Among General Spine Segmentation with Multi-scale and Discriminate Feature Fusion

Tingwei Wen, Yao Lu, Guangming Lu*
Harbin Institute of Technology, Shenzhen

Xiaosheng Chen, Xinhai Lu
Shenzhen Second People's Hospital

Abstract

Automatic spine segmentation from X-ray images is an important step for diagnosing spinal diseases like scoliosis. However, manual segmentation is time-consuming and prone to errors due to subjective judgments. Thesis proposes a supervised convolutional neural network for accurate and efficient spine segmentation based on X-ray images. The proposed network adopts DUCK-Net, a U-Net structure with six parallel convolution paths, as the backbone and introduces several improvements. To detect vertebrae with different sizes, we introduce Attention Gates between encoder-decoder layers to strengthen multi-scale feature fusion. Channel Interaction Attention block is proposed to enhanced feature fusion process for more discriminate feature representation. Additionally, a curvature loss is included as a regularization term during training to discourage connected vertebrae segmentation. We evaluate our method on a spine segmentation dataset and a polyp segmentation dataset, showing that it achieves reliable performance on Dice coefficient, Jaccard similarity, Precision and Recall. Our model have achieved state-of-the-art performance in spine segmentation from X-ray images and has been implemented in an automated scoliosis diagnosis system in hospital, which shows significant clinical application value and theoretical significance.

Keywords: Medical images, scoliosis, spine segmentatio, semantic segmentatio, convolutional neural networks.

1. Introduction

The spine is one of the most vital structures in human body. It serves numerous essential functions, including bearing the weight of body and protecting the spinal cord and nerves within it. In both anterior and posterior views, the normal spine should be upright and located in the center of the pelvis, while scoliosis is a pathological condition where the spine is abnormally curved to the left or right side. Around 2% to 5% of adolescents worldwide suffer from scoliosis, which typically emerges during the rapid growth of the spine and can cause physical deformities in body appearance. In severe cases, it can even lead to paralysis. Therefore, giving an accurate diagnosis of whether adolescent patients suffer from scoliosis is particularly important to provide treatment plans for patients. Clinically, the general method for diagnosing scoliosis is to capture X-ray images of the patient's spine, manually measuring the scoliosis severity. However, clinicians need to rely on a great wealth of experience, leading to a significant risk of diagnostic errors due to subjective judgments. In addition, with the popularization of medical imaging equipment such as X-rays, CT, MRI, etc., the number of medical images rapidly increases, putting enormous pressure on clinicians, further affecting the efficiency and accuracy.

Therefore, it is an urgent requirement to develop an automated technology to precisely measure scoliosis. In this study, we present a supervised convolutional neural network architecture for spine segmentation. Our model uses DUCK-Net [4] as backbone, which is a model with U-Net architecture, using six variations of convolutional blocks in parallel for better feature extract. DUCK-Net is evaluated on several benchmark datasets for polyp segmentation and achieved state-of-the-art results, however, performed poorly in spine X-ray image segmentation. The segmentation results were prone to problems such as adjacent vertebrae being connected, incomplete segmentation of vertebrae or incorrect segmentation of spinal boundaries. Compared with images of polyps, spine X-ray images have lower brightness and contrast, blurred boundaries, smaller vertebrae near neck while larger vertebrae near pelvis, and issues with rib interference. We considered that DUCK-Net used addition as the feature fusion method, which caused the neglect of edge features that should be paid attention to after addition. Thus, we introduce the Channel Interaction Attention block, changing the way of feature fusion to concatenation without feature loss, and selecting task relevant channel in-formation from the feature maps and reducing the number of channels. Besides, we use Attention Gates [13] on each skip connection between encoders and decoders, exchanging the information among layers, enhancing the information extracted from features at different scales to adapt

to different sizes of vertebrae. Applying such two modules, our model can utilize multi-scale information fusion to obtain more discriminate features. Lastly, we introduce curvature loss as a regularization term to punish the segmentation of connected vertebrae. Our main contributions are summarized as follows:

Our Attention Gates merge the relevant information from two encoders of adjacent layers to generate a fused feature map and send it to the decoder, effectively fusing the multi-scale information and alleviating the information gaps between layers, to make the model focus more on the vertebrae, even if the sizes are inconsistent.

Our CIA (Channel Interaction Attention) block can preserve important features such as edges during the feature fusion process, automatically learn the importance of channels and implicitly provide higher weights for these channels, driving the model to utilize these high weighted channels when reducing the number of channels, enabling a learnable feature fusion process.

Our curvature loss uses the curvature of discrete points in the segmentation results as a regularization loss, which produces the model looks to minimize curvature loss, aims to smooth the segmentation curvature, and makes the segmentation results closer to the real vertebrae.

