PAPER

Cvm-Unet: a spinal x-ray multi-lesion segmentation network based on convnext and vmamba

To cite this article: Zhilong Xue *et al* 2025 *Eng. Res. Express* **7** 025293

Engineering Research Express

PAPER

Cvm-Unet: a spinal x-ray multi-lesion segmentation network based on convnext and vmamba

RECEIVED
7 March 2025

REVISED
30 May 2025

ACCEPTED FOR PUBLICATION
4 June 2025

PUBLISHED
16 June 2025

Zhilong Xue , Shuangcheng Deng* , Zhiwu Li , Yang Yang, Yiqun Yue, Chenping Chen, Yubang Liu and Shilong Sun

Beijing Institute of Petrochemical Technology, Qingyuan North Road, No. 19, Daxing District, Beijing 102617, People's Republic of China
* Author to whom any correspondence should be addressed.

E-mail: 2023520074@bipt.edu.cn, dengshuangcheng@bipt.edu.cn, 2022520049@bipt.edu.cn, 2022520078@bipt.edu.cn, 2024520144@bipt.edu.cn, 2024520105@bipt.edu.cn, 2024520128@bipt.edu.cn and 2024520135@bipt.edu.cn

Keywords: UNet, spinal segmentation, multiple spinal diseases, vss block, attention mechanism

Abstract

With the rapid advancements in medical imaging and artificial intelligence, the early diagnosis and precise treatment of spinal disorders have emerged as critical priorities in clinical research. However, current diagnostic approaches predominantly rely on the subjective expertise of clinicians, which is inherently limited by individual knowledge and often time-intensive. Although various spinal segmentation networks have been proposed, their applicability and accuracy in handling multiple spinal pathologies remain suboptimal.To address these limitations, this study proposes a novel semantic segmentation model for spinal x-ray images, designed to enable accurate identification of diverse spinal lesions. The model adopts U-Net as the foundational architecture, integrates ConvNeXt as the backbone for enhanced feature representation, and incorporates the VSS Block from VMamba as the decoder to improve contextual understanding and feature extraction. Additionally, a Res-ReLU Block is introduced at the skip connections, while a spatial-channel cooperative attention (SCSA) mechanism is embedded in the bottleneck layer to further enhance the model's adaptability, precision, and robustness across varied spinal conditions.Extensive experiments conducted on our curated spinal x-ray dataset demonstrate that the proposed method achieves superior performance compared to existing models, with Dice, mIoU, and Hausdorff Distance (HD) scores reaching 91.1, 85.5, and 3.852, respectively. Furthermore, the model accurately segments a range of spinal abnormalities, including spondylolysis, vertebral wedge deformities, spondylolisthesis, and scoliosis, thereby offering strong support and guidance for clinical image analysis.

1. Introduction

The spine plays a crucial role in human movement [1]. With the aging global population [2], changes in lifestyle, increased sedentary habits, and the influence of traumatic factors, the incidence of spinal diseases such as scoliosis, spondylolysis, vertebral wedge deformity, and spondylolisthesis continues to rise [3], becoming a significant public health issue worldwide. These conditions not only cause pain and limited mobility but severely impact the quality of life for patients. If not addressed in a timely manner, they may lead to chronic pain and further deterioration [4]. Research has shown that early and accurate diagnosis of spinal diseases is critical for formulating effective treatment plans and improving patient prognosis [5].

For decades, due to the low cost and relatively low radiation dose of x-ray imaging [6], it has been widely used as an essential tool for the clinical evaluation and monitoring of spinal diseases [7]. However, the interpretation of spinal x-ray images presents several challenges. Traditional diagnostic methods rely on manual analysis by medical experts, which is not only time-consuming and labor-intensive but also prone to variability in assessments, even among experienced radiologists, despite adhering to the same diagnostic standards. Furthermore, interpretative errors in radiological images remain a significant issue in clinical practice [8]. The

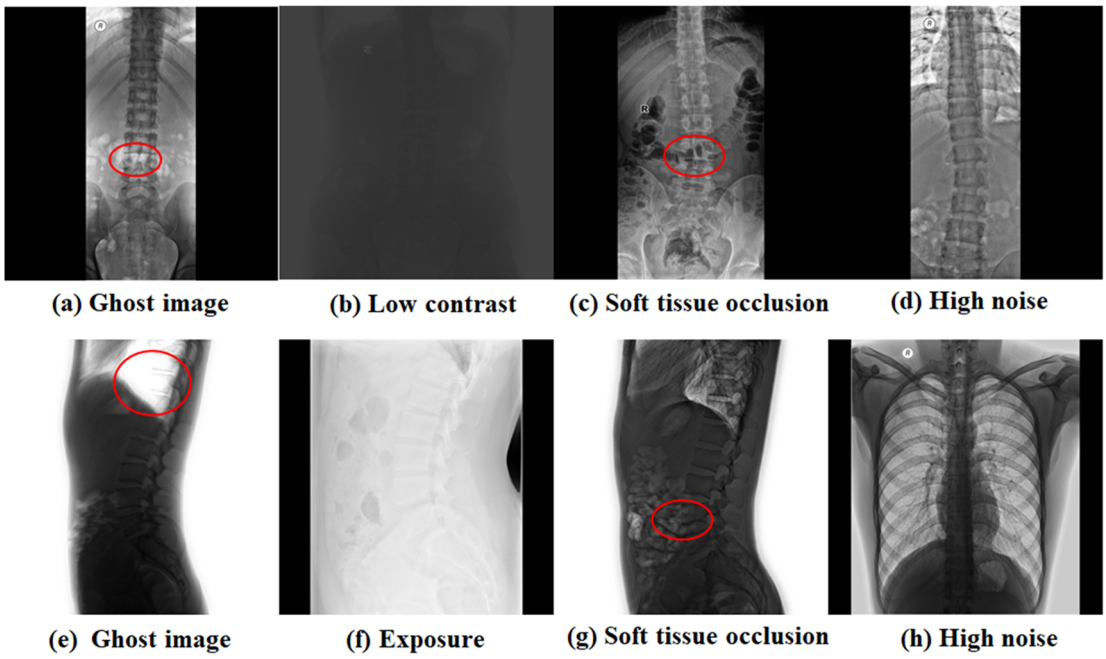

Figure 1. Quality issues in thoracolumbar spine x-ray images. The red-marked regions indicate areas with image quality issues, and the corresponding issue types are labeled beneath each image. Specifically: (a) and (e) show ghost images; (b) has low contrast; (c) and (g) are affected by soft tissue occlusion; (d) and (h) suffer from high noise; (f) is overexposed.

quality of x-ray images (as shown in figure 1) also affects diagnostic accuracy, with factors such as soft tissue shadows, artifacts, and the limitations of equipment and post-processing algorithms [9], potentially making subtle lesions difficult to identify. Therefore, achieving fast, automated, and accurate diagnosis of spinal diseases has become a key research direction in this field [10].

In recent years, the rapid development of artificial intelligence (AI), particularly deep learning techniques, has brought revolutionary advancements to medical image analysis [11]. In the field of medical image segmentation, deep learning algorithms have significantly improved the diagnostic efficiency of spinal diseases [12], offering a more objective and efficient means of analysis to overcome the limitations of traditional methods. These advanced segmentation technologies have become key to enhancing diagnostic performance [13].

Convolutional Neural Networks (CNNs), known for their exceptional feature extraction capabilities, have achieved remarkable success in medical image segmentation [14], driving the continuous development of segmentation techniques [2]. U-Net [15], a classic fully convolutional network (FCN), employs an encoder-decoder architecture that effectively captures and restores image details, while utilizing skip connections to preserve high-resolution information. This structure has been widely applied in medical image segmentation. U-Net++ [16] optimizes the skip connections to further improve segmentation accuracy, while Mask R-CNN [17] adds an instance segmentation branch to Faster R-CNN [18], enabling finer medical image segmentation. DenseNet [19] enhances information flow through a dense connectivity mechanism, improving feature reuse and demonstrating excellent performance in medical image classification and segmentation tasks. ResNet [20], leveraging residual learning, mitigates the gradient vanishing problem, making deeper network training feasible. However, the limitation of CNNs lies in their reliance on local receptive fields for convolution operations, making it challenging to effectively capture global information and model long-range dependencies [21].

To address this limitation, researchers have introduced Transformer-based architectures [22], with Vision Transformer (ViT) and Swin Transformer making significant progress. ViT [23] was the first to apply a standard Transformer encoder structure to computer vision tasks, effectively integrating global information through the Multi-Head Self-Attention (MSA) mechanism. Swin Transformer [24] utilizes hierarchical feature extraction and a sliding window attention strategy, achieving efficient medical image segmentation. Additionally, TransUNet [25] combines U-Net with Transformer, demonstrating superior generalization capabilities in medical image segmentation tasks.

In 2023, the Mamba model emerged as a new sequence modeling architecture [26]. Inspired by the classical state-space model (SSM), it is designed to efficiently capture complex dependencies within sequence data, becoming a formidable competitor to Transformers. The Mamba model maintains comparable modeling capabilities to Transformers while offering near-linear scalability for sequence lengths, demonstrating

exceptional performance in long-sequence modeling tasks. Furthermore, several variants based on the Mamba architecture have emerged, such as u-mamba [27], Weak-mamba-unet [28], and graph-mamba [29], bringing innovations across various fields. For instance, in biomedical image segmentation tasks, u-mamba adopts a hybrid CNN-SSM architecture, effectively capturing both local fine-grained features and long-range dependencies.

Regarding spinal datasets, several publicly available spinal x-ray image datasets, such as BUU-LSPINE [30] and VinDr-SpineXR [31], support spinal lesion detection and classification research. Among these, the VinDr-SpineXR dataset includes 13 common spinal abnormalities. However, for the field of spinal segmentation, directly usable datasets are scarce and often require some degree of preprocessing.

Despite the substantial progress achieved by existing methods in the field of medical image analysis, current models still exhibit limited applicability and insufficient accuracy in multi-lesion segmentation of spinal x-ray images. This limitation primarily stems from the considerable heterogeneity and morphological complexity of spinal pathologies, which vary significantly in type and pathological characteristics. Most existing models are optimized for single-lesion scenarios, thereby constraining their feature extraction capacity and impeding the effective integration of global contextual and local structural information. Furthermore, conventional encoder–decoder architectures suffer from inherent limitations in fusing low-level spatial details with high-level semantic features, making it challenging to cope with the intricate anatomical structures of diverse spinal disorders.

To address these issues, we propose a novel multi-lesion spinal segmentation model, Cvm-UNet, designed to enhance the precision of multi-lesion spinal segmentation while improving vertebral segmentation accuracy. The main contributions of this paper are as follows:

1. We reconstruct the VSS Block and use it as a feature extraction module for the decoder, enhancing noise resistance and improving the ability to capture complex structural features, optimizing the fusion of global and local features.
2. We introduce a spatial-channel cooperative attention (SCSA) mechanism at the bottleneck of the model, allowing the model to adaptively adjust attention distribution and focus on key regions and important feature channels, thereby addressing the issue of inaccurate key region information retrieval.
3. We introduce an improved residual block (Res-ReLU Block) in the skip connections between the encoder and decoder, facilitating the efficient transfer of low-level texture features (such as vertebral endplates) and high-level semantic features (such as spinal curvature), mitigating the gradient vanishing problem and enhancing feature representation for complex lesions.
4. We construct a spinal x-ray dataset encompassing various lesion types, including normal thoracolumbar vertebrae, spondylolysis, vertebral wedge deformity, spondylolisthesis, and scoliosis, with both anteroposterior and lateral views, totaling 1,174 images. This provides rich data support for model training.
5. We propose a novel Cvm-UNet architecture that integrates ConvNeXt, the VSS Block, the Res-ReLU Block, and a spatial-channel collaborative attention (SCSA) mechanism. This design enables more accurate extraction of critical features from complex pathological regions, even under the challenging conditions of low contrast and high noise in x-ray imaging. The model demonstrates superior capability in fusing global and local features, substantially enhancing segmentation stability and accuracy. By overcoming the limitations of existing methods that are often optimized for single-lesion scenarios, Cvm-UNet exhibits strong adaptability and delivers robust segmentation performance across diverse spinal pathologies.

The remainder of this paper is structured as follows: section 2 provides a brief overview of related work. In section 3, we introduce Cvm-UNet and its components. Section 4 details the dataset and experimental results. Finally, section 5 offers a concise summary and discusses future prospects.

2. Related work

2.1. Advances in deep learning for spinal x-ray segmentation

In recent years, deep learning techniques have made significant strides in the field of medical image segmentation. The U-Net model, proposed by Ronneberger *et al.*, has been widely applied to spinal x-ray image segmentation tasks due to its symmetric encoder-decoder structure. Building upon this, several optimized versions have been proposed, such as Spine-UNetx [32], which incorporates advanced feature extraction modules to enhance the segmentation accuracy of vertebral regions. Attention U-Net [33] integrates attention mechanisms, effectively improving segmentation performance for small lesion regions. Additionally, HRNet [34] employs a multi-scale parallel feature extraction structure, enhancing high-resolution feature

representation while retaining global information, and has demonstrated excellent performance in spinal x-ray segmentation tasks. However, both U-Net and its variants still have limitations in capturing global information when processing spinal x-ray images.

To address these issues, researchers have introduced Transformer architectures to optimize spinal x-ray segmentation by leveraging their superior global information modeling capabilities, as seen in models like ViT. However, Transformers still face challenges in fine-grained feature extraction for medical image segmentation tasks. As a result, recent research has explored hybrid architectures that combine CNNs and Transformers, such as Swin-Unet [35], MRATUNet [36], and ST-UNet [37]. These approaches merge the local feature extraction capabilities of CNNs with the long-range dependency modeling capabilities of Transformers, while incorporating attention mechanisms and multi-resolution aggregation strategies to enhance the interaction between global and local information in spinal x-ray image segmentation.

