IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 24 January 2024, accepted 27 May 2024, date of publication 31 May 2024, date of current version 7 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3407833

RESEARCH ARTICLE

Multi-Scale Hybrid Attention Convolutional Neural Network for Automatic Segmentation of Lumbar Vertebrae From MRI

JING LIU[1], YUEE ZHOU[1], XINXIN CUI[1], FENGQING JIN[1], GUODONG SUO[1], HAO XU[1], AND JIANLAN YANG[1,2]

[1]School of Information Engineering, Gansu University of Traditional Chinese Medicine, Lanzhou 730000, China
[2]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 Quanzhou Orthopedic Hospital, Fujian, China.

ABSTRACT Due to the high incidence of lumbar vertebral lesions causing spondylolisthesis and intervertebral disc protrusion, lumbar spine (vertebrae and intervertebral discs) MRI image segmentation can provide effective clinical information for the initial diagnosis and early treatment of current lumbar spine-related diseases. However, in MRI images, there is a significant overlap and similarity in features between the vertebral bones and intervertebral discs within the lumbar spine. Therefore, the effective identification and segmentation of each vertebra and intervertebral disc in the lumbar spine pose a significant challenge. We propose a lumbar spine MRI segmentation model based on the 3D Residual U-Net, incorporating boundary segmentation structures and a hybrid attention mechanism. The model achieves boundary-constrained segmentation of individual vertebrae and intervertebral discs by integrating the boundary segmentation module. Additionally, it utilizes a hybrid attention module based on convolutional attention and self-attention mechanisms for multiscale feature extraction in the lumbar spine. The proposed model is trained and validated using the publicly available datasets MRSpineSeg2021 and SpineSagT2Wdataset3. Experimental results demonstrate a significant improvement in segmentation performance, as measured by metrics such as the Dice similarity coefficient (DSC) and Hausdorff distance (HD). This validates the superiority and generalization performance of our proposed lumbar spine MRI.

INDEX TERMS Lumbar spine segmentation, convolutional neural network, vertebral bone boundary segmentation, hybrid attention mechanism.

I. INTRODUCTION

The lumbar spine is an essential component of the human skeletal system, serving as a foundational structure that carries out various critical functions, including protecting the nervous system, supporting body weight, and maintaining body balance [1]. It constitutes the axial skeleton of the human body, consisting of lumbar vertebrae (L1-L5), sacral vertebrae, and the coccyx. The lumbar spine is a complex structure comprising multiple vertebral bodies, intervertebral discs, vertebral arches, among other parts. The lumbosacral plexus forms a complex sensory and motor neural network around the lumbar and sacral vertebrae [2]. Spinal disorders resulting from lumbar vertebral and intervertebral disc pathologies include spinal trauma, spondylolisthesis, and neural foraminal stenosis [3]. Understanding the precise anatomical structure of the lumbar spine is a focal point in current medical research. In comparison to computed

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

tomography (CT) technology, which may display surrounding muscles and nerves indistinctly, magnetic resonance imaging (MRI) provides radiologists and clinical practitioners with higher spatial resolution and contrast pathology images [4], revealing both skeletal and soft tissue information of the lumbar spine. With the advancement of medical imaging informatics and computer-aided diagnosis (CAD) systems, lumbar spine segmentation plays a crucial role in the preliminary diagnosis of various spinal conditions. To assist radiologists and achieve rapid, stable, and accurate segmentation of lumbar spine MR images, providing valuable information for clinical pathological diagnosis, surgical planning, and postoperative assessment [5], it is necessary to address current challenges such as limited lumbar spine MRI datasets and slow segmentation model training speeds. Further in-depth research is needed to tackle the existing difficulties in lumbar spine MR segmentation. As shown in Figure 1, there is a high degree of similarity among the vertebrae and intervertebral discs in lumbar spine MRI, which makes it difficult to achieve comprehensive segmentation in the current segmentation task.

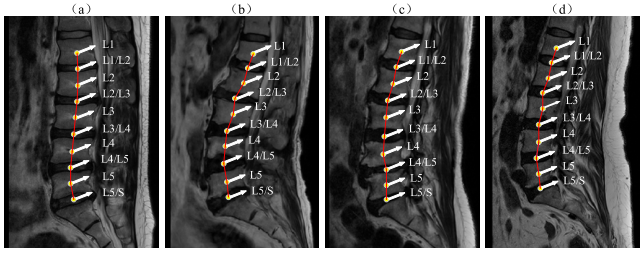


FIGURE 1. Sagittal MRI slices of the lumbar spine in Figures (a)-(d), in which the lumbar vertebrae L1-L5 and intervertebral bones overlap in the region of the L1/L2-L5/S junction and there is a high degree of similarity between the vertebrae and intervertebral discs.

A. RELATED RESEARCH WORK

Current existing research related to lumbar spine segmentation mainly includes segmentation methods based on digital images, segmentation methods based on machine learning theories, and research methods combining convolutional neural networks with Transformers.

1) SEGMENTATION BASED ON DIGITAL IMAGE TECHNOLOGY

Lumbar spine segmentation based on digital image technology involves fitting boundary curves for 2D or 3D regions of interest through signal analysis. Klinder et al. [6] proposed a fully automatic vertebral segmentation framework based on statistical shape models for spine image analysis. This framework is used for vertebral relative positioning and collecting morphological curve information. The segmented region includes lumbar vertebrae L1-L5 but lacks intervertebral discs and the vertebral canal. The framework utilizes spinal volume for curve reconstruction, achieves spine detection through generalized Hough transform, and identifies and segments vertebrae through multi-level processing. This method requires high-quality and extensive spine datasets for complex training.

In contrast, Korez et al. [7] introduced a statistical model-based automatic spine localization and vertebral segmentation model. They use interpolation techniques to effectively fill in pixel-deficient regions and employ morphological operations for initial spine segmentation. Individual vertebra detection is achieved through geometric positioning, and high-precision segmentation is ultimately realized using shape statistical models. This method can handle high-resolution spine CT images, providing segmentation results for the entire spine. However, pixel region interpolation techniques may introduce noise and false-positive regions, impacting segmentation accuracy in certain areas of interest. Ibragimov et al. [8] proposed a novel lumbar spine segmentation framework based on transportation theory and game theory. By incorporating the theory of target dominance during the object detection process, the computational cost on the cross-sectional plane of lumbar vertebrae and femoral heads was reduced by a factor of three. Castro-Mateos et al. [9] introduced an active contour model for overall spine boundary segmentation. This model utilizes the Statistical Image Model (SIM) to model intervertebral discs between adjacent vertebrae, calculating relative positional information of intervertebral discs and combining biological curve information to achieve vertebra segmentation. However, this method requires manual selection of the center of interest in the intervertebral disc and manual initialization of the spine contour curve, making it impractical for clinical application. Kadoury et al. [10] were the first to use Markov Random Fields to segment the entire spine into a set of intervertebral discs, introducing geometric characteristics between adjacent vertebrae. The measurement results are used as the intervertebral disc set to achieve consistent curve fitting for various regions of the vertebrae. Building on this research, Kadoury [11] employed manifold embedding and higher-order Markov Random Fields for the segmentation of multimodal spine vertebrae. This method not only maps foreground region pixels to a low-dimensional manifold space but also, compared to previous Markov Random Fields, locates and segments each vertebra through higher-order MRF. However, this model has a high complexity requiring a large number of high-order MRF models for fitting calculations and is sensitive to noisy data, making it prone to underfitting.

2) MACHINE LEARNING SEGMENTATION

With the maturation of machine learning theory, spine segmentation models based on machine learning theory have continuously evolved. Early segmentation models primarily focused on high-density bone regions, but for individual vertebrae recognition and segmentation of the spine, including the lumbar vertebrae L1-L5, they still lacked contrast in the less prominent intervertebral disc and vertebral canal regions. Glocker et al. [12] introduced a method for automatic

localization and recognition of spine vertebrae in CT scan images using regression forests and probabilistic graphical models for localization and detection. The visible part of the spine vertebrae is detected using regression forests, and precise localization and recognition are achieved by capturing the spine's geometric model. This method achieves an overall median localization error of less than 6mm and a recognition rate of 81% on 200 scan images. Building on previous research, Bromiley et al. [13] proposed the use of multiple sets of 2D locators for localization and recognition. The first set uses a random forest regressor to locate the spine in the coronal plane, and the second set of regressors identifies each vertebra in the coronal plane. In 2015, Suzani et al. [14] first introduced a lumbar spine segmentation region based on a multilayer perceptron. They conducted statistical analysis on voxel intensity, generated the original lumbar spine model, and iteratively approached the true lumbar spine parameters through local thresholds. Chu et al. [15] and colleagues used random forest regression to achieve vertebrae localization and detection. Simultaneously, they employed a hidden Markov model to generate voxel distribution probability maps, eliminating segmentation ambiguity caused by the high shape similarity between vertebrae.