Our model can accurately segment the spine in the X-ray images with low brightness and low contrast, which is conducive to the subsequent calculation of various spinal scoliosis parameters and the diagnosis of spinal scoliosis.

2. Related work

Giannoglou and Stylianidis [5] published a review on scoliosis measurement for spinal scoliosis and spinal X-ray image processing. They mentioned that the processing order in spinal X-ray images is generally: segment image to extract regions of interest, recognize individual vertebra, and predict spinal curvature degree. Image segmentation is mentioned as the first step in spine X-ray image processing and the core of the entire process. The segmentation result directly affects the subsequent calculation of parameters. The correct segmentation of images is an essential step in medical image processing, with the main task being to remove unimportant parts of the image and extract parts containing special meaning or targets for further analysis. Currently, the mainstream medical image segmentation methods are based on deep learning method.

In recent years, deep convolutional neural networks have shown great potential in medical image segmentation. Unlike traditional machine learning, convolutional neural networks do not require manual feature extraction during training and can perform end-to-end target segmentation. U-Net [14], an encoder-decoder model developed initially for biomedical image segmentation, combines shallow features of compressed path and deep features of expanded path through skip connections to achieve the trade-off between local features and contextual information, reducing the loss of edge features caused by downsampling operations to a certain extent. U-Net and its variants Unet++ [20], V-net [12], etc., have simple structures, small numbers of parameters, and require less data for network training, making them suitable for medical image segmentation. Horng et al. [7] extracted a rectangular region of the spine in the X-ray image based on pixel intensity, and then fed it into U-Net for segmentation. Imran et al. [8] made some improvements to U-Net. They used a convolution operation at each layer of the U-Net decoder to extract the outputs of that layer and fused them together as the final outputs. These progressive lateral outputs ensured that the deep features in the image were not lost by the decoder. The model achieved better results in spine segmentation than U-Net. Shen et al. [16] replaced the fully connected layers of VGG-Net with the decoder of U-Net and incorporated a wavelet decomposition module to enhance the detail information of images. Shao et al. [15]established a semi-supervised training framework that utilized Stable Diffusion to generate a large number of spinal X-ray images. They employed ResNet50 as the backbone network and used a feature pyramid network to extract features, predicting the spinal contour and four corner points to calculate the Cobb angle.

In medical image segmentation task, polyp segmentation datasets are widely recognized, where many models have demonstrated their performance. Srivastava et al. [17] added a global multi-scale feature fusion mechanism based on ResNet [6], combined with cross-scale attention and subsequent multi-scale feature selection modules, proposed GMSRF-Net to achieve accurate and generalized segmentation of polyps. Chen et al. [2] used Transformer to encode tokenized image patches from a convolution neural network (CNN) feature map as the input sequence for extracting global contexts, and then used decoder to upsample the encoded features which are then combined with the high-resolution CNN feature maps to enable precise localization. Tomar et al. [18] used residual blocks with ResNet-50 as the backbone and takes the advantage of transformer self-attention mechanism as well as dilated convolution, proposed TransResU-Net. Dumitru et al. [4] used their custom-built convolutional block, DUCK (Deep Understanding Convolutional Kernel), which allowed more in-depth feature selection, enabling the model to locate the polyp targets accurately and correctly to predict the borders, and residual downsampling, which allowed the model to use the initial image information at each resolution level in the encoder. DUCK-Net achieved the state-of-the-art in polyp segmentation task. We employed the DUCK-Net as the backbone and made some improvements and proposed our spine segmentation model.

3. Methodology

Fig. 1 shows the architecture of our proposed model. Our enhancement involves employing Attention Gates to enhance the information exchange from adjacent layers with different scales for vertebrae of different sizes, introducing Channel Interaction Attention block to strengthen the feature fusion process also facilitate a learnable feature integration process, and taking curvature loss as a regularization loss to punish the segmentation of connected vertebrae.

3.1. Backbone: DUCK-Net

DUCK-Net uses the encoder-decoder architecture of the U-Net, with two significant improvements: a novel convolutional block called DUCK block and residual downsampling. DUCK block contains six variations of convolutional blocks in parallel to locate the target precisely and detect the border accurately. The Residual block, which is first introduced in ResUNet++ [11], is one of the DUCK block components, aiming to understand the small details. There are combinations of one, two and three Residual blocks in parallel. The other two components are Midscope and Widescope blocks using dilated convolutions to understand the higher-level features. The last one is the Seperated block, which uses a 1 $\times$ N kernel and an N $\times$ 1 kernel to simulate an N $\times$ N kernel, enabling the model to capture the spatial connections in both vertical and horizontal directions. DUCK-Net replaces the 3 $\times$ 3 convolutional blocks used by U-Net with the DUCK block and adds a secondary downscaling layer that does no convolutional operations as the residual downsampling to handle the issue caused by DUCK block such as losing details. DUCK-Net evaluated on several benchmark datasets for medical image segmentation and achieved state-of-the-art results in terms of mean Dice coefficient, Jaccard index and other metrics, showing strong generalization abilities with limited training data. Therefore, we use DUCK-Net as the model backbone for spine X-ray image semantic segmentation.