2.2. Application of attention mechanisms in spinal x-ray segmentation

In deep learning-driven spinal x-ray image segmentation tasks, attention mechanisms have become a key technique for improving model performance. These mechanisms primarily include channel attention, spatial attention, and hybrid attention. The SE module is one of the earliest channel attention mechanisms proposed, which computes the importance of each channel through global average pooling and adaptively adjusts the weights, thereby enhancing feature representation capabilities [38]. This module has been widely applied in U-Net and its variants to improve the model's responsiveness to key regions. Additionally, CBAM combines channel and spatial attention, allowing the model to focus on important features at different scales [39]. ECA further optimizes the channel attention calculation method, maintaining strong feature selection capability while reducing computational overhead [40]. These methods have achieved good results in spinal x-ray segmentation tasks, particularly in enhancing the model's focus on lesion areas.

To further enhance the flexibility of attention mechanisms, researchers proposed the Self-Calibrated Channel and Spatial Cooperative Attention (SCSA) mechanism [41], which can adaptively adjust attention weights based on different spinal x-ray data, enabling the model to more accurately focus on critical regions. Compared to traditional attention mechanisms, SCSA performs better in reducing false positives and false negatives in segmentation. This paper fully leverages the advantages of SCSA in spinal x-ray segmentation and applies it to the bottleneck module of the network to improve segmentation performance in complex lesion areas.

Overall, although significant progress has been made in current spinal x-ray segmentation research, most studies focus on the analysis of single spinal lesion x-ray images, lacking systematic modeling of multiple complex lesions. In multi-lesion scenarios, the model's adaptability still requires improvement, and segmentation accuracy remains a challenge. These issues continue to be a key area of research.

3. Method

3.1. Overall Network Architecture

This paper proposes the CVM-UNet segmentation model, which integrates multiple advanced mechanisms, as shown in figure 2. The model is based on the U-Net architecture and incorporates several key components to enhance the automatic segmentation capability of multiple spinal diseases in x-ray images. The model first takes spinal x-ray images as input and utilizes an encoder based on ConvNeXt Block to extract multi-scale features. These features are then passed through the Self-Calibrated Channel and Spatial Cooperative Attention (SCSA) mechanism to improve the model's ability to focus on critical regions.

During the decoding process, multiple Up modules are designed for feature upsampling and fusion. Each Up module consists of bilinear interpolation upsampling, VSS Block, and convolution layers, progressively restoring the spatial resolution of feature maps while efficiently integrating skip connection information. Finally, feature channels are adjusted through the Res Block and convolution layers. Additionally, skip connections are processed using the Res-ReLU Block to enhance feature representation capabilities. The model outputs segmentation results that align with the annotations.

3.2. ConvNeXt block

In this study, the ConvNeXt Block is employed as the feature extraction module of the encoder, as shown in figure 3. The ConvNeXt Block is derived from the advanced ConvNeXt architecture, which systematically optimizes traditional convolutional neural networks (CNNs) to enhance both the model's feature extraction capability and computational efficiency.

From a structural design perspective, the ConvNeXt Block offers significant advantages. Its core utilizes a 7×7 large convolution kernel with depthwise separable convolutions, effectively capturing long-range

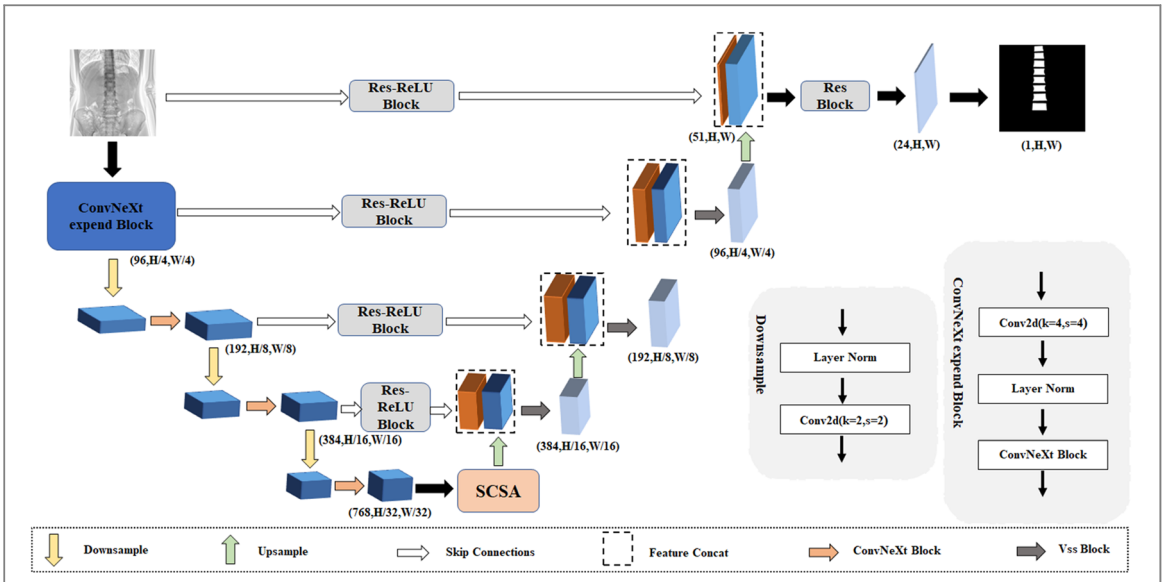

Figure 2. CVM-UNet overall network.

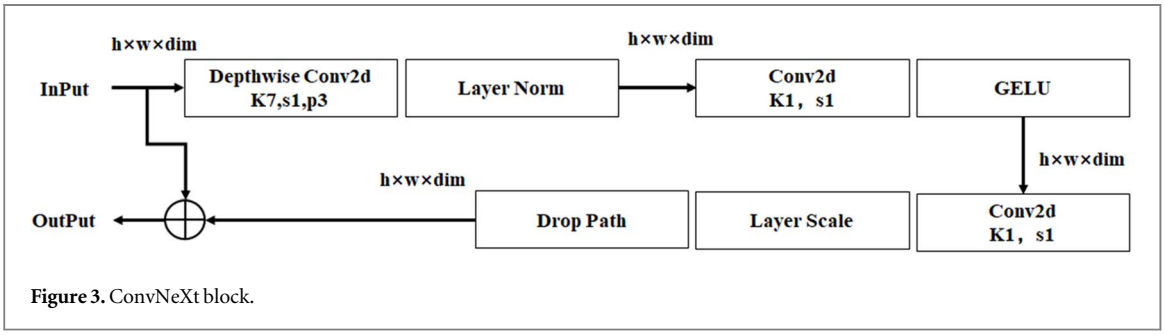

Figure 3. ConvNeXt block.

dependencies and global features. This design expands the receptive field while enhancing both the efficiency and accuracy of feature extraction. Furthermore, the introduction of depthwise separable convolutions significantly reduces the model's parameter count and computational complexity, allowing it to maintain exceptional performance even in resource-constrained environments. Additionally, the incorporation of residual connections strengthens gradient propagation, effectively mitigating both gradient vanishing and exploding issues, thereby improving the stability and trainability of the network. In medical image analysis tasks, residual connections further facilitate the model's ability to learn high-level features, enhancing the accuracy of disease detection.

3.3. VSS Block

This paper introduces an improved strategy to optimize the VSS Block from VMamba by combining it with CNN, making it the primary feature extraction module for the decoder in the overall network. This modification significantly enhances the feature extraction and information utilization capabilities.

In the VSS Block, the input feature map is first processed through layer normalization and then split into two branches. The first branch processes the features using a linear layer followed by the SiLU activation function. The second branch passes sequentially through a linear layer, depthwise separable convolution, SiLU activation function, and then enters the 2D Selective Scanning (SS 2D) module, as shown in figure 4. The SS 2D module, the core component of the VSS Block, consists of Scan Expanding, S6 Block, and Scan Merging. Its unique structure enables multi-directional scanning, effectively capturing rich contextual information.

To further enhance the feature extraction capability, we have explored the potential of the second branch. After passing through the linear layer, depthwise separable convolution, and SiLU activation, the features are fed into the SS 2D module for in-depth analysis. After outputting from the module, the features undergo normalization to stabilize their distribution, ensuring the stability of subsequent processing. The feature dimensions are then adjusted via a linear layer before being input into the ECA (Efficient Channel Attention) module, as shown in figure 5.

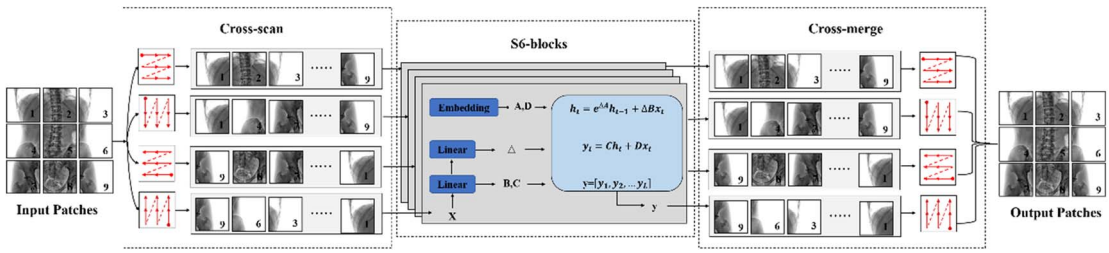

Figure 4. Schematic diagram of two dimensional selective scanning (SS 2D).

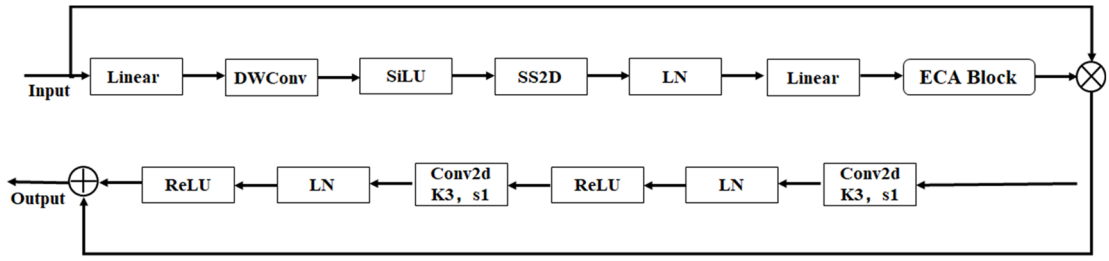

Figure 5. VSS block.

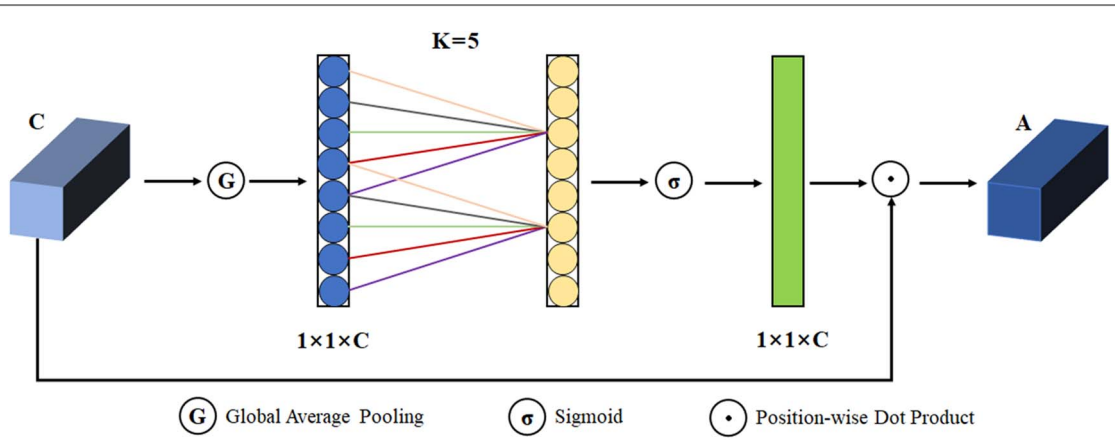

Figure 6. Efficient channel attention.

The ECA module, as an efficient and lightweight channel attention mechanism, adaptively computes the attention weights for each channel to emphasize key features and suppress redundant or noisy information. This optimization improves the model's output and enhances its feature representation capabilities. It first aggregates features using Global Average Pooling (GAP), and then performs a fast 1D convolution with a kernel size of K to generate the channel weights, where K is adaptively determined through a mapping of the channel dimension C [40]. After processing by the ECA, the output A is obtained, which contains rich key information, as shown in figure 6.

Subsequently, the output A is concatenated with the original input features (Concat), allowing the model to retain the original information while integrating high-level features. The concatenated features undergo two convolutional layers, normalization layers, and activation functions to further explore feature relationships, reduce the number of channels, and minimize redundancy. Finally, the output B is obtained and added (Add) to the concatenated output, consolidating features from different stages to enhance feature representation and model performance. As shown in figure 4, the optimized VSS Block improves feature extraction efficiency and information utilization, providing higher-quality feature representations for subsequent tasks.

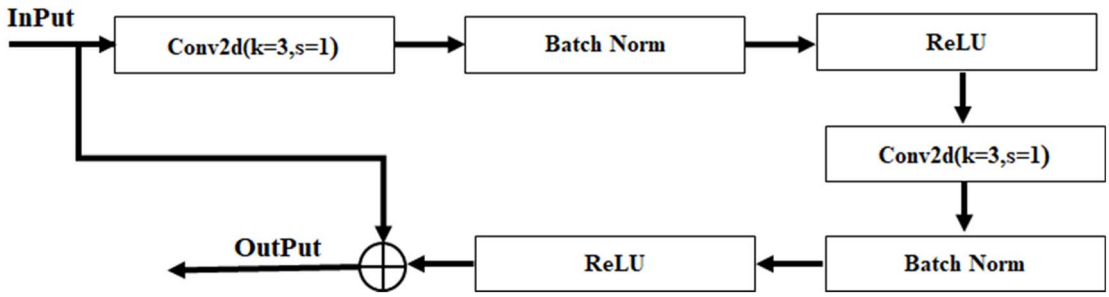

Figure 7. Res-ReLU block.

