



ARTICLE

Spinal Vertebral Fracture Detection and Fracture Level Assessment Based on Deep Learning

Yuhang Wang^{1,*}, Zhiqin He¹, Qinmu Wu¹, Tingsheng Lu², Yu Tang¹ and Maoyun Zhu¹

¹The Electrical Engineering College, Guizhou University, Guiyang, 550025, China

²Guizhou Provincial Orthopedics Hospital, Guiyang, 550025, China

*Corresponding Author: Yuhang Wang. Email: gs.wangyuhang21@gzu.edu.cn

Received: 03 November 2023 Accepted: 13 March 2024 Published: 25 April 2024

ABSTRACT

This paper addresses the common orthopedic trauma of spinal vertebral fractures and aims to enhance doctors' diagnostic efficiency. Therefore, a deep-learning-based automated diagnostic system with multi-label segmentation is proposed to recognize the condition of vertebral fractures. The whole spine Computed Tomography (CT) image is segmented into the fracture, normal, and background using U-Net, and the fracture degree of each vertebra is evaluated (Genant semi-qualitative evaluation). The main work of this paper includes: First, based on the spatial configuration network (SCN) structure, U-Net is used instead of the SCN feature extraction network. The attention mechanism and the residual connection between the convolutional layers are added in the local network (LN) stage. Multiple filtering is added in the global network (GN) stage, and each layer of the LN decoder feature map is filtered separately using dot product, and the filtered features are re-convolved to obtain the GN output heatmap. Second, a network model with improved SCN (M-SCN) helps automatically localize the center-of-mass position of each vertebra, and the voxels around each localized vertebra were clipped, eliminating a large amount of redundant information (e.g., background and other interfering vertebrae) and keeping the vertebrae to be segmented in the center of the image. Multilabel segmentation of the clipped portion was subsequently performed using U-Net. This paper uses VerSe'19, VerSe'20 (using only data containing vertebral fractures), and private data (provided by Guizhou Orthopedic Hospital) for model training and evaluation. Compared with the original SCN network, the M-SCN reduced the prediction error rate by 1.09% and demonstrated the effectiveness of the improvement in ablation experiments. In the vertebral segmentation experiment, the Dice Similarity Coefficient (DSC) index reached 93.50% and the Maximum Symmetry Surface Distance (MSSD) index was 4.962 mm, with accuracy and recall of 95.82% and 91.73%, respectively. Fractured vertebrae were also marked as red and normal vertebrae were marked as white in the experiment, and the semi-qualitative assessment results of Genant were provided, as well as the results of spinal localization visualization and 3D reconstructed views of the spine to analyze the actual predictive ability of the model. It provides a promising tool for vertebral fracture detection.

KEYWORDS

Deep learning; vertebral fracture detection; medical image processing



1 Introduction

Spinal fractures are a common type of fracture, with vertebral fractures being the most frequent. Such fractures may be caused by pathologic, traumatic, or nontraumatic causes [1]. This systemic skeletal disorder has serious implications for the elderly and chronically ill [2]. Globally, it is estimated that approximately 1.4 million new clinical vertebral fractures occur each year. Their CT (Computed Tomography) image reading is a huge workload and time-consuming. It also results in significant medical and social costs. Often, vertebral fractures may go undetected due to a variety of reasons, such as the asymptomatic nature of the fracture in its early stages, the failure of radiologists to detect the fracture in time due to inadvertence, and the difficulty in distinguishing between normal and pathologically deformed vertebrae because of the extreme similarity in their shapes. If omission or misdiagnosis occurs or patients are not treated in time, these vertebral fractures may lead to secondary fractures, causing permanent disability or even death [3]. Therefore, early detection of vertebral fractures is extremely important for the prevention of secondary fractures.

In this context, some researchers have proposed methods for intelligent diagnosis of vertebral fractures, and most of the initial studies have focused on machine learning. Valentinitich et al.'s study [4] employed a random forest (RF) classifier to identify vertebral compression fracture (VCF) based on texture features and regional vertebral bone mineral density (BMD) analysis and experimented with more than 200 patients with good results. Wang et al. [5] used a set of support vector machines (SVMs) to classify fractures and considered different sets of features, including measurement features, longitudinal features, and combinations of both. Burns et al. [6] used a water-shed algorithm to segment the vertebrae. And then extracted the vertebral heights and bone density. These features were used to train a support vector machine to grade compression fractures. It achieved a sensitivity of 95.7% and a false positive rate of 0.29.

Recently, deep learning has been used for automatic vertebral fracture detection in medical images [7,8]. Murata and other scholars [9] proposed a spinal fracture detection method based on a deep convolutional neural network (DCNN). It was trained using plain thoracolumbar radiography (PTLR) images of 300 patients (150 patients with VF and 150 patients without VF). The resultant accuracy, sensitivity, and specificity were 86.0%, 84.7%, and 87.3%, respectively. Bar et al. [10] first segmented the spine portion of a computed tomography (CT) scan and divided it into two-dimensional sagittal section slices. Then, the sagittal slices were classified using a (2D) CNN. Similarly, Tomita et al. [11] proposed a 2D CNN-based recognition method that first abstracts data from CT slices using a CNN feature extraction module. and then determines vertebral fractures using a Long Short-Term Memory (LSTM) network-based approach. Although these 2D CNN-based methods can judge vertebral fractures more effectively, their limitation is that they cannot fully utilize the information of the complete vertebral sequence, which may easily result in missing information.

A full 3D convolution-based approach can be effective in solving this problem. Nicolaes et al. [12] proposed a two-stage VCF detection method by firstly cropping the vertebral body using the vertebral body center of mass co-ordinates labeled by a physician and then 3D CNN predicting the category probability for each voxel (fractures and non-fractures). This study demonstrated the advantages of 3D CNN, but its main drawback was that only a small dataset containing 90 patients was used, and that the method was a semi-automated diagnostic approach. Chettrit et al. [13] proposed an end-to-end 3D CNN-based architecture for vertebral fracture detection, whereby the framework inputs axial CT images of the thorax and abdomen and 3D reconstructs a 3D model of the entire spine, which is used to locate the vertebral body. The YOLO detector was then used to localize the location of the vertebrae in each slice, and these slices were filtered and clipped based on the localization results.

Finally, these combined slices are used to make fracture judgments using a full 3D CNN. Although this method avoids the manual labeling of vertebral body center of mass, the YOLO detector may not be identical in each CT slice, which may lead to errors in vertebral body location clipping, and also cannot distinguish the vertebral body corresponding to each slice, which will have a certain impact on the subsequent fracture judgment.

To address the above problems, this study proposes a fully automated 3D CNN-based intelligent recognition system for vertebral fractures. The difference is that this paper adopts the M-SCN [14] key point regression model for the selection of a vertebral position, which can directly regress the vertebral center of mass position on the 3D structure without the need to be processed on each slice, and subsequently determine the vertebral type through multi-label segmentation, which adopts the segmentation in with the fracture degree of each vertebra can be evaluated.

The innovations of this article are summarised as follows:

1. In this paper, an M-SCN network is used to automatically locate the center of mass coordinates of each vertebra without the need for physicians to manually label vertebral center of mass coordinates or analyze individual slices. The category name of the current vertebra is then determined based on the location of the feature map where the heatmap is located.

2. Based on the localization results, local vertebrae are centrally cropped to eliminate a large amount of redundant information and to make the current vertebrae located in the center of the image, so that the vertebral body information in the image is more explicit. This cropped region is then segmented into individual vertebrae using the U-Net [15] network, and the result is divided into background voxels, fractured vertebrae and normal vertebrae.

3. Based on the segmentation results of individual vertebrae, a smart fracture-level assessment method was introduced. The sagittal 2D slices required for Genant assessment were selected using the vertebral bone center as a benchmark, and the key points of vertebral bone morphology were regressed again using the M-SCN network structure, and the final assessment results were obtained according to the Genant formula.

4. This paper improves and optimizes the SCN network. In the LN stage, this study introduces the Attention mechanism and the U-Net network structure, and incorporates residual connections in the convolution block to improve the model robustness and network output accuracy. In the GN stage, this paper adds a multiple filtering design, and takes the 4-layer feature map output from the decoder in the LN stage and the LN output heatmap as the input to the GN. By re-extracting the features of the LN heatmap, each layer of the feature map of the LN decoder is filtered separately using dot product, and finally the GN output heatmap is obtained by up-sampling and re-convolution.

In conclusion, the system proposed in this study can be used in clinical practice. It allows early detection of patients with suspected vertebral fractures. Targeted therapeutic measures can be performed, thus reducing the burden of fracture diagnosis on a wider scale.