3.2. Attention Gates

U-Net adopts an encode-decoder structure and directly sends the results of the encoders to the decoders through skip connections, alleviating the loss of features caused by downsampling to some extent. However, only connections between encoders and decoders within the same layer exist in U-Net, which may lead to ineffective utilization of features at different scales, while vertebrae are of different sizes that features from all scales matter. To address this issue, we introduce Attention Gates, which were originally used in natural image analysis, knowledge graphs, and natural language processing (NLP) for image captioning, machine translation, and classification tasks. Attention U-Net [13] proposed a novel self-attention gating module that can be utilized in U-Net for image segmentation tasks. The Attention Gate takes the output features of two adjacent encoder layers as inputs and generates attention coefficients. This allows the signals of the same region of interest in the two feature maps to be enhanced, while the different regions between them can also serve as auxiliary information, achieving the aggregation of information between two feature maps of different scales. As shown in Fig. 1, the Attention Gate performs a pixel-wise 1 $\times$ 1 convolution on the outputs of two adjacent encoder layers, adds them up as an attention coefficient map, to obtain more accurate results than multiplicative attention. After that, a ReLU activation function and a linear transformation through a 1 $\times$ 1 convolution are applied. Finally, we normalize the attention coefficients using sigmoid activation function. The Attention Gate is formulated as follows:

$$F_{out} = \sigma_2(W_3(\sigma_1(W_1F_1 + W_2F_2 + b_1)) + b_3) \quad (1)$$

where W_1,W_2 and W_3 represent the linear transformations, b_1 represents the sum of the bias terms corresponding to W_1 and W_2, b_3 represents the bias term corresponding to W_3, σ_1 corresponds to ReLU activation function and σ_2 corresponds to sigmoid activation function. The result attention map is a grid signal based on image spatial information, aggregating information from multiple imaging scales. After performing attention calculation with the encoder output, the result is superimposed on the decoder. Attention Gates are integrated into the skip connections of each layer, utilizing two feature maps from the current layer and the subsequent deeper layer. Through the Attention Gates, relevant information in the feature maps of two scales are merged to generate a fused feature map, which is then concatenated with the upsampled results, facilitating the exchange of information between layers within the U-Net structure.

3.3. Channel Interaction Attention Block

In the DUCK block [4], features extracted from six parallel paths are fused using the add operation, which is a simple and efficient way of feature fusion. Although the amount of computation is small, add operation has the disadvantage of losing features because after addition, the original features disappear directly and the number of features decreases, which cannot be learned or under controlled. Therefore, we changed the feature fusion method in the block to concatenation operation. Concatenation operation directly stacks the features of different paths on the channel dimension without causing feature loss, preserving important features like edges. However, after stacking, the number of features expands to 6 times than the original number, needed to be screened. Inspired by ECA-Net proposed by Wang et al. [19] and DACE framework proposed by Chen et al. [1], we propose a channel interaction attention mechanism that automatically learns the importance of channels,