3) BASED ON CONVOLUTIONAL NEURAL NETWORKS SEGMENTATION

Compared to traditional machine learning, convolutional neural networks (CNNs) based on deep learning possess stronger feature representation capabilities. They can automatically learn current regional features for classification and segmentation processing by designing improved structures according to the problem environment. The end-to-end medical segmentation model U-Net network [16], proposed in 2015 within convolutional neural networks, has strong feature learning abilities on small medical datasets. Therefore, it has been extended to a specialized segmentation network framework based on data type-driven modifications to the basic U-Net. In the field of lumbar spine segmentation, Fan et al. [17] utilized a 3D U-Net network for the automatic segmentation of the lumbosacral nerves, bones, and intervertebral discs in CT data from 31 patients. They employed 3D reconstruction techniques to simulate percutaneous endoscopic transforaminal discectomy (PETD) trajectory for intervertebral foramen shaping surgery simulation, assessing clinical surgical difficulty. Korez et al. [18] and colleagues employed a convolutional neural network (CNN) for feature learning and target segmentation in MR images of spinal vertebrae. They used combinations of convolutional and pooling layers for iteratively learning local features of spinal vertebrae. Through two fully connected layers, they mapped vertebral features to the mask on the spine for spinal vertebrae boundary segmentation. Sekuboyina et al. [19] proposed a two-stage network based on deep neural networks for labeling multi-label lumbar spines. In the first stage, the lumbar spine is defined through nonlinear regression using a multilayer perceptron (MLP). The second stage involves using a U-Net classification model for multi-classifying each vertebral bone in the localized lumbar spine region, achieving segmentation and labeling of the lumbar spine by learning local features in context. Nazir et al. [20] introduced ECSU-Net, a network combining embedded clustering slices with a fusion strategy, aiming to efficiently learn distinguishable features from a small amount of training data. This network, based on the 3D U-Net framework, introduces a fusion strategy to better capture the shape and structural information of intervertebral discs by merging information from multiple slices. However, this method results in lower training efficiency due to the need to calculate the loss for every two different vertebral segments. Pang et al. [21] proposed the DGMSNet network, which simultaneously trains and learns from both strong and weakly supervised datasets. This method uses a keypoint detection task to assist the lumbar spine segmentation task. It comprises a segmentation path and a detection path, where the segmentation path generates spine segmentation predictions, and the detection path generates heat map predictions for keypoints. Dynamic parameters in the detection path are generated by a detection-guided learning (DGL) and used as adaptive convolution kernels for extracting semantic information in the segmentation path. However, this method has some drawbacks, including not taking advantage of the low-cost benefits of keypoint detection annotated datasets and the fact that the auxiliary task only affects the main task in the feature space, not in the prediction space. Wu et al. [22] proposed a 3D lumbar spine localization and segmentation network (LVLS-HVPEE) based on 2D mixed visual projection image fusion envelopes. This network acquires the complete position of each lumbar vertebra in the coronal and sagittal planes by fusing visual projection images. Under the conditions of obtaining the 3D localization subspace of each vertebra, a 3D segmentation network based on spatial localization knowledge is introduced to achieve cervical spine segmentation. Zhao et al. [2] introduced the Residual-atrous Attention Network (RA2-Net) lumbar spine segmentation network model, which uses dilated convolutions to learn multiscale features with different dilation rates in the encoder. Additionally, it introduces scale attention blocks in the decoder to adaptively fuse features. Moreover, this method utilizes residual skip connections to combine features in the encoder with high-resolution spatial details with high-level contextual features, aiming to improve segmentation performance. Wang et al. [23] proposed a segmentation network MLKCA-Unet for spinal MRI images. In comparison to the traditional U-Net structure, which faces challenges in obtaining distant features due to the use of small convolutional kernels, the authors proposed improvements to the encoder's feature extraction capability by introducing multiscale large convolutional kernels and attention mechanisms. They captured features for different-sized feature maps using convolutional kernels of different sizes and used 1×1 convolutional kernels to reduce the computational load.

4) SEGMENTATION WITH COMBINATION OF CONVOLUTIONAL NEURAL NETWORKS AND TRANSFORMER

The emergence of the Transformer structure based on the self-attention mechanism signifies a revolutionary advancement in the field of natural language processing [24], but its impact is not confined to the text domain alone. Due to its outstanding sequential modeling capability and the flexibility of the self-attention mechanism, the Transformer structure has brought new insights and performance improvements to medical image segmentation tasks. In the context of locating and extracting features from regions of interest in medical images, ViT [25] introduces self-attention mechanism by transforming the image segmentation task into a sequence-to-sequence problem. This allows ViT to understand the context and relationships within the image globally, contributing to capturing complex relationships between different regions and enhancing the accuracy of medical image segmentation. However, the original ViT is sensitive to image sizes and cannot autonomously learn different-sized images. Given the diversity of medical datasets due to various collection devices and complex sample sources, the original ViT network cannot leverage its advantages in medical images. With the integration of window attention mechanism and hierarchical feature representation in the Swin Transformer network [26], there is a further improvement in performance in image classification and segmentation, while reducing the overall computational complexity of the network. The image segmentation model, combining features based on the Transformer structure and U-shaped network, has further advanced in the field of medical image. Tao et al. [27] propose a two-stage method for labeling and segmenting vertebrae. The first stage uses Spine-Transformers to treat the automatic labeling of vertebrae in any FOV spine CT scan as a one-to-one set prediction problem. This is achieved by designing a global loss and a lightweight Transformer architecture for unique prediction and learnable position embedding. The authors introduce the InSphere detector to replace traditional box detectors, providing better handling of volume direction changes. In the second stage, a single multi-task encoder-decoder network is trained for the refinement of central coordinates and segmentation of recognized vertebrae. You et al. [28] propose a single-stage network model (EG-Trans3DUNet) based on Transformer and U-Net for vertebral segmentation in spine CT images. By introducing an edge detection module and a Transformer-based global information module, the model addresses the issue of blurred vertebral boundaries in CT images while maintaining the segmentation consistency of each vertebra.

B. CONTRIBUTIONS

In summary, our main contributions can be summarized as follows:

(i) We propose a multi-scale feature extraction module inserted between the subsampling layers of the encoder to enhance the segmentation accuracy of lumbar vertebrae and intervertebral disc soft tissues. By introducing a three-branch structure for feature extraction, the upper branch is primarily for multi-scale feature extraction, the middle branch uses a residual structure to preserve shallow layer feature information, and the lower branch employs a mixed attention module. The mixed attention module leverages channel relationships to extract meaningful semantic information, while the spatial attention mechanism focuses on the position information of the features. The combination of these two attention mechanisms better captures crucial details and edge information in the lumbar vertebrae and intervertebral disc.

(ii) We propose a hybrid attention mechanism that combines channel attention, spatial attention, and self-attention based on a three-dimensional local volume. This mechanism inputs features suppressed by channel attention (irrelevant channel information) and features focused on the position information of the regions of interest, extracted by spatial attention, into the Transformer structure based on local three-dimensional volume self-attention. The combination of convolutional attention and self-attention effectively integrates fine-grained features from shallow layers and semantic features from deep layers, reducing the semantic gap between heterogeneous features and better capturing details and boundaries of small-scale variations.

(iii) To address the under-segmentation issue of lumbar vertebrae boundaries, we design an upper branch in the multi-branch structure for multi-scale edge feature extraction. This structure, based on convolutional neural networks, employs multiple upsampling deconvolution operations for scale normalization of features of different sizes. By controlling the number of feature channels using 1D special convolutional kernels, we concatenate the extracted multidimensional lumbar boundary feature maps to obtain features containing rich local and global information. The deep fusion of local features and overall information is achieved by constraining the boundary features of individual vertebrae, ensuring the consistency of lumbar segmentation.