2 Method

2.1 Method Overview

As shown in Fig. 1, the method proposed in this paper is divided into two steps: In the first step, the vertebral body center of mass is localized by M-SCN network, and in the second step, the surrounding voxels were clipped according to the vertebral center-of-mass coordinates, and the clipped vertebrae were segmented with multiple labels using the U-Net model (divided into three label masks: Background, fracture, and normal, respectively). Subsequently, the morphological key points required

for vertebral fracture grade assessment were regressed using the same network structure as in step one. Then the assessment results were obtained according to the Genant Semi-Qualitative Assessment [16] formula.

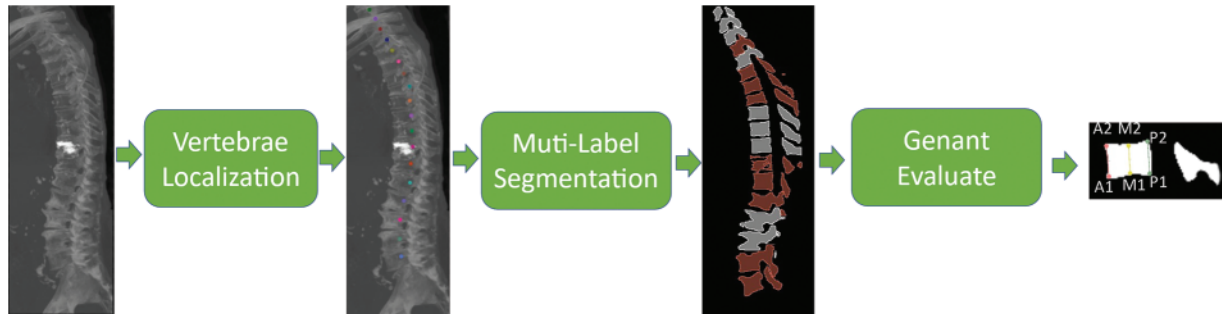


Figure 1: Vertebral fracture analysis process. Firstly, the vertebral center of mass in the original CT image was localized, followed by multi-label image segmentation, and the segmented vertebral fractures were evaluated according to the Genant principle

In Step 1, the entire M-SCN network framework process starts with the input CT image to be detected, which is first sampled to a uniform 1mm voxel interval. The Overlapping approach is adopted, as shown in Fig. 2 (patch size is 96, 96, 128), which demonstrates the heatmap regression results using the M-SCN network model to reason about the whole CT image. Subsequently, the coordinate information represented by the current heatmap was filtered based on the highest brightness in the center of the heatmap.

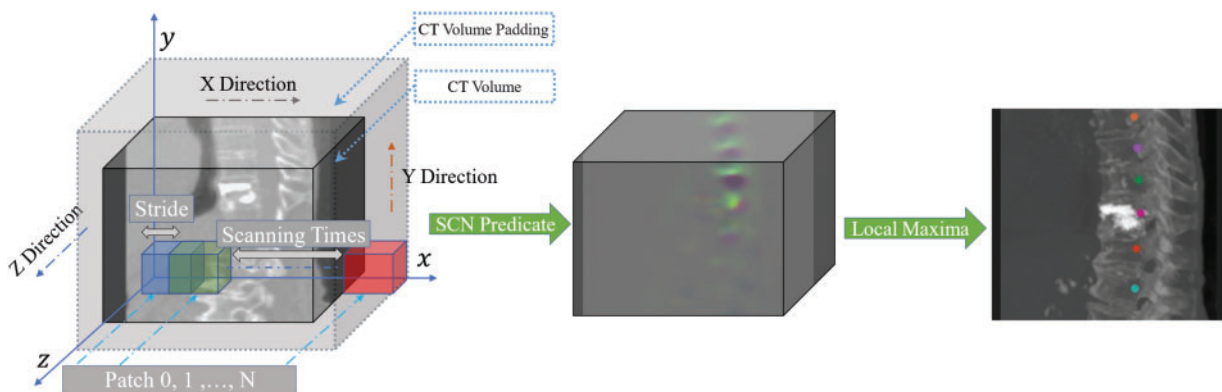


Figure 2: Overlapping patch inference. The original CT image is first filled to ensure that the original CT image is scanned in its entirety. The squares in the figure represent the selected patches, which are scanned along the xyz direction of the whole CT image. The blue squares represent the location of the first patch, the green squares represent the location of the second patch with stride as the offset, and the red squares represent the last patch scanned in the current x-direction. the process is repeated until the entire CT image is scanned, and the entire predicted heatmap is obtained and the coordinates are filtered out

Step 2 is shown in Fig. 3, based on the coordinate information regressed by the localization network, the surrounding voxels (patch size 128, 128, 128) centered on this coordinate are cropped to remove some redundant information and to make the vertebrae represented by the current point

located in the center of the cropped image. Use the model U-Net to reason about the multi-label segmentation results, classified into three labels: Background, normal vertebrae, and fractured vertebrae. And retaining the current Patch inference result, it is also necessary to resample the cropped Patch positions to the positions in the original CT image and combine them. This process is repeated until all coordinate traversals are finished.

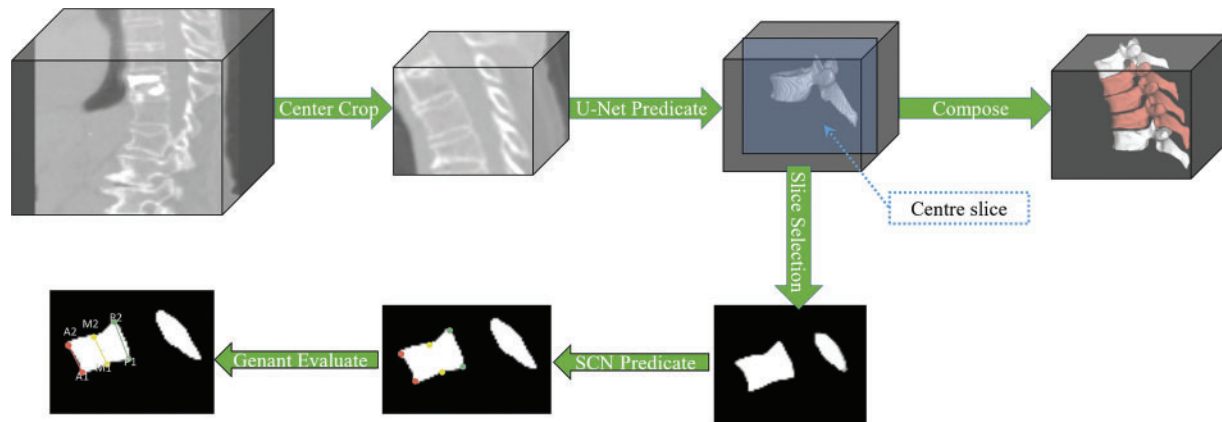


Figure 3: Segmentation and vertebral Genant evaluation process. Firstly, individual vertebrae were centrally clipped, the current vertebrae were segmented by U-Net, the current predicted patch was retained, and the patch was resampled and combined to obtain the whole spine segmentation image. Secondly, the center of the patch was selected and sliced for Genant semi-qualitative assessment

In the vertebral fracture grade assessment stage, as shown in Fig. 3, the Patch segmentation result obtained during the segmentation process was first adjusted to a binary mask. Secondly, the sagittal plane slice at the center of this segmented image was selected as the assessment object. Since it is difficult to obtain the coordinates of the points to be assessed by traditional methods, a fixed-point strategy similar to that of Step 1 is used to regress the required vertebral morphology coordinate points in the 2D dimension. Then the HA, HM, and HP vertebral bone morphology parameters were calculated to obtain the assessment results according to the Genant semi-qualitative assessment formula.

2.2 Vertebrae Localization

In this paper, the M-CSN network is used to localize the vertebral centers. Which is modeled by two parts, the LN and the GN. In this two-part structure, the role of its GN is to filter and noise-reduce the output of the LN to make the network output more accurate. The major difference between these two networks is the size of the receptive field [17], the LN is designed to extract local features that are mostly valid but can be noisy for the target heatmap, And the GN is designed as a network structure with a large convolutional kernel for extracting the global features. which has a larger receptive field and filters the output of the LN. Fig. 4 shows the vertebral localization network. Since the M-SCN model output is a target heatmap, no activation function is used in the last layer of the model [18]. Where LN convolution block and GN convolution block are designed as shown in Fig. 5.

