DOI: 10.53469/wjimt.2025.08(02).05

# SCAR-Net: Spine Segmentation in MRI based on Cross Attention and Recognition-assisted Label Fusion

**Xinyu Lei[1], Mingwei Wang[1,*], Rui Wang[2], Jicheng Tan[3], Weizhuo Wang[4]**

[1]Shaanxi university of science and technology, Xi'an 710021, China
[2]Xi'an Jiaotong University, Xi'an 710049, China
[3]Xi'an Lianhu Qinhua Hospital, Xi'an 710003, China
[4]The Second Affiliated Hospital, Xi'an Jiaotong University, Xi'an 710000, China
**Corresponding Author*

**Abstract:** ***The segmentation of multiple vertebrae and intervertebral discs in magnetic resonance images (MRI) plays a crucial role in diagnosing and treating spinal disorders. However, the inherent complexity of the spine, coupled with the challenges of balancing inter-class similarity and intra-class variety, complicates the task. Additionally, improving the generalization ability, learning rate, and accuracy of spine segmentation remains difficult. To address these challenges, this paper proposes a spine segmentation method based on cross attention and recognition-assisted label fusion (SCAR-Net). The approach introduces a multi-channel cross attention (MCCA) mechanism to generate a comprehensive spine description by fusing inter-class and intra-class features. Furthermore, a key-points recognition-assisted learner (KRAL) is designed, incorporating mixed-supervision recognition-assisted label fusion (RALF) to reduce reliance on a single dataset and enhance network generalization. Experimental results on T2-weighted volumetric MRI datasets demonstrate that SCAR-Net achieves outstanding performance, with a mean Dice similarity coefficient (DSC) of 96.12% for 5 vertebral bodies and 95.07% for 5 intervertebral discs. The proposed method proves to be highly effective for both the localization and segmentation of intervertebral discs in MRI spine images.***

## 1. INTRODUCTION

In spine pathologies locating, spine diseases diagnosis and surgical treatment planning, multi spine segmentation is crucial for experts to make correct judgments. However, manual spine segmentation is a labor-intensive and time-consuming task that requires the expertise of experienced radiologists or clinicians. This process can be prone to inter-observer variability, where different experts may produce varying results. Therefore, automated multiple spine segmentation is of great significance to mitigate these problems by reducing the time and effort required for segmentation, enhancing accuracy and consistency, and minimizing human error [1]. Additionally, since spine images are highly complex and lack simple linear features, achieving accurate multiple spine segmentation remains a challenge. Furthermore, most spine segmentation networks are fully supervised, which greatly limits the network's generalization ability. Therefore, it is of great significance to develop a network with high accuracy and strong generalization ability for multiple spine segmentation to assist professional physicians in making diagnoses.

Due to the strong learning ability and portability, deep learning has been widely used in spine segmentation. Chen et al. [2] proposed a method for vertebrae detection and recognition, which captured the shape and appearance of single spine. Nevertheless, their approach led to confusion in identifying the vertebral body and surrounding tissues, which limited its extension to segmenting spines consisting of multiple vertebrae. Sekuboyina et al. [3] designed two networks, one for spine segmentation and the other for spine positioning, which effectively distinguished the vertebral body and surrounding tissues. However, it ignored the inter-class similarity learning, leading to poor segmentation performance in the edge regions. Schlemper et al. [4] proposed an attention gate (AG) model to focus on target structures in the edge region, but it was only effective for segmenting single vertebrae but not for multiple spine, resulting in a significant reduction in segmentation ac- curacy. Using long short-term memory (LSTM), Han et al. [5] generated semantic features of the spine and increased the long-distance spatial correlation of pixels in order to segment multiple spines. However, their method showed poor segmentation performance in the internal details of the LSTM due to insufficient learning of intra-class variety of spine. Change et al. [6] employed the graph convolutional network (GCN) to capture the spatial correlations between spine structures and generate the semantic features for accurate spine segmentation. However, their approach failed to balance the intra-class variety of spine due to the neglect of local spinal information integration, leading to blurred edge details. Pang et al. [7] designed SpineParseNet, which was a semantic image representation framework consisting of two segmentation stages that focused on the internal details and edge details of the spine. How- ever, their method did not adequately balance the inter-class similarity and the intra-class variety of the spine because it ignored the correlation attention between different spinal structures. To sum up, the aforementioned segmentation methods cannot balance the inter-class similarity and intra-class variety of spine.