II. LUMBAR SPINE MRI SEGMENTATION MODEL

A. NETWORK ARCHITECTURE

In the study of lumbar spine segmentation, we designed a network framework based on convolutional neural networks and a hybrid attention mechanism, as shown in Figure 2, for segmenting the lumbar vertebrae (L1-L5) and intervertebral discs (L1/L2-L5/S) in the human spine. The overall network framework consists of four main parts, employing an asymmetric structure including a contraction path for feature extraction and an expansion path for image recovery. The first part is the Residual U-Net encoder, where the lumbar spine MRI dataset is preprocessed and input into the Residual U-Net encoder. The encoder comprises four layers of repeated downsampling layers connected by our specially designed three-branch structure. The second part consists of a hybrid attention structure designed by combining

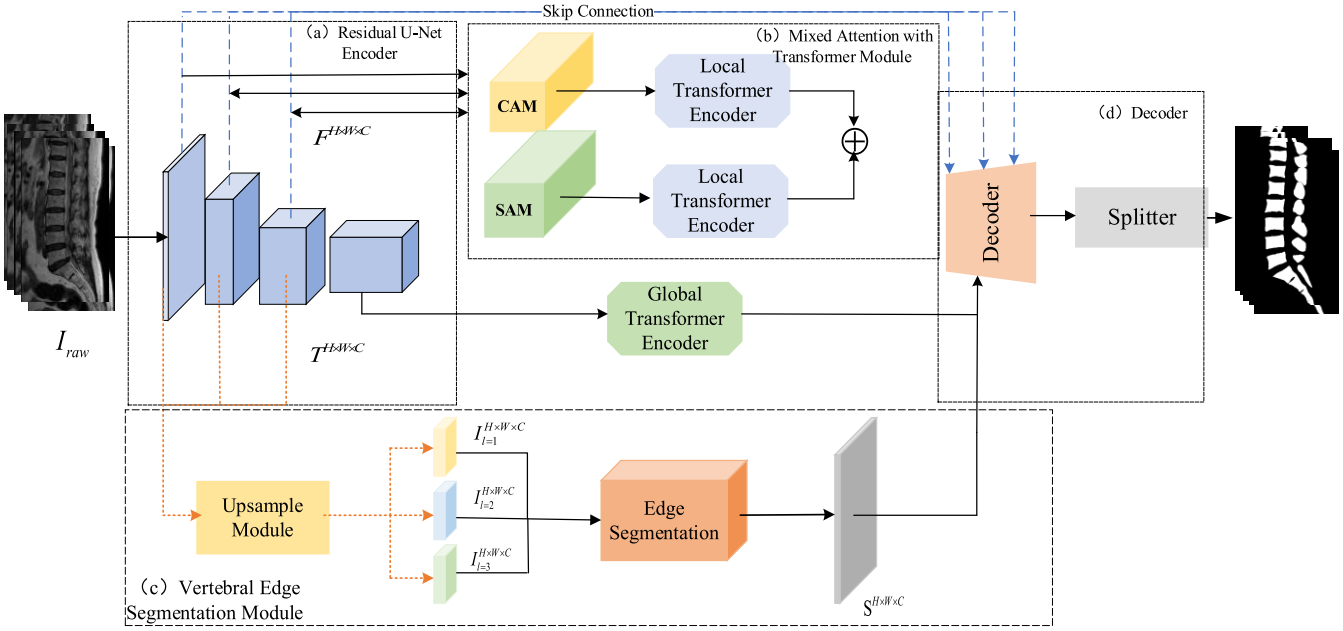


FIGURE 2. The overall segmentation network framework diagram is mainly divided into (a) Residual U-Net encoder, (b) Hybrid attention module, (c) Vertebral edge segmentation model, and (d) Decoder and classifier.

convolutional attention mechanism and a self-attention Transformer structure. The hybrid attention structure includes channel attention mechanism, spatial attention mechanism, and local volume-based self-attention mechanism. The channel and spatial attention mechanisms adopt a parallel dual-branch structure to suppress irrelevant channel information and enhance spatial position information from the extracted features of the downsampling layers. These two sets of features are then input into the local Transformer module based on local volume self-attention mechanism, which can better capture local information and details compared to traditional global attention. The third part is the decoder module, consisting of four layers of repeated upsampling layers that restore high-dimensional data to low-dimensional information. The features extracted by the downsampling layers in the encoder are input into each layer of the decoder through skip connections to achieve fusion of high-dimensional features and low-dimensional information. The fourth part is the boundary feature module, where a basic convolutional neural network is introduced. The features containing rich local details extracted by the first three downsampling layers are input separately, and boundary features are extracted using different convolutional kernels. Finally, these three sets of features are standardized in size through upsampling operations and fused with the decoder output features, containing rich local and global feature.

B. MULTI-SCALE FEATURE EXTRACTION MODULE

We have designed a multi-scale feature extraction module between the encoder's downsampling layers, as illustrated in Figure 3. This module consists of three branches. By copying the features $I_{l=i(i=1,2,3)}^{H\times W\times C}$ (where l represents the downsampling layer, $H \times W$ represents the height and width of the image, and C represents the feature channel dimension) extracted from the first three downsampling layers into three sets of features, they are simultaneously input into the three-branch structure. The upper branch structure is the boundary feature module. Since the features $I_{l=i(i=1,2,3)}^{H\times W\times C}$ extracted from the first three downsampling layers mainly focus on the shallow local details of the image, the first three layers of features are restored to the image size $H \times W \times C$ through upsampling operations to extract the boundary information of the lumbar vertebrae. Among the three branches, the middle branch and the lower branch input features are $F^{H\times W\times C}$ and $T^{H\times W\times C}$, respectively. The two branches consist of two different convolution attention mechanisms. The middle branch structure introduces a channel attention module based on channel attention, while the lower branch structure uses a spatial attention module based on spatial attention. To address the issue of lumbar spine feature loss due to repeated convolution operations in the downsampling layers, we introduce channel-domain and spatial-domain attention mechanisms to compensate for the detail loss caused by repeated convolution operations and further enhance the feature of the region of interest to obtain global positional information. The middle branch structure aggregates the features $F^{H\times W\times C}$ through max-pooling and global average pooling operations to generate two differentiated spatial context features F_{Gavg}^{C} and $F_{\max}^{C}$. The spatial features are compressed to obtain two spatial background information parameters, F_{Gavg}^{C} and $F_{\max}^{C}$, which are then input into a multi-layer perceptron (MLP) with shared channel weights to calculate the channel feature

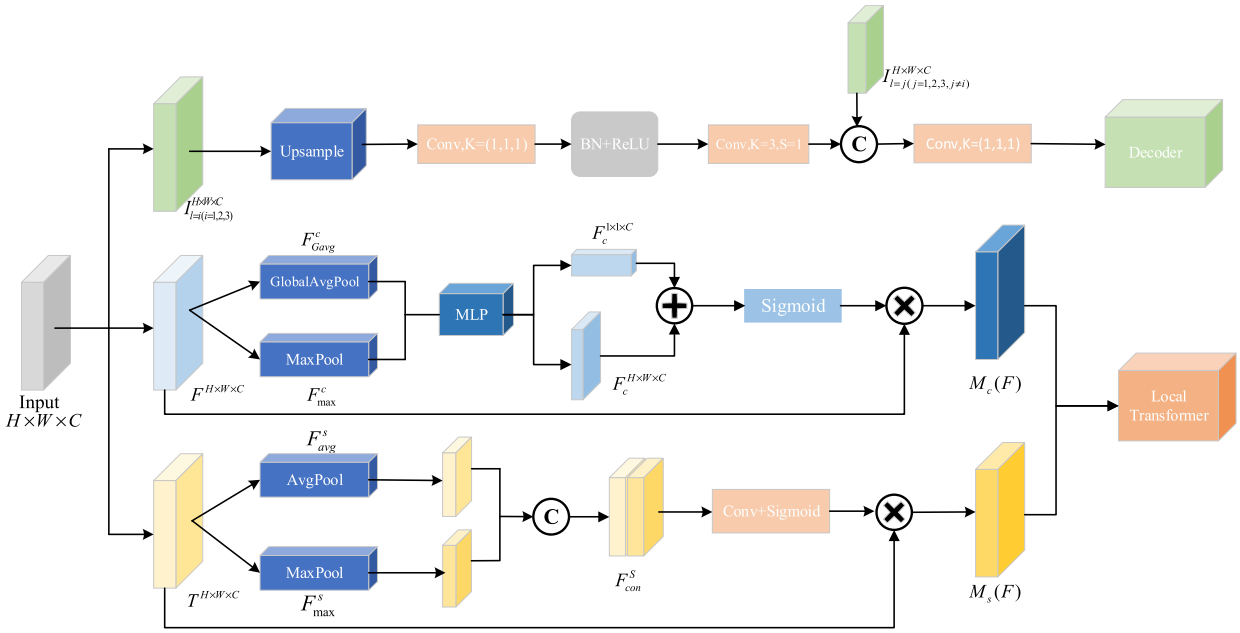


FIGURE 3. The multi-scale feature extraction module we propose is designed as a three-branch structure. The upper branch features are primarily input to the boundary segmentation module, while the middle and lower branch features are processed using a mixed attention mechanism.