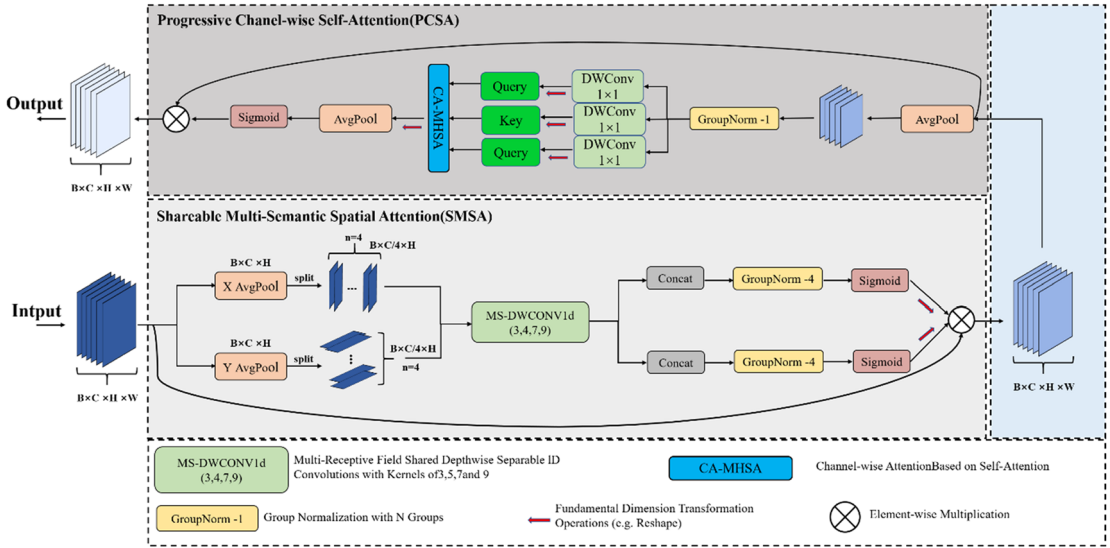

Figure 8. Space and channel collaborative attention (SCSA).

3.4. Res-relu block

The Res-ReLU Block consists of a main path and a skip connection. The main path includes convolutional layers, batch normalization (BN) layers, and ReLU activation functions, while the skip connection bypasses the main path and adds the input to the output of the main path. Unlike traditional residual blocks, this module innovatively uses ReLU as the activation function, thereby altering the network's learning objective to focus on learning the residual mapping. This effectively mitigates the gradient vanishing and explosion problems commonly encountered in deep networks. Furthermore, this design promotes the fusion of features from different layers, enhancing the model's ability to express features and adapt to various spinal diseases, as shown in figure 7.

In the task of spinal x-ray image segmentation, the Res-ReLU Block leverages the skip connection mechanism to fully utilize information across different scales. This approach effectively addresses challenges such as the diverse variety of datasets and low image quality, thereby enhancing the segmentation performance of the model.

3.5. Bottleneck attention mechanism

In the bottleneck section of the model, we introduce the SCSA (Spatial and Channel-wise Collaborative Attention) mechanism. This mechanism consists of Shareable Multi-Semantic Spatial Attention (SMSA) and Progressive Channel-wise Self-Attention (PCSA), as shown in figure 8.

In the SMSA module, the input feature map is first processed by average pooling, and the mean values are computed along the height and width directions, generating two feature maps, x_h and x_w. Next, xhx_hxh is split into local features lhx and three different scales of global features (g^h_{xs} g^h_{xm} g^h_{xl}), while x_w undergoes the same splitting process. Each part has a channel number of group_chans (4/C). Then, the features are processed using four 1D convolutions with shared weights (with kernel sizes of 3, 5, 7, and 9) to extract information at different

scales. The processed features are concatenated and further optimized through a normalization layer. Finally, the spatial attention weights x_{wattn} and x_{hattn} are calculated using a Sigmoid gating mechanism to assess the importance of features along the spatial dimension. This process enhances the expression of critical regions and suppresses background noise.

In the PCSA module, the feature map x ($B\times C\times H\times W$) processed by SMSA is first downsampled via a pooling layer, generating a smaller-dimensional feature map ($B\times C\times H'\times W'$). Next, the downsampled features undergo normalization, and a 1×1 depthwise convolution is applied to generate the query (q), key (k), and value (v), which are used to compute the inter-channel correlations. The attention matrix is computed via the dot product of q and k, with a scaling factor applied to prevent numerical overflow. Subsequently, channel attention weights are calculated by combining the Sigmoid gating mechanism and Dropout regularization, and matrix multiplication is performed with v to obtain the channel-enhanced feature map, thereby strengthening the expression of important information along the channel dimension.

The application of the SCSA module to the bottleneck section of spinal x-ray image processing significantly enhances model performance. In the spatial dimension, SCSA effectively focuses on key regions within the spinal images, enhancing feature representation while suppressing irrelevant background interference. In the channel dimension, SCSA adaptively adjusts the feature responses of each channel, emphasizing key information related to spinal disease diagnosis, thereby improving the model's feature sensitivity and discriminative capability. The incorporation of SCSA not only enables precise capture of important information within spinal x-ray images and effectively filters out irrelevant details, but also significantly enhances feature expression without increasing computational burden, thanks to its lightweight convolution and progressive compression strategy.

3.6. The loss Function

During the training process, to balance model accuracy and convergence speed, a loss function combining BCE loss and Dice loss is employed, with optimization performed using the Adam optimizer. The loss function is defined as follows:

$$loss = \alpha * \mathrm{BCE}\ loss + \beta\ * Dice\ loss$$

In this context, BCE loss refers to the Binary Cross Entropy loss, while Dice loss represents the Dice loss. For this experiment, the parameters are set as $\boldsymbol{\alpha = 1, \beta = 1}$

BCE Loss (Binary Cross Entropy loss) is a widely used loss function in machine learning, particularly for binary classification tasks. Typically, the class labels are either 0 or 1. Given that the true label y can only be 0 or 1, and the predicted value $\hat{y}$ is a probability value between 0 and 1, the formula for BCELoss is as follows:

$$\mathrm{L}(y,\hat{y}) = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

Where $y_i \in \{0,1\}$ represents the ground truth labels (0 for background and 1 for the spinal region), and $\hat{y} \in [0,1]$ denotes the predicted probability by the model. N is the total number of pixels.

Dice Loss is defined based on the Dice Coefficient, with the objective of transforming the Dice Coefficient into an optimizable loss function. Since the Dice Coefficient takes values in the range of [0, 1], with higher values indicating a greater similarity between the predicted results and the ground truth labels, Dice Loss is defined as follows:

$$\mathrm{DiceLoss} = 1 - \mathrm{DiceCoefficient}$$

$$\mathrm{DiceCoefficient} = \frac{2\ |X \cap Y|}{|X| + |Y|}$$

Here, |X| and |Y| represent the number of elements in sets A and B, respectively, while$|X|\cap|Y|$denotes the number of elements in the intersection of sets X and Y.

BCE Loss primarily ensures the classification accuracy of each pixel, particularly in suppressing background regions, while Dice Loss enhances the overall structural matching of the target region, alleviating the issue of class imbalance. These two losses complement each other. In x-ray images, where the contrast between the spine and surrounding tissues is low, Dice Loss improves the recognition ability of the target region through structural matching, promoting continuity, preventing breaks, and reinforcing the overall shape. On the other hand, BCE Loss primarily optimizes the classification of edge pixels. The combination of both significantly improves the quality of boundary segmentation. Additionally, the stable gradients provided by BCE Loss effectively mitigate the initial instability of Dice Loss, thereby accelerating the model's convergence.

3.7. Evaluation metrics

To evaluate the model's performance, we selected Dice Coefficient, mIoU (Mean Intersection over Union), and HD (Hausdorff Distance) as evaluation metrics.

The Dice Coefficient is used to measure the similarity between two sets. In image segmentation, it evaluates the overlap between the predicted segmentation result and the ground truth segmentation, also known as the Sørensen-Dice coefficient. Its calculation formula is as follows:

$$\text{Dice Coefficient} = \frac{2\ |X \cap Y|}{|X| + |Y|}$$

mIoU (Mean Intersection over Union) is an evaluation metric obtained by calculating the Intersection over Union (IoU) for each class and then averaging the values. IoU measures the ratio of the intersection to the union of two sets, which, in the context of image segmentation, typically refers to the pixel sets corresponding to a specific class in the predicted segmentation result and the ground truth label. It reflects the degree of overlap between the two sets. Its calculation formula is as follows:

$$\text{mIoU} = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FP_i + FN_i}$$

Here, True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) are defined.

HD (Hausdorff Distance) is a metric used to measure the distance between two point sets. In image segmentation, it represents the maximum mismatch distance between the predicted segmentation result and the ground truth label, or the maximum mismatch distance from the ground truth label to the predicted segmentation result, taking the larger of the two values. Its calculation formula is as follows:

$$HD(A,\ B) = \max(d_H(A,\ B), d_H(B,\ A))$$

4. Experiments

4.1. Dataset

The spinal x-ray dataset used in this study was provided by our collaborating institution. This dataset comprises 657 anteroposterior (AP) images, which include both normal thoracolumbar vertebrae and images depicting cases of scoliosis. These images are suitable for analyzing the structural features of the spine and potential lesions in the anteroposterior view. Additionally, the dataset contains 517 lateral (LA) images, which cover normal lumbar vertebrae, thoracolumbar wedge deformities, lumbar spondylolisthesis, and lumbar pars defects, making it highly relevant for the recognition and analysis of spinal lesions in lateral view.

The complete dataset, including both AP and LA views, consists of 1174 spinal x-ray images sourced from different patients aged between 18 and 65 years, with a gender ratio of approximately 9:1. The dataset contains 825 positive samples and 349 negative samples, encompassing a wide range of common spinal diseases. All images are stored in the DICOM format, a widely recognized international medical imaging standard. The image widths range from 1223 to 3072 pixels, and the heights range from 2840 to 3072 pixels, ensuring high resolution to support precise analysis of spinal structures and lesion detection.

The annotations for the dataset were performed collaboratively by the AIRobot research team and a specialized medical team from the partner hospital. To ensure the accuracy and consistency of the annotations, the widely accepted ITK-SNAP software [42] was employed. The dataset and its annotation details are shown in figures 9 and 10, with the experimental dataset parameters summarized in table 1.

4.2. Data Preprocessing

In the task of thoracic and lumbar vertebrae recognition and segmentation in spinal image analysis, several preprocessing and augmentation operations were performed. First, during the annotation phase, the thoracolumbar vertebrae were defined as the foreground, with the remaining areas designated as the background. The vertebral labels were then converted into binary images to simplify the structure and highlight the key information.

Subsequently, considering hardware limitations and model requirements, the 1174 annotated image samples were randomly split into a training set and a test set at a ratio of 8:2. To standardize the image dimensions, we adopted a padding strategy that preserved the original aspect ratio. Following this, the padded square images were uniformly scaled to 480×480 pixels, maximizing the preservation of the geometric integrity of the spinal anatomical structures. This approach ensured consistent image dimensions prior to input into the deep learning model, while avoiding any morphological distortion caused by forcing the aspect ratio.

During data preprocessing, we applied Contrast Limited Adaptive Histogram Equalization (CLAHE) with a grid size of 8×8 blocks. Each block underwent independent histogram equalization, and a contrast-limited

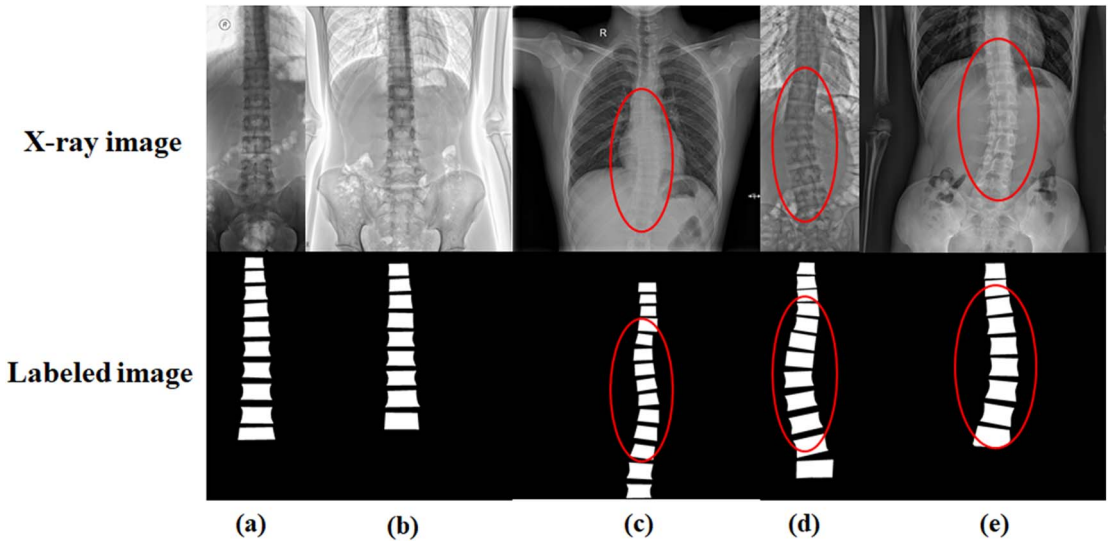

Figure 9. Orthopedic spine x-rays. The red-marked regions indicate lesion areas, and the labels beneath the thumbnails denote their corresponding image indices. Specifically, (a) and (b) show normal spines, while (c), (d), and (e) show cases of scoliosis.

