Hindawi
International Journal of Intelligent Systems
Volume 2023, Article ID 8686471, 16 pages
https://doi.org/10.1155/2023/8686471

Research Article

A Spine Segmentation Method under an Arbitrary Field of View Based on 3D Swin Transformer

Yonghong Zhang,[1,2] Xuquan Ji,[2,3] Wenyong Liu ,[3] Zhuofu Li ,[4,5,6] Jian Zhang,[1,2] Shanshan Liu,[4,5,6] Woquan Zhong,[4,5,6] Lei Hu ,[1,2] and Weishi Li [4,5,6]

[1]*Robotics Institute, School of Mechanical Engineering and Automation, Beihang University, Beijing, China*
[2]*Beijing Zoezen Robot Co., Ltd., Beijing, China*
[3]*School of Biological Science and Medical Engineering, Beihang University, Beijing, China*
[4]*Peking University Third Hospital, Department of Orthopaedics, Beijing, China*
[5]*Engineering Research Center of Bone and Joint Precision Medicine, Ministry of Education, Beijing, China*
[6]*Beijing Key Laboratory of Spinal Disease Research, Beijing, China*

Correspondence should be addressed to Lei Hu; hulei9971@buaa.edu.cn and Weishi Li; puh3liweishi@163.com

Received 20 November 2022; Revised 10 September 2023; Accepted 15 September 2023; Published 13 October 2023

Academic Editor: Mohammad R. Khosravi

High-precision image segmentation of the spine in computed tomography (CT) images is important for the diagnosis of spinal diseases and surgical path planning. Manual segmentation is often tedious and time consuming. Thus, an automatic segmentation algorithm is expected to solve this problem. However, because different areas are scanned, the number of spines in the original CT image and the coverage area are often different, making it extremely difficult to directly conduct a fully autonomous spine segmentation. In this study, we propose a two-stage automatic spine segmentation method based on 3D Swin Transformer. In the first stage, the 3D Swin-YoloX algorithm is used to achieve an accurate positioning of each spine segment in the CT images. In the second stage, 3D Swin-UNet is used to achieve a high-precision segmentation of the spine. Using an open dataset, the average Dice of our approach can reach 0.942 and the average Hausdorff distance can reach 6.24, indicating a higher accuracy in comparison with other published methods. Our proposed method can effectively eliminate any adverse effects of the different scanning areas on a spinal image segmentation and has a high application value.

1. Introduction

CT imaging is an important source of information, assisting doctors in the diagnosis of spinal diseases and their indicators [1]. Using CT sequence data, doctors can clearly understand the condition of the affected area and thereby provide a reference for subsequent treatment. However, CT images are made up of hundreds or even thousands of two-dimensional images, which do not provide doctors an intuitive three-dimensional view [2]. Therefore, in most cases, doctors use specialized medical image processing software, such as Mimics, to manually segment the spine in CT images and achieve 3D reconstruction [3]. Spinal image segmentation plays an important role in the diagnosis and surgical planning of spinal disorders. Examples include spinal fracture detection [4], spinal arthritis diagnosis [5], spinal surgical path planning [6], and the 3D printing of auxiliary surgical equipment [7]. However, there is an overlap between the vertebrae of the spine, and a manual segmentation relies heavily on doctor experience and is time consuming and laborious. In addition, as the number of patients with spinal diseases increases, the workload of radiologists significantly increases [8, 9]. Moreover, the manual segmentation of the spine is excessively subjective [10], and to solve this problem, an automatic spine image segmentation algorithm is required.

Owing to the uncertainty of the CT scanning area and the high similarity between adjacent spinal segments, locating the 3D spatial region of all spinal segments in a CT image with an arbitrary field of view (FOV) is the first difficult problem to overcome in achieving an automatic spine segmentation. Cootes et al. proposed a method for applying a statistical shape model to integrate global shape information [11]. Vrtovec et al. proposed a polynomial model for describing a spine curve detected in CT images [12]. Michael Kelm et al. proposed a novel approach combining efficient local object detection based on an iterative version of marginal space learning (MSL) with a global probabilistic prior model for the vertebral column [13]. Ebner et al. proposed a landmark localization algorithm that applies regression forests [14]. Lindner et al. used random forest regression voting to quickly generate response images, based on which shape model fitting can be realized [15]. Glocker et al. proposed a spine location and recognition algorithm based on a supervised classification forest to avoid using an explicit appearance-parameter model [16]. In recent years, with the rapid development of deep learning technology, localization methods achieving an excellent performance have been proposed. Chen et al. proposed a three-stage spine positioning method for generating a group of rough cone centroids based on a random forest classifier, used a convolutional neural network (CNN) algorithm to identify the cone types and eliminate the error detection process, and finally applied a shape regression model to refine the predicted centroid [17]. Suzani et al. adopted an effective method based on a depth feedforward neural network for predicting the position of each vertebral body using the context information in an image [18]. Based on heat map regression, Payer et al. proposed a method for achieving an accurate landmark positioning of each spinal segment [19, 20]. Liao et al. developed a multitask-based fully 3D CNN to effectively extract short-range contextual information around the target vertebrae, thereby realizing automatic recognition and positioning of the cone [21]. In our previous study, we proposed a two-stage spatial region-positioning algorithm for each lumbar segment. First, LRP-Net is used to locate the overall spatial region of the lumbar spine in the original CT image, and LDM-Net is used to achieve accurate positioning of each lumbar segment spatial region [22].

Owing to the limited resolution of CT images, the overlapping parts between adjacent vertebral bodies have blurred edges. Therefore, achieving an accurate pixel-level segmentation between adjacent vertebral bodies is the second difficulty in completing an automatic spine segmentation. In medical image segmentation, traditional machine learning methods have achieved certain achievements. Gauch proposed a Markov relaxation method for segmenting the internal structure of 3D MR brain images based on a watershed adjacency graph [23]. Kamiya et al. proposed an automatic segmentation method for the paraspinal muscles in 3D CT images of the trunk based on a multiscale iterative random forest classification [24]. In addition, Oktay et al. proposed a method for achieving a disc segmentation using a support vector machine (SVM) and an active appearance model [25]. Lecron et al. proposed a spine segmentation method based on a one-class support vector machine [26]. Zukić et al. proposed a method for segmenting the spine based on the Viola–Jones object detection framework [27]. With the rapid development of deep learning technology in recent years, efficient medical image segmentation algorithms have been proposed. Ronneberger et al. proposed U-Net at the ISBI Cell Tracking Challenge 2015 and achieved excellent results [28]. In addition, Milletari et al. proposed V-Net for 3D medical image segmentation [29], whereas Zhou et al. proposed a deeply supervised encoder-decoder network in which the encoder and decoder subnetworks are connected through a series of nested, dense skip pathways [30]. With the recent development of attention mechanisms, various image segmentation algorithms with combined attention mechanisms have been developed. Xiao et al. proposed Res-UNet with a weighted attention mechanism for retinal vascular segmentation [31]. Oktay et al. proposed a novel attention gate (AG) model for medical imaging that automatically learns to focus on target structures of varying shapes and sizes, based on which attention U-Net was developed [32]. Although both Res-UNet and attention U-Net combine attention mechanisms, the full potential of such mechanisms has yet to be shown. Vaswani et al. proposed a full attention-mechanism-based transformer and applied it to machine translation tasks with stunning results [33]. Dosovitskiy et al. also introduced a transformer into the field of computer vision and developed it for use in ViT, proving its significant potential [34]. Chen et al. combined a transformer and U-Net in the development of TransUNet, demonstrating the advantages of both [35]. Although both ViT and TransUNet use a transformer, they have not exploited its full potential. To solve the problem of insufficient local information extraction and a high computational complexity of a transformer, Liu et al. proposed Swin Transformer with a linear computational complexity. The algorithm has the capability to extract both global and local information and has achieved an outstanding performance in various computer vision tasks [36]. Cao et al. developed Swin-UNet with a U-Net-like structure based on Swin Transformer [37]. However, Swin-UNet does not have the ability to process 3D medical images. Moreover, the single-channel medical images are stacked into 3-channel images, similar to RGB images, and thus, Swin Transformer can be directly used with ImageNet to pretrain the weight, which has also been acknowledged by the author to be a suboptimal solution. At present, many scholars have proposed several improved algorithms for spine segmentation task based on the abovementioned algorithms [38–40].

Most of the abovementioned research studies only focus on one of the two tasks of target location or image segmentation and do not combine the two tasks. However, several studies have combined localization and segmentation algorithms to achieve an automatic segmentation of the spine within the region of a CT image. Because a disease is more likely to develop in the lumbar region, many scholars have developed segmentation methods for this region. Sekuboyina et al. used a multilayer perceptron to apply nonlinear regression to locate the lumbar region [41], based upon which 2D U-Net was used to segment the lumbar spine. Janssens et al. used a complete convolutional network to locate the lumbar region and applied 3D U-Net to segment the region [42]. Although the lumbar region has a higher incidence rate, spinal diseases can occur in various areas of the spine; therefore, the scan area of CT images is not fixed. This requires the spine image segmentation algorithm to accurately segment any portion of the spine in any scanning area. Cheng et al. proposed a three-stage approach to positioning each spinal segment through the first and second stages, and on this basis, finally realized the segmentation of each spinal segment. This method achieves spinal positioning through probabilistic reasoning and spinal segmentation through context-specific foreground and background constraints; thus, its accuracy and robustness are insufficient [43]. Lessmann et al. realized the segmentation of various spinal areas using a sliding window [44]; however, the realization process of this method is complicated, and the efficiency of the sliding window is low. Payer et al. realized landmark positioning of each spinal segment based on heatmap regression, and with each landmark as the center, intercepted local images in the space of a fixed size as the input of 3D U-Net, thus achieving segmentation of each spinal segment [45]. However, owing to the different sizes of the spine in different areas of different patients and the high similarity of adjacent spine structures, robustness is insufficient. Altini et al. automatically segmented a complete spine based on a CNN and combined it with the KNN algorithm to realize the separation of each segment [46], but the accuracy of this method is poor. Tao et al. developed an object detector with an internal sphere to locate the detection center of each spinal segment. Based on this, they used heat maps to refine the detection center. Finally, to achieve a segmentation of this spinal segment, with this detection center as the center, a fixed area of $[144, 144, 96]$ (dimensions according to ITK (z, y, x)) in size was cut as the input of the segmentation network [47]. The robustness of the algorithm is insufficient owing to the different sizes of the spine in each area and the high similarity of the structure of the adjacent spine.