In order to balance the inter-class similarity and intra-class variety of the spine, much efforts have been made to solve this thorny problem. Rasoulian et al. [8] developed a statistical multi-vertebrae shape pose model to simultaneously capture the variations in shape and posture in spinal images, which enhanced the ability to learn the inter-class similarity of the spine. However, this method neglected the intra-class variety learning of spine, resulting in confusions in spine detail segmentation. Chen et al. [9] demonstrated that the fully convolutional network (FCN) was effective in efficiently segmenting spine and supplementing the learning of intra-class variety. However, the method still exhibited an unbalanced learning of the inter-class similarity and intra-class variety of the spine, leading to a loss of boundary details. In order to segment continuous spine edge, Kolaˇrík et al. [10] proposed a supervised deep learning approach to segment 3D spine and to obtain segmented results at the original resolution. However, 3D CNN suffered from high computation cost, storage cost and the risk of over fitting for medical images with limited labeled data. To address these bottlenecks, Wang et al. [11] proposed a liver tumor segmentation network named CPAD-Net, which was built upon the traditional U-Net architecture while integrating contextual parallel attention and dilated convolution to narrow semantic gaps and increase detailed information. Zhang et al. [12] applied inter-slice attention (ISA) mechanism based on 2D convolutional network to acquire inter slice information for 3D segmentation task, which achieved high accuracy and efficiency. However, due to the small number of pixel-level datasets in the spine, the generalization ability of the network is weak. Pang et al. [13] presented a network that utilized key-points-level datasets to increase the generalization ability of the network and enhance the segmentation accuracy. However, this approach significantly reduced the learning efficiency of the network and carried the risk of losing shallow features.

In summary, most existing spine segmentation networks face the following problems. Firstly, due to the complex and highly similar structures of the spine, it is difficult to balance the inter-class similarity and intra-class variety, which results in an inability to distinguish between intervertebral discs (IVDs) and vertebral bodies (VBs), leading to segmentation ambiguity on IVDs and VBs. Secondly, most networks adopt full supervision methods that rely highly on the specific datasets, thereby weakening the generalization ability of the network. Thirdly, most CNN-based spine segmentation networks often aim to improve segmentation accuracy by increasing convolution depth, which will not only damage the learning rate of the network, but also increase the risk of losing superficial feature information with the deeper layers. In order to address these challenges, a multiple spine segmentation in MRI based on cross attention and recognition-assisted label fusion (SCAR-Net) has been proposed. The objective is to effectively extract the multiscale features from MRI, balance the inter-class similarity and the intra-class variety of the spine, and enhance the generalization capability of the network.

## 2. RELATED WORKS

The intricate spinal structure poses significant challenges for traditional methods to balance the inter-class similarity and the intra-class variety in multiple spine segmentation. Due to the subtle differences between spines, achieving accurate and reliable segmentation remains a complex task. In recent years, convolutional neural networks (CNNs) have emerged as a powerful tool for automatic recognition and segmentation of spinal structures, without requiring manual intervention. Ji et al. [14] presented an approach for IVD segmentation that employed a standard CNN to extracted plaques around each pixel using convolution operations. In a subsequent work, Zeng et al. [15] proposed an approach called DSMS-FCN, which incorporated multi-scale deep supervision to mitigate gradient disappearance during training. However, these methods are primarily applied for single mode scenes and the accuracy decreases significantly in multi-mode segmentation. Aygün et al. [16] investigated various approaches to combine multiple modes in the context of CNNs by treating each mode as an independent input and subsequently fusing them at different stages (early, middle, or late).