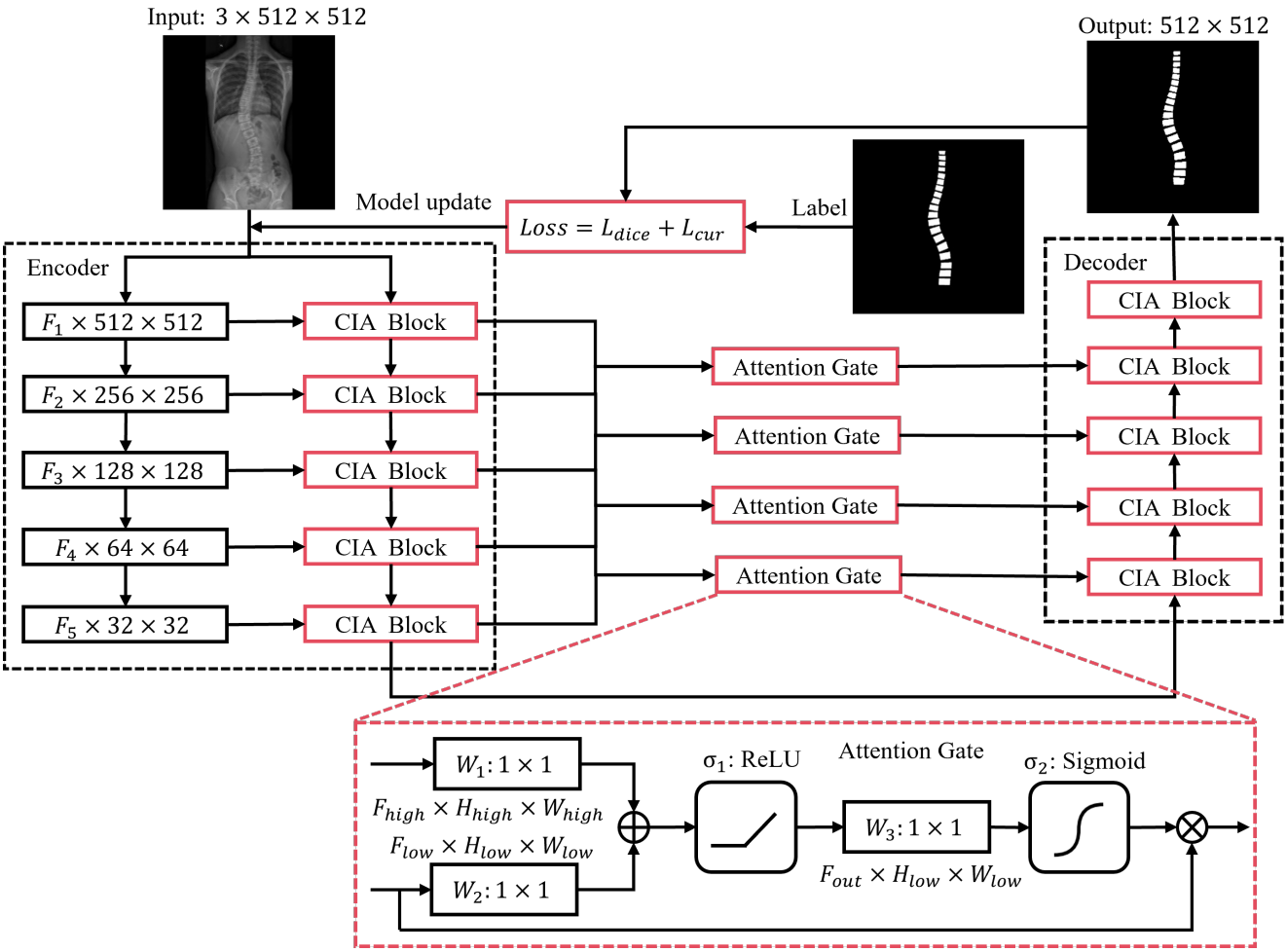


Figure 1. The architecture of the proposed model. Input images are resized as 3 × 512 × 512 and are downsampled for 5 times in the encoder. Attention Gates are integrated into the skip connections of each layer. CIA blocks are introduced in both encoder and decoder for feature extract. Curvature is introduced in loss function as a regularization loss.

to assign different weights to each channel, and finally reduce the number of channels with 1 × 1 convolution. Fig. 2 shows the architecture of the proposed Channel Interaction Attention block.

The main idea of Channel Interaction Attention block is to automatically select task relevant channel information in the feature map and provide higher weights for these channels. Specifically, given the output feature map F of each convolutional block as an input, a Global Average Pooling is firstly performed, describing the global information as a channel descriptor, and generating defined statistical information for each channel, which can be defined as:

$$F' = GAP(F) = \frac{1}{h \times w} \sum_{i=1}^{h} \sum_{j=1}^{w} F(i,j) \qquad (2)$$

where h and w denote the height and width of each feature map. The CIA block generates weights for each channel using one-dimensional convolution with a kernel size of α. Finally, the responses from each channel are re-calibrated using the corresponding weights, with the output defined as: which is formulated as follows:

$$\hat{F} = \sigma(C1D_{\alpha}(F')) \cdot F' \qquad (3)$$

where $C1D(\cdot)$ denotes a one-dimensional convolution operation, α denotes the kernel size of one-dimensional convolution, σ represents the generation of normalized weights using a simple gating mechanism with sigmoid activation.

After obtaining the channels with their importance recalibrated by attention weights, a 1 × 1 convolution is introduced to reduce the number of channels. During training, channels with high weight values contribute significantly to the results and have higher gradients. This drives the parameter updates of the 1 × 1 convolution to gradually favor preserving these channels, enabling a learnable feature fusion process. With the help of Interaction Attention block,