None of the abovementioned research schemes can accurately locate the spatial region adaptive to the shape of each spinal segment; therefore, the interference caused by unnecessary background information and the similarity of adjacent spines cannot be completely eliminated. To solve this problem, the corresponding spinal segment should be surrounded by a retractable area, which should completely surround the segment and contain as little information as possible regarding other segments. De Vos et al. used a CNN to determine whether a specific ROI exists in 2D CT sections in three orthogonal directions and finally synthesized the final 3D boundary box [48]. This method cannot be used to directly obtain the target region and is unsuitable for the spatial localization of each spinal segment. In recent years, research has been conducted on the application of object detection algorithms in the field of natural image processing for medical image processing. Krawczyk and Starzyński used the YOLO algorithm to detect the pelvic region in CT images [49]. Xu et al. proposed a target detection algorithm for organ location based on a faster RCNN and combined it with the CT image features [50]. However, the algorithm combined with prior knowledge of a fixed number of organs was incapable of locating each spinal segment within an arbitrary FOV.

Accurate positioning of the spatial region of the spine in an arbitrary FOV is an important basis for an accurate segmentation of spine segments. However, owing to the uncertainty of the FOV, high levels of noise in low-dose CT, and metal artifacts left over from surgery, it is extremely difficult to achieve an accurate localization and segmentation of the spine. To address these shortcomings in existing research and achieve an accurate spinal positioning and segmentation, we developed 3D Swin-YoloX and 3D Swin-UNet in combination with Swin Transformer, the excellent target detection algorithm YoloX [51], and the image segmentation algorithm U-Net.

The main innovations of this paper are as follows:

(1) We extended Swin Transformer to 3D space and applied it to 3D medical image processing

(2) We proposed an efficient 3D medical image target detection algorithm, 3D Swin-YoloX, which can realize accurate positioning of each spinal segment in CT images under arbitrary FOV, which is not only efficient but also accurate

(3) We proposed an efficient 3D medical image segmentation algorithm, 3D Swin-UNet, which has great advantages over other medical image segmentation algorithms and has excellent processing ability for serious interference such as metal artifacts

(4) We proposed the overall scheme of first using the target detection algorithm to achieve accurate positioning of each spinal segment under arbitrary FOV and then achieving accurate segmentation of each spinal segment on this basis, which is more efficient and accurate than the scheme proposed by other scholars

2. Methods

We present a two-stage method for an accurate segmentation of each spinal segment in an arbitrary FOV. In the first stage, 3D Swin-YoloX is used to achieve accurate positioning of each spinal segment. In the second stage, 3D Swin-UNet is used to achieve an accurate segmentation of each spinal segment in the local area. The overall flow of the solution is shown in Figure 1.

2.1. Spinal Positioning. As shown in Figure 1, we included two substeps in the first stage. First, the whole spine is positioned in the original CT image, based on which accurate positioning of each spinal segment is achieved. Both steps are conducted using 3D Swin-YoloX. Although 3D Swin-YoloX can directly locate each spinal segment in the original CT image, owing to the uncertainty of the CT scan area, some CT images contain useless background information, including air, as well as the head, lower limbs, and other soft tissues of the patient, with only a small amount of

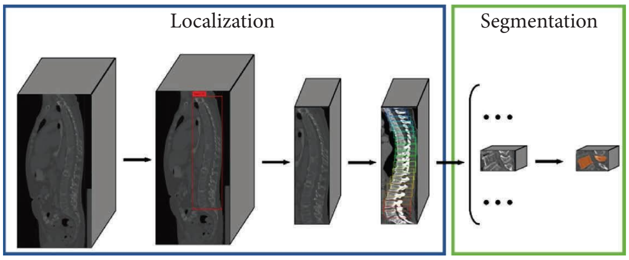

Figure 1: Overall scheme of the proposed process.

information about the spinal area being provided. However, owing to the limited hardware level, the 3D image size of input 3D Swin-YoloX is strictly limited, thereby requiring compression of these CT images, resulting in blurred and indistinguishable spinal segments. To solve this problem, many scholars have applied the sliding window method to their processing; however, this method requires at least two operations. We therefore used two 3D Swin-YoloX algorithms to locate the overall spine and the spatial regions of each spinal segment. To position the overall spatial region of the spine, it is unnecessary to obtain clear edges of each segment of the spine, and thus, the input image resolution can be relatively low. For the positioning of each spinal segment spatial area, the number of human spinal segments is limited; therefore, after realizing the positioning of the overall spinal spatial area, the image in this local area is scaled to a fixed size to ensure the clarity of each segment. Therefore, through these two substeps, we can minimize the number of calculations while ensuring positioning accuracy.

Owing to the excellent effect of Swin Transformer in the field of computer vision and the outstanding performance of YoloX in terms of target detection, we extended Swin Transformer and YoloX into three dimensions. On this basis, to build 3D Swin-YoloX, the backbone network of 3D YoloX is replaced with 3D Swin Transformer, the network structure of which is shown in Figure 2.

Similar to ViT and Swin Transformer, the input CT image is divided into nonoverlapping 3D patches of the same size, and the 3D patch is called a token. The 3D patch size $\mathrm{Ps} = (P_z, P_y, P_x)$ can be adjusted according to the actual needs. In order to achieve the initial extraction of shallow features from an image, a 3D convolution operation is adopted for processing CT images, and the feature dimension of each 3D patch is mapped to any dimension C. As the absolute position of each pixel or region in the image is not decisive to the information expressed by the image and when the image is divided into nonoverlapping tokens and expanded flat, the absolute position coding of these tokens will jump greatly. Therefore, when absolute position embedding is added into patch embedding, performance degrades [36]. Therefore, we do not add to patch embedding absolute position embedding representing the absolute position information. As shown in Figure 2, the dimensions of the input and output tensors of the patch embedding layer are $(B, 1, Z, Y, X)$ and (B, L, C), respectively, where B is the batch size, (Z, Y, X) is the size of the input image, C is the number of feature dimensions of the output, and $L = Z \times Y \times X / P_z \times P_y \times P_x$.

In the backbone network, we use the 3D Swin block to extract the image features, and a 3D Swin block is composed of multiple 3D basic Swin blocks in series. The 3D basic Swin block is obtained using a 3D extension based on the Swin Transformer block, and its specific structure is shown in Figure 3.

In view of the shortage of absolute position embedding and the lacking position information extraction ability in the standard transformer, Swin Transformer introduces a relative position bias to represent the relative position information among various tokens in the window. We adopted this method and obtained the relative position bias $B \in \mathbb{R}^{P_z \times P_y \times P_x}$ in a 3D space. The calculation method for attention is, therefore, as follows:

$$\mathrm{Attention}(Q, K, V) = \mathrm{Softmax}\left(\mathrm{QK}^T / \sqrt{d} + B\right)V, \quad (1)$$

where $Q, K, V \in \mathbb{R}^{P_z \times P_y \times P_x \times d}$ are the query, key, and value matrices and d is the dimension of Q. In each window, the relative position index between tokens is located between $[-P+1, P-1]$, where $P \in \{P_z, P_y, P_x\}$. We therefore set a learnable relative position bias table $\widehat{B} \in \mathbb{R}^{(2P_z-1)\times(2P_y-1)\times(2P_x-1)}$ and remove the corresponding relative position bias value in $\widehat{B}$ according to the relative position index used in the calculation process. Because the relative position index is located between $[-P+1, P-1]$, we add $P-1$ to the original relative position index to correctly obtain the corresponding value in $\widehat{B}$.

As shown in Figure 3, the 3D basic Swin block contains a 3D window multihead self-attention (3D W-MSA) module as well as a 3D shifted window multihead self-attention (3D SW-MSA) module. This is mainly because the 3D W-MSA module divides the feature image into nonoverlapping 3D windows of the same size and only calculates the relationship between each token in each window, thus reducing the computational complexity of the algorithm. However, in this way, the information interaction between tokens in different windows is isolated, and the ability to extract global information is lost. Therefore, the 3D SW-MSA module needs to be obtained by adding a shift window operation based on the 3D W-MSA module, and thus, an information interaction can be carried out among the tokens that are not originally in the same window but are actually connected within the image. Thus, the ability to extract global information is obtained. Therefore, to ensure that the 3D basic Swin block can extract both global and local information, we set it to contain both a 3D W-MSA module and a 3D SW-MSA module. To obtain the 3D SW-MSA module, we extended the shift window operation to a 3D space, as shown in Figure 4.

As shown in Figure 4, the 3D window is the same size in all three dimensions. We move the token in the window in three steps along the z, y, and x directions. The number of token layers that are moved each time is indicated by shifted_size. By default,