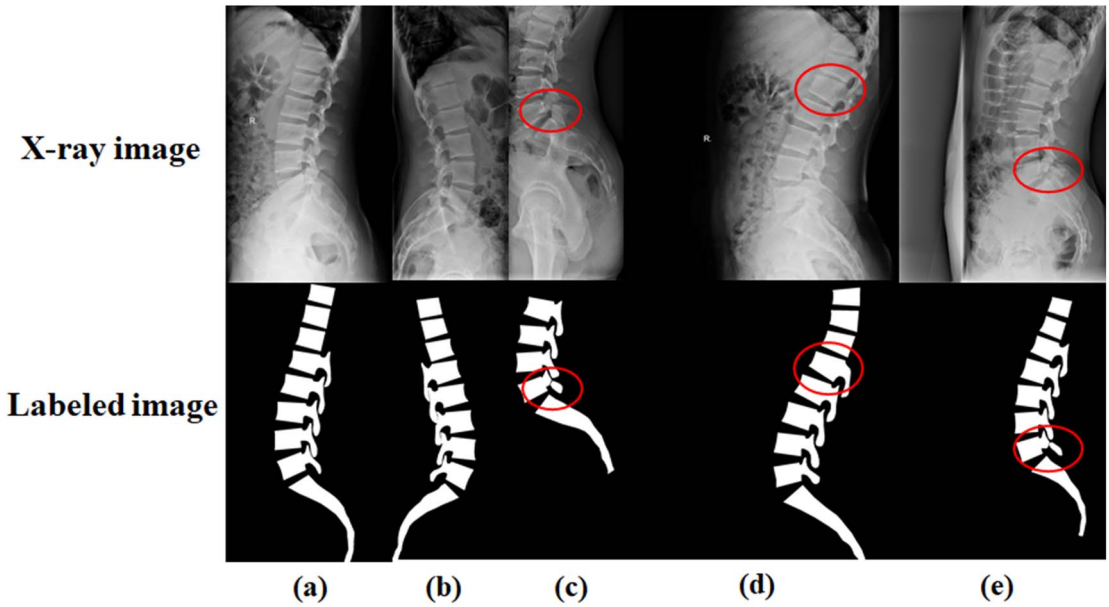

Figure 10. Lateral spine x-ray images. The red-marked regions indicate lesion areas, and the labels beneath the thumbnails denote their corresponding image indices. Specifically, (a) and (b) show normal spines; (c) shows lumbar spondylolysis; (d) shows wedge-shaped vertebral deformation; and (e) shows lumbar spondylolisthesis.

Table 1. Experimental data parameters.

Title	Value
Source	Air Force Characteristic Medical Center
Number of positive samples	657
Number of lateral samples	517
Total number of samples	1174
Resolution (pixels)	2840 × 3072 to 1223 × 3072
Proportion of training and testing samples	8:2

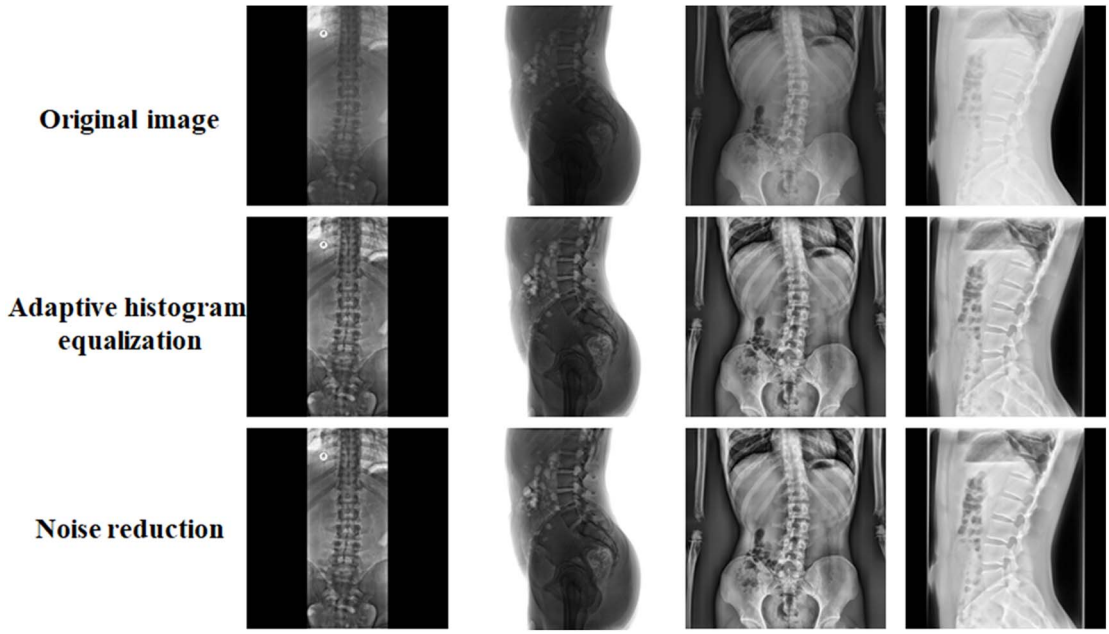

Figure 11. Data preprocessing rendering.

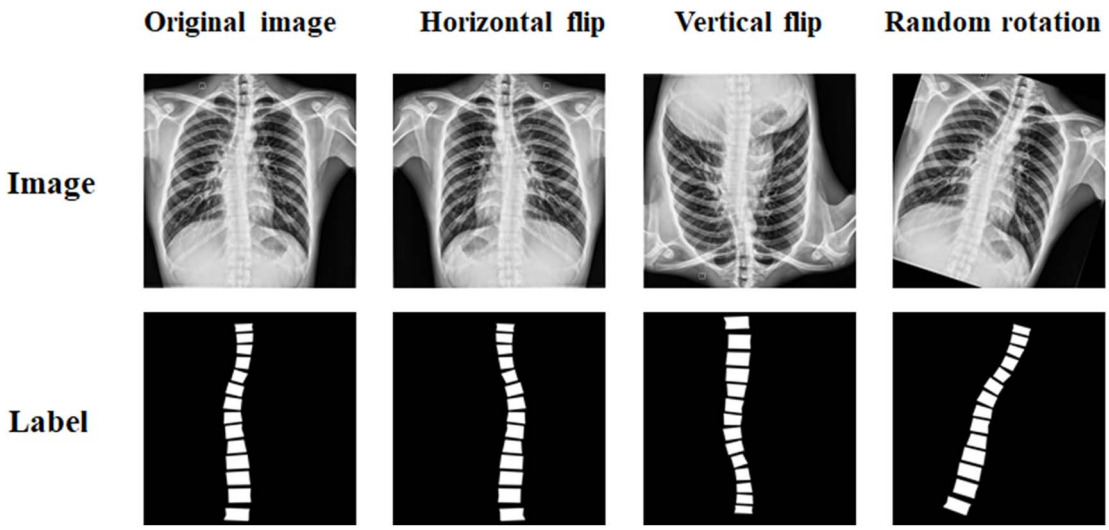

Figure 12. Data augmentation effect image.

threshold of 2.0 was set to prevent over-enhancement of noise. Additionally, the Fast Non-Local Means Denoising algorithm (fastNlMeansDenoising) was used for noise reduction, with a denoising strength parameter h=10, a template window size of 7×7 pixels, and a search window size of 21×21 pixels. This effectively reduced the speckle noise commonly found in medical images, improving the overall image quality and enhancing the contrast between the Region of Interest (ROI) and non-ROI areas, as shown in figure 11.

To enhance the robustness and generalization capability of the model, data augmentation strategies such as vertical flipping, random rotation, and horizontal flipping are employed to increase the diversity of the dataset. This enables the model to learn features from different perspectives, thereby improving its adaptability in real-world applications. The effect of data augmentation is illustrated in figure 12, which lays a solid foundation for subsequent model training and precise recognition.

4.3. Experimental Configuration and Hyperparameter Settings

The experimental workstation is equipped with an AMD Ryzen Threadripper PRO 5945WX CPU, 128 GB of RAM, and an NVIDIA RTX A6000 GPU with 48 GB of VRAM, running on Ubuntu 22.04. The experimental

Table 2. Training configuration parameters.

Hyper-parameters	Hyperparameter configuration
Image_size	480×480
Batch size	4
Initial learning rate	1e-3
Learning rate strategy	Cosine annealing function +WarmUp
epochs	300

environment is based on the PyTorch 2.0.1 framework, utilizing Python 3.10 and CUDA 11.8 for GPU parallel computation acceleration.

For model training, the aforementioned configuration parameters are used, with the following hyperparameter settings: batch size of 4, image size of 480×480, initial learning rate of 1e-3, and cosine annealing for learning rate adjustment. The training process employs a warm-up strategy, with a total of 300 epochs, and the learning rate is dynamically updated at each step to ensure the model converges to the optimal solution, as shown in table 2.

4.4. Comparative experiments

To validate the effectiveness of the proposed CVM-UNet in spinal x-ray segmentation tasks and ensure its clinical applicability, we conduct a comprehensive comparative analysis against mainstream medical segmentation models under fair experimental conditions. The evaluation is based on our in-house spinal x-ray dataset encompassing multiple disease types, using the same preprocessing pipeline, hyperparameter settings, and dataset split strategy as described in section 4.3. The comparison includes a diverse range of representative architectures: Dense-UNet, which leverages dense connections; Residual-UNet, which incorporates residual blocks; UNet++, which introduces nested multi-scale features; and attention-enhanced variants such as Attention-UNet and Channel-UNet [43]. In addition, we benchmark Transformer-based methods like Trans-UNet and Swin-UNet, lightweight designs such as LightM-UNet [44], UNeXt [45], and META-UNet [46], multimodal extensions like MD-UNet [47], and domain-specific architectures including Vm-UNetV2 [48] and ConDSeg [49]. We subsequently analyze the segmentation performance of each model on spinal anatomical structures, highlighting CVM-UNet's superior accuracy, enhanced lesion sensitivity, and significantly improved adaptability across diverse spinal pathologies through both quantitative metrics and qualitative visualization.

First, we perform a performance analysis of CVM-UNet. Given the large scale of the dataset, we evaluate the model every 10 epochs to show the average training loss. As shown in figure 13, the x-axis represents the evaluation group (Group/10 Epoch), and the y-axis represents the average loss value. In the early stages of training, the rapid decrease in loss indicates that CVM-UNet can quickly learn the initial features of the data. Particularly after the 25th group (Epoch 250), the loss curve becomes stable, indicating that the model is approaching convergence. The stability of the loss value further suggests that continuing training may not yield significant improvements. Additionally, the smoothness of the curve reflects that the learning rate and other hyperparameters have been properly adjusted.

Furthermore, in multiple comparative experiments, we analyze the training process over 300 epochs and compute and plot three trend graphs every 30 epochs, representing the three key evaluation metrics: Dice, mIoU, and HD (Hausdorff Distance). These metrics are used to measure the performance of the segmentation models, providing a more intuitive comparison. As shown in figure 14(A), the Dice coefficient reflects the degree of overlap between the model's segmentation result and the ground truth label. The higher the value, the more accurate the model's predictions. From the figure, it can be observed that as the epochs increase, the Dice scores of different segmentation networks generally show an upward trend, with most networks stabilizing within 30 epochs, indicating that they can learn effective features early on, and subsequent improvements are more gradual. In terms of specific model performance, CVM-UNet stands out throughout the entire training process, maintaining the highest Dice score, outperforming classic segmentation networks such as UNet, Dense-UNet, and UNet++. Additionally, Transformer-based networks like Swin-UNet and Trans-UNet perform relatively average in this task. In contrast, more complex models such as VM-UNetV2, META-UNet, and MD-UNet show slower convergence and ultimately achieve scores significantly lower than mainstream U-Net variants.

mIoU is an important evaluation metric in semantic segmentation, used to measure the intersection-over-union between the predicted and ground truth regions. A higher value indicates better segmentation accuracy. As shown in figure 14(B), during the training process, there are significant differences in the performance of different networks on the mIoU metric. However, the mIoU values of most networks gradually improve with the increase in training epochs and stabilize after the 6th group (Epoch 180), indicating model convergence. Compared to Attention-UNet, Channel-UNet, Dense-UNet, and Swin-UNet, CVM-UNet excels, not only

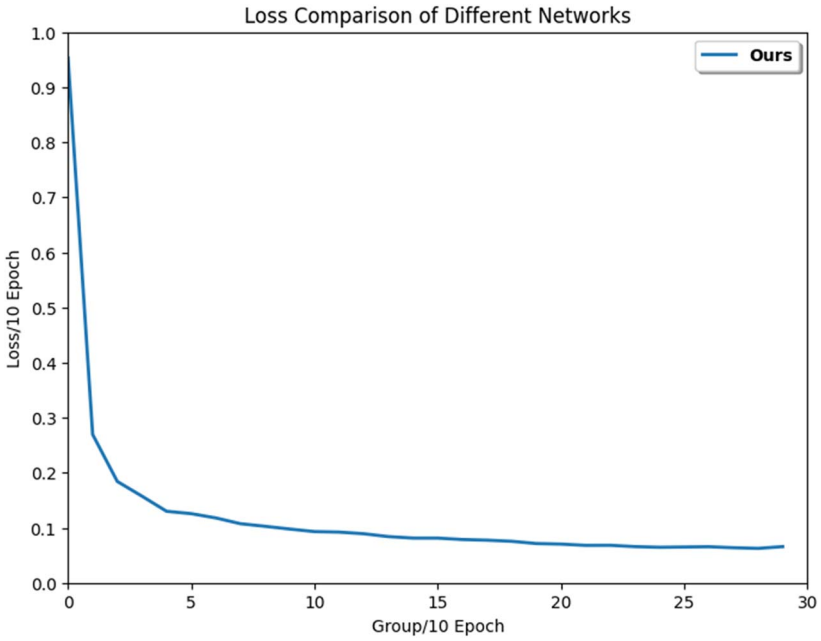

Figure 13. CVM-UNet training loss.

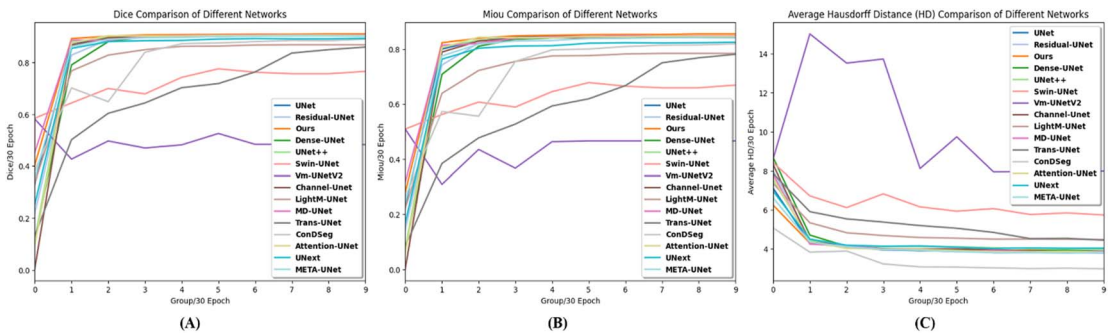

Figure 14. Compare Dice, MIoU, and Aver-HD training change curves of models.

improving the fastest but also achieving the highest value among all networks, demonstrating outstanding segmentation performance.