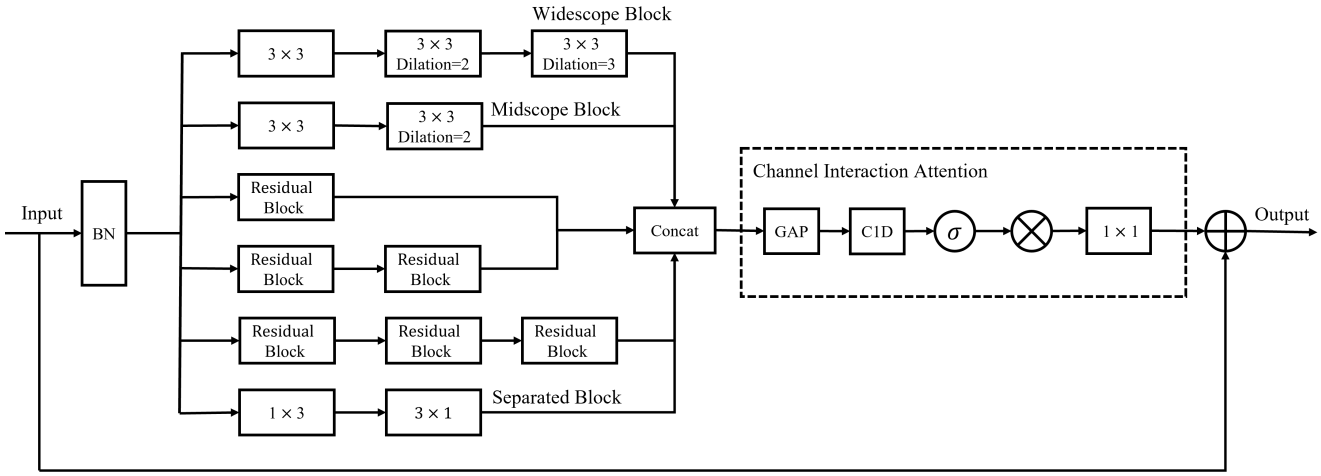


Figure 2. The architecture of the Channel Interaction Attention block, which extracts features through six parallel paths and then fuses feature by concatenation operation to avoid information loss. We introduce channel interaction attention mechanism to learn the importance of channels and reduce channels according to channel weights.

the proposed model can better utilize features and achieve discriminate feature fusion.

3.4. Loss function

Due to the relatively small gap between spinal blocks, it is easy for adjacent vertebrae to be connected on the left or right side in the segmentation result. The shapes of vertebrae are basically similar and can be approximated as rounded rectangles. However, in the segmentation result where the vertebrae are connected with each other, two vertebrae are connected on one side and there is a very obvious hollow on the other side. To address this phenomenon, we calculate the curvature of each piece of vertebra segmentation results and use the curvature as a regularization loss. The correct segmentation of the vertebrae can be approximated as many rounded rectangles, with small curvature at each point on edges, while the curvature of the concave parts in two connected vertebrae is very large. During the training process, the model seeks to minimize curvature loss, aiming to smooth the segmentation curvature, thereby penalizing the segmentation of vertebrae into a continuous entity. This can be interpreted as adding a prior knowledge to the model: the desired shape of the segmentation outcome is to achieve a more stable and smooth form. Given that pixel points are discrete, the curvature approximation of a discrete point is described in differential form. The curvature calculation formula is as follows:

$$curvature = \frac{d\theta}{ds} = \frac{arctan(y_{i+1} - y_i, x_{i+1} - x_i)}{\sqrt{(y_{i+1} - y_i)^2 + (x_{i+1} - x_i)^2}} \quad (4)$$

where (x_i, y_i) and (x_{i+1}, y_{i+1}) denote two neighboring points on a contour, θ represents the angle difference between two points to approximate the radian, s denotes the distance between two points to approximate the arc length. $d\theta/ds$ represents the rate of change of radians with respect to arc length, which is curvature. By representing the curvature using discrete values, the average curvature of the entire contour is taken as the regularization term loss L_{cur}.

The Dice loss is a loss function commonly used in medical image segmentation tasks. It uses the Dice coefficient, which measures the overlap between two sets. In image segmentation, the Dice loss can penalize the model for incorrect or incomplete segmentation of objects, and handle class imbalance, which is often a concern in medical image segmentation, where some classes may be much more prevalent than others. The whole loss function is defined as:

$$L = L_{dice} + \lambda \cdot L_{cur} \quad (5)$$

where λ is the coefficient for the curvature loss. Initially, we set it to a small value of 0.001, aiming for the Dice loss to dominate the overall loss function. As the number of training epochs increased, we gradually raised λ to 0.005, intending for the model to focus on the connected vertebrae in the segmentation results during the later stages of training.