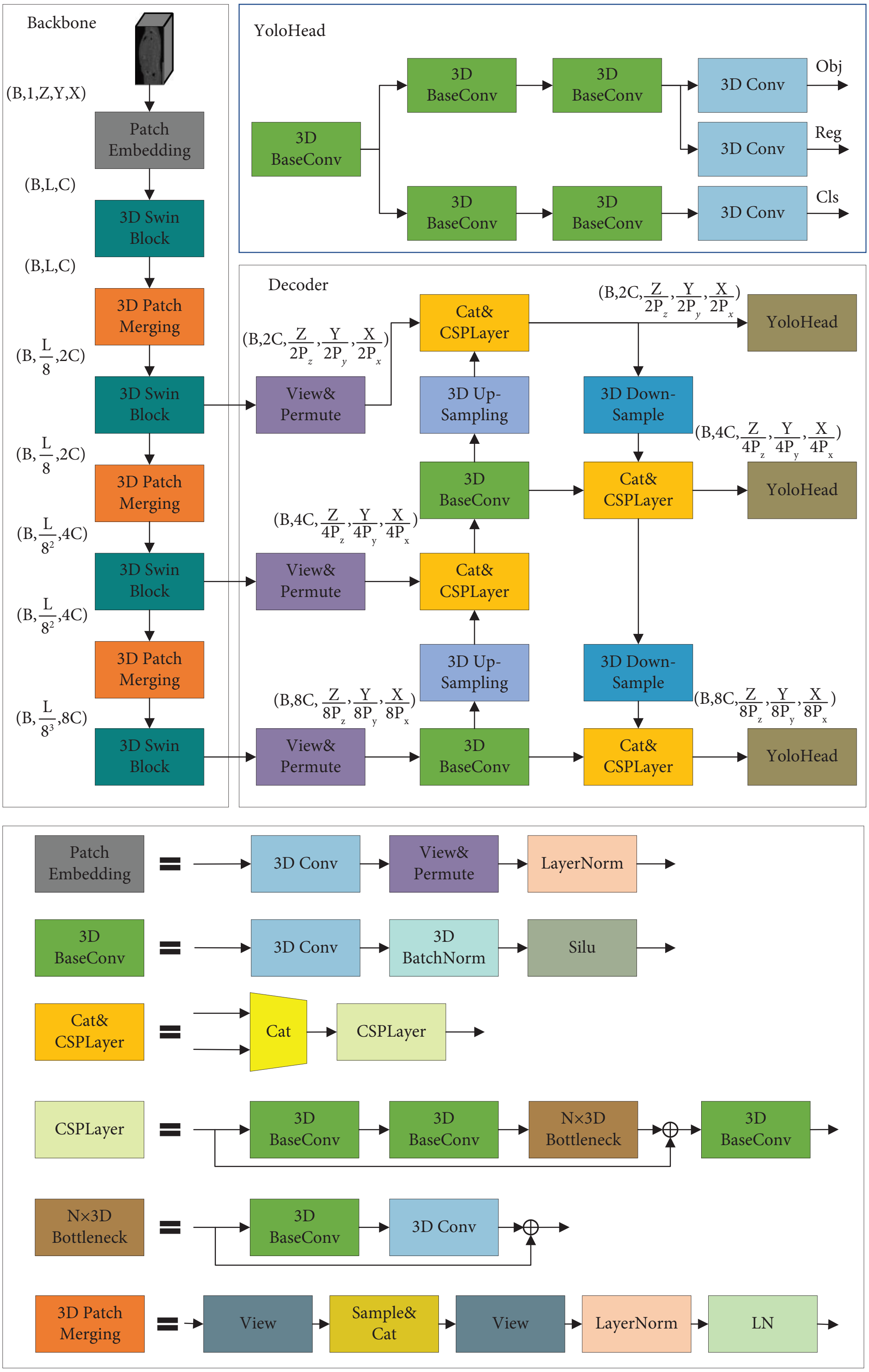
Backbone
(B,1,Z,Y,X)
Patch Embedding
(B,L,C)
3D Swin Block
(B,L,C)
3D Patch Merging
$(B,\frac{L}{8},2C)$
3D Swin Block
$(B,\frac{L}{8},2C)$
3D Patch Merging
$(B,\frac{L}{8^2},4C)$
3D Swin Block
$(B,\frac{L}{8^2},4C)$
3D Patch Merging
$(B,\frac{L}{8^3},8C)$
3D Swin Block
YoloHead
3D BaseConv
3D BaseConv
3D BaseConv
3D Conv
Obj
3D Conv
Reg
3D BaseConv
3D BaseConv
3D Conv
Cls
Decoder
$(B,2C,\frac{Z}{2P_z},\frac{Y}{2P_y},\frac{X}{2P_x})$
Cat& CSPLayer
$(B,2C,\frac{Z}{2P_z},\frac{Y}{2P_y},\frac{X}{2P_x})$
YoloHead
View& Permute
3D Up-Sampling
3D Down-Sample
$(B,4C,\frac{Z}{4P_z},\frac{Y}{4P_y},\frac{X}{4P_x})$
3D BaseConv
Cat& CSPLayer
YoloHead
$(B,4C,\frac{Z}{4P_z},\frac{Y}{4P_y},\frac{X}{4P_x})$
View& Permute
Cat& CSPLayer
3D Up-Sampling
3D Down-Sample
$(B,8C,\frac{Z}{8P_z},\frac{Y}{8P_y},\frac{X}{8P_x})$
$(B,8C,\frac{Z}{8P_z},\frac{Y}{8P_y},\frac{X}{8P_x})$
View& Permute
3D BaseConv
Cat& CSPLayer
YoloHead
Patch Embedding = 3D Conv View& Permute LayerNorm
3D BaseConv = 3D Conv 3D BatchNorm Silu
Cat& CSPLayer = Cat CSPLayer
CSPLayer = 3D BaseConv 3D BaseConv N×3D Bottleneck 3D BaseConv
N×3D Bottleneck = 3D BaseConv 3D Conv
3D Patch Merging = View Sample& Cat View LayerNorm LN

Figure 2: Structure diagram of 3D Swin-YoloX.

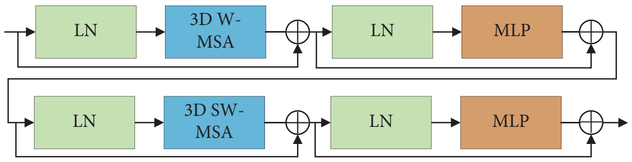

Figure 3: Structure of 3D basic Swin block.

$$\text{shifted}_{\text{size}} = \text{floor}\left(\frac{P_w}{2}\right), \tag{2}$$

where P_w represents the size of the 3D window in the z, y, and x dimensions. All multihead attention modules in the 3D Swin block have the same number of heads.

To ensure that the algorithm has a strong ability to extract global information, it is necessary to enlarge the receptive field of the 3D window through downsampling. This is the same principle as enlarging the receptive field of the convolution kernel through subsampling used with the CNN algorithm. We expanded the patch merging layer into the 3D space and obtained the 3D patch merging layer, replacing the pooling operation that will lose information to realize the downsampling of the feature image to enlarge the receptive field of the 3D window and assist SW-MSA in extracting global information. The size relationship between the input tensor and the output tensor of the 3D patch merging layer is as follows: $(B, L_P, C_P) \longrightarrow (B, L_P/8, 2C_P)$.

After resampling CT images with an arbitrary FOV to a fixed size, the size of both the whole spine and each spinal segment is uncertain. We therefore retain the design of the feature grid output of the three scales in YoloX such that it has the ability to accurately detect targets of different sizes. Moreover, because the multiscale feature fusion mechanisms of both the YoloX network and YoloHead, which can retrieve confidence Obj, coordinate information Reg, and target class Cls, are extremely effective, we retained these structures in YoloX. In the decoder part of 3D Swin-YoloX, we did not adopt a 3D patch merging layer to achieve downsampling and instead adopted an ordinary 3D max pooling module. We do not adopt this structure in the decoder mainly because limited by the hardware level, the algorithm needs to achieve a balance between precision and complexity and the 3D patch merging layer is quite complex. Moreover, the backbone network extracts sufficiently rich feature information, providing a good basis for the decoder to achieve accurate target identification.

2.2. Spinal Image Segmentation.

2.2. Spinal Image Segmentation. After achieving accurate positioning of each spinal segment, CT images are intercepted according to the local spatial area obtained through positioning, allowing a local CT image to be created that contains the complete spinal segment with as little information as possible about the other segments. Although a traditional CNN such as U-Net has been able to achieve an accurate segmentation of the spinal segments in local CT images, such approaches cannot cope with certain cases. For example, CT images of the spine may contain "feature holes" caused by osteoporosis and artifacts caused by metal objects left over from surgery. Although convolution has a strong local information extraction capability, the global information extraction capability is weak, and the abovementioned problems cannot be effectively solved based only on local information. Many scholars have proven that the introduction of global information plays a significant role in improving image segmentation [52–54]. Although most previous convolutional neural networks can enlarge the receptive field of the convolution kernel through downsampling, thereby extracting global information owing to the intrinsic locality of convolution operations, it is difficult for CNN-based approaches to learn explicit global and long-range semantic information interactions [35]. However, based on the shifted window mechanism, Swin Transformer retains the powerful global information extraction capability of a transformer, and the algorithm can effectively deal with information that needs to be focused on through the attention mechanism; it therefore has the potential to solve the abovementioned problems. In addition, Swin Transformer effectively draws on the advantages of the convolution operation and has the ability to extract local information by calculating the relationship between each token in the window. Therefore, Swin Transformer has powerful local and global information extraction capabilities. In conclusion, to develop 3D Swin-UNet, we chose Swin Transformer based on 3D extension, and we finally achieved an accurate segmentation of each spinal segment. The structure of 3D Swin-UNet is shown in Figure 5.

Patch embedding, 3D Swin block, and the 3D patch merging layer in the algorithm are all identical to those of 3D Swin-YoloX. The 3D patch expanding layer is the inverse operation of the 3D patch merging layer, which is achieved through a 3D expansion based on the patch expanding layer proposed by Cao et al. [37]. The size relationship between the input tensor and output tensor of the 3D patch expanding layer is as follows: $(B, L_P, C_P) \longrightarrow (B, 8L_P, C_P/2)$. The final expanding layer has the same structure as the 3D patch expanding layer, except that it has a higher amplification ratio of the feature tensor resolution and can realize feature integration and tensor size transformation. The size relationship between the input tensor and output tensor is as follows: $(B, L, C) \longrightarrow (B, P_z \times P_y \times P_x \times L, 1) \longrightarrow (B, 1, Z, Y, X)$.

3D Swin-UNet is the same as U-Net, including the encoder and decoder. The encoder expands the field of perception of each stereo window through the 3D patch merging layer and efficiently extracts global and local information with a 3D Swin block. The decoder restores the feature tensor obtained from the encoder to the same resolution as the input CT image to achieve an accurate pixel-level segmentation. Between the encoder and decoder, shallow detailed information from feature maps of different scales can be combined with deep semantic information by jumping connections. The number of heads of multihead attention modules in the corresponding 3D Swin block in the encoder and decoder is the same.

We adopted Dice loss as a loss function to optimize 3D Swin-UNet. This is calculated as follows:

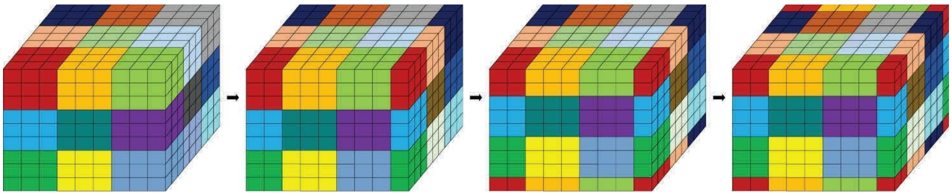

Figure 4: Shift window operation extended to a 3D space.