As shown in table 3, we evaluate the performance of different models using multiple metrics, including mean Intersection over Union (mIoU), averageneralizability Hausdorff Distance (aver_HD), Dice coefficient (Dice), number of parameters, and floating-point operations (FLOPs). The results from the final evaluation indicate that CVM-UNet achieves the best overall performance among all compared models, demonstrating superior results in terms of Dice, mIoU, and HD. Specifically, CVM-UNet attains a Dice coefficient of 91.1% and an mIoU of 85.5%, outperforming the classical UNet, its enhanced variants such as Dense-UNet and UNet++, as well as Transformer-based models including Swin-UNet and Trans-UNet. By integrating the SCSA attention mechanism, CVM-UNet enhances its focus on critical regions, while the VSS Block and Res-ReLU Block further optimize feature representation. These components enable robust fusion of global and local information, which is essential for spinal image segmentation tasks. Although its HD value of 3.852 is close to that of some other models, CVM-UNet maintains an optimal balance among Dice, mIoU, and HD, aligning well with clinical requirements for structural accuracy. This balance indicates that the model not only delivers high-precision segmentation but also effectively reduces boundary errors, thereby preserving the integrity of the predicted structures. From the perspectives of parameter count and FLOPs, CVM-UNet achieves high segmentation accuracy with a reasonable computational load, demonstrating strong feature extraction and adaptability. The comprehensive evaluation across Dice, mIoU, and HD confirms the reliability and accuracy of CVM-UNet in multi-lesion segmentation tasks on spinal x-ray images.

Table 3. Training results of each model.

Model	Years	Dice (%)	mIoU (%)	HD	FLOPs (G)	Params (M)	Infer (ms)
UNet	2015	90.5	84.5	3.799	141.28	17.26	27.61
Dense-UNet	2017	90.4	84.4	3.904	26.31	12.27	42.31
Attention-UNet	2018	90.6	84.6	3.828	234.25	34.88	39.51
Residual-UNet	2018	90.5	84.5	3.816	147.89	18.06	33.62
UNet++	2019	90.5	84.6	3.990	701.87	47.18	57.31
Channel-UNet [37]	2019	90.6	84.6	3.848	381.40	49.15	56.04
MD-UNet [38]	2021	90.6	84.7	3.840	30.15	11.81	57.16
Trans-UNet	2021	85.7	77.7	4.451	115.56	67.86	34.96
Swin-UNet	2022	76.7	66.9	5.727	23.64	27.14	25.28
Unext [39]	2023	89.4	82.5	4.018	2.01	1.47	22.20
META-UNet [40]	2023	90.2	84.0	3.814	15.74	21.70	32.54
LightM-UNet [41]	2024	86.9	78.4	4.500	0.61	0.05	52.23
Vm-UNetV2 [42]	2024	48.3	46.5	7.993	13.47	17.91	54.54
ConDSeg [43]	2025	89.0	81.9	2.980	357.07	45.55	129.39
Our	2025	91.1	85.5	3.852	74.80	52.36	54.33

Table 4. Comparison of accuracy, recall, and boundary IoU performance across different models.

Model	Years	Acc	Recall	Boundary IoU
UNet	2015	0.9913 ± 0.0013	0.9868 ± 0.0167	0.8909 ± 0.0135
Dense-UNet	2017	0.9917 ± 0.0018	0.9869 ± 0.0128	0.8855 ± 0.0136
Attention-UNet	2018	0.9913 ± 0.0022	0.9921 ± 0.0063	0.8904 ± 0.0172
Residual-UNet	2018	0.9914 ± 0.0020	0.9953 ± 0.0046	0.8912 ± 0.0148
UNet++	2019	0.9915 ± 0.0020	0.9941 ± 0.0048	0.8857 ± 0.0180
Channel-UNet	2019	0.9780 ± 0.0055	0.9940 ± 0.0123	0.7471 ± 0.0390
MD-UNet	2021	0.9915 ± 0.0022	0.9925 ± 0.0059	0.8832 ± 0.0216
Trans-UNet	2021	0.9824 ± 0.0154	0.8717 ± 0.1805	0.7778 ± 0.1411
Swin-UNet	2022	0.9448 ± 0.0215	0.4640 ± 0.3215	0.3324 ± 0.2269
Unext	2023	0.9748 ± 0.0106	0.9944 ± 0.0090	0.7319 ± 0.0600
META-UNet	2023	0.9912 ± 0.0029	0.9943 ± 0.0048	0.8823 ± 0.0255
LightM-UNet	2024	0.9864 ± 0.0047	0.9528 ± 0.0248	0.8229 ± 0.0437
Vm-UNetV2	2024	0.9087 ± 0.0465	0.1236 ± 0.1656	0.0813 ± 0.1180
ConDSeg	2025	0.8527 ± 0.0549	0.5238 ± 0.1491	0.1949 ± 0.0605
Ours	2025	0.9919 ± 0.0021	0.9979 ± 0.0030	0.8904 ± 0.0173

To comprehensively evaluate model performance, we use the saved optimal weights for inference and compute additional metrics, including Accuracy, Recall, and Boundary IoU, as shown in table 4. The results indicate that our proposed model achieves a high overall accuracy (0.9919 ± 0.0021) and recall (0.9979 ± 0.0030), while also delivering strong performance on the Boundary IoU metric (0.8904 ± 0.0173), matching or exceeding that of classical models such as Attention-UNet and Residual-UNet. In contrast, some models, including Channel-UNet, Swin-UNet, and Trans-UNet, though exhibiting acceptable Dice and mIoU scores, perform significantly worse in terms of boundary IoU. This suggests a limitation in their ability to capture fine boundary details and small anatomical structures.

To ensure the stability of performance evaluation and the scientific validity and significance of statistical analysis, we conduct 11 evaluations during the converged and stable phase of a single training process (epochs 200–299; see tables 5 and 6), which serves as the primary basis for our statistical analysis. This stage is characterized by stable training behavior and low metric variability, better satisfying the assumptions of independence and stationarity required for valid statistical inference, and thereby offering a more accurate reflection of the model's true performance. To further enhance analytical robustness, we additionally report the mean, standard deviation (SD), and 95% confidence interval (CI) over 30 evaluations spanning epochs 10–299, as well as statistical summaries over all 31 evaluations from epochs 0–299 (appendix tables 10 and 11). For each model, table 5 presents the average Dice, mean Intersection over Union (mIoU), and Hausdorff Distance (HD), together with their corresponding SDs and CIs. Our model achieves a Dice of 0.911 ± 0.001 and an mIoU of 0.855 ± 0.001, both significantly outperforming most baseline models. The narrow CIs further indicate high consistency in segmentation accuracy. Regarding the boundary accuracy metric HD, our model attains a mean of 3.891 ± 0.030 within the core evaluation range. The extended results from appendix table 10 (3.998 ± 0.166)

Table 5. Mean, standard deviation (SD), and 95% confidence interval (CI) of the training results for epochs 200–299.

Model	Mean+ SD			95%CI		
	Dice	mIoU	HD	Dice	mIoU	HD
UNet	0.905 ± 0.000	0.845 ± 0.000	3.803 ± 0.014	[0.905, 0.905]	[0.845, 0.845]	[3.795, 3.810]
Dense-UNet	0.905 ± 0.000	0.845 ± 0.001	3.897 ± 0.020	[0.905, 0.905]	[0.844, 0.845]	[3.887, 3.907]
Attention-UNet	0.906 ± 0.001	0.846 ± 0.001	3.855 ± 0.022	[0.905, 0.906]	[0.845, 0.846]	[3.843, 3.867]
Residual-UNet	0.905 ± 0.000	0.845 ± 0.001	3.819 ± 0.019	[0.905, 0.905]	[0.845, 0.845]	[3.808, 3.830]
UNet++	0.906 ± 0.001	0.846 ± 0.001	4.003 ± 0.014	[0.905, 0.906]	[0.846, 0.846]	[3.995, 4.011]
Channel-UNet	0.905 ± 0.000	0.845 ± 0.001	3.880 ± 0.030	[0.905, 0.906]	[0.845, 0.845]	[3.864, 3.896]
MD-UNet	0.906 ± 0.001	0.847 ± 0.001	3.857 ± 0.015	[0.906, 0.907]	[0.847, 0.848]	[3.849, 3.866]
Trans-UNet	0.851 ± 0.008	0.769 ± 0.011	4.501 ± 0.044	[0.847, 0.855]	[0.763, 0.774]	[4.477, 4.528]
Swin-UNet	0.763 ± 0.007	0.665 ± 0.007	5.759 ± 0.040	[0.758, 0.766]	[0.661, 0.668]	[5.739, 5.782]
Unext	0.893 ± 0.001	0.824 ± 0.001	4.035 ± 0.023	[0.892, 0.894]	[0.823, 0.824]	[4.023, 4.048]
META-UNet	0.902 ± 0.001	0.840 ± 0.001	3.804 ± 0.024	[0.902, 0.903]	[0.840, 0.841]	[3.791, 3.819]
LightM-UNet	0.869 ± 0.001	0.784 ± 0.001	4.493 ± 0.015	[0.869, 0.870]	[0.784, 0.785]	[4.485, 4.502]
Vm-UNetV2	0.484 ± 0.001	0.466 ± 0.001	7.968 ± 0.025	[0.483, 0.484]	[0.466, 0.466]	[7.953, 7.982]
ConDSeg	0.885 ± 0.006	0.814 ± 0.009	3.002 ± 0.032	[0.881, 0.888]	[0.808, 0.817]	[2.988, 3.022]
Ours	0.911 ± 0.001	0.855 ± 0.001	3.891 ± 0.030	[0.910, 0.911]	[0.854, 0.855]	[3.875, 3.908]

exhibit a consistent trend, with no notable increase in SD, thereby validating the long-term stability of the model throughout training.

Moreover, for statistical significance testing, we employ paired *t*-tests for metrics that follow a normal distribution, accompanied by Cohen's *d* to assess effect size and practical relevance. For metrics that do not meet the normality assumption, we apply the Wilcoxon signed-rank test for non-parametric inference, along with Cliff's delta to estimate the magnitude and direction of differences. By integrating both parametric and non-parametric approaches, we ensure that performance improvements are evaluated with both statistical validity and practical significance. All significance tests are corrected using the Benjamini-Hochberg method, and statistically significant results are clearly indicated in the tables (table 6 and appendix table 11), providing robust evidence of the superiority of our method across multiple evaluation metrics. "**" represents the significance level.

To further verify the reliability of CVM-UNet in the segmentation task of various spinal diseases, we integrated the segmentation results of all comparison networks, as shown in figure 15. The red circles highlight areas with segmentation errors. Each column corresponds to a comparison network, and each row represents a type of spinal disease, including normal spine, scoliosis, spondylolysis, wedge deformity, and spondylolisthesis.

Despite the reasonable model parameter design and effective image preprocessing, most models perform well during the training phase, yet their actual segmentation results remain suboptimal. This can primarily be attributed to the complexity and challenges of the spinal segmentation task. Pathological features across different spinal diseases vary significantly, and medical images exhibit considerable variations in quality, resolution, and individual anatomical structures. Furthermore, the grayscale values and texture features of different spinal structures in medical images are often similar, making it difficult for the model to accurately distinguish between them, thereby increasing the segmentation difficulty. This issue is common in the field of medical image processing and may impact actual clinical diagnoses.

The results indicate that CVM-UNet demonstrates significant advantages in the segmentation of multiple spinal diseases. It not only accurately segments lesion regions, reduces errors, and improves diagnostic reliability, but also overcomes the limitations of existing models, which are only capable of handling a single spinal disease. Comparative experimental results show that CVM-UNet exhibits superior performance in the automatic segmentation task of various spinal diseases in x-ray images, providing a reliable technical solution for clinical medical applications.

4.5. Ablation Study

This section further validates the positive contribution of each module in the CVM-UNet model to its overall performance through a series of ablation experiments. We trained the baseline model, ConvNext-UNet, and used it as a comparison model. Subsequently, we performed a comparative analysis of the impact of the SCSA module in the bottleneck, the improved VSS block, and the Res-ReLU block in the skip connections on the model's performance.

To ensure the fairness of the experiments, it is essential to maintain consistency in the upsampling and downsampling channel numbers, upsampling feature extraction modules, and experimental parameter configurations. This consistency guarantees that the quantitative comparison of each module's contribution is convincing. Therefore, in this experiment, the upsampling and downsampling channel numbers,

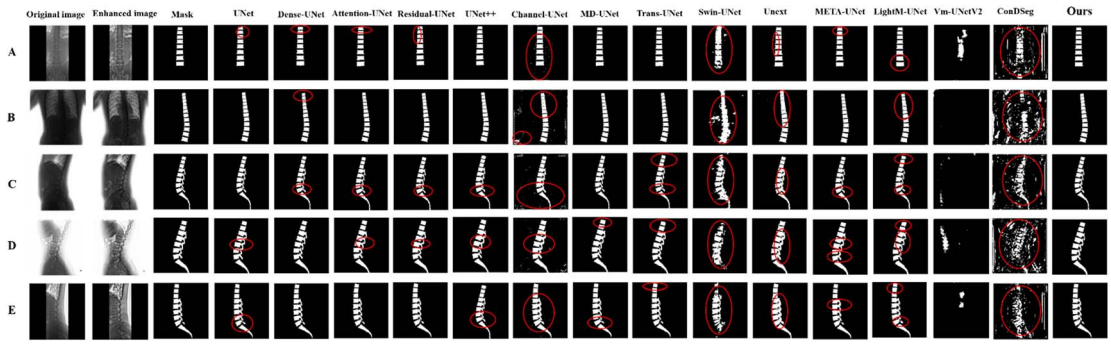

Figure 15. Segmentation results of each network. The red-marked regions indicate incorrectly segmented lesion areas. Specifically, A shows a normal spine; B shows scoliosis of the spine; C shows spondylolysis of L5; D shows a wedge-shaped deformation of the L1 vertebral body; and E shows spondylolisthesis of the L5 vertebral body.