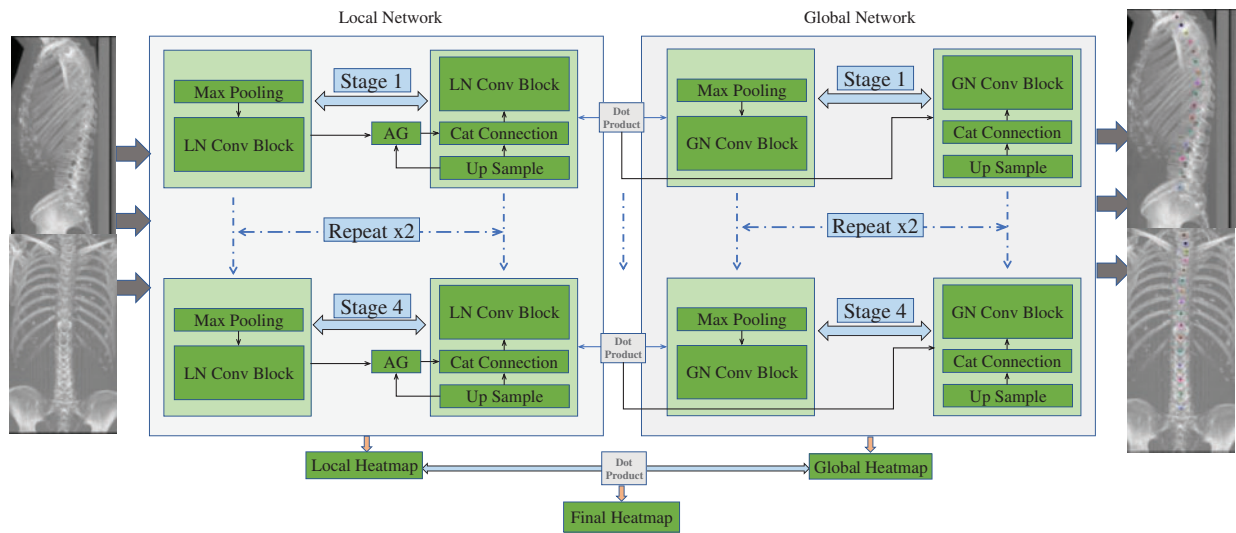


Figure 4: M-SCN network structure, the model consists of two parts, the local network and the global network. The local network is a U-shaped structure with an attention mechanism. The global network is a U-shaped structure containing a large convolutional kernel, which is mainly used for feature filtering. This network inputs the features of each layer of the LN decoder as well as the LN heatmap. The LN heatmap is re-featured and dot-multiplied with the output of each layer of the decoder, and finally the GN heatmap is obtained by up-sampling and re-convolution. The final output of the model is the result of dot product of LN heatmap and GN heatmap

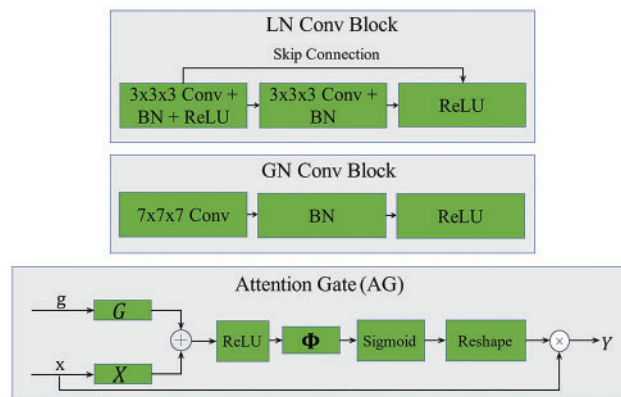


Figure 5: Convolutional block and attention mechanism. LN convolutional block is designed with double-layer convolution and residual joins are added. GN convolutional block is designed with single-layer convolution, which reduces the amount of model computation. Also $7 \times 7 \times 7$ large convolution kernel is used instead of small convolution kernel

Assuming that the input image is $X_i \in R^{C_i \times Z_i \times H_i \times W_i}$ and its corresponding vertebral center of mass coordinates are $l_{gt} = \{(x_{ic'}, y_{ic'}, z_{ic'})\}$, and the labeled heatmap $Y_{ik} \in R^{C_i \times Z_i \times H_i \times W_i}$ for the

k^{th} ($k \in [1, 2, \dots, C_i]$) image by using a Gaussian distribution function:

$$Y_{ik} = g(x_i, \sigma_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(x - l_{gt})^2}{2\sigma_i^2}\right) \quad (1)$$

As a result, thermal pixels with target coordinates near l_{gt} will have higher values, and these values will decrease smoothly but quickly as they move away from l_{gt} . The standard deviation σ_i is used to determine the peak width of the Gaussian function in the thermogram image. where is the number of channels on the input image (for CT images, $C_i = 1$), C'_i is the number of channels on the output thermogram (representing the number of vertebrae), Z_i is the value of the image X_i along the cross-section. H_i is the value of the image X_i along the coronal plane. W_i is the value of the image X_i along the sagittal plane.

Similarly, coordinate positions can be determined by obtaining the maximum value in the heatmap.

$$l_{coord} = \operatorname{argmax}(g_i(x; \sigma_i)) \quad (2)$$

The M-SCN network regresses the heatmaps of N target vertebrae by minimizing the Euclidean distance between the predicted and labeled heatmaps so that the predicted values gradually approximate the true values, which can be described as:

$$L = \min_{w,b} \sum_{i=1}^N \sum_x \|h_i(x; w, b) - g_i(x; \sigma_i)\|_2^2 + \lambda \|w\|_2^2 \quad (3)$$

where w is the network weight, b is the network bias, and λ controls the effect of the L2 paradigm of the network weight w .

2.2.1 Local Network

The purpose of LN is to extract local features and generate local heatmaps. In this paper, U-shaped structure with Attention Gates (AG) mechanism is used as the LN structure and residual connectivity is added to the LN convolutional block to improve the robustness of the model. The AG (Attention Gates) module [19] automatically learns and distinguishes the shape of the target and automatically recognizes the region of interest, suppresses irrelevant regions and focuses on features in useful regions. AG can be described as:

$$f(x_i) = W_g^T x_i \quad (4)$$

$$\operatorname{attn}(x_i, g_i) = \sigma_1(\sigma_2(f(x_i) + f(g_i) + b_1) + b_2) \quad (5)$$

where the function $f(x_i)$ denotes the linear embedding of x_i , σ_1 denotes the ReLU activation function, and σ_2 denotes the Sigmoid activation function. The LN model extracts features from X_i to generate a local heatmap \hat{F}_i^L :

$$\hat{F}_i^L(x_i) = h_i(x_i; w_{atn}^L, b_{atn}; w_s^L, b_s) \quad (6)$$

where W_{atn}^L is the weight parameter corresponding to the AG module, w_s^L is the parameter corresponding to the ordinary convolution, b_{atn} and b_s represent the bias parameters of AG and normal convolution.

2.2.2 Global Network

The GN aims to filter the output features of the LN. Its structure differs in the use of a large 7×7 convolution kernel to extract the features, which provides a larger sensory field for this network, facilitates the capture of important global information [20] and reduces the computational effort by using a single layer of convolution. The inputs to this network are the output heatmap of the LN network and the 4-layer decoded feature map output by the LN decoder, and these 4 layers of decoded features are filtered separately.

$$\hat{F}_i^G(x_i) = h_i(\hat{F}_i^L, E_i^L; w_s^L, b_s) \quad (7)$$

Finally, the LN heatmap and the GN heatmap are dot-multiplied to obtain the predicted heatmap.

$$\hat{F}_i = \hat{F}_i^L \cdot \hat{F}_i^G \quad (8)$$

2.2.3 Loss Function

In this regression task, this paper choose smooth L1 [21] as the loss function, which can be seen from the smooth L1 formula, the gradient is constant when x is a large value, solving the problem of large gradient destroying the training parameters in L2 loss, and when x is a small value, the gradient will be dynamically reduced solving the problem of difficult convergence in L1 loss. It can be defined as:

$$\text{Smooth-L1}(p_r, y_r) = \begin{cases} 0.5(\sigma x)^2, & |x| < \frac{1}{\sigma} \\ |x| - \frac{0.5}{\sigma^2}, & \text{otherwise} \end{cases} \quad (9)$$

where p_r and y_r are the predicted and true values, σ is the smoothing factor. In the experiment, it was set to 1.0.

2.3 Vertebrae Fracture Recognition and Genant Semi-Qualitative Assessment

2.3.1 Identification of Vertebral Fractures

In order to be able to recognize the pathological features of the vertebrae, this paper uses an image segmentation method. The center-of-mass coordinates of each vertebra have been predicted in Task 1 (Vertebral Localization), so the region of interest can be cropped according to the center-of-mass coordinates so that the vertebrae are located in the center of the cropped image, ensuring that the network will only process the current vertebrae. Based on the localization results from Task 1, this process is repeated from top to bottom across the spine until the last vertebra is iterated.

In this paper, U-Net is chosen as the segmentation model, and the prediction results are divided into three state labels, which are background, fractured vertebrae and normal vertebrae. The overall process of vertebral fracture detection is shown in Fig. 6 below.

Let $V = \{v_1, v_2, \dots, v_n\}$ be the set of coordinates of all vertebrae in the CT image of the spine. The input image is $X_i \in R^{C_i \times Z_i \times H_i \times W_i}$. Then there are:

$$x_i = \text{crop}(X_i; v_i), v_i \in V \quad (10)$$

$$\hat{F}_i = h_i(x_i; w, b) \quad (11)$$

where x_i denotes the image cropped according to the center of mass coordinates v_i . \hat{F}_i denotes the predicted voxel class results.