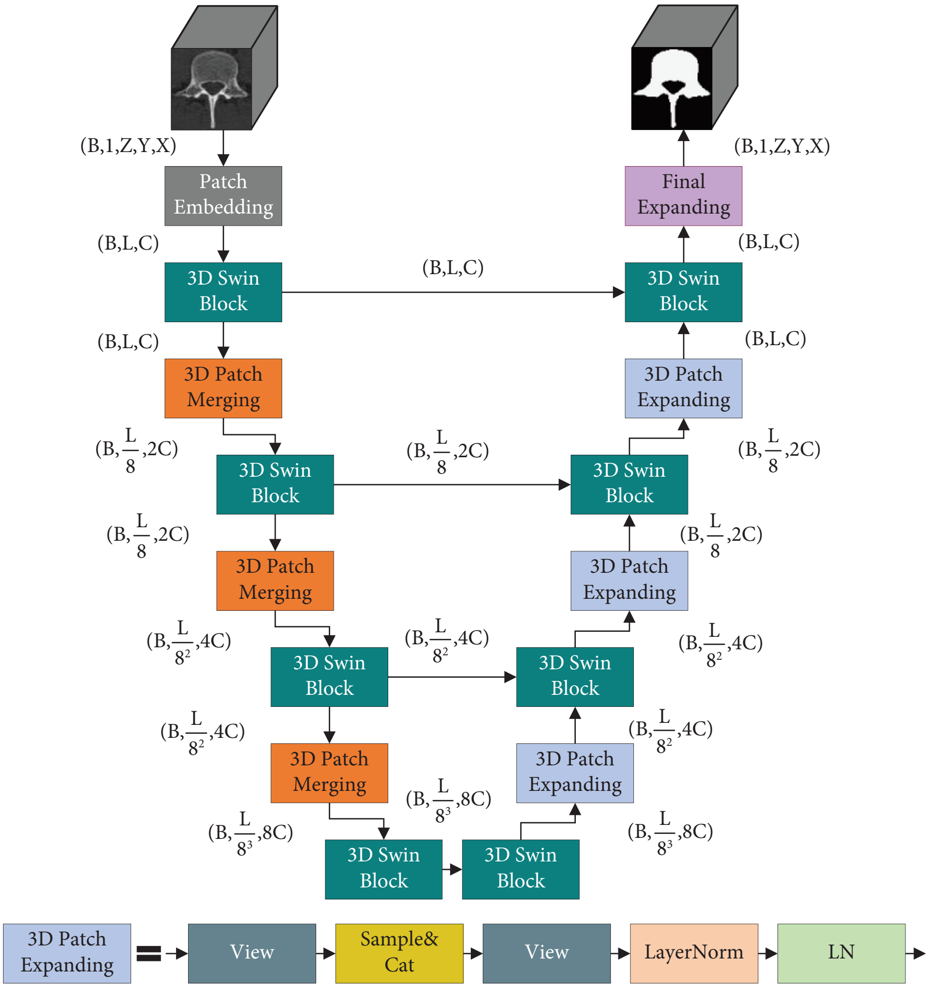

Figure 5: Structure diagram of 3D Swin-UNet.

$$\text{Dice Loss} = 1 - \text{Dice} = 1 - \frac{2\left|\text{Seg}_{\text{pre}} \cap \text{Seg}_{\text{tar}}\right|}{\left|\text{Seg}_{\text{pre}}\right| + \left|\text{Seg}_{\text{tar}}\right|}, \quad (3)$$

where Seg_{pre} represents the segmentation result output by the network and Seg_{tar} represents the real segmentation label.

2.3. Implementation Details. As shown in Figure 6, image A is the input image of first 3D Swin-YoloX, image B is the input image of second 3D Swin-YoloX, and images *C*1, *C*2, *C*3,. . .*Cn* are the input images of 3D Swin-UNet. Image A is resampled to a fixed size of (256, 160, 160), image B is resampled to a fixed size of (224, 128, 128), and images *C*1, *C*2, *C*3,. . .*Cn* are resampled to a fixed size of (96,128,128). In the process of image resampling, it is necessary to resample the segmentation label corresponding to the image to ensure that the image and label have the same size. Image resampling is conducted based on cubic spline interpolation, and label resampling is achieved based on the nearest-neighbor interpolation.

For 3D Swin-YoloX, we set $C = 32$, $Ps = (P_z, P_y, P_x) = (4, 2, 2)$, and the number of 3D basic Swin blocks in all 3D Swin blocks is 2. Dimensions: $P_w = 3$ of the 3D window in the 3D W-MSA module and the 3D SW-MSA module. The number of heads of multihead attention modules in the four 3D Swin blocks is $N_H = (4, 4, 8, \text{and } 8)$.

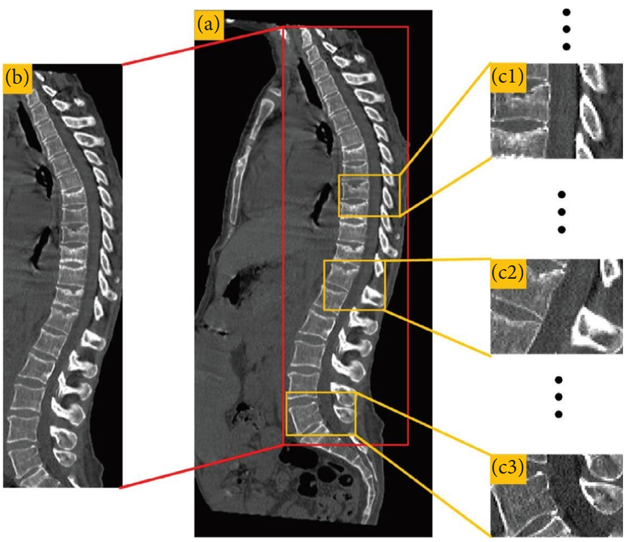

Figure 6: Input image data from each stage of the algorithm.

For 3D Swin-UNet, we set $C = 96$, $Ps = (P_z, P_y, P_x) = (3, 4, 4)$, and the number of 3D basic Swin blocks in all 3D Swin blocks is 2. Dimensions: $P_w = 3$ of the 3D window in the 3D W-MSA module and the 3D SW-MSA module. In addition, the number of heads of the multihead attention modules in the four 3D Swin blocks in the encoder and decoder is $N_H = (8, 12, 12, \text{and}\, 24)$.

Network training and verification were conducted on a server equipped with four NVIDIA Quadro RTX6000 graphic cards (we only used two of them), and the algorithm was implemented under the Pytorch framework. In the training process of 3D Swin-YoloX and 3D Swin-UNet, the initial learning rate was $1e-4$, and for every 40 rounds of training, the learning rate was reduced by half. 3D Swin-YoloX and 3D Swin-UNet were trained for 200 rounds each, and the batch size of first 3D Swin-YoloX was 4, whereas the batch size of the other two algorithms was 2. During the training process, the Adam optimizer was used to optimize the iterative training of the algorithm. Owing to the limited number of CT images in the dataset, we enhanced the random data during the training process. The intensity values of the images were multiplied randomly using $[0.5, 1.5]$, and the images were randomly rotated using $[-15^\circ, 15]$. To increase the contrast between the spine and soft tissue, the window width and position of the image must be adjusted before it is input into the algorithm. The calculation method is as follows:

$$P_{\text{Des}} = \begin{cases} 0, & P_{\text{Src}} \le \text{Min}_{\text{Bound,}} \\ \dfrac{P_{\text{Src}} - (\text{Min}_{\text{Bound}})}{\text{Max}_{\text{Bound}} - (\text{Min}_{\text{Bound}})}, & \text{Min}_{\text{Bound}} < P_{\text{Src}} < \text{Max}_{\text{Bound,}} \\ 1, & P_{\text{Src}} \ge \text{Max}_{\text{Bound}}. \end{cases} \tag{4}$$

In this paper, $\text{Min}_{\text{Bound}} = -200$ and $\text{Max}_{\text{Bound}} = 600$. By adjusting the window width and position, CT images can show the spine more prominently, and the image pixels are compressed to a range of $[0, 1]$.

3. Experiment

We trained and tested 3D Swin-YoloX and 3D Swin-UNet on the large-scale vertebrae segmentation challenge (VerSe) data [55]. The VerSe 2019 challenge dataset includes 160 CT image series and contains 220 cervical, 884 thoracic, and 621 lumbar segments for a total of 1,725 spinal segments. All data labels in the VerSe 2019 dataset were annotated by five trained medical students and modified and refined by three trained radiologists with 30 years of experience. As a result, the dataset is highly accurate. Because the VerSe 2019 challenge dataset contains a voxel-wise segmentation (Seg_{tar}) for each spinal segment, we used Seg_{tar} to compute the regional coordinates for the entire spine, as well as the individual spinal segments. These regions are represented as Roi: $(z_{\min}, y_{\min}, x_{\min}, z_{\max}, y_{\max}, x_{\max})$ and are used as labels for 3D Swin-YoloX positioning. In addition, we intercepted the local CT image information in the corresponding space area of each spinal segment according to Roi, and the corresponding segmentation labels were intercepted. All regions corresponding to incomplete spinal segments in the segmentation label were filled with 0 s, and all regions corresponding to complete spinal segments were filled with 1 s. Finally, a binary image highlighting the complete spinal segment was obtained. The processing flow is illustrated in Figure 7.

We used the local CT image containing the complete intercepted spinal segment as the input of 3D Swin-UNet, as shown in Figure 7, and used the processed binary image as the corresponding label. We used the Dice coefficient (Dice), Hausdorff distance (HD), boundary $F1$ score [56], and

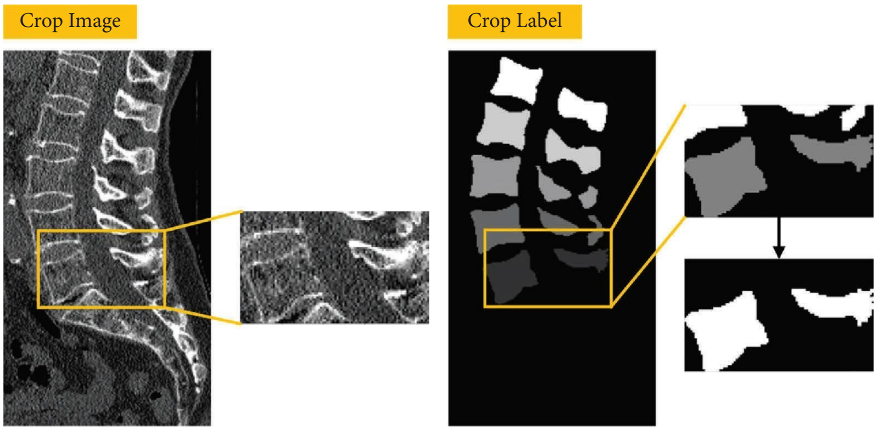

Figure 7: Diagram of CT image and label processing flow (taking L5 as an example).

boundary IoU [57] as indicators to evaluate the segmentation effect. We projected the segmentation results $\mathrm{Seg}_{\mathrm{pre}}^{\mathrm{sub}}$ of each spinal segment obtained through positioning and segmentation into the all-0 image $\mathrm{Seg}_{\mathrm{pre}}$ with the same size as the original CT image and finally calculated these indicators based on $\mathrm{Seg}_{\mathrm{pre}}$ and $\mathrm{Seg}_{\mathrm{tar}}$. As shown in Figure 8, to describe the overall surgical process, including spinal positioning, segmentation, and projection, we selected a set of CT images containing only five lumbar segments as examples.