Table 6. Statistical comparison between baseline models and ours (epochs 200–299).

Baseline model	Metric	Effect Size Type	Effect Size	Test	P_adj	Conclusion
UNet	Dice	Cliff's delta	+1.000	Wilcoxon	0.0010^{***}	Ours Better
	mIoU	Cliff's delta	+1.000	Wilcoxon	0.0010^{***}	Ours Better
	HD	Cohen's d	+3.746	t-test	0.0000^{***}	Baseline Better
Dense-UNet	Dice	Cliff's delta	+1.000	Wilcoxon	0.0015^{**}	Ours Better
	mIoU	Cohen's d	+14.419	t-test	0.0000^{***}	Ours Better
	HD	Cohen's d	-0.216	t-test	0.5787	Ours Better
Attention-UNet	Dice	Cliff's delta	+1.000	Wilcoxon	0.0010^{*****}	Ours Better
	mIoU	Cliff's delta	+1.000	Wilcoxon	0.0010^{***}	Ours Better
	HD	Cliff's delta	+0.711	Wilcoxon	0.0010^{***}	Baseline Better
Residual-UNet	Dice	Cliff's delta	+1.000	Wilcoxon	0.0010^{***}	Ours Better
	mIoU	Cliff's delta	+1.000	Wilcoxon	0.0010^{***}	Ours Better
	HD	Cohen's d	+2.866	t-test	0.0000^{***}	Baseline Better
UNet++	Dice	Cliff's delta	+1.000	Wilcoxon	0.0010^{***}	Ours Better
	mIoU	Cohen's d	+11.518	t-test	0.0000^{***}	Ours Better
	HD	Cohen's d	-4.675	t-test	0.0000^{***}	Ours Better
Channel-UNet	Dice	Cliff's delta	+1.000	Wilcoxon	0.0015^{**}	Ours Better
	mIoU	Cohen's d	+12.253	t-test	0.0000^{***}	Ours Better
	HD	Cliff's delta	+0.231	Wilcoxon	0.4648	Baseline Better
MD-UNet	Dice	Cliff's delta	+1.000	Wilcoxon	0.0015^{**}	Ours Better
	mIoU	Cliff's delta	+1.000	Wilcoxon	0.0015^{**}	Ours Better
	HD	Cohen's d	+1.419	t-test	0.0015^{**}	Baseline Better
Trans-UNet	Dice	Cohen's d	+10.585	t-test	0.0000^{***}	Ours Better
	mIoU	Cohen's d	+11.334	t-test	0.0000^{***}	Ours Better
	HD	Cohen's d	-16.219	t-test	0.0000^{***}	Ours Better
Swin-UNet	Dice	Cliff's delta	+1.000	Wilcoxon	0.0010^{***}	Ours Better
	mIoU	Cliff's delta	+1.000	Wilcoxon	0.0010^{***}	Ours Better
	HD	Cohen's d	-52.880	t-test	0.0000^{***}	Ours Better
Unext	Dice	Cohen's d	+22.268	t-test	0.0000^{***}	Ours Better
	mIoU	Cohen's d	+27.433	t-test	0.0000^{***}	Ours Better
	HD	Cohen's d	-5.329	t-test	0.0000^{***}	Ours Better
META-UNet	Dice	Cohen's d	+14.063	t-test	0.0000^{***}	Ours Better
	mIoU	Cliff's delta	+1.000	Wilcoxon	0.0010^{***}	Ours Better
	HD	Cohen's d	+3.179	t-test	0.0000^{***}	Baseline Better
LightM-UNet	Dice	Cliff's delta	+1.000	Wilcoxon	0.0010^{***}	Ours Better
	mIoU	Cliff's delta	+1.000	Wilcoxon	0.0010^{***}	Ours Better
	HD	Cohen's d	-24.961	t-test	0.0000^{***}	Ours Better
Vm-UNetV2	Dice	Cohen's d	+646.266	t-test	0.0000^{***}	Ours Better
	mIoU	Cohen's d	+564.982	t-test	0.0000^{***}	Ours Better
	HD	Cohen's d	-145.680	t-test	0.0000^{***}	Ours Better
ConDSeg	Dice	Cliff's delta	+1.000	Wilcoxon	0.0010^{***}	Ours Better
	mIoU	Cliff's delta	+1.000	Wilcoxon	0.0010^{***}	Ours Better
	HD	Cohen's d	+28.549	t-test	0.0000^{***}	Baseline Better

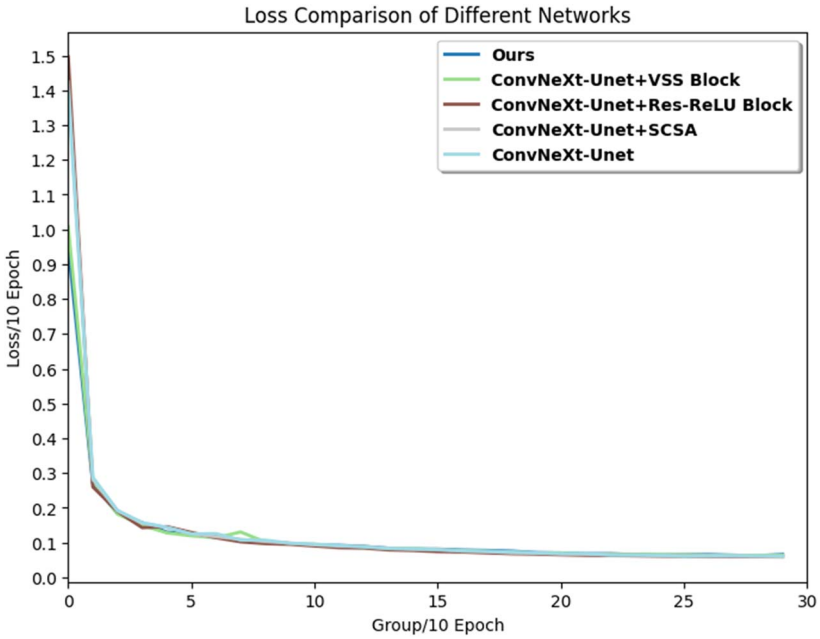

Figure 16. Network loss curve of ablation experiment.

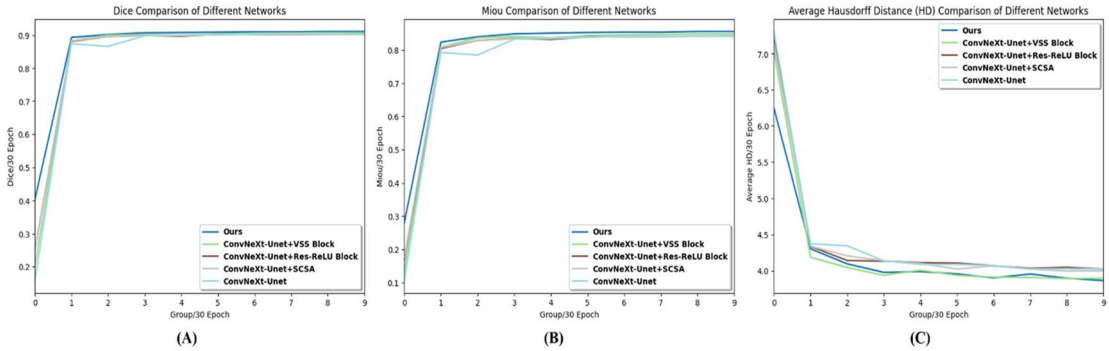

Figure 17. Dice, MIoU, Aver HD training change curve for the ablation experiment model.

downsampling modules, and experimental configurations of ConvNext-UNet are kept consistent with those of CVM-UNet. Additionally, we use widely adopted evaluation metrics in medical image segmentation to objectively assess the accuracy of the algorithms. These metrics include the Dice coefficient, mean Intersection over Union (mIoU), and Hausdorff Distance (HD).

In the ablation study, we continue to use the method of calculating training loss every 10 epochs. After training for 300 epochs, the average training loss is calculated and the loss curve is plotted. As shown in figure 16, the loss of all models (including ConvNext-UNet) decreases rapidly in the early stages of training and then gradually stabilizes. Among them, the CVM-UNet model achieves the lowest loss value in the early stages of training and maintains a relatively low level throughout the training process, indicating its superior performance over ConvNext-UNet. This result highlights the critical role of the VSS Block, SCSA module, and Res-ReLU Block in medical spinal image segmentation tasks.

Figures 17(A)–(C) present the comparative results of the Dice coefficient, mean Intersection over Union (mIoU), and Hausdorff Distance (HD), respectively, evaluated and plotted every 30 epochs. The figures clearly demonstrate the influence of incorporating different modules on the performance metrics across training stages. Specifically, the performance curves of ConvNeXt-UNet integrated with the Res-ReLU Block, SCSA, and VSS Block gradually stabilize and consistently surpass the baseline ConvNeXt-UNet. Notably, the CVM-UNet exhibits the most favorable performance upon convergence across all evaluated metrics.

Table 7 and appendix table 12 report the mean values, standard deviations, and 95% confidence intervals of the evaluation metrics. A consistent improvement in performance is observed with the progressive integration

Table 7. Mean, standard deviation (SD), and 95% confidence interval (CI) of the training results for epochs 200–299.

Model	Mean+ SD			95%CI		
	Dice	mIoU	HD	Dice	mIoU	HD
ConvNext-UNet	0.903 ± 0.000	0.840 ± 0.001	4.010 ± 0.010	[0.903, 0.903]	[0.840, 0.841]	[4.004, 4.015]
ConvNext-UNet+ Res-ReLU Block	0.903 ± 0.000	0.841 ± 0.001	4.032 ± 0.013	[0.903, 0.903]	[0.841, 0.842]	[4.025, 4.039]
ConvNext-UNet +SCSA	0.904 ± 0.001	0.842 ± 0.000	3.993 ± 0.015	[0.903, 0.904]	[0.842, 0.842]	[3.986, 4.002]
ConvNext-UNet +VSS Block	0.907 ± 0.000	0.848 ± 0.000	3.880 ± 0.012	[0.907, 0.907]	[0.847, 0.848]	[3.873, 3.886]
Ours	0.911 ± 0.001	0.855 ± 0.001	3.891 ± 0.030	[0.910, 0.911]	[0.854, 0.855]	[3.875, 3.908]

Table 8. Comparison of accuracy, recall, and boundary IoU.

Model	Acc	Recall	Boundary IoU
ConvNext-UNet	0.9910 ± 0.0024	0.9942 ± 0.0034	0.8891 ± 0.0189
ConvNext-UNet+ Res-ReLU Block	0.9911 ± 0.0030	0.9941 ± 0.0017	0.8812 ± 0.0196
ConvNext-UNet +SCSA	0.9909 ± 0.0028	0.9953 ± 0.0036	0.8883 ± 0.0226
ConvNext-UNet +VSS Block	0.9917 ± 0.0026	0.9949 ± 0.0039	0.8871 ± 0.0232
Ours	0.9919 ± 0.0021	0.9979 ± 0.0030	0.8904 ± 0.0173

Table 9. Statistical comparison between baseline models and conv next-unet (epochs 200–299).

Baseline model	Metric	Effect Size Type	Effect Size	Test	p_adj	Conclusion
ConvNext-UNet+ Res-ReLU Block	Dice	Cliff's delta	+0.091	Wilcoxon	1.0000	ConvNext-Unet Better
	mIoU	Cohen's d	−1.252	t-test	0.0243^{*}	Baseline Better
	HD	Cohen's d	−1.886	t-test	0.0009^{***}	ConvNext-Unet Better
ConvNext-UNet +SCSA	Dice	Cliff's delta	−0.587	Wilcoxon	0.0156^{*}	Baseline Better
	mIoU	Cliff's delta	−0.975	Wilcoxon	0.0029^{**}	Baseline Better
	HD	Cohen's d	+1.350	t-test	0.0136^{*}	Baseline Better
ConvNext-UNet +VSS Block	Dice	Cliff's delta	−1.000	Wilcoxon	0.0010^{***}	Baseline Better
	mIoU	Cliff's delta	−1.000	Wilcoxon	0.0010^{***}	Baseline Better
	HD	Cohen's d	+11.561	t-test	0.0000^{***}	Baseline Better
Ours	Dice	Cliff's delta	−1.000	Wilcoxon	0.0010^{***}	Baseline Better
	mIoU	Cliff's delta	−1.000	Wilcoxon	0.0010^{***}	Baseline Better
	HD	Cohen's d	+5.225	t-test	0.0000^{***}	Baseline Better

of each module. In particular, the addition of the VSS Block leads to a marked enhancement, with the mean Dice and mIoU increasing to 0.907 and 0.848, respectively, while the HD significantly decreases to 3.880. These results indicate that the introduced modules substantially enhance model performance, further validating the critical roles of the Res-ReLU Block, SCSA, and VSS Block within the CVM-UNet architecture.

Furthermore, we evaluate the saved optimal weights of each model on additional metrics, including Accuracy, Recall, and Boundary IoU, as presented in table 8. All models achieve strong results in terms of accuracy and recall. Of particular note, Boundary IoU—which reflects the precision of lesion boundary delineation—shows that our model performs exceptionally well (0.8904 ± 0.0173), further confirming its superior segmentation capability. These findings highlight the essential contributions of the proposed modules in enhancing the model's overall segmentation performance.

In addition, we apply rigorous statistical significance testing to determine whether these module combinations lead to meaningful performance improvements. Specifically, we adopt the non-parametric Wilcoxon signed-rank test (with Cliff's delta for effect size) and the parametric paired t-test (with Cohen's d). The ablation variants consistently exhibit statistically significant gains (Wilcoxon test $p < 0.01$ or t-test $p < 0.001$), with generally large effect sizes (Cliff's delta/Cohen's d), underscoring the practical impact of the proposed modules. Detailed statistical results are presented in table 9 and appendix table 13, clearly demonstrating both the effectiveness and statistical significance of each module. ″**″ represents the significance level.