4. Experiment

4.1. Settings

We used a dataset of 1367 high-resolution spine X-Ray images with evidence of scoliosis to various extents. We split the dataset into training (1147), testing (110), and validation (110) sets. To keep the original aspect ratio of images, the training images were zero-padded into square and then resized to 512 × 512. We trained the model for 200 epochs using Adam optimizer, with learning rate set to 1e-4. In addition to the spine X-ray image dataset, we also val-

idated our model on a polyp segmentation dataset, Kvasir-SEG [10], to assess its performance on other medical imaging segmentation datasets. We evaluated the segmentation results generated by the model using four metrics to measure the extent of similarity between the predicted mask and the ground truth: the Dice coefficient (Dice), Jaccard similarity (JS), precision (PRE), and recall (REC).

4.2. Experiments on spine segmentation

From Table 1, we have the following observations: 1) Attention U-Net exhibits a significant performance improvement over U-Net, indicating that the Attention Gate structure can play a vital role in spine segmentation. 2) DUCK-Net outperforms both U-Net and ResUNet, demonstrating the feasibility of the DUCK block and residual downsampling. 3) Our model achieves the highest scores on all metrics and beats DUCK-Net, GMSRF-Net and TransResUnet, which were the state-of-the-arts on the Kvasir-SEG dataset, confirming the viability of our proposed CIA block and curvature loss. Especially in terms of the precision metric, the experimental results show a more significant improvement, which demonstrates the effectiveness of curvature loss. This is because the introduction of curvature loss inhibits the connection between spines, and the connected parts are false positives, which would reduce precision. Thus, these comparative results demonstrate the effectiveness of the proposed model for spine segmentation. Fig. 3 shows an examples of spine segmentation results compared to other models, from which we can see that Unet++ and TransResU-Net are affected by ribs or other factors that segmentation results contain incorrect parts, incomplete vertebrae segmentation results exist in Attention U-Net, ResUNet and TransUNet, and there are connected vertebrae in UNet, nnU-Net, GMSRF-Net and DUCK-Net. The results of our model were not affected by interference, with a relatively complete spine segmentation, and there is no phenomenon of vertebrae that connect with each other, which proves the effectiveness of our improvement. Fig. 4 shows the changes in loss function and Dice coefficient of the proposed model during the training process, from which we can see that as the number of training epochs increases, the Dice coefficient gradually rises while the loss consistently decreases, indicating an improvement in the model's performance. Furthermore, after the 160th training epoch, the trends of the Dice coefficient and the loss stabilize, signifying that the model has converged and reached a relatively stable level of performance.

4.3. Experiments on Kvasir-SEG dataset

Table 2 shows the experiments conducted on the polyp segmentation dataset Kvasir-SEG, aiming to evaluate the performance of our model on other medical image segmentation datasets, compared with the other state-of-the-art methods. The experiment results indicate that our model outperforms other models such as U-Net, nnU-Net, TransResU-Net and GMSRF-Net, and is closely comparable to DUCK-Net, demonstrating a robust polyp segmentation capability. The empirical evidence indicates that the proposed model demonstrates robust generalization abilities, which not only excels in the domain of spine segmentation but also shows efficacy when applied to diverse datasets pertaining to medical image segmentation.

Method	Dice	JS	PRE	REC
U-Net [14]	0.7764	0.6345	0.7827	0.7703
ResUNet [3]	0.7835	0.6441	0.7898	0.7773
Unet++ [20]	0.7860	0.6474	0.7922	0.7799
Attention U-Net [13]	0.7961	0.6613	0.8026	0.7897
TransUNet [2]	0.8029	0.6707	0.8094	0.7965
GMSRF-Net [17]	0.8097	0.6802	0.8028	0.7988
TransResU-Net [18]	0.8279	0.7064	0.8540	0.8034
nnU-Net [9]	0.8293	0.7085	0.8360	0.8229
DUCK-Net [4]	0.8379	0.7211	0.8372	0.8387
Ours	**0.8725**	**0.7739**	**0.8801**	**0.8651**

Table 1. Segmentation results on the spine segmentation dataset. Best model results are in bold.

Method	Dice	JS	PRE	REC
U-Net [14]	0.8125	0.6842	0.8126	0.8124
ResUNet [3]	0.8158	0.6890	0.8042	0.8278
Unet++ [20]	0.8320	0.7124	0.8125	0.8524
Attention U-Net [13]	0.8340	0.7153	0.8283	0.8398
TransUNet [2]	0.8706	0.7709	0.8769	0.8645
GMSRF-Net [17]	0.9286	0.8667	0.9321	0.9251
TransResU-Net [18]	0.8884	0.8214	0.9022	0.9106
nnU-Net [9]	0.9341	0.8763	0.9315	0.9367
DUCK-Net [4]	**0.9502**	**0.9501**	**0.9628**	0.9379
Ours	0.9483	0.9016	0.9555	**0.9412**

Table 2. Segmentation results on Kvasir-SEG dataset. Best model results are in bold.