After above processing, to observe the segmentation and projection effect more directly, VTK was used to process $\mathrm{Seg}_{\mathrm{pre}}$ and realize the three-dimensional reconstruction of the spine. The reconstruction results are shown in Figure 9.

Since there are no data on the boundary $F1$ score and boundary IoU in the relevant data disclosed by other scholars, we only list Dice and HD data of the overall scheme for comparison here. We counted the Dice and HD mean and median values calculated according to $\mathrm{Seg}_{\mathrm{pre}}$ and $\mathrm{Seg}_{\mathrm{tar}}$. The final mean values (median values) of the experimental results are listed in Table 1.

It can be observed from Table 1 that our method achieved better results than the most advanced method. For the public test data, the method proposed by Sekuboyina et al. achieved the best results before [55]. The average Dice and HD of Sekuboyina et al.'s method were 0.930 and 6.39, respectively, whereas the average Dice and HD of our method were 0.942 and 6.24, respectively, both of which are better than those of the method proposed by Sekuboyina et al. However, in the hidden test data, the method proposed by Tao et al. previously obtained the best results before [47]. The mean Dice and HD of Tao et al.'s method were 0.901 and 6.68, respectively, whereas the mean Dice and HD of our method were 0.941 and 6.00, which were also better than those of Tao et al.'s method. This is closely related to the excellent performance of 3D Swin-UNet. To prove this, we compared 3D Swin-UNet with various excellent previously published medical image segmentation algorithms. To eliminate other influences, we conducted the experiment directly in the local CT image sequence containing the complete spinal segment intercepted according to $\mathrm{Seg}_{\mathrm{tar}}$ instead of the local CT image provided by 3D Swin-YoloX. In addition, we did not reproject $\mathrm{Seg}_{\mathrm{pre}}^{\mathrm{sub}}$ segmented using various algorithms and instead calculated each indicators directly according to $\mathrm{Seg}_{\mathrm{pre}}^{\mathrm{sub}}$ and the segmentation truth value $\mathrm{Seg}_{\mathrm{tar}}^{\mathrm{sub}}$ of the corresponding local CT image. The final experimental results are shown in Table 2.

Table 2 shows that 3D Swin-UNet achieved a better performance than the other algorithms. For the public test data, attention U-Net previously achieved the best results [32]. The average Dice, HD, boundary $F1$ score, and boundary IoU of attention U-Net were 0.948, 7.68, 0.942, and 0.892, respectively, whereas the average Dice, HD, boundary $F1$ score, and boundary IoU of 3D Swin-UNet were 0.951, 5.57, 0.943, and 0.899, respectively, which are both better than those of attention U-Net. However, in the hidden test data, attention U-Net still obtained the previous best result. The average values of attention U-Net's test indicators were 0.944, 8.91, 0.933, and 0.896, respectively, whereas the average values of 3D Swin-UNet's test indicators were 0.949, 11.2, 0.942, and 0.898, respectively. This is mainly due to the excellent global information extraction capability achieved by Swin Transformer, which can help the algorithm effectively improve its ability to cope with interference such as high levels of noise and metal artifacts. To visualize the excellent performance of 3D Swin-UNet, we provide some examples of actual segmentation effects in Figure 10.

As shown in Figure 10, compared with other algorithms, 3D Swin-UNet has a significant advantage in processing metal artifacts and high-noise CT images. In particular, owing to its powerful capability to extract global information, it has an extremely strong processing capacity for metal artifacts. In addition, by comparing the 3D Swin-UNet segmentation results listed in Tables 1 and 2, it can be seen that 3D Swin-YoloX achieves a high positioning accuracy for the spatial regions of each spinal segment, and the positioning error causes a small decrease in the segmentation accuracy, which can provide an extremely good basis for spinal image segmentation. To prove this point, we used 3D_IOU [22] as the evaluation index and tested the comprehensive positioning accuracy of two 3D Swin-YoloX models for each spinal segment space region under an

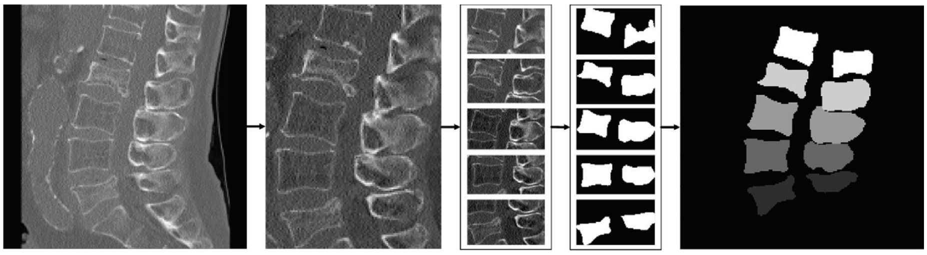

Figure 8: Schematic diagram of the overall processing flow.

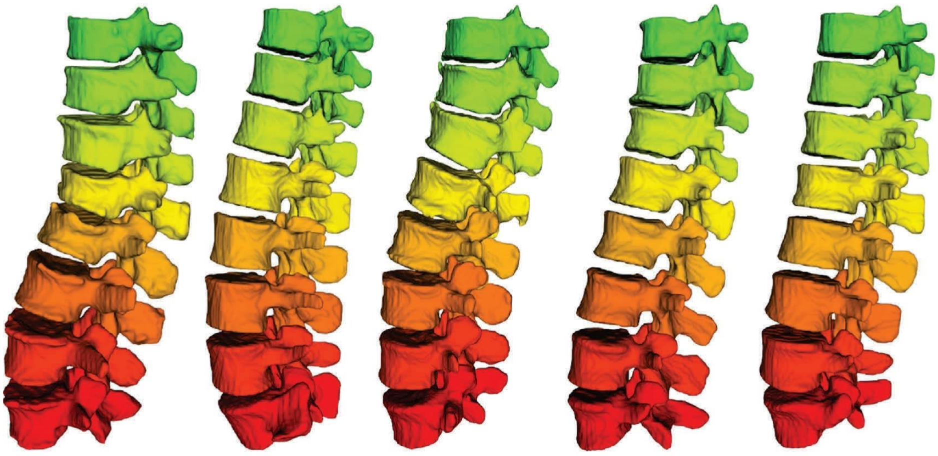

Figure 9: Schematic diagram of three-dimensional spine reconstruction.

Table 1: Vertebral segmentation results (mean and median (in parentheses)) from the VerSe 2019 challenge dataset were evaluated.

Ref. authors	Evaluated on public test data		Evaluated on hidden test data	
	Dice	HD	Dice	HD
Tao et al. [47]	0.911 (0.950)	6.34 (4.12)	0.901 (0.939)	6.68 (4.12)
Payer et al. [45]	0.910 (0.955)	6.35 (4.62)	0.898 (0.955)	7.08 (4.45)
Sekuboyina et al. [55]	0.930 (0.960)	6.39 (4.88)	0.826 (0.965)	9.98 (5.71)
Lessmann et al. [44]	0.851 (0.943)	8.58 (4.62)	0.858 (0.939)	8.20 (5.38)
Our method	**0.942 (0.945)**	**6.24 (5.99)**	**0.941 (0.945)**	**6.00 (5.58)**

The best average results are highlighted in bold. The data for the methods proposed by Sekuboyina et al., listed in the table, are derived from the VerSe official [55].

Table 2: Vertebral segmentation results (mean and median (in parentheses)) were evaluated against those of other segmentation algorithms on the VerSe 2019 challenge dataset.

		Dice	HD	Boundary *F*1 score	Boundary IoU
Evaluated on public test data	U-Net [28]	0.944 (0.950)	8.05 (6.16)	0.936 (0.939)	0.890 (0.904)
	V-Net [29]	0.934 (0.948)	18.1 (11.3)	0.927 (0.935)	0.872 (0.898)
	U-Net++ [30]	0.947 (0.952)	6.68 (5.10)	0.942 (0.947)	0.889 (0.902)
	Res-UNet [31]	0.941 (0.948)	10.7 (7.35)	0.929 (0.946)	0.890 (0.901)
	Attention U-Net [32]	0.948 (0.952)	7.68 (4.90)	0.942 (0.947)	0.891 (0.901)
	Our method	**0.951 (0.956)**	**5.57 (4.12)**	**0.942 (0.948)**	**0.892 (0.901)**
Evaluated on hidden test data	U-Net [28]	0.942 (0.948)	**7.66 (5.74)**	0.940 (0.946)	0.887 (0.899)
	V-Net [29]	0.936 (0.944)	15.7 (12.4)	0.937 (0.945)	0.882 (0.895)
	U-Net++ [30]	0.942 (0.949)	**8.33 (5.92)**	0.942 (0.949)	0.892 (0.903)
	Res-UNet [31]	0.941 (0.947)	9.04 (6.71)	0.926 (0.944)	0.866 (0.895)
	Attention U-Net [32]	0.944 (0.950)	8.91 (5.83)	0.933 (0.942)	0.896 (0.905)
	Our method	**0.949 (0.956)**	11.21 (6.00)	**0.942 (0.951)**	**0.898 (0.907)**

The best average results are highlighted in bold.