attention maps $F_c^{1\times1\times C}$ and $F_c^{H\times W\times C}$. The two features are element-wise added, and the final channel feature $M_c(F)$ is obtained through the Sigmoid activation function. The channel attention calculation process is shown in Equation (1):

$$M_c(F) = \sigma(MLP(\text{GlobalAvgPool}(F)) + MLP(MaxPool(F)))$$
$$M_c(\text{F}) = \sigma\left(W_1\left(W_0\left(F_{Gavg}^c\right)\right) + W_1\left(W_0\left(F_{Gavg}^c\right)\right)\right) \quad (1)$$

where $W_0 \in \mathbb{R}^{C/r\times C}$ and $W_1 \in \mathbb{R}^{C\times C/r}$. The feature $T^{H\times W\times C}$ is input into the spatial attention model, and the structure of the spatial attention model is shown in Fig. 3. Compared with the channel attention mechanism that focuses on the channel information, the spatial attention mechanism concentrates on learning the pixel position information in the feature map, and the features F_{avg}^s and $F_{\max}^s$ are obtained by the serial computation of the maximal pooling and the average pooling. The features F_{avg}^s and $F_{\max}^s$ are spliced into F_{con}^s by channel splicing operation, and then the spatial attention feature F_{con}^s is dimensionalized to 1 dimension by convolution operation, and the spatial attention score map $M_s(F)$ is obtained by the Sigmoid activation function, and the computation process is shown in Eq. (4):

$$M_s(F) = \sigma\left(f^{7\times7}\left([AvgPool(F); MaxPool(F)]\right)\right)$$
$$M_s(F) = \sigma\left(f^{7\times7}\left(\left[F_{avg}^s; F_{\max}^s\right]\right)\right) \quad (2)$$

Due to the semantic differences between heterogeneous features, directly concatenating and fusing features cannot fit well. The introduction of channel-space attention effectively combines the features extracted from the first three layers of the encoder, removing noise and irrelevant organizational information. This allows the model to better differentiate and utilize features in both channel and spatial dimensions, enhancing the network's focus on channel-dimensional feature information. Finally, the extracted channel attention features and spatial attention features are input into the Local Transformer structure based on local volume self-attention mechanism. The self-attention mechanism captures relationships between different positions in sequences or images, the channel attention mechanism focuses on features between different channels, and the spatial attention mechanism captures important information in spatial dimensions. The hybrid use of these three attention mechanisms enables the acquisition of crucial information at multiple scales, contributing to a more comprehensive understanding of input data.

C. EDGE SEGMENTATION MODULE

The structure of the boundary segmentation module, as shown in Figure 4, addresses the challenge of accurate segmentation of the lumbar vertebral structure in medical images due to its complexity. The lumbar vertebrae exhibit a multi-level, intricate structure in the lumbar region, with interwoven structures such as intervertebral discs and intervertebral foramina. This complexity makes accurate segmentation of the lumbar vertebrae a challenging task in medical imaging. The lumbar vertebral structure is a challenge in the field of medical image processing due to its complexity. Vertebrae present a multi-level, interlaced longitudinal structure in the lumbar region, and small structures such as intervertebral discs and intervertebral foramina are intertwined with each other, which makes accurate segmentation of lumbar vertebral bones in medical images a rather difficult task. In order to solve the

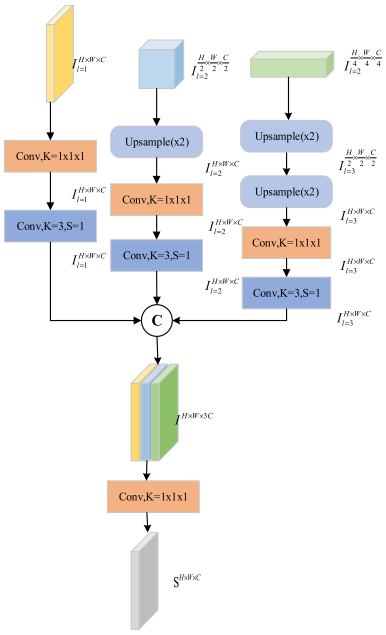


FIGURE 4. Boundary Segmentation Module, where three features ($I_{l=1}^{H\times W\times C}$, $I_{l=2}^{\frac{H}{2}\times\frac{W}{2}\times\frac{C}{2}}$, $I_{l=3}^{\frac{H}{4}\times\frac{W}{4}\times\frac{C}{4}}$) are derived from the first three downsampling layers of the Residual U-Net encoder. The features undergo convolutional scale transformation through upsampling.

problem of under-segmentation of lumbar vertebral bone boundaries, we propose a CNN-based boundary segmentation module. The structure of the boundary segmentation module is shown in Fig. 4, where the different scale features $I_{l=i(i=1,2,3)}^{H\times W\times C}$ extracted from the first three layers of downsampling of the encoder containing rich local details of the image are input into the module, and according to the different sizes of the features the size is unified as $H \times W \times C$ using multiple inverse convolution operations, the number of the feature channels is controlled by the special convolution kernel of $1 \times 1 \times 1$,, and the convolution kernel of $3 \times 3 \times 3$ is taken to the extraction of the localized information of the shallow vertebral bone containing the different scales, and the extracted The multi-dimensional spine vertebrae boundary feature map is spliced with feature channels, and the spliced multi-dimensional features are fused again with the $1 \times 1 \times 1$ convolution kernel, due to the relatively large image resolution $H \times W$, the first downsampling layer has less loss of details, so the feature $I_{l=1}^{H\times W\times C_1}$ channel C_1 is controlled to be 32, while the second $I_{l=2}^{H\times W\times C_2}$ and third $I_{l=1}^{H\times W\times C_3}$ downsampling layers contain relatively less information, so the channel dimensions C_2 and C_3 are set to be 64 and 128, and the lumbar vertebrae boundary features are extracted with the final output of the decoder. The extracted lumbar spine boundary features are fused with the final output features of the decoder, and the boundary features are utilized as potential constraints to ensure the consistency of segmentation of each vertebra in the spine. The calculation process is shown in Equation (5). Where $f^{3\times3\times3}$ denotes a convolution operation with a convolution kernel size of $3 \times 3 \times 3$. The calculation process is shown in Equation (3).

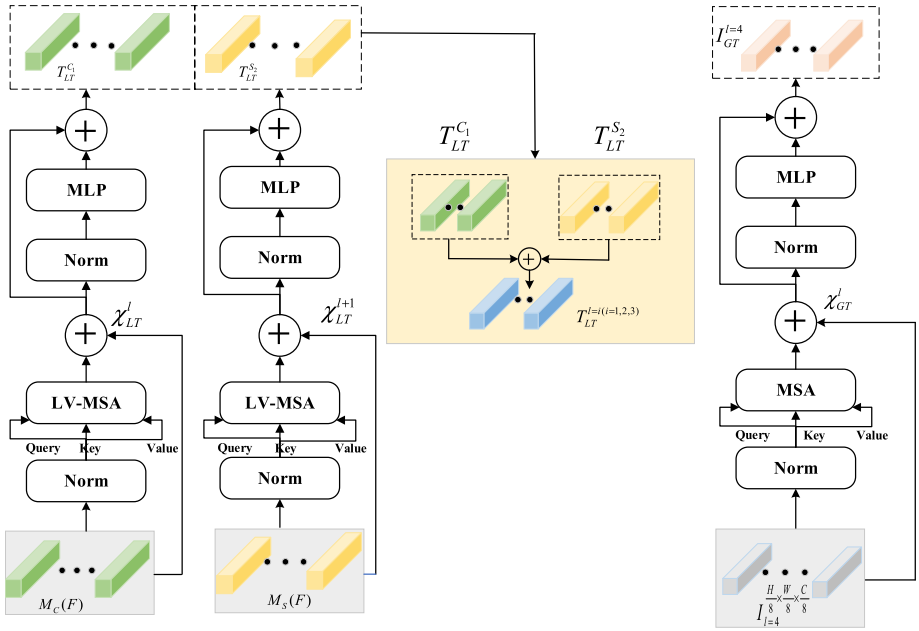
$$\begin{aligned} &I_{l=i(i=1,2,3)}^{H\times W\times C_{i(i=1,2,3)}} \\ &= f^{3\times3\times3}(f^{1\times1\times1}(Upsample^{n=j(j=0,2,4)}(I_{l=i(i=1,2,3)}^{H_i\times W_i\times C_i}))) \\ &S^{H\times W\times C} \\ &= f_{c=C}^{1\times1\times1}(Concatenate(I_{l=1}^{H\times W\times C_1}, I_{l=2}^{H\times W\times C_2}, I_{l=3}^{H\times W\times C_3})) \end{aligned} \tag{3}$$

D. MIXED ATTENTION MECHANISM

In the transverse section of lumbar spine MRI images, the area of the foreground region is relatively small compared to the background region. This leads to deficiencies in traditional U-Net networks and their variant network architectures in extracting target features, particularly in terms of multi-scale information and segmentation attention. In the U-Net encoder, as the number of downsampling layers increases, the extracted low-level and high-level semantics of the lumbar spine are gradually diluted. The low-level semantics of the lumbar spine contain spatial information about the boundaries of lumbar vertebrae, while high-level semantics contribute to vertebral localization and identification. Currently, mainstream lumbar spine datasets exhibit characteristics of sample anisotropy. To overcome the limitations of traditional convolutional neural networks, we propose using a mixed attention mechanism to capture multiscale information in features. The mixed attention structure consists of channel attention models, spatial attention models, and Local Transformer and Global Transformer structures based on self-attention, as illustrated in Figure 5. The features are filtered for irrelevant channel information and enhanced for spatial position information by the channel attention module and spatial attention module.