However, this method relied on single-layer fusion to approximate the complex inter- modal relationships, leading to limited generalization performance of the network. Dolz et. al [17] addressed multi-modal IVD localization and segmentation by proposing a U-Net based architecture that leveraged multi-modal data to create a coding path, allowing for modeling the inter-modal relationships. However, the direct fusion strategy adopted by this approach cannot fully capture the comprehensive spinal information required for accurate segmentation. In summary, the above methods have improved feature fusion to some extent. However, attention mechanism needs to be integrated for more effective feature fusion.

# 3. METHODS

Overall, the proposed SCAR-Net method effectively balances the inter-class similarity and the intra-class variety of the spine, achieving high integrity and detail accuracy in spine segmentation. Its combination of global and local contexts, attention mechanism, unsupervised learning component, and multi-scale feature fusion make it a promising approach for multiple spine segmentation. The SCAR-Net architecture consists of several modules that cooperate together to perform automated spine segmentation. The input to the network is a 3D MRI volume, and the output is the segmented spinal cord. The proposed SCAR-Net framework is illustrated in Figure 1, comprised of two components: a segmentation network (the main path) and a recognition network (the branch path), where the branch path $R_\beta$ is parameterized by $\beta$ and the main path $S_{[\alpha,\beta]}$ is parameterized by $\alpha$ and $\beta$. To simplify notation, in the subsequent discussion, $R_\beta$ and $S_{[\alpha,\beta]}$ are replaced with R and S. The primary objective of the branch path is to extract semantic features that aid in producing precise segmentation results through an encoder-decoder architecture. By incorporating the branch path, SCAR-Net is able to capture multi-scale contextual information and enrich the feature representation of the main path, which improves the accuracy and robustness of the seg- mentation results. Additionally, the implementation of skip connections and feature fusion allows the main path and branch path to influence each other, ensuring that both pathways are optimized for the accurate spine segmentation.

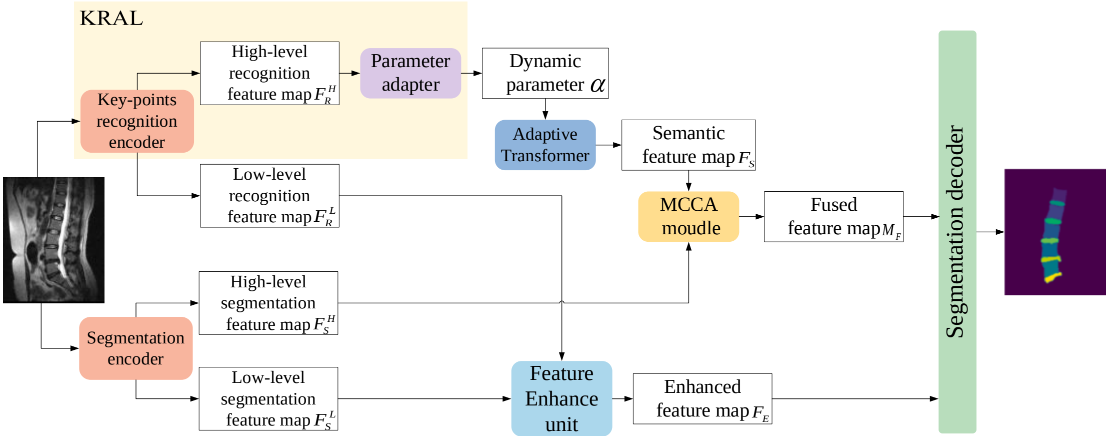
KRAL
High-level recognition feature map $F_R^H$
Parameter adapter
Dynamic parameter $\alpha$
Key-points recognition encoder
Adaptive Transformer
Semantic feature map $F_S$
Low-level recognition feature map $F_R^L$
MCCA moudle
Fused feature map $M_F$
Segmentation decoder
High-level segmentation feature map $F_S^H$
Segmentation encoder
Low-level segmentation feature map $F_S^L$
Feature Enhance unit
Enhanced feature map $F_E$

**Figure 1:** The basic framework of the proposed SCAR-Net, which consists of the main path for spine segmentation and the branch path for spine recognition.