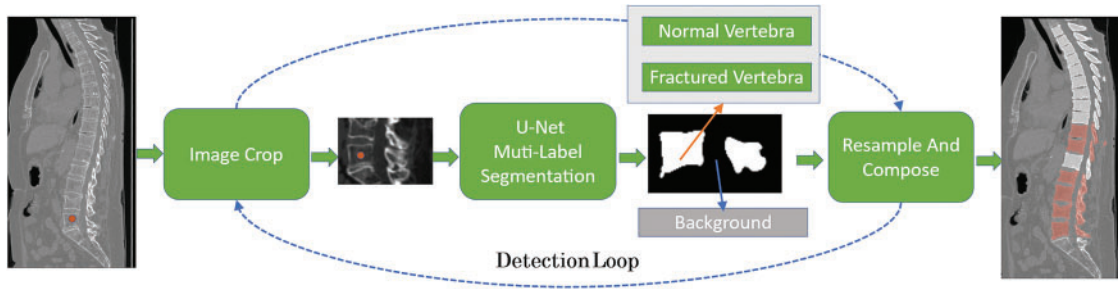


Figure 6: The overall process of vertebral fracture recognition, using U-Net as the segmentation model, based on the vertebral body center of mass predicted in Task 1, clipping the ROI region and performing multi-label task segmentation, which is divided into fracture, normal, and background, and repeating this process until the end of vertebral. Finally, the vertebral segmentation results are combined to obtain an analysis of the entire spine

Similar to Task 1, the vertebral segmentation network approximates the predicted value to the true value by minimizing the difference between the probability distribution of the predicted outcome and the true probability distribution. This can be described as:

$$L = \min_{w,b} \sum_{i=1}^N \sum_x 1 - \frac{2|\hat{F}_i \cdot Y_i|}{|\hat{F}_i|^2 + |Y_i|^2} + Y_i \log(\hat{F}_i) + \lambda \|w\|_2^2 \quad (12)$$

Finally, to create the final multi-label prediction results, the results of the individual predicted images need to be merged. The cropped vertebrae x_i also need to be resampled and transformed to their position in the original input image.

2.3.2 Loss Function

Considering the accuracy of predictive segmentation and the matching accuracy of category distribution, this paper uses a combination of dice loss [22] and cross-entropy loss. It can be a more comprehensive measure of the model’s performance in multi-label segmentation tasks. It can be defined as follows:

$$Loss(p_r, p_y) = 1 - \frac{2|p_r \cdot p_y| + \varepsilon}{|p_r|^2 + |p_y|^2 + \varepsilon} + p_r \log(p_y) \quad (13)$$

where p_r and p_y are the predicted and true values, respectively. ε is the smoothing coefficient, which was set to 10^{-5} in the experiment.

2.3.3 Genant Semi-Qualitative Assessment

Prior to Genant assessment, the sagittal plane needs to be selected from the vertebral fracture segmentation results. For each vertebral image the sagittal plane where the vertebral body center of mass was located was used as a benchmark for Genant assessment. To obtain the key points of vertebral body morphology required for Genant assessment, two problems were faced in this study:

1. As shown in Fig. 7, the traditional Harris corner point detector, although it can detect Genant's evaluation of key points, still produces a high number of redundant points, does not accurately obtain the corner coordinates, and is extremely dependent on image quality.

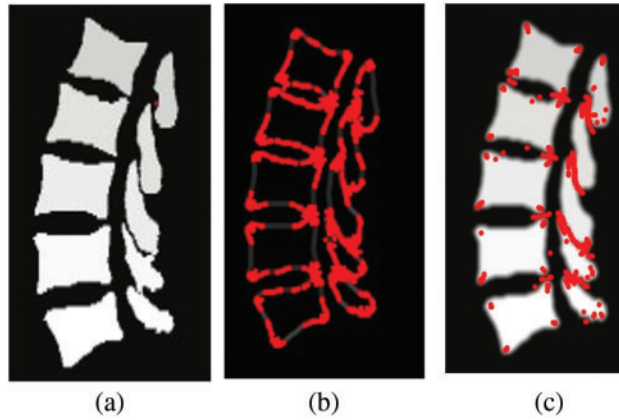


Figure 7: Figure a is the sagittal plane slice, and figure b and c are the Harris corner point detector results after Canny edge detection and smoothing filtering, respectively

2. As shown in Fig. 8, the selection of the M point is more difficult, the M point location does not necessarily contain the corner similar features, so it can not be detected by the Harris corner point detector, if the vertebrae are relatively smooth, the coordinates of the M point can be the average value of the points A, P. If the vertebrae number of biconcave-type (or irregularly shaped) fracture, the location of the M point will occur in the y-axis position of the deviation, Yousefi et al. [23] used the strategy of moving the M point up and down to solve this problem, but there is no universal value for the distance of moving M up and down, and the distance of moving is different for different cases of vertebrae. Therefore this method is more limited.

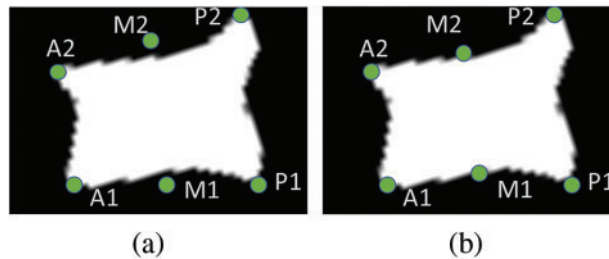


Figure 8: For irregular vertebrae, the selection of M-points produces different degrees of bias among them, figure a is the case when there is a deviation of M point, and figure b is the case when the M point is adjusted by moving up and down

Therefore, this paper uses the same strategy as in Step 1 to regressively localize the desired vertebral morphological keypoints using M-SCN. Fig. 9 shows the overall assessment process.

It can be described as:

$$s_i = Y_i[:, : y_{coord} : ,] \quad (14)$$

$$\hat{F}_i = h_i(s_i; w, b) \quad (15)$$

$$l_{coord} = \arg \max (\hat{F}_i) \quad (16)$$

$$HP = |P_1 - P_2|, HA = |A_1 - A_2|, HM = |M_1 - M_2| \quad (17)$$

$$G = \frac{\min (HP, HA, HM)}{\max (HP, HA, HM)} \quad (18)$$

where Y_i is the input vertebral segmentation mask and s_i denotes the sagittal slice where the vertebral center of mass is located. l_{coord} is the key point coordinates, \hat{F}_i is the predicted heatmap result. HA , HP , HM indicate the morphological heights of the three parts of the vertebrae, Moderately and severely fractured vertebrae ($G < 0.74$), and mildly fractured ($0.74 < G < 0.8$) [23].



Figure 9: First, the sagittal slice where its center of mass was located was selected based on the segmented vertebrae CT images, and its six points were regressed by M-SCN, and then the Genant value was calculated

3 Experiments and Discussions

3.1 Experimental Environment and Dataset

The experimental environment was conducted under the WSL-Ubuntu22.04 system, and the programming environment was python3.10.8, which was completed based on the Pytorch deep learning framework. The relevant hardware environment is CPU Intel(R) Core(TM) i9-12900K, GPU NVIDIA GeForce RTX 4090x2, Cuda version 11.7. 128 GB of RAM and 64-bit operating system.

The dataset used was VerSe'19 and VerSe'20 [24], which is a large dataset of vertebral segmentation and identification organized by MICCAI, where VerSe'19 and VerSe'20 include 160 and 355 CT spine scans, respectively. Spine types included vertebral fractures, metal implants, bone cement, and transitional vertebrae. In this paper, images of vertebral fractures were screened as the training dataset for this study, and 74 CT data provided by Guizhou Orthopedic Hospital were also collected. Fig. 10 a and b shows the vertebral localization data and the vertebral segmentation data, respectively.

3.2 Training Parameter Settings

Raw CT images cannot be used directly for model training. In this paper, SimpleITK is used as a CT data processing toolkit to preprocess and enhance real-time CT data. For each CT image, it was first resampled to a uniform voxel spacing of 1 mm. For each voxel value in the CT volume, it was cropped to between $[-2048, 8192]$ and the overall data was normalized by Eq. (19). Where x is the input image, μ is the image mean, and σ^2 is the image variance, a very small constant set to $2.220e-16$.

$$x' = \frac{x - \mu}{\sqrt{\sigma^2 + \varepsilon}} \quad (19)$$

A total of 30,000 iterations were set for training, the optimizer was SGD, The ratio of training set, validation set and test set was 7:2:1, the learning rate was set to 10^{-4} , the L2 weight regularisation

factor was set to 5^{-4} , and the BatchSize was set to 4. In addition, in order to improve the convergence efficiency of the model, gradient clipping was used in the training of the model, and the value of the gradient clipping was set to 2. The details of the experiment are shown in Table 1 below.