The Local Transformer module proposed in this paper employs a Volume-based Multi-scale Self-Attention mechanism (V-MSA) to process the channel attention feature $M_c(F)$ and spatial attention feature $M_s(F)$. Its structure is illustrated in Figure 5(a). Since $M_c(F)$ and $M_s(F)$ belong to the downsampling shallow-layer features with a relatively larger resolution compared to deep-layer features, we use the local volume self-attention mechanism instead of the traditional global self-attention feature. The computational complexity formulas for both self-attention mechanisms are shown in formula (4), where $\{S_H \times S_W \times S_D\}$ represents the size of the local volume. By restricting each position of the voxel to focus only on the information within the local volume concerning the global position, we achieve a significant improvement in computational efficiency. This approach better captures the local features of lumbar vertebrae and, through the consideration of both channel and spatial dimensions, facilitates a more comprehensive focus on the local features and global position information of lumbar vertebrae.

$$\begin{aligned} \Omega(MSA) &= 4hwC^2 + 2(hw)^2C \\ \Omega(LV-MSA) &= 4hwdC^2 + 2S_HS_WS_Dhwd C \end{aligned} \tag{4}$$



(a) Local Transformer Structure (b) Global Transformer Structure

FIGURE 5. Local transformer structure and global transformer structure. Where local transformer uses the V-MAS based mechanism for long range dependency modeling and global transformer uses the traditional MSA mechanism for feature spatial location information acquisition.

The features $T_{LT}^{C_1}$ and $T_{LT}^{S_2}$ based on channel and spatial dimensions are further enhanced by element-wise addition after the output. The computational process of the Local Transformer is illustrated in formula (5), where the features from different dimensions are added element-wise, strengthening the feature relationships.

$$\begin{aligned}\chi_{l+1}^{\mathrm{LT}} &= MLP(Norm(LV \\ &\quad - MSA(Norm\left(\chi_{l}^{\mathrm{LT}}\right) + \chi_{l}^{\mathrm{LT}}))) \quad l = 0, 1, 2 \ldots\ldots \mathrm{L} \\ \chi_{l+2}^{\mathrm{LT}} &= MLP(Norm(LV - MSA\left(Norm\left(\chi_{l+1}^{\mathrm{LT}}\right)\right) + \chi_{l+1}^{\mathrm{LT}}))\end{aligned} \tag{5}$$

where l represents the number of layers in the Transformer structure and Norm represents the regularization, this paper adopts layer normalization to ensure the overall stability and effectiveness of the model. Compared with the shallow features extracted by downsampling, the fourth layer of downsampling has lost too much local detail information after repeated convolution operations, and the feature $I_{l=4}^{\frac{H}{8}\times\frac{W}{8}\times\frac{C}{8}}$ contains more deep local features and lower resolution, so we introduce the traditional global-based multi-head self-attention mechanism Global Transformer structure to capture the global location information, and establish global correlation through MSA. correlation. The structure is shown in Fig. 5(b), and the calculation process is shown in Eq. (6).

$$\begin{aligned}\chi_{l+1}^{\mathrm{GT}} &= MLP(Norm(MSA(Norm\left(\chi_{l}^{\mathrm{GT}}\right) \\ &\quad + \chi_{l}^{\mathrm{GT}}))) \quad l = 0, 1, 2 \ldots\ldots \mathrm{L} \\ \chi_{l+2}^{\mathrm{LT}} &= MLP(Norm(MSA\left(Norm\left(\chi_{l+1}^{\mathrm{GT}}\right)\right) + \chi_{l+1}^{\mathrm{GT}}))\end{aligned} \tag{6}$$

E. LOSS FUNCTION

For lumbar spine MRI images where the lumbar vertebrae and intervertebral disc region belongs to a typical multi-class segmentation problem, for which we use the Dice loss function for training, in the training process we found that the lumbar vertebrae and intervertebral discs and other prospective areas are small and inconsistent with the proportion of each sample, and the Dice loss function in the narrow target loss will lead to dramatic changes in the model gradient. To address this problem, this paper uses the Dice loss function combined with cross-entropy loss for training and learning, the loss function as a whole $\mathcal{L}$ as shown in Equation (7), $\mathcal{L}$ is mainly composed of two parts of the loss function, the overall labeling loss function $\mathcal{L}_{Dice}$ and the category segmentation loss function $\mathcal{L}_{CE}$.

$$\mathcal{L} = \alpha\mathcal{L}_{Dice} + \beta\mathcal{L}_{\mathrm{CE}} \tag{7}$$

By introducing a hybrid loss function that integrates the Dice loss function with the cross-entropy loss function, the loss function gives different weights to the Dice loss values of each feature class while calculating the cross-entropy of each feature class, and at the same time, we introduce the balancing factors $\alpha = 0.6, \beta = 0.4$ to reduce the impact of the overall loss function $\mathcal{L}_{Dice}$ on the training process, and utilize the cross-entropy loss function $\mathcal{L}_{CE}$ to compensate for the imbalance of the foreground region that occupies a

small proportion. The loss function $\mathcal{L}_{Dice}$ and $\mathcal{L}_{CE}$ calculation process is shown in Equation (8):

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{i,k}\log\left(\hat{y}_{i,k}\right)$$
$$\mathcal{L}_{\text{Dice}} = 1 - \frac{1}{N}\sum_{k}^{N}\frac{2\sum_{i}^{N^*} y_{i,k}\hat{y}_{i,k}}{\sum_{i}^{N^*} y_{i,k} + \sum_{i}^{N^*}\hat{y}_{i,k}} \tag{8}$$

where N represents the number of samples, K represents the number of labeling categories, $y_{i,k}$ represents the actual labeling of the real sample i belonging to category k, and $\hat{y}_{i,k}$ represents the probability of predicting that sample i belongs to category k.

III. LUMBAR SPINE SEGMENTATION EXPERIMENTATION

A. DATA

Medical datasets mainly suffer from the problems of small data volume size and variable quality of data samples compared to natural image datasets, in order to evaluate the robustness of our proposed model, we select two publicly available lumbar spine MRI image datasets MRSpineSeg2021 and SpineSegT2Wdataset3.

Among them, the MRSpineSeg2021 dataset was published by the second Chinese Society for Image Graphics (CSIG) Graphics Technology Challenge. A total of 215 samples, including 6 cases of healthy people, 204 cases of disc degeneration, 177 cases of vertebral body degeneration, 91 cases of spinal stenosis, each sample included T2-weighted maps and labels manually labeled by an expert, with 20 labeling categories including 10 vertebrae and 9 intervertebral discs mainly focusing on the thoracic vertebral end and lumbar vertebral area, with an average resolution ranging from to, and slices ranging from 12 to 18.

The SpineSegT2Wdataset3 dataset contains 195 samples, which are mainly composed of patients with lumbar disc degeneration and lumbar disc herniation, and each patient data sample is a sagittal T2-weighted MR three-dimensional data, and the acquisition equipment is the same MR equipment, and the magnetic field strength is 3.0 T. It contains a total of 2,460 slices in 195 samples, with a background label of 0 and vertebral label of 2,460, and an average resolution of 12 to 18 slices. was 0 and vertebrae labeling was 1.

The two types of lumbar spine MRI datasets have a large 2D resolution, with an average resolution of. Therefore, in this paper, the experimental environment was selected as Xeon(R) Platinum 8352 + NVIDIA RTX 4090 GPU, with 24GB of video memory and 80GB of RAM, and the experimental system was Ubuntu 18.04, and the environment was configured with Python 3.8 + Pytorch 1.9.0 framework.

B. EVALUATION METRICS

The goal of this study is mainly for the lumbar vertebrae part of the spine, including the vertebrae and intervertebral discs, and the foreground regions such as vertebrae and intervertebral discs account for a relatively small percentage of the vertebral body and intervertebral discs, for this reason the use of conventional segmentation metrics such as the precision rate and the accuracy can not be accurately measure the performance of the experimental model. Because the experiments in this paper use Dice similarity coefficient, Hausdorff distance(HD), and recall evaluation indexes to assess the segmentation accuracy of the lumbar spine image segmentation results.DSC similarity coefficient index is a positive index that measures the similarity and overlap between the model prediction set and the real set of two sample pieces, with a value range of 0-1, which is mainly used to assess the segmentation effect of the lumbar vertebrae in the internal smooth region of the spine. The principle of calculation is shown in equation (9):

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

where P represents the model prediction result, T represents the expert segmentation result, and i represents the index of spinal vertebrae.

