, D. Štern, H. Bischof, and M. Urschler, “Integrating spatial configuration into heatmap regression based CNNs for landmark localization,” *Med. Image Anal.*, vol. 54, pp. 207–219, May 2019.

[23] A. Nazir, M. N. Cheema, B. Sheng, P. Li, H. Li, G. Xue, J. Qin, J. Kim, and D. D. Feng, “ECSU-Net: An embedded clustering sliced U-Net coupled with fusing strategy for efficient intervertebral disc segmentation and classification,” *IEEE Trans. Image Process.*, vol. 31, pp. 880–893, 2022.

[24] Y. Huang, A. Uneri, C. Jones, X. Zhang, M. D. Ketcha, N. Aygun, P. A. Helm, and J. H. Siewerdsen, “3D vertebrae labeling in spine CT: An accurate, memory-efficient (Ortho2D) framework,” *Phys. Med. Biol.*, vol. 66, no. 12, Jun. 2021, Art. no. 125020.

[25] S. Pang, C. Pang, Z. Su, L. Lin, L. Zhao, Y. Chen, Y. Zhou, H. Lu, and Q. Feng, “DGMSNet: Spine segmentation for MR image by a detection-guided mixed-supervised segmentation network,” *Med. Image Anal.*, vol. 75, Jan. 2022, Art. no. 102261.

[26] Z. Wu, G. Xia, X. Zhang, F. Zhou, J. Ling, X. Ni, and Y. Li, “A novel 3D lumbar vertebrae location and segmentation method based on the fusion envelope of 2D hybrid visual projection images,” *Comput. Biol. Med.*, vol. 151, Dec. 2022, Art. no. 106190.

[27] J. Zhao, L. Sun, X. Zhou, S. Huang, H. Si, and D. Zhang, “Residual-atrous attention network for lumbosacral plexus segmentation with MR image,” *Comput. Med. Imag. Graph.*, vol. 100, Sep. 2022, Art. no. 102109.

[28] D. Meng, E. Boyer, and S. Pujades, “Vertebrae localization, segmentation and identification using a graph optimization and an anatomic consistency cycle,” *Comput. Med. Imag. Graph.*, vol. 107, Jul. 2023, Art. no. 102235.

[29] 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*.

[30] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[31] S. F. Qadri, H. Lin, L. Shen, M. Ahmad, S. Qadri, S. Khan, M. Khan, S. S. Zareen, M. A. Akbar, M. B. B. Heyat, and S. Qamar, “CT-based automatic spine segmentation using patch-based deep learning,” *Int. J. Intell. Syst.*, vol. 2023, pp. 1–14, Mar. 2023.

[32] R. Tao, W. Liu, and G. Zheng, “Spine-transformers: Vertebra labeling and segmentation in arbitrary field-of-view spine CTs via 3D transformers,” *Med. Image Anal.*, vol. 75, Jan. 2022, Art. no. 102258.

[33] X. You, Y. Gu, Y. Liu, S. Lu, X. Tang, and J. Yang, “EG-Trans3DUNet: A single-staged transformer-based model for accurate vertebrae segmentation from spinal ct images,” in *Proc. IEEE 19th Int. Symp. Biomed. Imag. (ISBI)*, Mar. 2022, pp. 1–5.

[34] Y. Deng, C. Wang, Y. Hui, Q. Li, J. Li, S. Luo, M. Sun, Q. Quan, S. Yang, Y. Hao, P. Liu, H. Xiao, C. Zhao, X. Wu, and S. K. Zhou, “CTSpine1K: A large-scale dataset for spinal vertebrae segmentation in computed tomography,” 2021, *arXiv:2105.14711*.

[35] H.-Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, and Y. Yu, “NnFormer: Interleaved transformer for volumetric segmentation,” 2021, *arXiv:2109.03201*.

[36] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, “NnU-Net: A self-configuring method for deep learning-based biomedical image segmentation,” *Nature Methods*, vol. 18, no. 2, pp. 203–211, Feb. 2021.

HAO XU was born in Jiangsu, China, in 1999. He received the B.S. degree in the Internet of Things engineering from Xuzhou Medical University, China, in 2021. He is currently pursuing the M.S. degree in biomedical engineering with the Gansu University of Traditional Chinese Medicine, China. His main research interests include medical image processing, convolutional neural networks, and spine CT segmentation.

XINXIN CUI was born in Gansu, China, in 2000. She received the B.S. degree in medical information engineering from the Gansu University of Traditional Chinese Medicine, in 2022, where she is currently pursuing the M.S. degree in biomedical engineering. Her main research interests include medical image processing and deep learning.

CHAOFAN LI received the B.S. and M.S. degrees in medical informatics from Xuzhou Medical University, in 2019 and 2022, respectively. Since 2022, he has been with the Yancheng Third People’s Hospital, Yancheng, as a Data Manager. His research interests include medical informatics and artificial intelligence.

ZHENYU TIAN was born in Hebei, China in 1997. She received the bachelor’s degree in communication engineering from Tangshan University, in 2020. She is currently pursuing the master’s degree in biomedical engineering with the School of Information Engineering, Gansu University of Traditional Chinese Medicine, China. Her research interests include medical image processing and deep learning.

JING LIU was born in Shandong, China, in 1997. She received the B.S. degree in electronic information engineering from Qingdao Agricultural University, in 2021. She is currently pursuing the M.S. degree in biomedical engineering with the Gansu University of Traditional Chinese Medicine. Her main research interests include medical information and intelligent medicine. Her awards and honors include the National Motivation Scholarship and the Outstanding Student Scholarship of Qingdao Agricultural University.

JIANLAN YANG was born in Fujian, China, in 1974. He received the degree in health career management from the School of Public Health, Latrobe University, Australia. He is an associate professor and the master’s degree supervisor. He serves as the Deputy Director of the Chinese Medicine and Health Informatics Committee, Chinese Society of Medical Informatics; an Executive Member of the Chinese Medicine Informatics Committee, Chinese Society of Health Informatics; and the Vice President of the Cloud Health Branch, Chinese Society of Chinese Medicine Informatics. His research interests include health information data mining and medical image recognition and application.