4.4. Ablation experiments

To evaluate the effectiveness of each component added in the proposed model, we conduct comprehensive ablation experiments by removing each component. The experiment results are shown in Table 3, where CIA represents Channel Interaction Attention block, AGs represents Attention Gates, and DUCK-Net is the baseline of our work. All the w/o curvature loss model, w/o CIA model and w/o AGs model outperform the backbone DUCK-Net, verifying the effectiveness of the three improvements. When the curvature loss was removed, precision decreased even more significantly compared to dice coefficient, indicating an increase in the number of false positives in the segmentation results. Before the introduction of curvature loss, there was a phenomenon of vertebrae connected with each other in the segmentation results, where the connected parts correspond

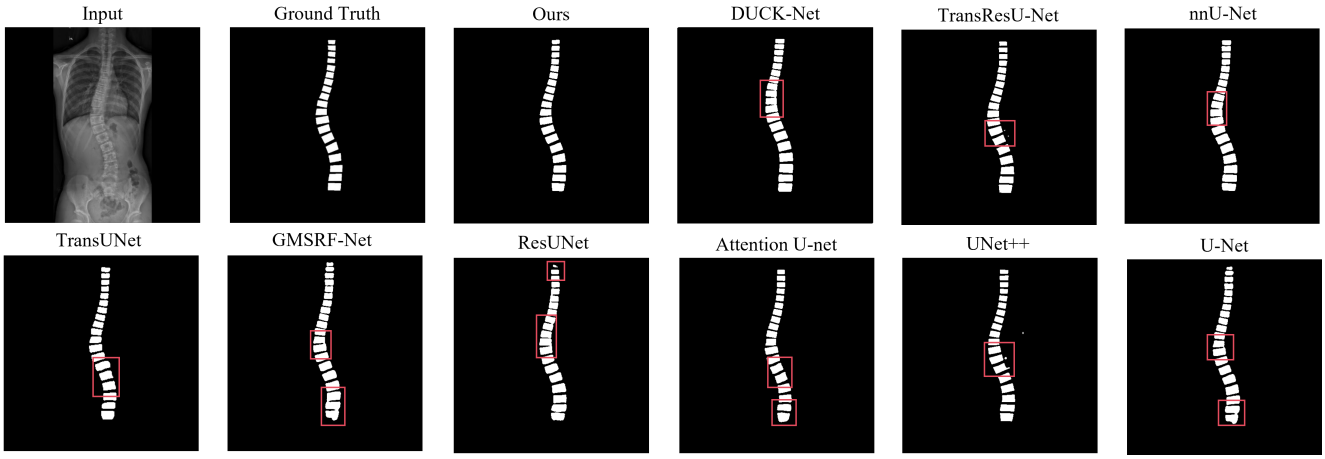
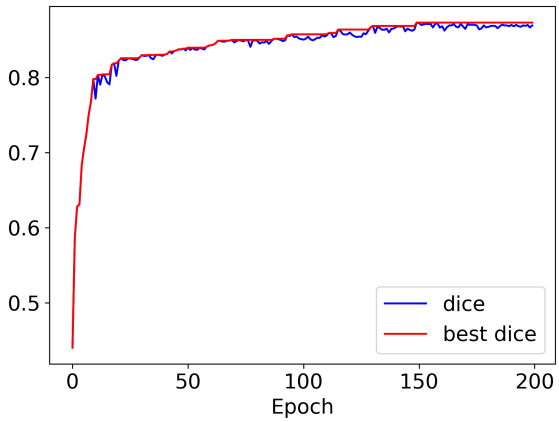
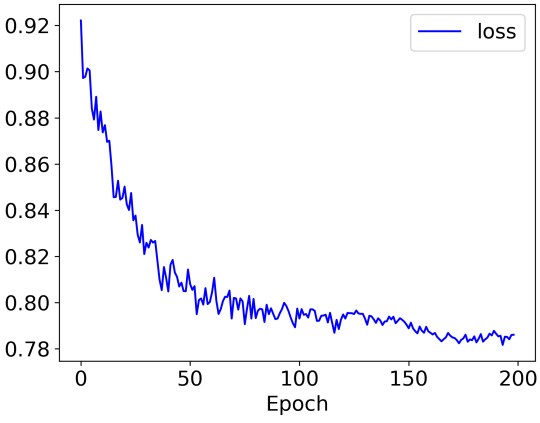


Figure 3. Comparison of spine segmentation results predicted by different models.