HD (mm) evaluation index measures the shortest distance between the farthest points between two sets, through the lumbar spine segmentation of the spatial distance between the prediction results and the labeled set in order to determine the degree of discrepancy between the predicted value and the true value, the asymmetric HD distance measures the maximum distance between the predicted set and the true set, the maximum distance quartile we set 95% to rule out the interference caused by the outlier, the calculation process shown in Eq. (10).

$$HD(P,T) = \max(\sup_{p\in P_i}\inf_{t\in T_i} d(p,t), \sup_{t\in T_i}\inf_{p\in P_i} d(p,t)) \tag{10}$$

where P_i represents the set of surface distances from the predicted segmentation labels of vertebrae, T_i is the set of surface distances from the vertebrae, and $d\,(p,t)$ represents the Euclidean distance between the points P_i and T_i in the set of P and the set of T. The HD metrics are used to measure the gap between the predicted segmentation results of the boundaries of lumbar vertebrae and intervertebral discs and the boundaries of the true labels. In addition to the above two metrics, we also used the recall rate Recall to assess the model's ability to detect true positive samples, where the Recall metric is calculated using the formula shown in (11), where the True Positive (TP) parameter represents the number of positive samples that the model correctly predicted as positive samples, and False Negative (FN) represents the number of positive samples that the model incorrectly predicted as the number of negative samples.

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

The Recall metric measures the degree of model coverage for each vertebra of the lumbar spine and represents the proportion of the total number of vertebrae that the model successfully detected as true vertebrae.

TABLE 1. DSC (%) scores of lumbar spine and intervertebral DISC MRI images segmented by different algorithms.

Methods	L1	L2	L3	L4	L5	L1/L2	L2/L3	L3/L4	L4/L5	L5/S
3D U-Net	79.91	80.42	80.94	82.86	85.43	81.17	80.77	83.21	83.42	84.70
3D ResUNet	80.49	85.61	86.82	86.77	86.22	84.26	87.71	87.86	85.13	84.69
3D DeepLabv3+	87.56	86.70	86.65	86.80	86.55	87.84	87.09	87.48	85.60	86.20
3D Graphonomy	88.57	88.22	88.19	87.86	87.48	89.18	88.54	88.77	86.42	86.33
3D GCSN	89.26	88.89	88.89	88.54	88.11	89.91	**89.34**	89.37	87.19	86.76
nnUnet	88.34	88.17	86.91	86.03	87.01	89.94	88.03	88.98	87.11	85.47
Our Models	**90.29**	**88.95**	**90.27**	**89.01**	**89.79**	**91.18**	89.32	**90.48**	**87.32**	**87.07**

TABLE 2. Recall scores of lumbar spine and intervertebral disc MRI images segmented by different algorithms.

Methods	L1	L2	L3	L4	L5	L1/L2	L2/L3	L3/L4	L4/L5	L5/S
3D U-Net	79.23	81.97	82.32	82.24	83.72	78.73	79.96	81.42	81.12	81.73
3D ResUNet	81.96	82.32	82.49	83.53	83.86	80.34	81.56	82.33	82.71	81.35
3D DeepLabv3+	82.53	84.41	84.38	84.72	85.97	82.93	83.78	84.28	83.96	83.81
3D Graphonomy	82.17	82.76	83.74	83.86	85.04	83.24	83.85	85.76	85.14	84.39
3D GCSN	84.74	84.05	86.41	86.69	86.25	84.16	84.36	85.92	85.78	85.49
nnUnet	79.96	80.15	81.63	81.58	82.39	79.86	79.98	81.02	81.17	81.96
Our Models	**85.67**	**87.29**	**87.58**	**87.49**	**88.72**	**85.03**	**86.64**	**86.38**	**86.10**	**86.22**

C. EXPERIMENTAL RESULT

1) DATA PREPROCESSING

The dataset selected for this experiment contains two types of datasets, MRSpineSeg2021 and SpineSegT2Wdataset3, for the differences between the two types of datasets we performed specific preprocessing operations. For all the samples in the MRSpineSeg2021 dataset, we performed the cropping, padding, and normalization preprocessing steps, regionally cropping the lumbar spine irrelevant regions, controlling the size of the image to $256 \times 256 \times 18$ using zero padding for the cropped image, and finally normalizing the image by voxel normalization.

For the SpineSegT2Wdataset3 dataset, the gray scale of the MR image changes slowly and is not uniformly distributed, resulting in the existence of different gray scale values in the same vertebrae or intervertebral disc region, we firstly corrected the offset field of the voxels in the MR image, and then extracted the three-dimensional T2W1 sagittal plane slices, and in order to reduce the effect of the background region on the model, we maintained the integrity of the lumbar vertebrae by randomizing the voxels. integrity, we augmented the dataset by random flipping, and the remaining preprocessing operations were kept the same as MRSpineSeg2021.

2) IMPLEMENTATION DETAILS

For the dataset MRSpineSeg2021, it contains 215 instance samples, each sample contains 12-18 slices. During model training 70% of the samples in the dataset are used for training, 10% of the samples are used for testing during training, and the remaining 20% of the data is used as a test set for evaluating the segmentation performance of the model, with an input image size of 256 × 256, a Batch-size set to 2, the use of the AdamW optimizer, and an initial learning rate set to 0.1%. For the dataset SpineSegT2Wdataset3, a total of 195 samples contain 2460 slices, of which 135 samples are used for model training, 40 samples are used for the validation set during model training, and 20 samples are used as a test set to evaluate the segmentation performance of the model, and the rest of the hyper-parameters are the same as those of the MRSpineSeg2021 dataset in order to maintain consistency of the experiments. The remaining hyperparameters are the same as the MRSpineSeg2021 dataset for consistency.

3) MRSpineSeg2021 SEGMENTATION RESULT

The segmentation results on the MRSpineSeg2021 dataset are shown in Tables 1 and 2, and in order to validate the sophistication of our proposed model we selected six mainstream segmentation models for the comparison of the experimental results, including 3D U-Net, 3D ResUNet, 3D DeepLabv3+, 3D Graphonomy, 3D GCSN, and nnUnet Table 1 shows the DSC scores of different segmentation models on the MRSpineSeg2021 dataset, in the segmentation of the lumbar spine L1-L5 and the DSC scores of the five intervertebral discs L1/L2-L5/S, all of our proposed models achieved the highest scores, and the overall average DSC score of our model was 89.37, which is the highest score among the many mainstream segmentation models, in comparison with the basic Compared with the basic 3D U-Net model, our proposed model improves the DSC coefficient by 8.6%, and compared with the 3D GCSN model, which has the best segmentation effect, the model in this paper also improves the overall segmentation accuracy by 0.8%, which proves that the segmentation performance of our proposed model has been further improved compared with mainstream segmentation models.

In terms of the recall index, since the lumbar vertebrae and intervertebral discs are close to each other, it is more difficult to identify and segment the lumbar vertebrae and intervertebral discs as a whole than to segment the spine as a whole,

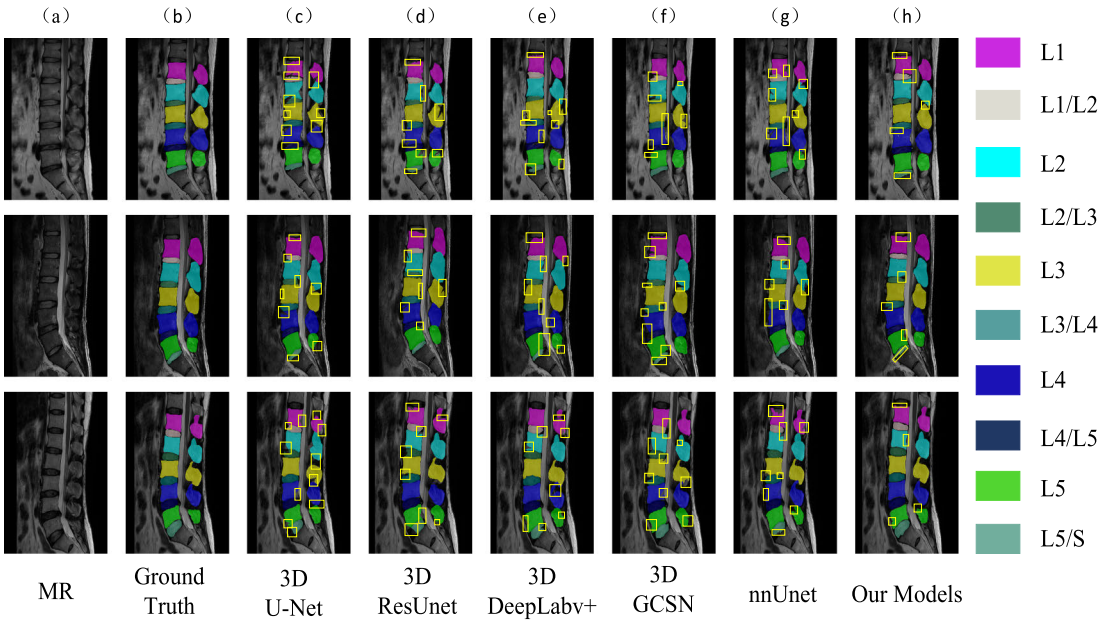


FIGURE 6. Each segmentation model segmented L1-L5 and L1/L2-L5/S for a total of ten current regions, where each row is a sagittal slice of the sample,yellow rectangles are used to mark the prediction results and the real results of different regions.

and when analyzing the sample data in the MRSpineSeg2021 dataset, the voxel space is, and the resolution within the slices is much higher than that between slices which belongs to the anisotropic data, so the recall of the model we propose is limited compared to the mainstream segmentation models. The performance improvement of our proposed model is limited compared to the recall of mainstream segmentation models, in which our model improves 6.7% compared to the traditional segmentation network 3D U-Net, and 3.1%, 3.2%, and 1.5% compared to the top three segmentation models in terms of segmentation performance (3D DeepLabv3+, 3D Graphonomy, and 3D GCSN), respectively., indicating that our model outperforms the current mainstream models in vertebrae and intervertebral disc detection.