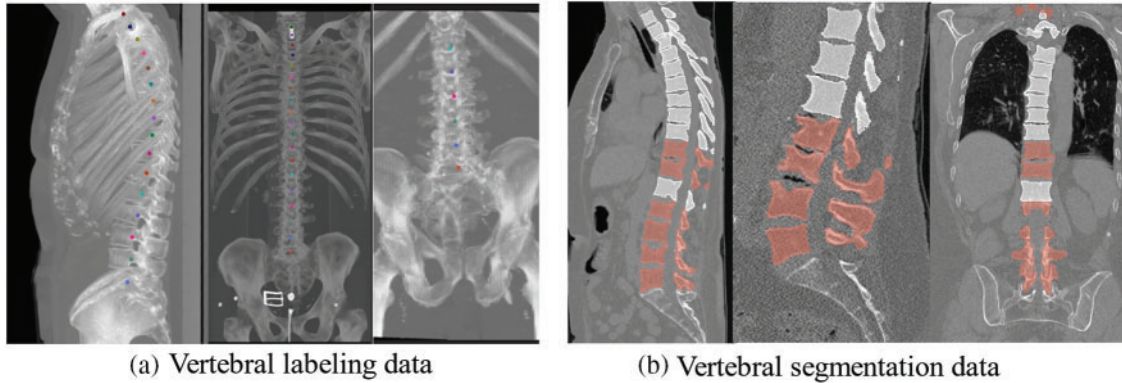


Figure 10: Experimental data

Table 1: Experimental environment configuration

Configure	Parameters
CPU	Intel(R) Core(TM) i9-12900K
GPU	GPU NVIDIA GeForce RTX 4090x2
Cuda version	11.7
Python	3.10.8
Pytorch	2.0.1
System	WSL-Ubuntu22.04
Data allocation ratio	7:2:1
Optimization	SGD
Lr_rate	10^{-4}
Lr_scheduler	Plateau
Gradient clipping	2.0
BatchSize	4
L2 decay	5^{-4}
Image preprocessing	Resample, Clamp image, Gaussian filter, Z-score normalization, Random rotation, Random flipping, Random translation
Processing toolkit	SimpleITK, Scipy
Loss function	Dice loss, Cross entropy loss

3.3 Evaluation of Indicators

In order to evaluate the performance of the network, this paper uses commonly used evaluation methods for studying segmentation [25–27], which can be divided into two categories:

1. The degree of overlap or similarity based on volume. For example, Dice Similarity Coefficient (DSC), Precision, and Recall. Such metrics are used as a way to calculate the similarity between predicted and true values.

2. Measurements based on the distance between the predicted surface and the real surface. For example, maximum symmetric surface distance (MSSD) and average symmetric surface distance (ASSD).

DSC, precision, and recall can be expressed in terms of true positives (TP), false negatives (FN), and false positives (FP). They can be defined as follows:

$$DSC = \frac{2TP}{2TP + FP + FN} \quad (20)$$

$$Precision = \frac{TP}{TP + FP} \quad (21)$$

$$Recall = \frac{TP}{TP + FN} \quad (22)$$

Let's define the Average Symmetric Surface Distance ASSD, assuming that P is the predicted volume and G is the true volume, as:

$$ASSD(P, G) = \frac{\left(\sum_{s_p \in S(P)} d(s_p, S(G)) + \sum_{s_G \in S(G)} d(s_G, S(P)) \right)}{|S(P) + S(G)|} \quad (23)$$

where d is the distance metric, and $d(s_p, S(G))$, $d(s_G, S(P))$ are the distances between each point on the predicted surface and the real surface (each point on the real surface and the predicted surface). Where the distance d between the point and the surface can be defined as:

$$d(l_c, S(C)) = \min_{l'_c \in S(C)} \|l_c - l'_c\| \quad (24)$$

The maximum symmetric surface distance MSSD can be defined as:

$$MSSD(P, G) = \max \{h(S(P), S(G)), h(S(G), S(P))\} \quad (25)$$

where $h(S(P), S(G))$ denotes the one-way Hausdorff distance between the surface of the predicted volume $S(P)$ and the surface of the real volume $S(G)$, defined as follows:

$$h(S(P), S(G)) = \sup_{s_p \in S(P)} \left\{ \inf_{s_G \in S(G)} d(s_p, s_G) \right\} \quad (26)$$

The definition of $h(S(G), S(P))$ is obtained by the same reasoning.

3.4 Results

3.4.1 Results Overview

A total of 30,000 iterations were set up for this experiment, and the experimental data were recorded every 6,000 iterations due to the long time consumed for each test. [Table 2](#) gives the evaluation results of U-Net vertebral segmentation, it can be seen that the model starts to converge after 12,000 iterations, the maximum value of Dice obtained is 93.50%, the classification performance precision

and recall both reach 95.82%, and the recall rate reaches 91.70%, respectively. The ASSD reaches a minimum of 0.262 mm. Fig. 11 shows the model gradient values and network loss values after clipping. In most cases, the model gradient remains between 0 and 1. In contrast using the gradient clipping training method is faster for training and its Loss minimum fluctuates around 0.01. Fig. 12 shows the curve changes of all the indicators, it can be seen that with the increase in the number of training, the model of the various test indicators gradually converge to saturation, and the curve float is small, which also shows that the model prediction output is relatively smooth.

Table 2: U-Net vertebral segmentation results

Iter	DSC (%)	ASSD (mm)	MSSD (mm)	Precision (%)	Recall (%)
6000	90.11	0.505	9.223	90.03	90.44
12000	92.64	0.314	6.062	94.81	90.73
18000	93.02	0.289	5.350	95.82	90.51
24000	93.50	0.263	5.022	95.52	91.70
30000	93.50	0.262	4.962	95.76	91.46

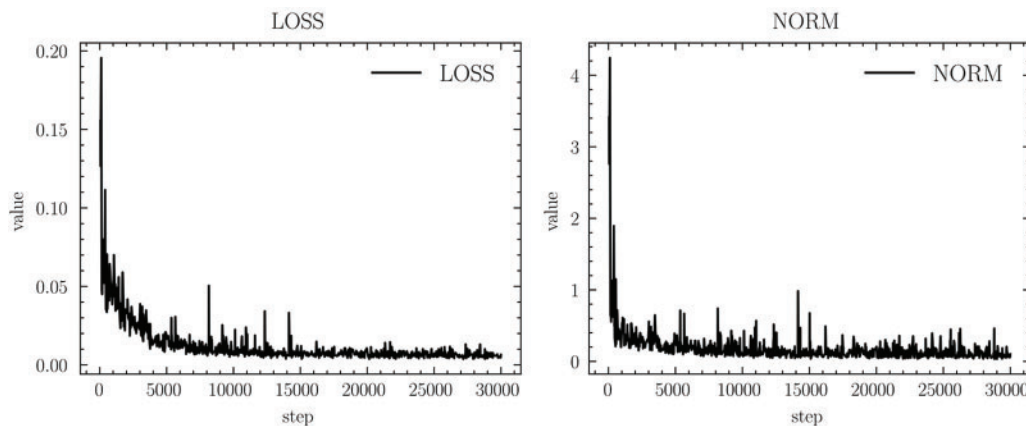


Figure 11: Gradient clipping results in norm and network loss

3.4.2 3-Fold Cross Validation

In the current experiments, due to the insufficient number of available datasets, to minimize random results due to a single division of the training and validation sets. Therefore, multiple divisions were performed using the available datasets, thus avoiding the selection of “chance models” that cannot generalize their capabilities due to specific divisions. In this study, the training dataset was divided into three similar subsets, three rounds of cross-validation were performed, and the average of the metrics for the three rounds of training was calculated. The experimental data are shown in Fig. 13, indicating that in the three rounds of cross-validation, the evaluation metrics of the model do not fluctuate much and are relatively smooth, which also indicates that the model has good generalization performance.