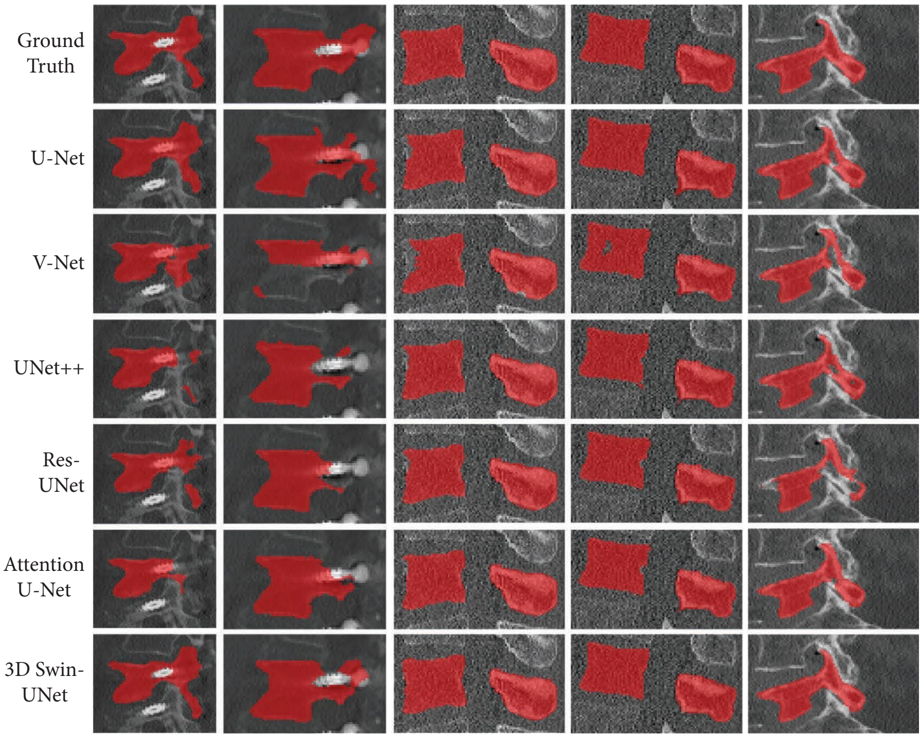

Figure 10: Examples of segmentation effects of various algorithms.

arbitrary FOV on the test dataset of the VerSe 2019 challenge. The final statistical results are presented in Table 3.

As shown in Table 3, 3D Swin-YoloX achieves a high positioning accuracy for each spinal segment space. As shown in Figure 11, boxes were used to display the spatial area obtained through positioning on the CT section, and different colors were applied to distinguish different boxes to more intuitively observe the positioning effect of the spatial area of each spinal segment.

As indicated in Figure 11, the positioning scheme proposed in this study can effectively deal with problems such as an uncertain scanning area, high noise, metal artifacts, and divergence of the bone cementing agent, with strong robustness.

In our previous relevant research, we have compared the scheme proposed by us with that proposed by other scholars, proving the superiority of our positioning method at that time [22]. The average 3D_IOU for spatial region localization of each spine segment in the previous method can reach 0.8559, while the method proposed in this paper can reach 0.8962. Therefore, the new localization method proposed by us is excellent.

In order to test the effect of different parameter settings on the 3D basic Swin block, we conducted more experiments. The configurable parameters of the 3D basic Swin block are mainly N_H and P_w. We, respectively, tested the effects of different N_H when P_w is fixed and different P_w when N_H is fixed on 3D Swin-YoloX and 3D Swin-UNet.

When $P_w = 3$, the test results of the influence of different N_H on 3D Swin-YoloX are shown in Table 4.

As shown in Table 4, when P_w remains unchanged, the larger N_H is, the better the 3D Swin-YoloX effect will be. However, with the further increase of N_H, the increase of 3D Swin-YoloX positioning accuracy becomes smaller and smaller. Considering the positioning accuracy and complexity of 3D Swin-YoloX, we finally set $N_H = (4, 4, 8, \text{and } 8)$.

Table 3: Positioning accuracy of each spinal segment space under an arbitrary FOV.

	Evaluated on public test data	Evaluated on hidden test data
Mean	0.896	0.898
Median	0.917	0.917
Std	0.082	0.079

When $N_H = (4, 4, 8, \text{and } 8)$, the test results of the influence of different P_w on 3D Swin-YoloX are shown in Table 5.

As shown in Table 5, when N_H is fixed, the smaller P_w is, the higher the positioning accuracy of 3D Swin-YoloX is. This is mainly because the smaller the 3D window, the stronger the 3D basic Swin block's ability to extract detailed features. In addition, the smaller P_w, the smaller the complexity of 3D Swin-YoloX, so we set $P_w = 3$.

When $P_w = 3$, the test results of the influence of different N_H on 3D Swin-UNet are shown in Table 6.

As shown in Table 6, when P_w remains unchanged, the larger N_H is, the better the 3D Swin-UNet effect will be. However, with the further increase of N_H, the increase of 3D Swin-UNet segmentation accuracy is also decreasing. Considering the segmentation accuracy and complexity of 3D Swin-UNet, we finally set $N_H = (8, 12, 12, \text{and } 24)$.

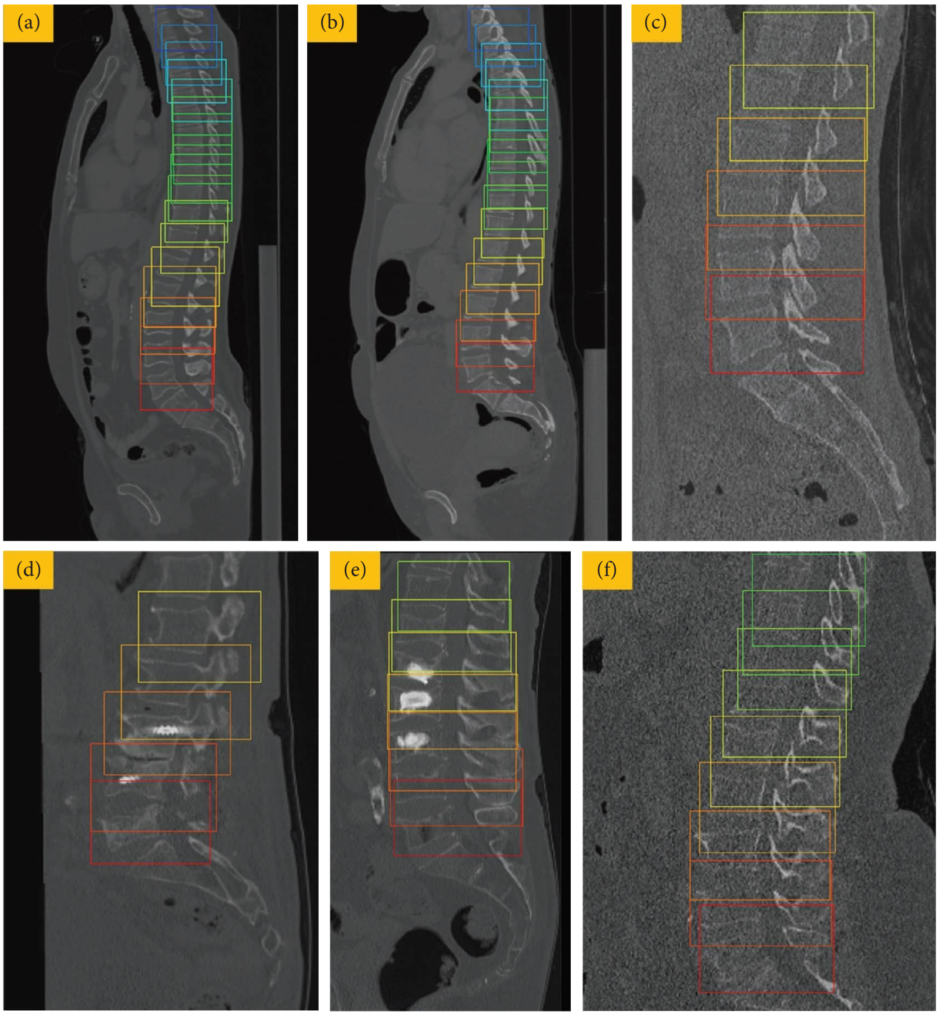

Figure 11: Spatial localization effect of each spinal segment.

Table 4: The influence of N_H on 3D Swin-YoloX.

	N_H	3D_IOU
Evaluated on public test data	$(4, 4, 4, 4)$	0.873 (0.904)
	$(4, 4, 4, 8)$	0.881 (0.903)
	$(4, 4, 8, 8)$	0.896 (0.917)
	$(4, 8, 8, 8)$	0.902 (0.921)
Evaluated on hidden test data	$(4, 4, 4, 4)$	0.867 (0.901)
	$(4, 4, 4, 8)$	0.887 (0.899)
	$(4, 4, 8, 8)$	0.898 (0.917)
	$(4, 8, 8, 8)$	0.904 (0.915)

Table 5: The influence of P_w on 3D Swin-YoloX.

	P_w	3D_IOU
Evaluated on public test data	3	0.896 (0.917)
	4	0.892 (0.905)
	5	0.884 (0.897)
	6	0.860 (0.881)
Evaluated on hidden test data	3	0.898 (0.917)
	4	0.891 (0.913)
	5	0.881 (0.901)
	6	0.869 (0.883)

Table 6: The influence of N_H on 3D Swin-UNet.

	N_H	Dice	HD	$F1$	IOU
Evaluated on public test data	$(8, 12, 12, 12)$	0.948 (0.951)	6.04 (5.66)	0.940 (0.945)	0.890 (0.897)
	$(8, 12, 12, 24)$	0.951 (0.956)	5.57 (4.12)	0.942 (0.948)	0.892 (0.901)
	$(8, 24, 24, 24)$	0.953 (0.958)	5.26 (4.97)	0.944 (0.949)	0.894 (0.902)
	$(16, 24, 24, 24)$	0.954 (0.961)	5.09 (5.03)	0.943 (0.948)	0.892 (0.901)
Evaluated on hidden test data	$(8, 12, 12, 12)$	0.947 (0.953)	11.41 (6.20)	0.940 (0.950)	0.897 (0.906)
	$(8, 12, 12, 24)$	0.949 (0.956)	11.21 (6.00)	0.942 (0.951)	0.898 (0.907)
	$(8, 24, 24, 24)$	0.950 (0.958)	10.57 (5.72)	0.943 (0.951)	0.900 (0.908)
	$(16, 24, 24, 24)$	0.951 (0.963)	9.38 (5.41)	0.944 (0.952)	0.901 (0.906)

Table 7: The influence of P_w on 3D Swin-UNet.