In the main path, an encoder-decoder architecture is adopted for multiple spine seg- mentation. The purpose of the segmentation encoder is to extract both the low-level seg- mentation feature map $F_S^L$ and the high-level segmentation feature map $F_S^H$. Specifically, an improved DeepLabv3+ encoder错误!未找到引用源。 is implemented to deepen the convolution network while enhancing the precision of semantic segmentation results. The segmentation encoder is consisting of an initial convolution layer followed by a series of high-level and low-level feature extraction modules. The high-level feature extraction module includes an atrous spatial pyramid pooling (ASPP) layer to extract the multi-scale features at different dilation rates. The low-level feature extraction module contains depthwise separable convolution layers to capture the low-level features at different levels of abstraction. Notably, the segmentation encoder is a 2D network compared to the depth convolution network, containing $F_S^L$ generated in the third separable convolution layer to conserve memory. In the branch path, the recognition encoder extracts the low-level recognition feature map $F_R^L$ and the high-level recognition feature map $F_R^H$, sharing the same structure as the segmentation encoder for convenience.

The decoder is consisting of two inputs: a high-level feature after fusion, MF, and a low-level feature map after enhancement, FE. The architecture of the segmentation decoder consists of a series of up-sampling layers that increase the resolution of the feature maps and a concatenation layer that combines the low-level and the high-level features. The concatenated feature map is then fed to a series of SeConv and Conv layers to refine the segmentation prediction. Finally, a Softmax layer is applied to obtain the segmentation probability maps.

### 3.1 Multi-channel Cross Attention

To effectively fuse the inter-class features and the intra-class features to generate a com- prehensive description of the spine that incorporates their complementary information, we propose a multi-channel cross attention

(MCCA) feature fusion method that combines attention mechanisms with feature fusion to capture both the local and the global features of spinal structures. An MCCA consists of three attention blocks: inter-class, intra-class and channel attention block. Although each operation generates dis- tinct feature mappings, the operations within each attention block are similar. In general, the MCCA method integrates the high-level segmentation features and the semantic features of the spine to capture the spinal information from multiple perspectives. Specifically, the features are processed by a convolution layer to generate the corresponding feature map. Thereafter, different segmented features perform corresponding attention mechanisms to reshape the feature map. The detailed steps are described as follows: Firstly, the features are fed into their corresponding attention block. Assume that the inputted feature is represented as $\tilde{f}$, which is processed by three convolution layers to generate three feature mappings: $\widetilde{f^a}, \widetilde{f^b}$ and $\widetilde{f^c}$. Subsequently, $\widetilde{f^a}$ is reshaped and transposed into a feature map, denoted as $\widetilde{f^{a'}}$, while $\widetilde{f^b}$ and $\widetilde{f^c}$ are reshaped into two additional feature maps, denoted as $\widetilde{f^{b'}}$ and $\widetilde{f^{c'}}$, respectively. Matrix multiplication is performed between $\widetilde{f^{a'}}$ and$\widetilde{f^{b'}}$ followed by application of a SoftMax layer to calculate the attention weight A. After- wards, matrix multiplication is calculated between A and $\widetilde{f^{c'}}$, and the result is reshaped to obtain the attention feature, denoted as $\tilde{f}'$. Secondly, the inter-class attention feature $F_S^{H'}$ and the intra-class attention feature $F_S{}'$ are generated through the inter-class and intra-classattention block, respectively. Thereafter, $F_S^{H'}$ is multiplied by a scale parameter γ to effectively fuse the inter-class and the intra-class information. Thirdly, element-wise summation is performed on the results obtained from step 2 to produce F1. Similarly, $F_S{}'$ undergoes the same procedures as $F_S^{H'}$ to generate F2. Afterward, F1 and F2 are combined to form FC. Finally, FC is inputted into the channel attention block to filter redundant information and obtain the final cross-fused feature MF . Since MF contains sufficient inter-class information and intra-class information, the proposed MCCA can focus on important regions and as- sign higher weights to informative features while suppressing noise and irrelevant features, thereby significantly balance the inter-class similarity and the intra-class variation of the spine.