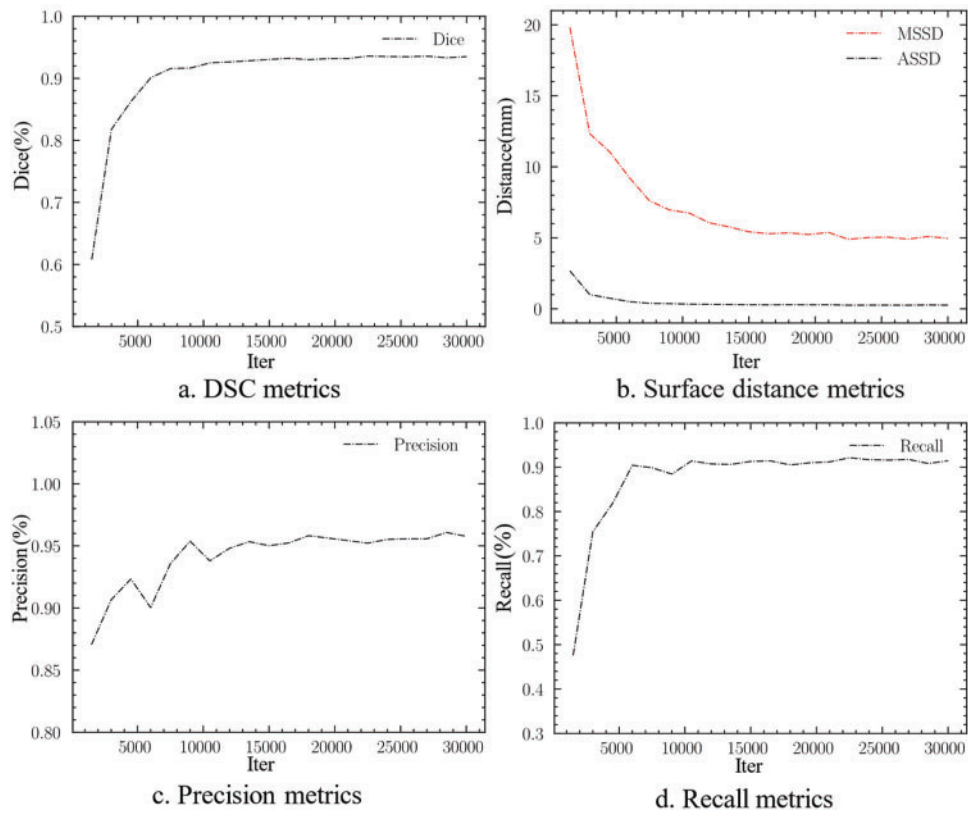


Figure 12: Segmentation metrics chart

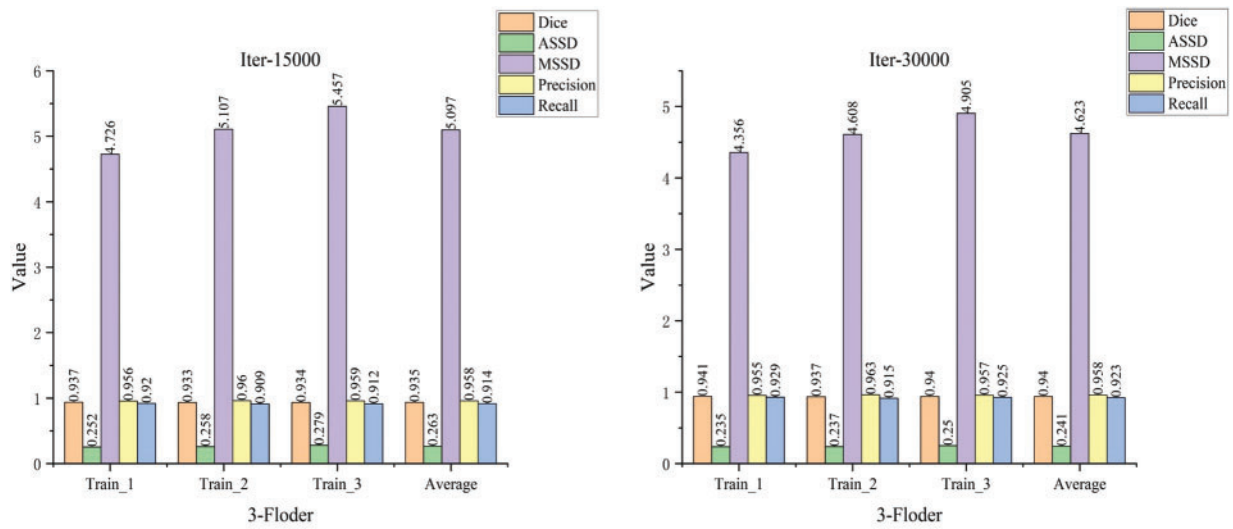


Figure 13: Cross-validation results

3.4.3 Compare with Other Segmentation Models

In this paper, a comparison is made with other SOTA models (TransUNet and SwinUNet) and the experimental results are shown in Table 3. In this study, the size of the dataset used is small, and the models that used a Transformer did not have a metric improvement in the evaluation metrics. Whereas the performance of Transformer models usually improves with the size of the model, larger models require more data for training, otherwise they are prone to overfitting. At the same time Transformer also brings a larger amount of computation. In the CT image Overlapping patch inference experiment with size [164, 164, 185], its results show that UNet inference takes 9.570 s, TransUNet takes 51.678 s, and SwinUNet takes 78.112 s. TransUNet and SwinUNet are respectively slower than units by 5.4 times and 8.163 times. Taken together, UNet reasoning is faster and also has higher accuracy, which is more suitable for this task.

Table 3: Model comparison experiment

Algorithm	DSC (%)	ASSD (mm)	MSSD (mm)	Precision (%)	Recall (%)	Inference time (s)
3D-UNet	93.50	0.263	5.022	95.52	91.70	9.570
3D-TransUNet	93.11	0.223	4.921	96.52	90.90	51.678
3D-SwinUNet	94.50	0.202	4.862	95.86	91.46	78.112

3.4.4 M-SCN Network Model Ablation Experiments

In the localization model, an important evaluation metric is the distance between the predicted coordinates and the true coordinates, which is considered to be correctly predicted if the distance is less than 20 mm [28] and can be defined as:

$$\#ID = \{\|x_k - x'_j\|_2 = PE \leq 20\} \quad (27)$$

In this paper, the ablation validation of the M-SCN model is carried out to prove the effectiveness of the model improvement. The experimental results are shown in Table 4, where the first column represents the localization results of the SCN model whose prediction error rate is 1.45%. The second and third columns represent the prediction results when only LN or GN is used, and its prediction error rate of 7.78% for LN is better than the prediction error rate of 9.31% for GN, which also highlights that the LN network plays a major role in the localization prediction. The fourth column indicates that LN is used together with GN and its prediction results are close to the SCN network model. Finally in the case of using AG, Multiple Filtering and Residual Connection respectively, their prediction results have been improved accordingly. The last column represents the experimental results of the M-SCN network, which has a prediction error rate of 0.45%, which is 1.09% lower compared to the SCN prediction error rate.

Table 4: Ablation experiment

AG	Multiple filtering	Residual connection	LN	GN	SCN	#O _r (%)		
						r > 4 mm	r > 10 mm	r > 20 mm
×	×	×	×	×	✓	6.36	2.90	1.54
×	×	×	✓	×	×	9.16	7.78	7.78
×	×	×	×	✓	×	13.15	9.86	9.31
×	×	×	✓	✓	×	7.65	2.12	1.63
✓	×	×	✓	✓	×	6.95	2.34	1.85
×	✓	×	✓	✓	×	4.95	1.65	1.65
×	×	✓	✓	✓	×	5.63	2.30	1.69
✓	✓	✓	✓	✓	×	5.31	0.91	0.45

3.5 Visualization and Discussion of Experimental Results

Fig. 14 shows the individual vertebrae segmentation process. In Fig. 14a, the localization results of the M-SCN model are marked using red dots, and Fig. 14b represents the results of cropping the surrounding voxels, whose background and other interfering information have been eliminated. Fig. 14c represents the segmentation result of the current vertebra (the sagittal part of this vertebra is shown), and Fig. 14d represents the result of resampling the position of the vertebra cropped at the current position and restoring it to the initial position. Fig. 14e shows the mixing of the current segmentation result with the original picture.