The segmentation visualization results of the MRSpineSeg2021 dataset are shown in Fig. 6, from Fig. 6 we can see that in the case of segmentation of anisotropic data, the results of segmentation of the lumbar spine by the 3D U-Net network show under-segmented regions relative to other models, and the under-segmented regions are mainly concentrated in the region of the lumbar spine vertebrae with highly similar heights, which suggests that the 3D U-Net network has a relatively low effect in the segmentation of the vertebrae and the disc boundaries and the This indicates that the 3D U-Net network has semantic difficulties in the segmentation of vertebrae and intervertebral disc boundaries and in the localization of lumbar vertebrae and lacks effective learning. nnU-Net model has a relatively low segmentation effect between vertebral bodies and intervertebral discs in the same sample because the spatial positions of neighboring vertebral intervals and intervertebral discs are very close to each other to the extent of overlap, which results in the semantic confusions that exist in the segmentation process at the moment and leads to the segmentation performance of the nnU-Net segmentation model being lower than the segmentation model of 3D GCSN. GCSN and other segmentation models. When segmenting the intervertebral disc and vertebral bone region, our proposed model can separate the vertebral bone and other similar regions to achieve excellent segmentation results by constraining the boundary range of the vertebral bone with a specially designed boundary segmentation model. It is verified that our proposed model has good generalization ability in segmenting the smooth region of the lumbar spine with part of the overlapping region of vertebrae and intervertebral discs.

4) SpineSegT2Wdataset3 SEGMENTATION RESULT

In order to verify the robustness and generalization of our model, we train on the SpineSegT2Wdataset3 dataset, which mainly contains five vertebrae of the lumbar spine L1-L5, with the labels are divided into two types of labels 0 and 1 For this purpose, we evaluate the segmentation performance of the model by using the two metrics of 95-HD distance and DSC coefficient. Experimental results we organize in Table 3 and Table 4, analyzing the data in Table 3, we can see that the model proposed in this paper in L1-L5 a total of five vertebrae on the segmentation performance of the DSC score have achieved the highest score, the average score of the five categories of vertebrae reached 90.04, compared with the traditional segmentation model based on convolution operation, the segmentation model based on the self-attention mechanism (TransU-Ne and Swin-Unet) are lower than the segmentation model represented by 3D U-Net in the

TABLE 3. DSC (%) scores of lumbar spine MRI images segmented by different algorithms.

Methods	L1	L2	L3	L4	L5
3D U-Net	88.13	88.97	89.92	89.89	89.74
U-Net++	88.95	89.58	89.71	89.48	88.37
TransU-Net	74.97	76.28	76.19	76.94	75.93
Swin-Unet	74.83	74.31	74.82	74.48	75.77
RAR-Unet	88.93	89.95	90.03	89.54	89.04
Our Models	**89.34**	**90.28**	**90.12**	**90.33**	**90.14**

TABLE 4. Segmentation of lumbar spine images by different algorithms 95-HD distance (mm).

Methods	L1	L2	L3	L4	L5
3D U-Net	4.02	3.96	3.41	3.07	3.74
U-Net++	3.05	2.78	2.71	2.21	2.07
TransU-Net	4.24	4.17	3.98	3.79	3.72
Swin-Unet	4.19	4.14	4.06	3.92	3.90
RAR-Unet	2.31	1.85	1.43	1.55	1.43
Our Models	1.33	1.63	1.71	1.67	1.24

segmentation performance of vertebrae instances, while the RAR-Unet segmentation model based on a large convolutional kernel achieves the highest score second only to our proposed model, which indicates that convolutional neural networks still have strong segmentation performance and competitiveness in the task of anisotropic data. For similar example segmentation tasks, more complete feature semantics can be learned under the condition of improving the sensory field. In Table 4, our proposed model achieves an average distance of 1.51 mm in the 95-HD distance, which is 2.14 mm shorter compared to the traditional segmentation model 3D U-Net, and 11.5% shorter compared to the optimal segmentation model RAR-Unet, combining a convolutional neural network with a hybrid attentional mechanism, which can be useful in extracting the vertebrae smooth region and suppressing vertebral boundary segmentation.

The visualization results of the SpineSegT2Wdataset3 dataset are shown in Fig. 7, where we selected 3D U-Net, U-Net++, Trans-Unet, Swin-Unet, and RAR-Unet to demonstrate with our proposed model. Since this dataset belongs to the two-class segmentation task therefore we use binarized labels for the demonstration, from Fig. 7 relative to other segmentation models, our proposed model can clearly segment between similar vertebrae in lumbar spine, which proves that our model avoids the problem of semantic confusion in the process of training similar vertebrae.

D. ABLATION EXPERIMENT

In order to verify the impact of our proposed improvement modules on the segmentation performance of the model, ablation experiments are used to verify the performance of the segmentation model. We categorize three modules for the improvement modules, the first one is boundary segmentation module EGM, the second one is CAM with SAM, and the third one is mixed attention module MATM. To ensure the validity of the experiments, we use a pure 3D U-Net network as a blank control group for the experiments. Meanwhile, in order to test the model robustness, we conduct the same experiments on two datasets (MRSpineSeg2021 and SpineSegT2Wdataset3) at the same time, and the data pairs of the whole ablation experiments are shown in Table 3. The ablation module is divided into four main parts: boundary segmentation module EGM, channel attention module CAM, spatial attention module SAM, and hybrid attention module MATM module. The six categories of the experimental group are 3D U-Net, 3D U-Net + EGM, 3D U-Net + CAM, 3D U-Net + SAM, 3D U-Net + MAMT, and our proposed model.

Analyzing the data in Table 5, we first compare the impact of the EGM module on the overall segmentation, comparing the DSC coefficient scores of the pure 3D U-Net network and the 3D U-Net + CAM and other networks in the MRSpineSeg2021 dataset and the SpineSagT2Wdataset3 dataset, and we can see that in the MRSpineSeg2021 dataset It can be seen that in the MRSpineSeg2021 dataset, the segmentation accuracy is only 1.4% higher than that of the pure 3D U-Net network with the addition of the EGM module, and the segmentation accuracy in the SpineSagT2Wdataset3 dataset is similar to that of the pure 3D U-Net, which verifies that the effect of the EGM module in improving the segmentation accuracy is not obvious. In the HD distance, the experimental group with the addition of the EGM module improves 39.7% and 60.1% compared to the basic segmentation network 3D U-Net, while the experimental group with the addition of the other modules shortens 1.44 mm and 1.74 mm on average, respectively, which proves that our proposed boundary segmentation module improves the effectiveness of vertebral bone boundary segmentation.

In the experimental group where CAM and SAM modules are added respectively, comparing the three groups of data in terms of DSC score, HD distance and Recall indexes, we can understand that the addition of the convolution-based attention mechanism achieves significant improvement in segmentation accuracy and target detection, and due to the small number of channel dimensions in the two datasets that we selected ( ), the addition of the CAM module does not have a significant effect on the segmentation performance of the DSC score. score is not significantly improved, while the experimental group of 3D U-Net + SAM module has an average improvement of 4.95% in DSC score and Recall score than 3D U-Ne t + CAM.

By adding the hybrid attention module MATM to the base 3D U-Net model, compared to the 3D U-Net + CAM experimental group and the 3D U-Net + SAM experimental group, the 3D U-Net + MATM experimental group on the MRSpineSeg2021 dataset did not have a significant improvement in the boundary detection metric HD Distance, but did not have a significant improvement in the segmentation accuracy metric DSC score and the Recall, a detection rate metric, achieved significant improvement. The hybrid attention mechanism based on convolutional attention and self-attention focuses more on the extraction of overall lumbar vertebrae position

TABLE 5. Comparison of ablation experiment results.