### 3.2 Recognition-assisted Label Fusion

he recognition-assisted label fusion (RALF) is designed to adjust the weights of the inter- mediate layers based on the detected key-points during training, which allows the network to learn robust features invariant to the orientation and the position of spinal cord. The RALE consists of two primary components: a key-points recognition-assisted learner (KRAL) mod- ule and a dynamic parameter generator for AT. For key-points recognition-assisted learner module, the KRAL adopts a key-point detection network to identify the landmark points on the vertebral bodies, which are essential to define the local regions of interest (ROIs) around each vertebra and to extract features for training the segmentation network. For dynamic parameter generator, it first computes the distance between each pixel and the key-points using Euclidean distance. Thereafter, a Gaussian kernel is performed to weight the distances, giving higher weights to pixels closer to the key-points. The resulting weight map modulates the attention weights in the self-attention mechanism of the AT, allowing the network to selectively focus on features close to the key-points. This component helps the model to learn highly robust and discriminative features from the key-point-annotated data, leading to improved performance on SCAR-Net. In the RALE, a recognition encoder and a parameter adapter operate in conjunction. The parameter adapter comprises an adaptive average pooling layer with a size of 7 × 7 and a convolution with a size of 1 × 1. It generates a dynamic parameter $\alpha \in R^{7\times7\times128}$ that serves as the convolution kernel for prediction. The kernel size is 7 × 7 and there are 128 AT channels. RALE seeks to generate the dynamic parameter α of the adaptive transformer (AT) to unify the information contained in the recognition feature and the segmentation feature, which improves the generalization ability of the network.

## 4. EXPERIMENTS AND RESULTS

### 4.1 Datasets Description

The datasets used in our study consist of both a fully-annotated and a weakly-annotated dataset. The fully-annotated dataset includes midsagittal T2-weighted MRI scans from 695 subjects. In this dataset, the masks for the vertebral bodies (VBs) of the five lumbar vertebrae and the corresponding masks for the five intervertebral discs (IVDs) were manually delineated by junior experts and then reviewed and corrected by senior experts using the open-source tool ITK-SNAP. The corrected masks, validated by senior experts, serve as the ground truth for spine segmentation. For each subject in the fully-annotated dataset, a T2-weighted MRI and its corresponding accurate standard mask are provided, with unique labels assigned to each VB and its corresponding IVD. The weakly-annotated dataset is derived from the Spark Digital Body AI Challenge, which includes the Spinal Disease

Dataset (available at: https://tianchi.aliyun.com/dataset/79463). This dataset comprises midsagittal T2-weighted MRI scans from 201 subjects, accompanied by manual key-point annotations for the centers of the five VBs and five IVDs. The training and testing datasets are constructed using 5-fold cross-validation. The fully-annotated dataset is randomly divided into five groups, each containing 43 subjects. Four of these groups, along with the entire weakly-annotated dataset, are used for training, while the remaining group is reserved for testing.

### 4.2 Mplementation Details

Our proposed methodology is implemented using Pytorch26 and the codes are executed on an NVIDIA GeForce RTX 3060 Ti. Due to equipment limitations, the SCAR-Net is trained with a batch size of 2, employing Adam optimizer27. The dataset is trained for a total of 600 epochs, with an initial learning rate of 0.0005. The learning rate is reduced to 1/5 of the initial value at 1/3 and 2/3 of the total epoch duration. During each iteration, four samples comprising two from the fully-annotated dataset and two from weakly-annotated dataset are randomly selected and merge to form a batch dataset. This process enables to calculate the segmentation loss and the recognition loss values during each iteration. The values of $\lambda$ in (1) are varied across the range of $\lambda \in [40, 50, 60, 70, 80, 90, 100]$.

### 4.3 Ablation Studies

The ablation study is conducted to evaluate the efficacy of specific components within the proposed model. First, SCAR-Net is adopted as the backbone network to quantitatively analyze the impact of MCCA and AT on the segmentation performance. Subsequently, the values of parameter $\lambda$ in the loss function are discussed and analyzed. Finally, the results are visually demonstrated in order to effectively understand and interpret the effects of the proposed network.