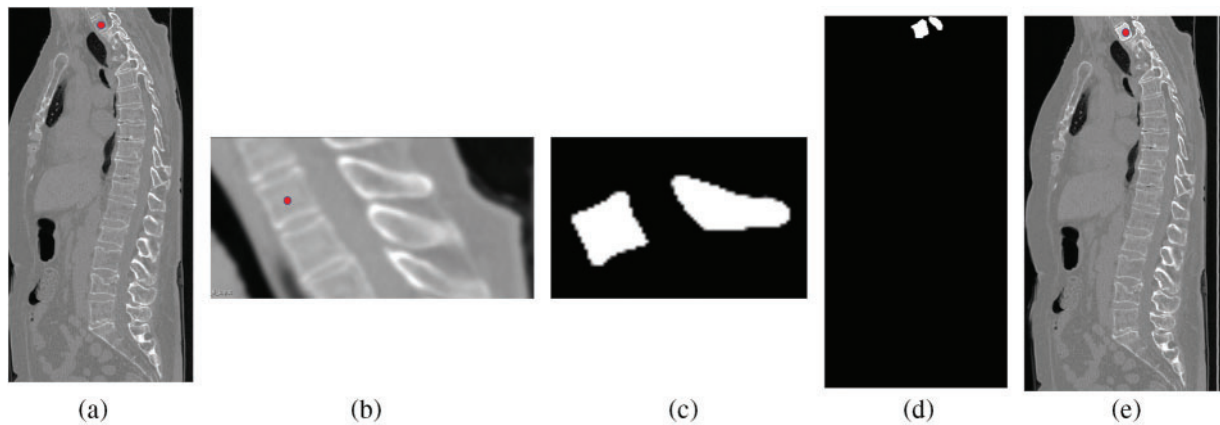
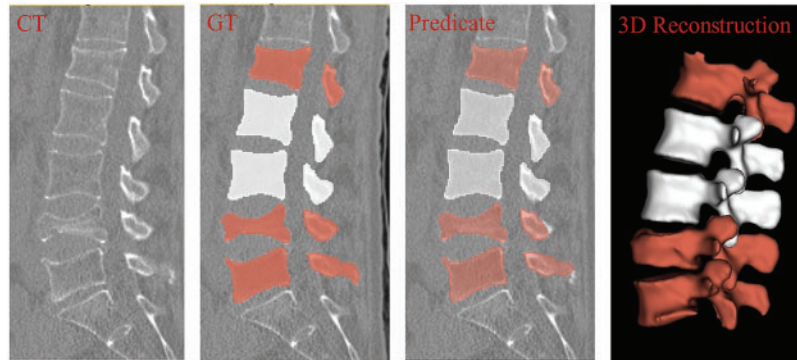
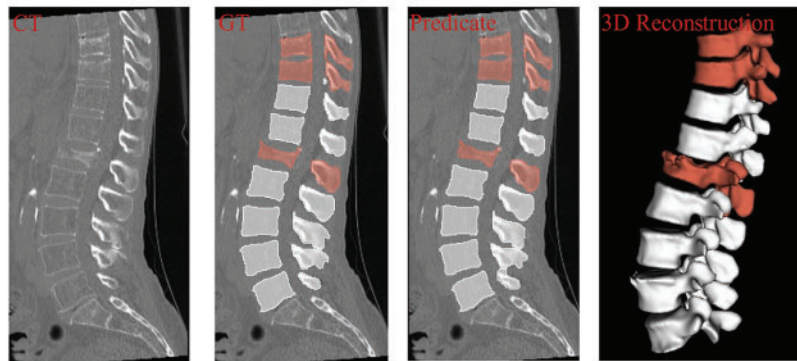


Figure 14: Visualization of the vertebral segmentation process, figure a is the original CT image, figure b is the clipped image, figure c is the result of individual vertebral segmentation, and figure d and figure e are the results of resampling the clipped position to the original position

Fig. 15 shows 2 sets of segmentation results for the entire spine, using white to mark normal vertebrae and red to mark fractured vertebrae. In Fig. 15a, the segmentation results for the L4 and L5 vertebrae show incomplete edge segmentation compared with the actual values, whereas in Fig. 15b, there is redundancy in the segmentation results for the L5 vertebrae. The overall segmentation results meet expectations. From the 3D reconstruction results, it can be seen that the red vertebrae show significant irregularities or depressions compared to the white vertebrae, which also suggests that these vertebrae may have vertebral fractures.



(a) Segmentation of predictive data 1.



(b) Segmentation of predictive data 2.

Figure 15: Vertebral segmentation results, where red represents a possible fracture of the vertebral body and white is normal

Fig. 16 illustrates the Genant assessment process. Fig. 16a represents a slice of the original CT image, Fig. 16b represents a sagittal slice of the vertebral body where the center of mass is located, and Fig. 16c represents the results of the Genant assessment. This experiment visualizes the M-SCN prediction heatmap. As shown in Figs. 17, 17a represents the GN output heatmap and Fig. 17b represents the LN output heatmap. Due to the inclusion of $7 \times 7 \times 7$ large convolutional kernel in GN, it has a larger perceptual field than LN, and the final prediction result Fig. 17c also becomes a more prominent data feature due to the filtering effect of GN. Fig. 18 shows the final prediction results of

the heatmap after extracting the highlighted areas (maximum values). It can be seen that among the four sets of prediction results, the M-SCN network model's prediction is more accurate for the middle region, and produces errors for the uppermost and lowermost vertebral marker points. In combination with the segmentation results. This may have contributed to the deterioration of segmentation results in the L4 and L5 vertebrae results.

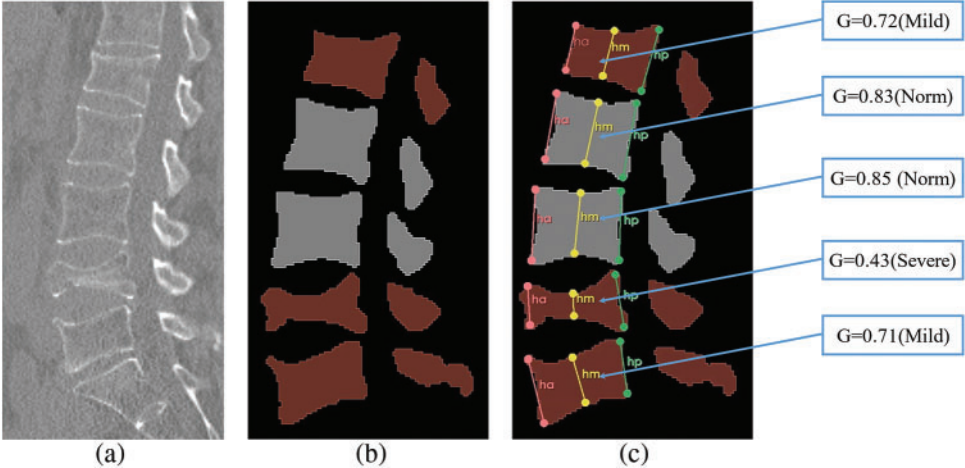


Figure 16: Vertebral Genant semi-qualitative assessment results, a shows the original spinal section image, b shows the segmented image, and c shows the Genant semi-qualitative assessment results

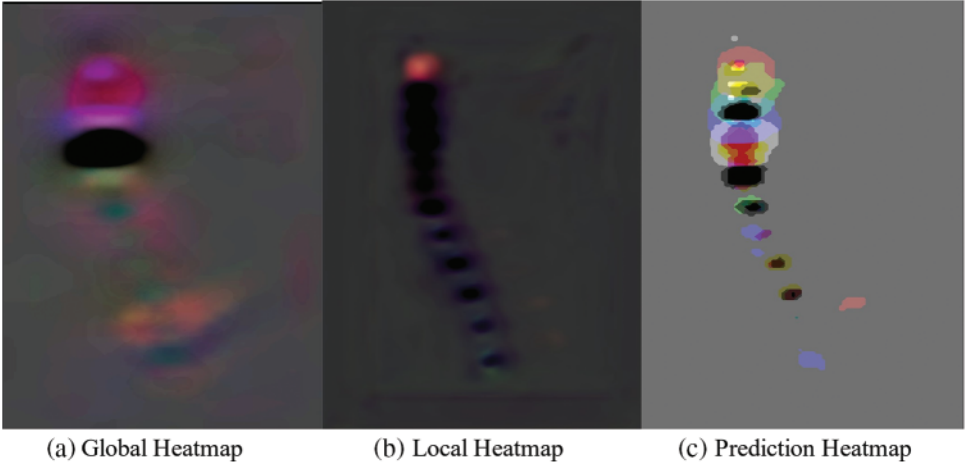


Figure 17: M-SCN prediction heatmap visualization

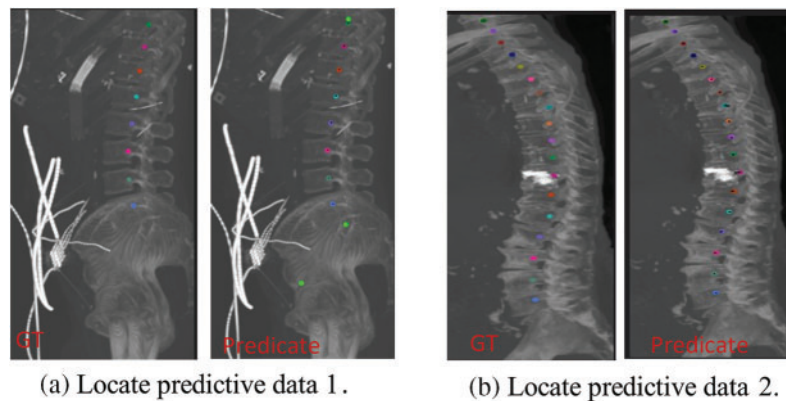


Figure 18: Vertebral localization results

4 Conclusion

This paper describes a fully automated 3D CNN-based system for the intelligent recognition of vertebral fractures. The system performs vertebral fracture detection in two steps (vertebral localization and vertebral segmentation) and subsequently evaluates the fracture grade based on sagittal slices (Genant Semi-Qualitative Assessment).