To further substantiate the theoretical analysis, we performed a visual comparison of segmentation results across different models, where the red circles highlight regions of incorrect segmentation. As illustrated in figure 18, substantial differences are observed in the quality of the predicted masks. Cvm-UNet consistently achieves superior performance across various spinal disease segmentation tasks, with its predicted masks accurately aligning with vertebral structures. These findings collectively demonstrate the outstanding effectiveness of Cvm-UNet and further validate its applicability and precision in multi-lesion spinal segmentation.

The ablation study confirms the applicability and accuracy of Cvm-UNet across a range of disease segmentation tasks in the medical domain, while also validating the positive contribution of each integrated module to the model's overall performance. Moreover, the qualitative segmentation results further demonstrate the practical effectiveness of Cvm-UNet in clinical applications.

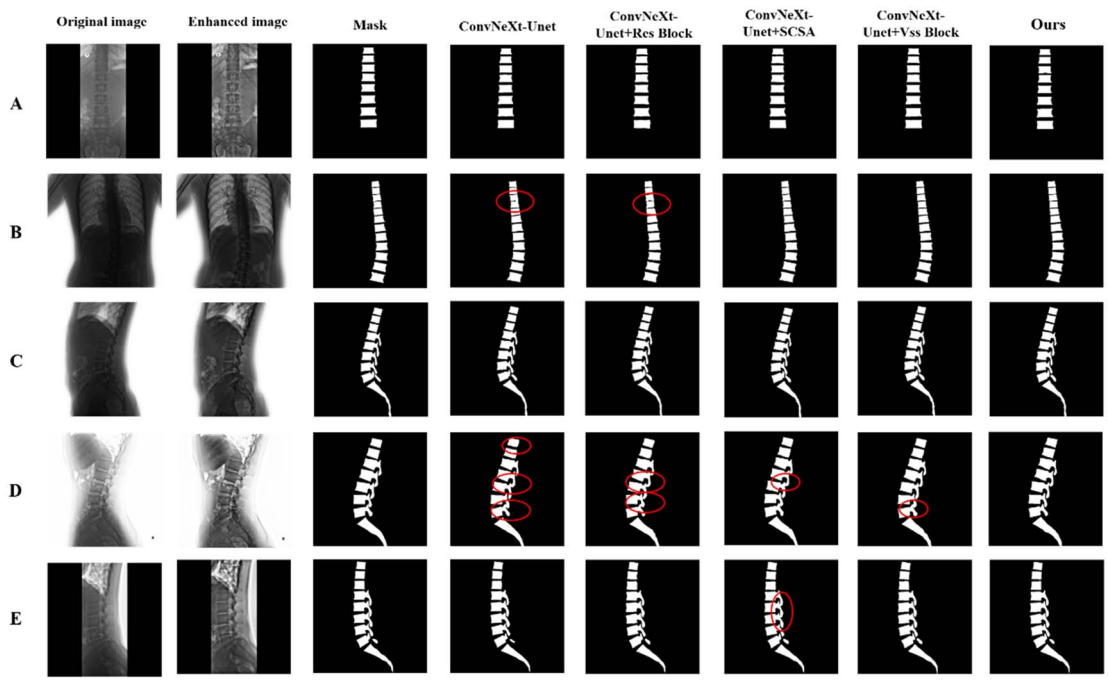

Figure 18. Segmentation results of each model. The red-marked regions indicate incorrectly segmented lesion areas. Specifically, A shows a normal spine; B shows scoliosis of the spine; C shows spondylolysis of L5; D shows a wedge-shaped deformation of the L1 vertebral body; and E shows spondylolisthesis of the L5 vertebral body.

5. Discussion

5.1. Limitations and challenges

Despite the demonstrated superiority of CVM-UNet in multi-lesion segmentation of spinal x-ray images, this study has several limitations. First, the experimental data are sourced from a single medical institution, and although the dataset includes common spinal pathologies such as scoliosis and spondylolisthesis, it suffers from imbalances in patient demographics (e.g., age and sex) and lacks diversity across external datasets. These limitations may restrict the model's generalizability to more complex cases or datasets from different clinical environments.

Second, while CVM-UNet integrates ConvNeXt, VSS Block, and the SCSA module to enhance segmentation accuracy, its relatively large parameter size (52.36M) and computational demand (74.80G FLOPs) may pose challenges for deployment in resource-constrained clinical settings, such as mobile devices or real-time diagnostic systems.

Third, ground truth annotations rely on manual delineation by experienced radiologists. Although annotation variability is mitigated through the use of ITK-SNAP software and multi-expert verification, subtle inconsistencies in lesion boundary interpretation may still impact the robustness of model training.

Lastly, the current statistical significance analysis is based on high-frequency evaluations from a single training cycle. Future work should incorporate multiple independent training runs to further validate the model's stability and reproducibility under varied initialization and training conditions.

5.2. Data and code availability

Dataset: Due to patient privacy considerations and institutional collaboration agreements, the original spinal x-ray images used in this study are not fully publicly available. Researchers interested in accessing de-identified subsets of the data for non-commercial academic research may submit a formal request to the authors. Access will be granted following institutional ethical review and approval.

Code: The core algorithm (CVM-UNet) is subject to intellectual property agreements and confidentiality clauses with partnering institutions; therefore, the complete source code is not publicly released at this time. However, to facilitate academic collaboration, non-sensitive modules of the implementation may be made available upon request by contacting the authors.

5.3. Future directions

In light of the aforementioned limitations and the clinical translation potential of CVM-UNet, future research will focus on the following directions to enhance the model's practicality and generalizability:

1. Multi-center Collaboration and Data Augmentation: We aim to collaborate with medical institutions across different regions to construct a large-scale and diverse spinal imaging dataset that covers various demographic characteristics (e.g., age, gender, ethnicity), and incorporates external datasets. Furthermore, we plan to explore synthetic data generation techniques based on Generative Adversarial Networks (GANs) to mitigate data distribution bias. This will enable a more rigorous assessment of model generalizability on external datasets and complex clinical cases, providing a comprehensive data foundation for training large-scale, general-purpose segmentation models.
2. Model Compression and Efficient Deployment: To facilitate real-time clinical applications in resource-constrained settings, future work will focus on compressing the model using techniques such as knowledge distillation, dynamic pruning, and neural architecture search (NAS). Inference speed and efficiency will be further optimized using hardware acceleration tools like TensorRT, enabling deployment on mobile devices and edge-computing platforms.
3. Weak Supervision and Annotation Efficiency: To alleviate the burden of exhaustive pixel-wise annotation, weakly supervised learning frameworks based on bounding boxes, point annotations, or image-level labels will be developed. These approaches will be combined with semi-supervised learning strategies to leverage unlabeled data and enhance model robustness, addressing both annotation cost and inter-observer variability.
4. Multimodal Integration: Future extensions will focus on adapting the model to process multimodal imaging data such as CT and MRI, facilitating cross-modality spinal lesion analysis and enabling integrated diagnostic and severity grading systems.
5. Clinical Validation and Interpretability: Large-scale, multi-center clinical validation will be conducted to evaluate the real-world performance of the model. Furthermore, interpretability will be enhanced through techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM) and feature visualization, providing insights into the model's decision-making process and fostering greater clinician trust in AI-assisted outcomes. These efforts aim to promote the seamless integration of CVM-UNet into routine clinical workflows.

6. Conclusions

To address the limited applicability and suboptimal accuracy of existing models in multi-lesion segmentation of spinal x-ray images, we propose CVM-UNet, a novel network tailored for automatic segmentation of diverse spinal pathologies. CVM-UNet is designed to overcome the constraint of prior models that focus primarily on single-lesion segmentation, thereby enhancing the precision and diagnostic reliability of spinal x-ray image analysis. The network incorporates ConvNeXt Blocks for efficient feature extraction, integrates a spatial-channel cooperative attention (SCSA) mechanism to strengthen regional awareness, and employs an enhanced VSS Block as the decoder to progressively reconstruct feature maps. Additionally, the Res-ReLU Block is utilized in skip connections to facilitate effective integration of global and local information. These architectural innovations significantly improve feature representation, fusion, and collaborative perception, offering a robust technical reference for generalizable multi-lesion spinal segmentation models.

Comprehensive evaluations, including ablation studies and comparisons with state-of-the-art (SOTA) methods, demonstrate that CVM-UNet achieves superior performance across key metrics such as Dice (91.1), mean IoU (85.5), Hausdorff Distance (3.852), accuracy, recall, and boundary IoU. The model not only ensures high segmentation accuracy but also exhibits strong stability and adaptability across various spinal disorders. Beyond average performance gains, we further validate the model's reliability through statistical analysis. By introducing standard deviation, 95% confidence intervals, and significance testing, we confirm that the observed improvements are both consistent and statistically meaningful. Comparative experiments against 16 representative baseline models using t-tests and Wilcoxon tests, along with effect size analysis, reveal significant advantages in most metrics. These findings remain consistent when extended to a broader training range (epochs 10–299), underscoring the robustness and generalizability of the proposed approach across different learning stages.

Furthermore, we construct a comprehensive spinal x-ray dataset covering a wide spectrum of spinal conditions—including normal thoracolumbar structures, spondylolysis, vertebral wedge deformity,

spondylolisthesis, and scoliosis—providing a large volume of annotated medical images for deep learning training and evaluation. This dataset not only supports our current research but also lays a solid foundation for future studies.

In summary, the innovative architectural design of CVM-UNet, together with the construction of a comprehensive multi-lesion spinal x-ray dataset, advances the development of automated multi-pathology segmentation in spinal imaging and effectively addresses the limitations of existing models in handling diverse spinal disorders. Clinically, the model's capacity for simultaneous multi-lesion segmentation demonstrates substantial potential in scenarios such as degenerative spine disease screening and rapid trauma assessment. Looking ahead, future work will focus on lightweight architecture exploration, model and evaluation optimization, and external dataset expansion, with the aim of integrating CVM-UNet into clinical workflows for broader real-world application. We believe this study offers a meaningful contribution to the field of medical image segmentation and will accelerate the translation of intelligent image analysis technologies into practical healthcare settings.

Data availability statement

The data cannot be made publicly available upon publication because they contain sensitive personal information. The data that support the findings of this study are available upon reasonable request from the authors.

Appendix

Table 10. Mean, standard deviation (SD), and 95% confidence interval (CI) of the evaluation metrics over epochs 10–299.

		Mean+ SD		95%CI		
Model	Dice	mIoU	HD	Dice	mIoU	HD
UNet	0.890 ± 0.043	0.825 ± 0.055	4.012 ± 0.340	[0.873, 0.902]	[0.802, 0.841]	[3.909, 4.136]
Dense-UNet	0.883 ± 0.055	0.817 ± 0.067	4.070 ± 0.366	[0.862, 0.900]	[0.790, 0.836]	[3.958, 4.213]
Attention-UNet	0.895 ± 0.026	0.831 ± 0.036	4.011 ± 0.247	[0.886, 0.903]	[0.818, 0.842]	[3.931, 4.104]
Residual-UNet	0.889 ± 0.038	0.824 ± 0.047	4.003 ± 0.298	[0.874, 0.899]	[0.803, 0.838]	[3.914, 4.110]
UNet++	0.897 ± 0.021	0.833 ± 0.029	4.131 ± 0.186	[0.889, 0.903]	[0.820, 0.842]	[4.072, 4.200]
Channel-UNet	0.892 ± 0.032	0.825 ± 0.046	4.108 ± 0.369	[0.879, 0.901]	[0.807, 0.838]	[3.995, 4.253]
MD-UNet	0.900 ± 0.015	0.837 ± 0.024	3.967 ± 0.187	[0.894, 0.905]	[0.827, 0.844]	[3.911, 4.039]
Trans-UNet	0.714 ± 0.149	0.614 ± 0.160	5.094 ± 0.643	[0.663, 0.765]	[0.558, 0.670]	[4.887, 5.325]
Swin-UNet	0.735 ± 0.043	0.639 ± 0.037	6.135 ± 0.517	[0.720, 0.749]	[0.626, 0.652]	[5.969, 6.322]
Unext	0.880 ± 0.028	0.805 ± 0.039	4.172 ± 0.221	[0.871, 0.888]	[0.789, 0.816]	[4.104, 4.259]
META-UNet	0.887 ± 0.033	0.817 ± 0.049	3.989 ± 0.304	[0.873, 0.897]	[0.798, 0.833]	[3.897, 4.110]
LightM-UNet	0.847 ± 0.039	0.752 ± 0.055	4.680 ± 0.292	[0.833, 0.859]	[0.731, 0.768]	[4.588, 4.783]
Vm-UNetV2	0.464 ± 0.077	0.415 ± 0.090	10.757 ± 3.595	[0.432, 0.483]	[0.381, 0.441]	[9.521, 12.031]
ConDSeg	0.833 ± 0.090	0.748 ± 0.108	3.247 ± 0.376	[0.799, 0.862]	[0.707, 0.784]	[3.126, 3.394]
Ours	0.905 ± 0.014	0.844 ± 0.022	3.998 ± 0.166	[0.899, 0.908]	[0.835, 0.850]	[3.948, 4.059]

Table 11. Statistical comparison between baseline models and ours (epochs 0–299).