4.3.1 Quantitative analysis

**Table 1:** The Ablation study of the proposed model is conducted utilizing baseline, with DSC (%) serving as the evaluation metric.

| Methods | L1 | L2 | L3 | L4 | L5 | L1/L2 | L2/L3 | L3/L4 | L4/L5 | L5/S1 | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Baseline | 93.16 | 94.42 | 94.49 | 94.32 | 95.19 | 93.37 | 94.81 | 91.91 | 93.42 | 94.85 | 93.99 |
| Baseline+MCCA | 94.77 | 95.15 | 94.79 | 95.84 | 96.07 | 94.41 | 93.57 | 93.56 | 95.52 | 93.88 | 94.76 |
| Baseline+AT | 94.65 | 95.20 | 95.36 | 96.45 | 95.78 | 94.26 | 93.27 | 92.53 | 93.13 | 95.74 | 94.64 |
| Baseline+MCCA +AT | **96.05** | **96.43** | **95.78** | **96.39** | **96.55** | **95.78** | **96.05** | **93.88** | **93.70v** | **95.93** | **95.65** |

Table 1 presents a comprehensive quantitative analysis of each module within the proposed network, highlighting their individual contributions to improving segmentation performance. The results show that the integration of the MCCA and AT modules outperforms the baseline indicators, indicating their significant role in enhancing overall segmentation accuracy. The MCCA method effectively highlights informative features, which boosts the accuracy and robustness of vertebral body segmentation, even in challenging scenarios where neighboring vertebrae appear similar. The AT module improves both the learning rate and accuracy by increasing the network's flexibility and adaptability to different types of spinal images, thereby enhancing generalization ability. When combined, these modules deliver superior segmentation performance compared to their individual use. Specifically, the DSC achieved in L1-L5 segmentation is 1% higher than the baseline, while L1/L2-L5/S1 segmentation shows a 0.5% increase. However, when all three modules are integrated, the improvements for L1 to L5 segmentation range from 1.76% to 2.89%, and for L1/L2 to L5/S1 segmentation, they range from 0.28% to 2.63%. These results strongly suggest that the proposed modules work synergistically, complementing each other to deliver exceptional segmentation outcomes. The average improvement rate across all components is calculated at 1.66%, confirming the effectiveness of the proposed modules in enhancing the overall segmentation framework. In conclusion, incorporating the MCCA and AT modules presents a promising approach to achieving robust and accurate spine segmentation.