In this paper, we greatly improved the efficiency and accuracy of vertebral segmentation detection by localizing the vertebrae and clipping the ROI region, and used the segmentation masks obtained from the segmentation model (vertebral fracture mask, normal vertebral mask, and background mask) as the detection results of vertebral fracture. Subsequently, for the segmentation mask, the center sagittal slice was selected according to the center of mass coordinates, the six evaluation key points were regressed by M-SCN, and finally the vertebrae were classified into normal vertebrae and fractured vertebrae (moderate, mild, and severe) according to the Genant semiquantitative assessment. In the experiment, the vertebral segmentation Dice index, ASSD index and MSSD index reached 93.50%, 0.262 and 4.962 mm, respectively, and the vertebral bone localization also achieved relatively excellent results. The comprehensive experimental visualization results show that for some CT images, the localization model and segmentation model still have some defects. In the future, we expect to introduce graph optimization [29] into the keypoint regression model to improve the localization accuracy. And more datasets are collected to improve the accuracy of network training. We expect that the proposed method can effectively improve the diagnostic efficiency of vertebral fracture and can be applied in the clinic.

Acknowledgement: The authors would like to thank the editorial department and reviewers for their suggestions on this article, which have helped us greatly improve the quality of the article.

Funding Statement: This research was funded by the Guizhou Provincial Key Technology R&D Program [2022] General 264 and by the Guizhou Provincial Key Technology R&D Program [2023] General 096.

Author Contributions: The authors confirm contribution to the paper as follows: Study conception and design: Yuhang Wang, Zhiqin He; data collection: Yu Tang, Maoyun Zhu; analysis and interpretation

of results: Qinmu Wu, Tingsheng Lu; draft manuscript preparation: Yuhang Wang. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

- [1] P. Sambrook and C. Cooper, “Comparative statistical analysis of osteoporosis treatment based on Hungarian claims data and interpretation of the results in respect to cost-effectiveness,” *Lancet*, vol. 367, pp. 2010–2018, 2006. doi: [10.1016/S0140-6736\(06\)68891-0](https://doi.org/10.1016/S0140-6736(06)68891-0).
- [2] A. Klibanski *et al.*, “Osteoporosis prevention, diagnosis, and therapy,” *J. Amer. Med. Assoc.*, vol. 285, no. 6, pp. 785–795, 2001.
- [3] G. Ballane, J. A. Cauley, M. M. Luckey, and G. E. Fuleihan, “Worldwide prevalence and incidence of osteoporotic vertebral fractures,” *Osteoporos. Int.*, vol. 28, pp. 1531–1542, 2017. doi: [10.1007/s00198-017-3909-3](https://doi.org/10.1007/s00198-017-3909-3).
- [4] A. Valentinitich *et al.*, “Opportunistic osteoporosis screening in multi-detector CT images via local classification of textures,” *Osteoporos. Int.*, vol. 30, pp. 1275–1285, 2019. doi: [10.1007/s00198-019-04910-1](https://doi.org/10.1007/s00198-019-04910-1).
- [5] Y. Wang, J. Yao, J. E. Burns, and R. Summers, “Osteoporotic and neoplastic compression fracture classification on longitudinal CT,” in *Proc. ISBI*, Prague, Czech Republic, 2016, pp. 1181–1184.
- [6] J. E. Burns, J. Yao, and R. M. Summers, “Vertebral body compression fractures and bone density: Automated detection and classification on CT images,” *Radiol.*, vol. 284, no. 3, pp. 788–797, 2017. doi: [10.1148/radiol.2017162100](https://doi.org/10.1148/radiol.2017162100).
- [7] S. Hamidian, B. Sahiner, N. Petrick, and A. Pezeshk, “3D convolutional neural network for automatic detection of lung nodules in chest CT,” in *Proc. SPIE*, vol. 10134, 2017, pp. 54–59.
- [8] J. Laserson *et al.*, “TextRay: Mining clinical reports to gain a broad understanding of chest x-rays,” in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, vol. 11071, 2018, pp. 553–561. doi: [10.1007/978-3-030-00934-2](https://doi.org/10.1007/978-3-030-00934-2).
- [9] K. Murata *et al.*, “Artificial intelligence for the detection of vertebral fractures on plain spinal radiography,” *Sci. Rep.*, vol. 10, no. 1, pp. 20031, 2020. doi: [10.1038/s41598-020-76866-w](https://doi.org/10.1038/s41598-020-76866-w).
- [10] A. Bar, L. Wolf, O. B. Amitai, E. Toledano, and E. Elnekave, “Compression fractures detection on CT,” in *Proc. SPIE 10134, Medical Imaging 2017: Computer-Aided Diagnosis*, vol. 10134, 2017, pp. 1036–1043.
- [11] N. Tomita, Y. Y. Cheung, and S. J. Hassanpour, “Deep neural networks for automatic detection of osteoporotic vertebral fractures on CT scans,” *Comput. Biol. Med.*, vol. 98, pp. 8–15, 2018.
- [12] J. Nicolaes *et al.*, “Detection of vertebral fractures in CT using 3D convolutional neural networks,” in *Computational Methods and Clinical Applications for Spine Imaging (CSI 2019)*, vol. 11963, 2020, pp. 3–14.
- [13] D. Chetrit *et al.*, “3D convolutional sequence to sequence model for vertebral compression fractures identification in CT,” *Med. Image Comput. Comput. Assist. Interv.–MICCAI 2020*, vol. 12266, pp. 743–752, 2020.
- [14] C. Payer, D. Stern, H. Bischof, and M. Urschler, “Integrating spatial configuration into heatmap regression based CNNs for landmark localization,” *Med. Image Anal.*, vol. 54, pp. 207–219, 2019.
- [15] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, vol. 9351, 2015, pp. 234–241.
- [16] H. K. Genant, C. Y. Wu, and V. Kuijk, “Vertebral fracture assessment using a semiquantitative technique,” *J. Bone Miner. Res.*, vol. 8, no. 9, pp. 1137–1148, 2009.

- [17] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Proc. NIPS*, Barcelona, Spain, 2016, vol. 29.
- [18] C. Lian *et al.*, "Multi-task dynamic transformer network for concurrent bone segmentation and large-scale landmark localization with dental CBCT," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, vol. 12264, 2020, pp. 807–816.
- [19] O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, and K. Misawa, "Attention U-Net: Learning where to look for the pancreas," arXiv preprint arXiv:1804.03999, 2018.
- [20] P. Wang, P. Chen, Y. Yuan, D. Liu, and Z. Huang, "Understanding convolution for semantic segmentation," in *Proc. WACV*, Lake Tahoe, NV, USA, 2018, pp. 1451–1460.
- [21] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, New York, NY, USA, 2015, pp. 1440–1448.
- [22] F. Milletari, N. Navab, and S. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 3DV*, Stanford, CA, USA, 2016, pp. 565–571.
- [23] H. Yousefi, E. Salehi, O. S. Sheyjani, and H. Ghanaati, "Lumbar spine vertebral compression fracture case diagnosis using machine learning methods on CT images," in *Proc. IPRIA*, Tehran, Iran, 2019, pp. 179–184.
- [24] A. Sekuboyina *et al.*, "VerSe: A vertebrae labelling and segmentation benchmark for multi-detector CT images," *Med. Image Anal.*, vol. 73, pp. 102166, 2021. doi: [10.1016/j.media.2021.102166](https://doi.org/10.1016/j.media.2021.102166).
- [25] Y. Lu, and H. Quan, "Adversarial multi-sample interpolation for medical image segmentation," in *Proc. BIBM*, Istanbul, Turkiye, 2023, pp. 1337–1342.
- [26] P. Bilic *et al.*, "The liver tumor segmentation benchmark (lits)," *Med. Image Anal.*, vol. 84, pp. 1–24, 2021.
- [27] B. Prencipe, N. Altini, G. D. Cascarano, A. Guerriero, and A. Brunetti, "A novel approach based on region growing algorithm for liver and spleen segmentation from CT scans," in *Intelligent Computing Theories and Application*, vol. 12463, 2020, pp. 398–410. doi: [10.1007/978-3-030-60799-9](https://doi.org/10.1007/978-3-030-60799-9).
- [28] Y. C. Yeh, C. Weng, Y. J. Huang, C. J. Fu, T. T. Tsai, and C. Y. Yeh, "Deep learning approach for automatic landmark detection and alignment analysis in whole-spine lateral radiographs," *Sci. Rep.*, vol. 7618, pp. 479, 2021. doi: [10.1038/s41598-021-87141-x](https://doi.org/10.1038/s41598-021-87141-x).
- [29] D. Meng, E. Mohammed, E. Boyer, and S. Pujades, "Vertebrae localization, segmentation and identification using a graph optimization and an anatomic consistency cycle," in *Int. Workshop Mach. Learn. Med. Imaging*, vol. 13582, 2022, pp. 307–317.