Dataset	Backbone +Ablation Module	DSC	HD	Recall
MRSpineSeg2021	3D U-Net	82.28	4.93	81.73
	3D U-Net+EGM	83.47	2.97	81.92
	3D U-Net+CAM	83.32	4.32	82.56
	3D U-Net+SAM	88.31	4.89	85.94
	3D U-Net+MATM	88.25	4.04	85.12
	Our Model	**89.37**	**6.67**	**86.71**
SpineSagT2Wdataset3	3D U-Net	89.33	3.64	89.94
	3D U-Net+EGM	89.21	**1.45**	89.01
	3D U-Net+CAM	89.39	3.52	90.43
	3D U-Net+SAM	89.72	2.98	91.87
	3D U-Net+MATM	89.64	3.07	92.78
	Our Model	**90.04**	1.51	**92.14**

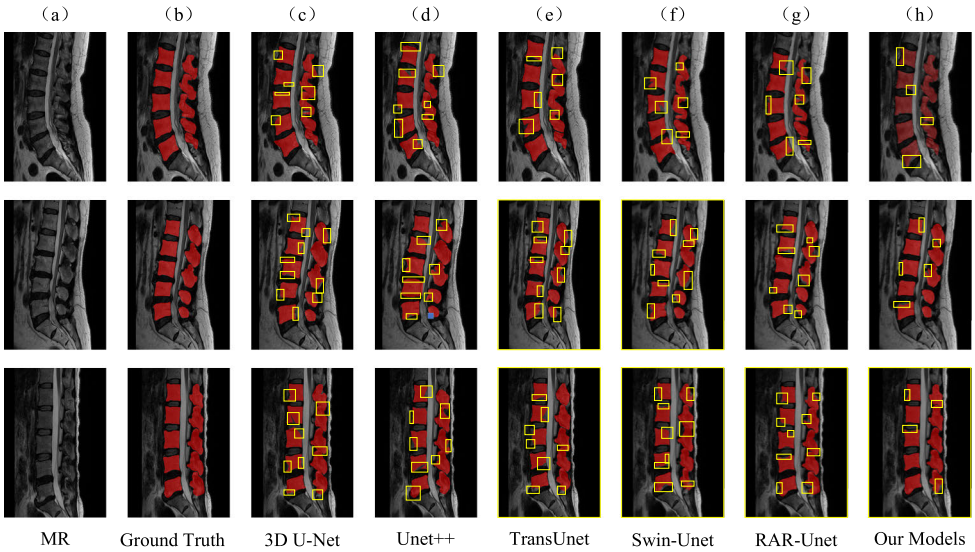


FIGURE 7. Sagittal section segmentation results of L1-L5 vertebrae.

information, and has obvious positive effects in reducing both the under-segmentation problem of spine smooth region segmentation and the spinal vertebrae localization ambiguity. Our proposed model achieves the optimal DSC scores of 89.37 and 90.04 and the optimal recall scores of 86.71 and 92.14, and the experimental results prove the effectiveness and superiority of our proposed improved module by stacking different modules and conducting multiple experiments.

IV. ANALYSIS AND CONCLUSION

We propose a lumbar spine MRI image segmentation model, based on 3D U-Net network by improving the downsampling layer of the encoder, designing a multi-scale feature extraction structure and boundary segmentation structure for boundary constraints and enhanced feature extraction of the lumbar vertebrae, and by mixing the use of convolution-based attention mechanism and local volume-based self-attention mechanism for multi-dimensional feature enhancement and spatial location information acquisition. By conducting experiments on different lumbar spine MRI datasets, our proposed lumbar spine segmentation model is compared with the mainstream segmentation models today. The experimental results demonstrate that our lumbar spine segmentation model achieves obvious constraint effects on the boundary segmentation of lumbar vertebrae, and most of the similar vertebrae and intervertebral discs can be clearly segmented to connect the overlapping places. In addition, our lumbar spine segmentation model also achieves some improvement in segmentation accuracy, which proves the generalization performance and robustness of our proposed model. However, during the experimental process, we found that the computational complexity of the model is relatively high, and due to the nested use of multiple attention models, our future research mainly focuses on the lightweighting and parameter optimization of the model.

ACKNOWLEDGMENT

(Jing Liu and Jianlan Yang contributed equally to this work.)

DATA AVAILABILITY

The lumbar spine MRIs used in the experiments were all publicly available datasets, including the MRSpineSeg2021 dataset and the SpineSegT2Wdataset3 dataset.

REFERENCES

[1] M. M. Panjabi and A. A. White, "Basic biomechanics of the spine," *Neurosurgery*, vol. 7, no. 1, pp. 76–93, Jul. 1980.
[2] J. Zhao, L. Sun, X. Zhou, S. Huang, H. Si, and D. Zhang, "Residual-atrous attention network for lumbosacral plexus segmentation with MR image," *Computerized Med. Imag. Graph.*, vol. 100, Sep. 2022, Art. no. 102109.
[3] S. Y. Lee, T.-H. Kim, J. K. Oh, S. J. Lee, and M. S. Park, "Lumbar stenosis: A recent update by review of literature," *Asian Spine J.*, vol. 9, no. 5, p. 818, 2015.
[4] K. Alsaleh, D. Ho, M. P. Rosas-Arellano, T. C. Stewart, K. R. Gurr, and C. S. Bailey, "Radiographic assessment of degenerative lumbar spinal stenosis: Is MRI superior to CT?" *Eur. Spine J.*, vol. 26, no. 2, pp. 362–367, Feb. 2017.
[5] A. L. Williams, F. R. Murtagh, S. L. G. Rothman, and G. K. Sze, "Lumbar disc nomenclature: Version 2.0," *Amer. J. Neuroradiol.*, vol. 35, no. 11, p. 2029, Nov. 2014.
[6] 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.
[7] R. Korez, B. Ibragimov, B. Likar, F. Pernus, 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.
[8] B. Ibragimov, B. Likar, F. Pernus, 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.
[9] I. Castro-Mateos, J. M. Pozo, M. Pereanez, 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.
[10] 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.
[11] 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.
[12] B. Glocker, J. Feulner, A. Criminisi, D. R. Haynor, and E. Konukoglu, "Automatic localization and identification of vertebrae in arbitrary field-of-view CT scans," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*, Nice, France. Berlin, Germany: Springer, 2012, pp. 590–598.
[13] P. A. Bromiley, E. P. Kariki, J. E. Adams, and T. F. Cootes, "Fully automatic localisation of vertebrae in CT images using random forest regression voting," in *Proc. Int. Workshop Comput. Methods Clinical Appl. Spine Imag.* Cham, Switzerland: Springer, 2016, pp. 51–63.
[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, pp. 269–275, Mar. 2015.
[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] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*, Munich, Germany. Springer, 2015, pp. 234–241.
[17] G. Fan, H. Liu, D. Wang, C. Feng, Y. Li, B. Yin, Z. Zhou, X. Gu, H. Zhang, Y. Lu, and S. He, "Deep learning-based lumbosacral reconstruction for difficulty prediction of percutaneous endoscopic transforaminal discectomy at L5/S1 level: A retrospective cohort study," *Int. J. Surg.*, vol. 82, pp. 162–169, Oct. 2020.
[18] R. Korez, B. Likar, F. Pernus, and T. Vrtovec, "Model-based segmentation of vertebral bodies from MR images with 3D CNNs," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2016, pp. 433–441.
[19] 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*.
[20] 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.
[21] 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.
[22] 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.
[23] B. Wang, J. Qin, L. Lv, M. Cheng, L. Li, D. Xia, and S. Wang, "MLKCA-UNet: Multiscale large-kernel convolution and attention in unet for spine MRI segmentation," *Optik*, vol. 272, Feb. 2023, Art. no. 170277.
[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[25] 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*.
[26] 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.
[27] 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.
[28] 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.

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 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.

YUEE ZHOU was born in Shandong, China, in 2000. She received the B.S. degree in software engineering from Beihua University, in 2022. She is currently pursuing the M.S. degree in biomedical engineering with Gansu University of Traditional Chinese Medicine. Her main research interests include medical image processing and deep learning. Her awards and honors include the National Motivation Scholarship and the First Class Scholarship of Beihua University.

XINXIN CUI was born in Gansu, China, in 2000. She received the B.S. degree in medical information engineering from 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.

FENGQING JIN was born in Gansu, China, in 2000. She received the B.S. degree in medical information engineering from Gansu University of Traditional Chinese Medicine, in 2023, where she is currently pursuing the M.S. degree in biomedical engineering. Her main research interests include medical image processing and deep learning.

GUODONG SUO was born in Henan, China, in 2000. He received the B.S. degree in medical information engineering from Gansu University of Traditional Chinese Medicine, in 2023, where he is currently pursuing the M.S. degree in biomedical engineering. His main research interests include medical image processing and deep learning.

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

JIANLAN YANG was born in Fujian, China, in 1974. He received the degree in health career management from the School of Public Health, La Trobe University, Australia. He is currently an Associate Professor and the Master's Degree Supervisor. His research interests include health information data mining and medical image recognition and application. He is an Executive Member of Chinese Medicine Informatics Committee, Chinese Society of Health Informatics. He is the Deputy Director of Chinese Medicine and Health Informatics Committee, Chinese Society of Medical Informatics. He is the Vice President of the Cloud Health Branch, Chinese Society of Chinese Medicine Informatics.