4.3.2 Choice of λ in the mixed-supervised loss

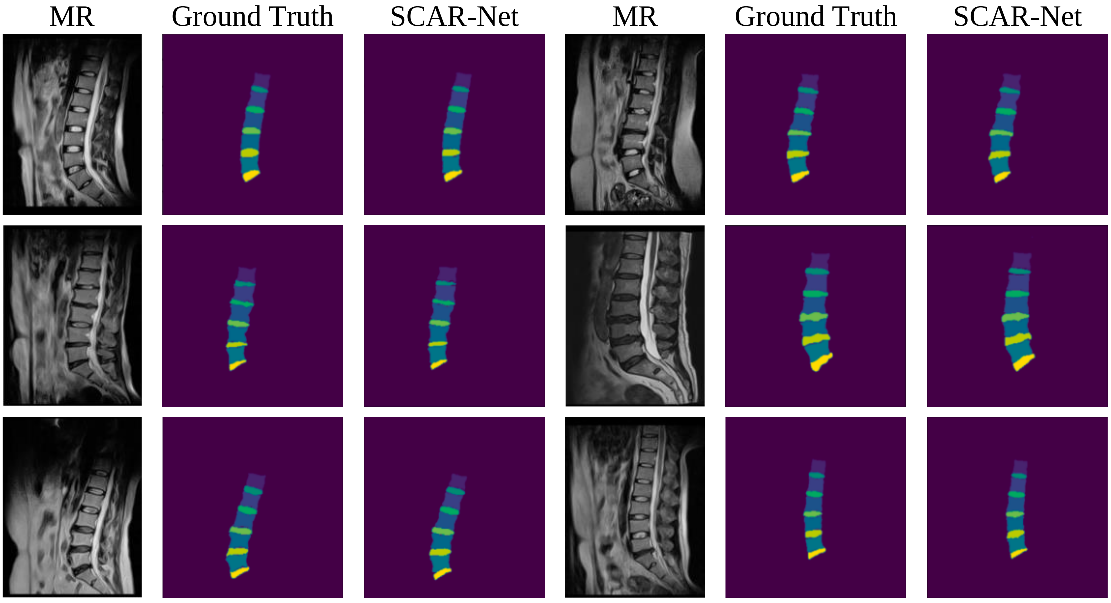

**Figure 2:** Six typical examples to show the SCAR-Net achieves excel- lent performance for spine segmentation.

Figure 2 demonstrate that the proposed SCAR-Net achieves highly accurate spine segmentation, with a mean Precision of 95.50% and a mean DSC of 95.65%, which indicate a significant overlap with the manually delineated mask, striking a balance be- tween under-segmentation and over-segmentation. A comprehensive comparison against state-of-the-art methods confirms the efficacy and the superiority of our approach, with superior quantitative performance measures and enhanced visual accuracy and precision. These findings underscore the potential utility of our proposed network in clinical settings for automating the spine segmentation process with high accuracy.

**Table 2:** The SCAR-Net achieves the highest mean DSC (%) of the most individual spinal structures segmentation.

| Methods | L1 | L2 | L3 | L4 | L5 | L1/L2 | L2/L3 | L3/L4 | L4/L5 | L5/S1 | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|
| MRLN [27] | 83.11 | 83.70 | 98.29 | 92.65 | 94.23 | 79.79 | 82.20 | 88.49 | 89.97 | 89.93 | 88.24 |
| SWDN [20] | 86.45 | 89.03 | 92.02 | 93.49 | 93.41 | 83.21 | 88.05 | 90.49 | 89.42 | 88.87 | 89.44 |
| GCSN [7] | 92.90 | 94.22 | 95.11 | 94.09 | 94.57 | 90.95 | 91.43 | 92.53 | 92.89 | 91.02 | 92.97 |
| SSHS [28] | 90.20 | 89.64 | 89.56 | 89.60 | 88.66 | 91.82 | 90.42 | 92.87 | 90.38 | 90.67 | 90.38 |
| SIMs [29] | 94.05 | 94.54 | 94.28 | 94.21 | 93.06 | 92.77 | 93.37 | 93.42 | 91.87 | 92.65 | 93.42 |
| DGMS [13] | 93.16 | 94.42 | 94.49 | 94.32 | 95.19 | 93.37 | 93.42 | 91.91 | 94.81 | 94.85 | 93.99 |
| SAM [30] | 92.13 | 92.09 | 91.92 | 91.91 | 90.86 | 92.30 | 91.90 | 93.15 | 91.13 | 91.66 | 91.91 |
| CDAN [31] | 92.90 | 91.86 | 92.74 | 92.85 | 92.43 | 89.89 | 90.82 | 92.33 | 91.14 | 90.80 | 91.78 |
| **SCA**R-Net | **96.05** | **96.43** | **95.78** | **96.39** | **96.55** | **95.78** | **96.05** | **93.88** | **93.70** | **95.93** | **95.65** |

As shown in Table 2, compared with other methods, SCAR-Net has significant advantages in DSC. A 2% higher DSC score means that SCAR-Net is more accurate in evaluating similarity indicators. These results demonstrate that the proposed approach effectively balances the inter-class similarity and the intra-class variety of the spine, achieving high integrity and detail accuracy in spine segmentation. By leveraging the strengths of global and local contexts, our model is able to capture fine-grained details and accurately distinguish between the spine and surrounding tissue. Overall, our findings highlight the effectiveness and potential of the proposed SCAR-Net as a valuable tool for clinical spine segmentation.