	P_w	Dice	HD	$F1$	IOU
Evaluated on public test data	3	0.951 (0.956)	5.57 (4.12)	0.942 (0.948)	0.892 (0.901)
	4	0.949 (0.954)	6.30 (5.20)	0.940 (0.941)	0.889 (0.897)
	5	0.947 (0.956)	6.57 (5.32)	0.938 (0.940)	0.882 (0.891)
	6	0.940 (0.945)	7.34 (5.39)	0.932 (0.934)	0.873 (0.884)
Evaluated on hidden test data	3	0.949 (0.956)	11.21 (6.00)	0.942 (0.951)	0.898 (0.907)
	4	0.947 (0.952)	11.32 (6.83)	0.938 (0.946)	0.894 (0.901)
	5	0.945 (0.948)	11.41 (7.10)	0.932 (0.941)	0.887 (0.897)
	6	0.937 (0.942)	11.94 (8.92)	0.923 (0.929)	0.874 (0.882)

When $N_H = (8, 12, 12, \text{and } 24)$, the test results of the influence of different P_w on 3D Swin-UNet are shown in Table 7.

As shown in Table 7, when N_H is fixed, the segmentation accuracy of 3D Swin-UNet with smaller P_w is higher, mainly because the smaller the 3D window, the stronger the ability of the 3D basic Swin block to extract detailed features. In addition, the smaller P_w, the smaller the complexity of the 3D Swin-UNet. So, we set $P_w = 3$.

In summary, the smaller P_w is, the stronger the feature extraction ability of the 3D basic Swin block is. The larger N_H is, the stronger the feature extraction ability of the 3D basic Swin block is.

Due to the potential of Hausdorff loss [58] (HD loss) in improving the capture of boundary details by segmentation algorithms, we also tested the effect of HD loss on 3D Swin-UNet. The test results are shown in Table 8.

As shown in Table 8, HD Loss does improve the ability of 3D Swin-UNet to capture boundaries, but it has certain side effects on the overall effect. Therefore, in the final scheme, we did not use HD loss for supervised training of 3D Swin-UNet.

After completing the design and development of the overall program, we invited experienced doctors to test and evaluate our program. The test data were collected from multiple hospitals, and sensitive patient information was removed. A total of 60 CT images, including 258 spinal segment information. We divided the evaluation levels into A (satisfied) and B (unsatisfied). Some evaluation cases are shown in Figure 12.

The test results are shown in Table 9.

As shown in Table 9, the practical value of our method in clinical application is very high, and in most cases, it is sufficient to meet clinical needs.

4. Discussion

Autonomous spine segmentation under an arbitrary FOV has a strong application value in medical robotic surgical path planning, intelligent medical diagnosis, and other methods. Although many solutions have been developed, they have been based on the use of a CNN. Owing to the limited ability of convolutional operations to extract global information, their effect has certain defects. However, Swin Transformer has an excellent global information extraction capability and a local information extraction capability equal to that of a convolution operation, thus achieving a better effect.

We therefore propose a two-stage, fully autonomous spinal segmentation method under an arbitrary FOV based on 3D Swin Transformer. The first stage consists of two steps. First, accurate positioning of the overall spine region is achieved based on 3D Swin-YoloX, based on which accurate positioning of each segment within the overall spine region is achieved. In the second stage, a local CT image containing the complete spinal segment obtained in the first stage is used as the input for 3D Swin-UNet, thus achieving an accurate segmentation of the spine.

Although our method already has a lower computational cost, particularly in the first stage, the precise positioning of each spinal segment is achieved through two substeps, which avoids multiple calculations through the sliding window. However, at present, we cannot quickly or accurately locate each spinal segment in the original CT image in a single step. In the future, we will develop a more efficient network structure allowing a larger sized image to be input into the network, and the precise location of each spinal segment can be determined in a single step.

Table 8: The influence of HD loss on 3D Swin-UNet.

	Loss function	Dice	HD	$F1$	IOU
Evaluated on public test data	Dice loss	0.951 (0.956)	5.57 (4.12)	0.942 (0.948)	0.892 (0.901)
	HD loss + Dice loss	0.946 (0.948)	5.37 (5.23)	0.937 (0.941)	0.885 (0.888)
	HD loss	0.938 (0.943)	5.23 (5.17)	0.932 (0.937)	0.881 (0.883)
Evaluated on hidden test data	Dice loss	0.949 (0.956)	11.21 (6.00)	0.942 (0.951)	0.898 (0.907)
	HD loss + Dice loss	0.944 (0.952)	7.01 (5.03)	0.939 (0.941)	0.891 (0.901)
	HD loss	0.942 (0.950)	6.83 (5.01)	0.935 (0.939)	0.887 (0.893)

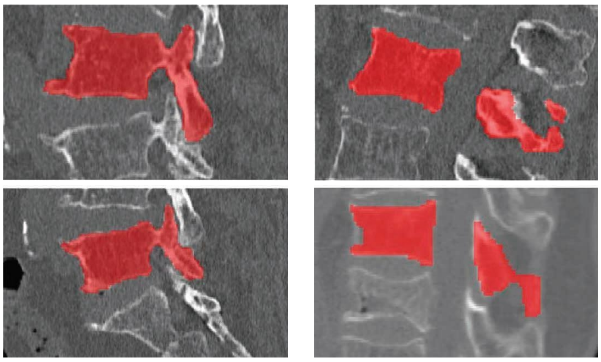
(a) (b)

Figure 12: Evaluation case.

Table 9: Doctor's evaluation of the system.

Satisfaction	A	B
Count	252	6
Ratio	97.67%	2.33%

Although we used as much data as possible when training the algorithm, relatively few of the CT data included cervical vertebral segments. Moreover, the spatial overlap rate of the adjacent cervical vertebral segments is relatively high, and it is therefore more difficult to achieve an accurate differentiation and segmentation of each cervical vertebra segment. Some defects in the segmentation effect on the cervical vertebrae remain when applying our method. Therefore, in the future, to improve the segmentation effect of our method when applied to cervical segments, we will add CT data on the cervical segments and corresponding segmentation labels.

5. Conclusion

In this paper, we propose a two-stage arbitrary FOV segmentation method for the spine based on 3D Swin Transformer. This method requires fewer computations and achieves a better segmentation effect than various arbitrary FOV segmentation methods proposed by other scholars. This is mainly due to the following: (1) in the first stage, as much of the excess soft tissue and other spinal segment interference as possible is removed. (2) In the second stage, 3D Swin-UNet has both powerful global and local information extraction capabilities and has stronger processing capabilities compared with other image segmentation algorithms based on convolutional operation for special cases including high noise and metal artifacts. Through the training and testing of the algorithm, average Dice of our method for the public test data of the VerSe 2019 challenge dataset is 0.942 and the average HD is 6.24. For the hidden test data, average Dice is 0.941, whereas the average HD is 6.00. The reason why our proposed method can achieve such a high accuracy is linked to the excellent performance of 3D Swin-YoloX and 3D Swin-UNet. The average 3D_IOU of 3D Swin-YoloX on the public and hidden test data reached 0.896 and 0.898, respectively, whereas Dice of 3D Swin-UNet reached 0.951 and 0.949, respectively. In short, our method achieves an excellent test effect on the dataset and a certain reference value. In the next step, we will test, apply, and promote clinical data to enhance the application value and further benefit most doctors.

Data Availability

The CT image data used to support the findings of this study have been uploaded to https://osf.io/nqjyw/.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Yonghong Zhang, Xuquan Ji, and Wenyong Liu contributed equally to this work.

Acknowledgments

This work was supported by the Beijing Natural Science Funds-Haidian Original Innovation Joint Fund: L202010 and the Natural Science Foundation of China: U20A20199.

References

[1] G. U. Kim, M. C. Chang, T. U. Kim, and G. W. Lee, "Diagnostic modality in spine disease: a review," *Asian Spine J*, vol. 14, no. 6, pp. 910–920, 2020.

[2] P. Xue, X. Chen, S. Chen, and Y. Shi, "The value of CT 3D reconstruction in the classification and nursing effect evaluation of ankle fracture," *Journal of Healthcare Engineering*, vol. 2021, Article ID 9596518, 5 pages, 2021.

[3] D. Chen, C. H. Chen, L. Tang et al., "Three-dimensional reconstructions in spine and screw trajectory simulation on 3D digital images: a step by step approach by using Mimics software," *Journal of Spine Surgery*, vol. 3, no. 4, pp. 650–656, 2017.

[4] J. Yao, J. E. Burns, H. Munoz, and R. M. Summers, "Detection of vertebral body fractures based on cortical shell unwrapping," in *Proceedings of the Medical Image Computing and Computer-Assisted Intervention--MICCAI 2012*, pp. 509–516, Springer, Berlin, Germany, October 2012.

[5] S. Liu, D. Gai, Q. Lu et al., "Application of CT image based on three-dimensional image segmentation algorithm in diagnosis of osteoarthritis," *Journal of Medical Imaging and Health Informatics*, vol. 11, no. 1, pp. 230–234, 2021.

[6] Q. Zhang, M. Li, X. Qi, Y. Hu, Y. Sun, and G. Yu, "3D path planning for anterior spinal surgery based on CT images and reinforcement learning," in *Proceedings of the 2018 IEEE International Conference on Cyborg and Bionic Systems (CBS)*, pp. 317–321, Shenzhen, China, October 2018.

[7] A. Kanawati, R. J. Rodrigues Fernandes, A. Gee et al., "The development of novel 2-in-1 patient-specific, 3D-printed laminectomy guides with integrated pedicle screw Drill guides," *World Neurosurgery*, vol. 149, pp. e821–e827, 2021.

[8] A. Saffari, S. Kölker, G. F. Hoffmann, M. Weiler, and A. Ziegler, "Novel challenges in spinal muscular atrophy–How to screen and whom to treat?" *Ann. Clin. Transl. Neurol.*, vol. 6, no. 1, pp. 197–205, 2019.

[9] X. Yang, R. Guo, X. Lv et al., "Challenges in diagnosis of spinal epidural abscess: a case report," *Medicine*, vol. 98, no. 5, Article ID e14196, 2019.

[10] A. P. Brady, "Error and discrepancy in radiology: inevitable or avoidable?" *Insights Into Imaging*, vol. 8, no. 1, pp. 171–182, 2017.

[11] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[12] T. Vrtovec, B. Likar, and F. Pernus, "Automated curved planar reformation of 3D spine images," *Physics in Medicine and Biology*, vol. 50, no. 19, pp. 4527–4540, 2005.