(a) Dice and best dice



(b) Loss

Figure 4. Training process of the proposed model.

to false positives. These experiment results demonstrate the effectiveness of curvature loss in reducing the number of

Method	Dice	JS	PRE	REC
DUCK-Net [4]	0.8379	0.7211	0.8372	0.8387
w/o curvature loss	0.8492	0.7379	0.8487	0.8497
w/o CIA	0.8482	0.7364	0.8641	0.8328
w/o AGs	0.8519	0.7421	0.8678	0.8367
Ours	**0.8737**	**0.7757**	**0.8937**	**0.8545**

Table 3. Ablation studies results on spine segmentation dataset. Best model results are in bold.

connected vertebrae. After removing CIA model, the evaluation scores drop significantly, and the complete proposed model apparently outperform the w/o AGs model, indicating that under the guidance of the two modules, our proposed model can better detect boundaries and distinguish vertebrae of different sizes. The ablation experiments confirmed the effectiveness of the components, with no conflicts observed between them.

5. Conclusion

In thesis, we have proposed an effective model for X-ray images spine segmentation. Our key contributions include introducing Attention Gates to enhance the information extracted from features at different scales to detect different sizes of vertebrae, designing a Channel Interaction Attention block for better feature fusion to reduce the loss of important features such as edges, introducing curvature as a regularization term in loss function to punish segmentation with connected vertebrae. Extensive experiments on spine segmentation and polyp segmentation datasets demonstrate the superiority of our method over previous state-of-the-arts. The proposed network provides an effective solution for automated spine segmentation based on X-ray images, which can benefit scoliosis diagnosis.

Acknowledgement

This work was supported in part by the NSFC fund (NO. 62206073, 62176077), in part by the Shenzhen Key Technical Project (NO. JCYJ20241202123728037, JSGG20220831092805009, JSGG20220831105603006, JSGG20201103153802006), in part by the Guangdong International Science and Technology Cooperation Project (NO. 2023A0505050108), in part by the Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies (NO. 2022B1212010005), in part by the Guangdong Shenzhen joint Youth Fund under Grant 2021A151511074, and in part by the Natural Science Foundation of Guangdong Province under Grant 2023A1515010893.

References

[1] B. Chen, Y. Liu, Z. Zhang, Y. Li, Z. Zhang, G. Lu, and H. Yu. Deep active context estimation for automated covid-19 diagnosis. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(3s):1–22, 2021. 3

[2] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 2, 6

[3] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162:94–114, 2020. 6

[4] R.-G. Dumitru, D. Peteleaza, and C. Craciun. Using duck-net for polyp image segmentation. *Scientific reports*, 13(1):9803, 2023. 1, 2, 3, 6, 7

[5] V. Giannoglou and E. Stylianidis. Review of advances in cobb angle calculation and image-based modelling techniques for spinal deformities. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 3:129–135, 2016. 2

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[7] M.-H. Horng, C.-P. Kuok, M.-J. Fu, C.-J. Lin, and Y.-N. Sun. Cobb angle measurement of spine from x-ray images using convolutional neural network. *Computational and mathematical methods in medicine*, 2019(1):6357171, 2019. 2

[8] A.-A.-Z. Imran, C. Huang, H. Tang, W. Fan, K. Cheung, M. To, Z. Qian, and D. Terzopoulos. Analysis of scoliosis from spinal x-ray images. *arXiv preprint arXiv:2004.06887*, 2020. 2

[9] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021. 6

[10] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. De Lange, D. Johansen, and H. D. Johansen. Kvasir-seg: A segmented polyp dataset. In *MultiMedia modeling: 26th international conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, proceedings, part II 26*, pages 451–462. Springer, 2020. 6

[11] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. De Lange, P. Halvorsen, and H. D. Johansen. Resunet++: An advanced architecture for medical image segmentation. In *2019 IEEE international symposium on multimedia (ISM)*, pages 225–2255. IEEE, 2019. 3

[12] F. Milletari, N. Navab, and S.-A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 2

[13] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018. 1, 3, 6

[14] 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: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 2, 6

[15] Z. Shao, Y. Yuan, L. Ma, D.-Y. Yeung, and X. Zhu. Sg-lra: Self-generating automatic scoliosis cobb angle measurement with low-rank approximation. *arXiv preprint arXiv:2411.12604*, 2024. 2

[16] X. Shen, Y. Zhang, R. Zhang, Q. Shi, Y. Song, and Q. Zhang. Segmentation method of x-ray whole spine coronal image based on vgg-net. *Foreign Electronic Measurement Technology*, 43(01):135–140, 2024. 2

[17] A. Srivastava, S. Chanda, D. Jha, U. Pal, and S. Ali. Gmsrf-net: An improved generalizability with global multi-scale residual fusion network for polyp segmentation. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 4321–4327. IEEE, 2022. 2, 6

[18] N. K. Tomar, A. Shergill, B. Rieders, U. Bagci, and D. Jha. Transresu-net: Transformer based resu-net for real-time colonoscopy polyp segmentation. *arXiv preprint arXiv:2206.08985*, 2022. 2, 6

[19] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11534–11542, 2020. 3

[20] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018. 2, 6