Baseline model	Metric	Effect size type	Effect size	Test	p_adj	Conclusion
UNet	Dice	Cliff's delta	+0.597	Wilcoxon	0.0000***	Ours Better
	mIoU	Cliff's delta	+0.583	Wilcoxon	0.0000***	Ours Better
	HD	Cliff's delta	+0.248	Wilcoxon	0.4048	Baseline Better
Dense-UNet	Dice	Cliff's delta	+0.618	Wilcoxon	0.0000***	Ours Better
	mIoU	Cliff's delta	+0.606	Wilcoxon	0.0000***	Ours Better
	HD	Cliff's delta	+0.022	Wilcoxon	0.1525	Baseline Better
Attention-UNet	Dice	Cliff's delta	+0.563	Wilcoxon	0.0000***	Ours Better
	mIoU	Cliff's delta	+0.548	Wilcoxon	0.0000***	Ours Better
	HD	Cliff's delta	+0.137	Wilcoxon	0.4684	Baseline Better
Residual-UNet	Dice	Cliff's delta	+0.623	Wilcoxon	0.0000***	Ours Better
	mIoU	Cliff's delta	+0.609	Wilcoxon	0.0000***	Ours Better
	HD	Cliff's delta	+0.247	Wilcoxon	0.4160	Baseline Better
UNet++	Dice	Cliff's delta	+0.592	Wilcoxon	0.0000***	Ours Better
	mIoU	Cliff's delta	+0.575	Wilcoxon	0.0000***	Ours Better
	HD	Cliff's delta	−0.610	Wilcoxon	0.0000***	Ours Better
Channel-UNet	Dice	Cliff's delta	+0.605	Wilcoxon	0.0000***	Ours Better
	mIoU	Cliff's delta	+0.595	Wilcoxon	0.0000***	Ours Better
	HD	Cliff's delta	−0.119	Wilcoxon	0.0007***	Ours Better
MD-UNet	Dice	Cliff's delta	+0.492	Wilcoxon	0.0000***	Ours Better
	mIoU	Cliff's delta	+0.495	Wilcoxon	0.0000***	Ours Better
	HD	Cliff's delta	+0.265	Wilcoxon	0.0064**	Baseline Better
Trans-UNet	Dice	Cliff's delta	+0.923	Wilcoxon	0.0000***	Ours Better
	mIoU	Cliff's delta	+0.919	Wilcoxon	0.0000***	Ours Better
	HD	Cliff's delta	−0.919	Wilcoxon	0.0000***	Ours Better
Swin-UNet	Dice	Cliff's delta	+0.935	Wilcoxon	0.0000***	Ours Better
	mIoU	Cliff's delta	+0.935	Wilcoxon	0.0000***	Ours Better
	HD	Cliff's delta	−0.954	Wilcoxon	0.0000***	Ours Better
Unext	Dice	Cliff's delta	+0.819	Wilcoxon	0.0000***	Ours Better
	mIoU	Cliff's delta	+0.816	Wilcoxon	0.0000***	Ours Better
	HD	Cliff's delta	−0.659	Wilcoxon	0.0000***	Ours Better
META-UNet	Dice	Cliff's delta	+0.689	Wilcoxon	0.0000***	Ours Better
	mIoU	Cliff's delta	+0.681	Wilcoxon	0.0000***	Ours Better
	HD	Cliff's delta	+0.315	Wilcoxon	0.0687	Baseline Better
LightM-UNet	Dice	Cliff's delta	+0.890	Wilcoxon	0.0000***	Ours Better
	mIoU	Cliff's delta	+0.890	Wilcoxon	0.0000***	Ours Better
	HD	Cliff's delta	−0.896	Wilcoxon	0.0000***	Ours Better
Vm-UNetV2	Dice	Cliff's delta	+0.938	Wilcoxon	0.0000***	Ours Better
	mIoU	Cliff's delta	+0.938	Wilcoxon	0.0000***	Ours Better
	HD	Cliff's delta	−1.000	Wilcoxon	0.0000***	Ours Better
ConDSeg	Dice	Cliff's delta	+0.873	Wilcoxon	0.0000***	Ours Better
	mIoU	Cliff's delta	+0.865	Wilcoxon	0.0000***	Ours Better
	HD	Cliff's delta	+0.810	Wilcoxon	0.0000***	Baseline Better

Table 12. Ablation study: mean, standard deviation (SD), and 95% confidence interval (CI) of the evaluation metrics over epochs 10–299.

	Mean+ SD			95%CI		
Model	Dice	mIoU	HD	Dice	mIoU	HD
ConvNext-UNet	0.892 ± 0.022	0.824 ± 0.034	4.137 ± 0.187	[0.884, 0.899]	[0.809, 0.835]	[4.076, 4.205]
ConvNext-UNet+ Res-ReLU Block	0.896 ± 0.015	0.830 ± 0.024	4.120 ± 0.140	[0.891, 0.901]	[0.821, 0.837]	[4.077, 4.177]
ConvNext-UNet +SCSA	0.895 ± 0.020	0.828 ± 0.032	4.112 ± 0.193	[0.886, 0.901]	[0.816, 0.837]	[4.057, 4.187]
ConvNext-UNet +VSS Block	0.899 ± 0.020	0.835 ± 0.031	3.983 ± 0.196	[0.891, 0.904]	[0.822, 0.843]	[3.929, 4.057]
Ours	0.905 ± 0.014	0.844 ± 0.022	3.998 ± 0.166	[0.899, 0.908]	[0.835, 0.850]	[3.948, 4.059]

Table 13. Ablation study: statistical comparison between baseline models and convnext-unet (epochs 0–299).

Baseline model	Metric	Effect size type	Effect size	Test	P_adj	Conclusion
ConvNext-UNet+ Res-ReLU Block	Dice	Cliff's delta	−0.036	Wilcoxon	0.0098**	Baseline Better
	mIoU	Cliff's delta	−0.161	Wilcoxon	0.0015**	Baseline Better
	HD	Cliff's delta	−0.068	Wilcoxon	0.4908	ConvNext-Unet Better
ConvNext-UNet +SCSA	Dice	Cliff's delta	−0.176	Wilcoxon	0.0017**	Baseline Better
	mIoU	Cliff's delta	−0.247	Wilcoxon	0.0009***	Baseline Better
	HD	Cliff's delta	+0.165	Wilcoxon	0.0017**	Baseline Better
ConvNext-UNet +VSS Block	Dice	Cliff's delta	−0.464	Wilcoxon	0.0000***	Baseline Better
	mIoU	Cliff's delta	−0.498	Wilcoxon	0.0000***	Baseline Better
	HD	Cliff's delta	+0.652	Wilcoxon	0.0000***	Baseline Better
Ours	Dice	Cliff's delta	−0.667	Wilcoxon	0.0000***	Baseline Better
	mIoU	Cliff's delta	−0.667	Wilcoxon	0.0000***	Baseline Better
	HD	Cliff's delta	+0.612	Wilcoxon	0.0000***	Baseline Better

ORCID iDs

Zhilong Xue https://orcid.org/0009-0006-0284-5463
Shuangcheng Deng https://orcid.org/0000-0002-7311-1401
Zhiwu Li https://orcid.org/0009-0005-7810-4082

References

[1] Pereira, Martins J F, Mari J F and Silva L H F P 2025 Exploiting data augmentation strategies to improve the classification of spinal disorders in x-ray images *Revista de Informática Teórica e Aplicada* **32** 257–64
[2] Trinh G M *et al* 2022 Detection of lumbar spondylolisthesis from x-ray images using deep learning network *Journal of Clinical Medicine* **11** 5450
[3] Fraiwan M *et al* 2022 Using deep transfer learning to detect scoliosis and spondylolisthesis from x-ray images *PLoS One* **17** e0267851
[4] Thangaleela S *et al* 2025 Nanoparticle-based imaging techniques in neurological disorders *Nanoparticles in Modern Neurological Treatment*. (Springer Nature) 43–107
[5] Hosseini H *et al* 2025 Bone tumors: a systematic review of prevalence, risk determinants, and survival patterns *BMC cancer* **25** 1–11
[6] Bezabh Y *et al* 2024 Classification of cervical spine disease using convolutional neural network *Multimedia Tools Appl.* 1–17
[7] Chen Y *et al* 2024 VertXNet: an ensemble method for vertebral body segmentation and identification from cervical and lumbar spinal x-rays *Sci. Rep.* **14** 3341
[8] Pal C K and Kumar S Investigating common sources and types of errors in radiological image interpretations and reportings *Ultrasound* **20** 20
[9] Zhang *et al* 2023 Diagnostic error and bias in the department of radiology: a pictorial essay *Insights into Imaging* **14** 163
[10] Qu B *et al* 2022 Current development and prospects of deep learning in spine image analysis: a literature review *Quantitative Imaging in Medicine and Surgery* **12** 3454
[11] Khalid H *et al* 2020 A comparative systematic literature review on knee bone reports from mri, x-rays and ct scans using deep learning and machine learning methodologies *Diagnostics* **10** 518
[12] Azimi P *et al* 2020 A review on the use of artificial intelligence in spinal diseases *Asian Spine Journal* **14** 543
[13] Simons S J and Papież B W 2024 SpineFM: leveraging foundation models for automatic spine x-ray segmentation *preprint* arXiv:2411.00326
[14] Sagar A S M S *et al* 2025 Uncertainty-aware adaptive multiscale u-net for low-contrast cardiac image segmentation *Appl. Sci.* **15** 2222
[15] Ronneberger O, Fischer P and Brox T 2015 U-net: Convolutional networks for biomedical image segmentation *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference Proceedings, Part III 18* 2015 *(Munich, Germany, October 5-9, 2015)* (Springer International Publishing) 234–41
[16] Zhou Z *et al* 2019 Unet++: Redesigning skip connections to exploit multiscale features in image segmentation *IEEE Trans. Med. Imaging* **39** 1856–67
[17] He K, Gkioxari G, Dollár P and Girshick R 2017 Mask R-CNN *IEEE International Conference on Computer Vision (ICCV), 2017* 2980–8
[18] Ren S *et al* 2015 Faster r-cnn: towards real-time object detection with region proposal networks *Advances in Neural Information Processing Systems* **28**
[19] Jégou S *et al* 2017 The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* 11–9
[20] Zhang Z, Liu Q and Wang Y 2018 Road extraction by deep residual u-net *IEEE Geosci. Remote Sens. Lett.* **15** 749–53
[21] Dutta P, Mitra S and Roy S K Wavelet-infused convolution-transformer for efficient segmentation in medical images *IEEE Transactions on Systems, Man, and Cybernetics: Systems* (https://doi.org/10.1109/TSMC.2025.3539573)
[22] Vaswani A *et al* 2017 Attention is all you need *Advances in Neural Information Processing Systems* 30
[23] Dosovitskiy A *et al* 2020 An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale *preprint* arXiv:2010.11929
[24] Liu Z *et al* 2021 Swin transformer: Hierarchical vision transformer using shifted windows *Proc. of the IEEE/CVF International Conference on Computer Vision*
[25] Chen J *et al* 2021 Transunet: Transformers Make Strong Encoders for Medical Image Segmentation *preprint* arXiv:2102.04306
[26] Gu A and Dao T 2023 Mamba: Linear-Time Sequence Modeling with Selective State Spaces *preprint* arxiv:2312.00752
[27] Ma J, Li F and Wang B 2024 U-mamba: Enhancing Long-Range Dependency for Biomedical Image Segmentation *preprint* arXiv:2401.04722

[28] Wang Z and Ma C 2024 Weak-mamba-unet: Visual Mamba Makes cnn and vit Work better for Scribble-Based Medical Image Segmentation *preprint* arxiv:2402.10887
[29] Wang C *et al* 2024 Graph-mamba: Towards Long-Range Graph Sequence Modeling with Selective State Spaces *preprint* arXiv:2402.00789
[30] Klinwichit P *et al* 2023 BUU-LSPINE: a thai open lumbar spine dataset for spondylolisthesis detection *Applied Sciences* **13** 8646
[31] Pham H H, Trung H N and Nguyen H Q 2021 VinDr-SpineXR: a large annotated medical image dataset for spinal lesions detection and classification from radiographs *PhysioNet* RRID:SCR_007345
[32] Deng S *et al* 2024 Efficient spineunetx for x-ray: a spine segmentation network based on convnext and UNet *J. Visual Commun. Image Represent.* **103** 104245
[33] Oktay O *et al* 2018 Attention u-net: Learning Where to look for the Pancreas *preprint* arXiv:1804.03999
[34] Huang G *et al* 2017 Densely connected convolutional networks *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition* 4700–8
[35] Cao H *et al* 2022 Swin-unet: unet-like pure transformer for medical image segmentation *European Conference on Computer Vision* (Springer Nature) 205–18
[36] Chen S *et al* 2022 Multiresolution aggregation transformer UNet based on multiscale input and coordinate attention for medical image segmentation *Sensors* **22** 3820
[37] Yu B, Yin H and Zhu Z 2019 St-unet: A spatio-Temporal u-Network for Graph-Structured Time Series Modeling *preprint* arXiv:1903.05631
[38] Hu J, Shen L and Sun G 2018 Squeeze-and-excitation networks *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition* 7132–41
[39] Woo S *et al* 2018 Cbam: convolutional block attention module *Proceedings of the European Conference on Computer Vision (ECCV)* 3–19
[40] Wang Q *et al* 2020 ECA-Net: Efficient channel attention for deep convolutional neural networks *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 11534–42
[41] Si Y *et al* 2024 SCSA: Exploring the Synergistic Effects between Spatial and Channel Attention *preprint* arXiv:2407.05128
[42] Yushkevich P A, Piven J, Hazlett H C, Smith R G, Ho S, Gee J C and Gerig G 2006 User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability *Neuroimage* **31** 1116–28
[43] Chen Y *et al* 2019 Channel-Unet: a spatial channel-wise convolutional neural network for liver and tumors segmentation *Frontiers in genetics* **10** 1110
[44] Liao W, Zhu Y, WangX, Pan C, Wang Y and Ma L 2024 Lightm-unet: mamba assistsin lightweight unet for medical image segmentation *preprint* arXiv:2403.05246
[45] Valanarasu J M J and PatelV M 2022 Unext: mlp-based rapid medical image segmentation network *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer) 23–33
[46] Wu H, Zhao Z and Wang Z 2023 Meta-unet: multi-scale efficient transformer attention unet forfast and high-accuracy polyp segmentation *IEEE Trans. Autom. Sci. Eng.* 1–12
[47] Ge R, Cai H, Yuan X, Qin F, Huang Y, Wang P and Lyu L 2021 Md-unet: multi-input dilated u-shape neural network forsegmentation of bladder cancer *Comput. Biol. Chem.* **93** 107510
[48] Zhang M, Yu Y, Gu L, Lin T and Tao X 2024 Vm-unet-v2 rethinking vision mamba unet for medical image segmentation arXiv:2403.09157
[49] Lei M *et al* 2024 ConDSeg: a general medical image segmentation framework via contrast-driven feature enhancement *preprint* arXiv:2412.08345