[13] B. Michael Kelm, M. Wels, S. Kevin Zhou et al., "Spine detection in CT and MR using iterated marginal space learning," *Medical Image Analysis*, vol. 17, no. 8, pp. 1283–1292, 2013.

[14] T. Ebner, D. Stern, R. Donner, H. Bischof, and M. Urschler, "Towards automatic bone age estimation from MRI: localization of 3D anatomical landmarks," *IJBD*, vol. 17, no. Pt 2, pp. 421–428, 2014.

[15] C. Lindner, P. A. Bromiley, M. C. Ionita, and T. F. Cootes, "Robust and accurate shape model matching using random forest regression-voting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1862–1874, 2015.

[16] B. Glocker, D. Zikic, E. Konukoglu, D. R. Haynor, and A. Criminisi, "Vertebrae localization in pathological spine ct via dense classification from sparse annotations," in *Proceedings of the Medical Image Computing and Computer-Assisted Intervention--MICCAI 2013*, pp. 262–270, Springer, Berlin, Germany, September 2013.

[17] H. Chen, C. Shen, J. Qin et al., "Automatic localization and identification of vertebrae in spine ct via a joint learning model with deep neural networks," in *Proceedings of the Medical Image Computing and Computer-Assisted Intervention--MICCAI 2015*, pp. 515–522, Springer, Berlin, Germany, October 2015.

[18] A. Suzani, A. Seitel, Y. Liu, S. Fels, R. N. Rohling, and P. Abolmaesumi, "Fast automatic vertebrae detection and localization in pathological ct scans-a deep learning approach," in *Proceedings of the Medical Image Computing and Computer-Assisted Intervention--MICCAI 2015*, pp. 678–686, Springer, Berlin, Germany, October 2015.

[19] C. Payer, D. Štern, H. Bischof, and M. Urschler, "Regressing heatmaps for multiple landmark localization using cnns," in *Proceedings of the International conference on medical image computing and computer-assisted intervention*, pp. 230–238, Springer, Berlin, Germany, October 2016.

[20] C. Payer, D. Štern, H. Bischof, and M. Urschler, "Integrating spatial configuration into heatmap regression based CNNs for landmark localization," *Medical Image Analysis*, vol. 54, pp. 207–219, 2019.

[21] H. Liao, A. Mesfin, and J. Luo, "Joint vertebrae identification and localization in spinal CT images by combining short- and long-range contextual information," *IEEE Transactions on Medical Imaging*, vol. 37, no. 5, pp. 1266–1275, 2018.

[22] Y. Zhang, N. Hu, Z. Li et al., "Lumbar spine localisation method based on feature fusion," *CAAI Transactions on Intelligence Technology*, vol. 8, no. 3, pp. 931–945, 2022.

[23] J. M. Gauch, "Image segmentation and analysis via multiscale gradient watershed hierarchies," *IEEE Transactions on Image Processing*, vol. 8, no. 1, pp. 69–79, 1999.

[24] N. Kamiya, J. Li, M. Kume, H. Fujita, D. Shen, and G. Zheng, "Fully automatic segmentation of paraspinal muscles from 3D torso CT images via multi-scale iterative random forest classifications," *International Journal of Computer Assisted Radiology and Surgery*, vol. 13, no. 11, pp. 1697–1706, 2018.

[25] A. B. Oktay, N. B. Albayrak, and Y. S. Akgul, "Computer aided diagnosis of degenerative intervertebral disc diseases from lumbar MR images," *Computerized Medical Imaging and Graphics*, vol. 38, no. 7, pp. 613–619, 2014.

[26] F. Lecron, J. Boisvert, S. Mahmoudi, H. Labelle, and M. Benjelloun, "Three-dimensional spine model reconstruction using one-class SVM regularization," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 11, pp. 3256–3264, 2013.

[27] D. Zukić, A. Vlasák, J. Egger, D. Hořínek, C. Nimsky, and A. Kolb, "Robust detection and segmentation for diagnosis of vertebral diseases using routine MR images," *Computer Graphics Forum*, vol. 33, no. 6, pp. 190–204, 2014.

[28] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *Proceedings of the Medical Image Computing and Computer-Assisted Intervention--MICCAI 2015*, pp. 234–241, Springer, Berlin, Germany, October 2015.

[29] F. Milletari, N. Navab, S. A. Ahmadi, and V. Net, "Fully convolutional neural networks for volumetric medical image segmentation," in *Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571, Stanford, CF, USA, October 2016.

[30] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: a nested u-net architecture for medical image segmentation," *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, Springer, Berlin, Germanypp. 3–11, 2018.

[31] X. Xiao, S. Lian, Z. Luo, and S. Li, "Weighted res-unet for high-quality retina vessel segmentation," in *Proceedings of the 2018 9th International Conference on Information Technology in Medicine and Education (ITME)*, pp. 327–331, Hangzhou, China, October 2018.

[32] O. Oktay, J. Schlemper, L. L. Folgoc et al., "Attention u-net: learning where to look for the pancreas," 2018, https://arxiv.org/abs/1804.03999.

[33] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[34] A. Dosovitskiy, L. Beyer, A. Kolesnikov et al., "An image is worth 16 × 16 words: transformers for image recognition at scale," 2020, https://arxiv.org/abs/2010.11929.

[35] J. Chen, Y. Lu, Q. Yu et al., "Transunet: transformers make strong encoders for medical image segmentation," 2021, https://arxiv.org/abs/2102.04306.

[36] Z. Liu, Y. Lin, Y. Cao et al., "Swin transformer: hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, Montreal, BC, Canada, October 2021.

[37] H. Cao, Y. Wang, J. Chen et al., "Swin-unet: unet-like pure transformer for medical image segmentation," 2021, https://arxiv.org/abs/2105.05537.

[38] M. Kolařík, R. Burget, V. Uher, K. Říha, and M. K. Dutta, "Optimized high resolution 3D dense-u-net network for brain and spine segmentation," *Applied Sciences*, vol. 9, no. 3, p. 404, 2019.

[39] P. Cheng, Y. Yang, H. Yu, and Y. He, "Automatic vertebrae localization and segmentation in CT with a two-stage Dense-U-Net," *Scientific Reports*, vol. 11, no. 1, pp. 22156–22213, 2021.

[40] H. Tang, X. Pei, S. Huang, X. Li, and C. Liu, "Automatic lumbar spinal CT image segmentation with a dual densely connected U-Net," *IEEE Access*, vol. 8, pp. 89228–89238, 2020.

[41] 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, https://arxiv.org/abs/1703.04347.

[42] R. Janssens, G. Zeng, and G. Zheng, "Fully automatic segmentation of lumbar vertebrae from CT images using cascaded 3D fully convolutional networks," in *Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 893–897, Washington, DC, USA, April 2018.

[43] E. Cheng, Y. Liu, H. Wibowo, and L. Rai, "Learning-based spine vertebra localization and segmentation in 3D CT image," in *Proceedings of the 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pp. 160–163, Prague, Czech Republic, April 2016.

[44] N. Lessmann, B. van Ginneken, P. A. de Jong, and I. Isgum, "Iterative fully convolutional neural networks for automatic vertebra segmentation and identification," *Medical Image Analysis*, vol. 53, pp. 142–155, 2019.

[45] C. Payer, D. Stern, H. Bischof, and M. Urschler, "Coarse to fine vertebrae localization and segmentation with spatialconfiguration-net and u-net," *VISIGRAPP*, pp. 124–133, 2020.

[46] N. Altini, G. De Giosa, N. Fragasso et al., "Segmentation and identification of vertebrae in ct scans using cnn, k-means clustering and k-nn," *Informatics*, Multidisciplinary Digital Publishing Institute, Basel, Switzerland, 2021.

[47] R. Tao, W. Liu, and G. Zheng, "Spine-transformers: vertebra labeling and segmentation in arbitrary field-of-view spine CTs via 3D transformers," *Medical Image Analysis*, vol. 75, Article ID 102258, 2022.

[48] B. D. De Vos, J. M. Wolterink, P. A. De Jong, M. A. Viergever, and I. Išgum, "2d image classification for 3d anatomy localization: employing deep convolutional neural networks," *Medical imaging 2016: Image processing*, SPIE, France,pp. 517–523, 2016.

[49] Z. Krawczyk and J. Starzyński, "Bones detection in the pelvic area on the basis of YOLO neural network," in *Proceedings of the 19th International Conference Computational Problems of Electrical Engineering*, pp. 1–4, Banska Stiavnica, Slovakia, September 2018.

[50] X. Xu, F. Zhou, B. Liu, D. Fu, and X. Bai, "Efficient multiple organ localization in CT image using 3D region proposal network," *IEEE Transactions on Medical Imaging*, vol. 38, no. 8, pp. 1885–1898, 2019.

[51] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: exceeding yolo series in 2021," 2021, https://arxiv.org/abs/2107.08430.

[52] J. Ma, J. He, and X. Yang, "Learning geodesic active contours for embedding object global information in segmentation CNNs," *IEEE Transactions on Medical Imaging*, vol. 40, no. 1, pp. 93–104, 2021.

[53] J. Xu, J. Gong, J. Zhou, X. Tan, Y. Xie, and L. Ma, "Sceneencoder: scene-aware semantic segmentation of point clouds with a learnable scene descriptor," 2020, https://arxiv.org/abs/2001.09087.

[54] W. Liu, A. Rabinovich, and A. C. Berg, "Parsenet: looking wider to see better," 2015, https://arxiv.org/abs/1506.04579.

[55] A. Sekuboyina, M. E. Husseini, A. Bayat et al., "VerSe: a Vertebrae labelling and segmentation benchmark for multi-detector CT images," *Medical Image Analysis*, vol. 73, Article ID 102166, 2021.

[56] G. Csurka, D. Larlus, F. Perronnin, and F. Meylan, "What is a good evaluation measure for semantic segmentation?" *British Machine Vision Conference*, vol. 10, p. 5244, 2013.

[57] B. Cheng, R. Girshick, P. Dollár, A. C. Berg, and A. Kirillov, "Boundary iou: improving object-centric image segmentation evaluation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15334–15342, Nashville, TN, USA, June 2021.

[58] D. Karimi and S. E. Salcudean, "Reducing the hausdorff distance in medical image segmentation with convolutional neural networks," *IEEE Transactions on Medical Imaging*, vol. 39, no. 2, pp. 499–513, 2020.