## 5. CONCLUSIONS

In recent years, automated spine segmentation has garnered significant attention due to the complexity and variability of spinal structures. This task remains challenging because of the need to balance inter-class similarity and intra-class diversity while maintaining high accuracy and generalization ability. To address these challenges, we propose SCAR-Net, a highly efficient and accurate network for automatic segmentation of MR spinal images. SCAR-Net incorporates a multi-channel cross attention (MCCA) module to fuse complementary features of the spine and provide a comprehensive representation of spinal structures. Additionally, a recognition-assisted label fusion (RALF) strategy is introduced to improve generalization by leveraging weakly annotated datasets during training. The proposed SCAR-Net outperforms state-of-the-art methods on several benchmark datasets, demonstrating its effectiveness and generalizability. This method is well-suited for a range of medical applications, including disease diagnosis, preoperative planning, and postoperative evaluation. By providing automated and accurate spine segmentation, SCAR-Net enables clinicians to obtain more reliable measurements, ultimately improving decision-making and patient outcomes.

## ACKNOWLEDGEMENTS

This work was supported by some research projects: Major Projects of Xianyang Science and Technology Bureau, Shaanxi Province, China(L2022-JBGS-GY-01); QinChuangYuan 'Scientist + Engineer' Team Building Project by Shaanxi Provincial Department of Science and Technology, China(2024QCY-KXJ-181); Key Research and Development Plan of Shaanxi Province, China, 2023(2023-YBGY-215), Technology Innovation Guidance Special Project of Shaanxi Province, China(2023GXLH-072).Thanks for the contribution to discussions about these researches.

## REFERENCES

[1] Z. Wu, G. Xia, X. Zhang, F. Zhou, J. Ling, X. Ni, and Y. Li, A novel 3D lumbar vertebrae location and segmentation method based on the fusion envelope of 2D hybrid visual projection images, Computers in Biology and Medicine 151, 106190 (2022).

[2] A. Sekuboyina, J. Kukačka, J. S. Kirschke, B. H. Menze, and A. Valentinitsch, Attention- Driven Deep Learning for Pathological Spine Segmentation, in Computational Meth- ods and Clinical Applications in Musculoskeletal Imaging - 5th International Work- shop, MSKI 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 10, 2017, Revised Selected Papers, edited by B. Glocker, J. Yao, T. Vrtovec, A.F. Frangi, and G. Zheng, volume 10734 of Lecture Notes in Computer Science, pages 108–119, Springer, 2017.

[3] H. Chen, C. Shen, J. Qin, D. Ni, L. Shi, J. C. Cheng, and P.-A. Heng, Automatic localization and identification of vertebrae in spine CT via a joint learning model with deep neural networks, in Medical Image Computing and Computer-Assisted Intervention– MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part I 18, pages 515–522, Springer, 2015.

[4] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueck- ert, Attention gated networks: Learning to leverage salient regions in medical images, Medical image analysis 53, 197–207 (2019).

[5] Z. Han, B. Wei, A. Mercado, S. Leung, and S. Li, Spine-GAN: Semantic segmentation of multiple spinal structures, Medical image analysis 50, 23–35 (2018).

[6] H. Chang, S. Zhao, H. Zheng, Y. Chen, and S. Li, Multi-vertebrae segmentation from arbitrary spine MR images under global view, in Medical Image Computing and Com- puter Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23, pages 702–711, Springer, 2020.

[7] S. Pang, C. Pang, L. Zhao, Y. Chen, Z. Su, Y. Zhou, M. Huang, W. Yang, H. Lu, and Q. Feng, SpineParseNet: spine parsing for volumetric MR image by a two-stage segmentation framework with semantic image representation, IEEE Transactions on Medical Imaging 40, 262–273 (2020).

[8] A. Rasoulian, R. Rohling, and P. Abolmaesumi, Lumbar spine segmentation using a statistical multi-vertebrae anatomical shape+ pose model, IEEE transactions on medical imaging 32, 1890–1900 (2013).

[9] Y. Chen, Y. Gao, K. Li, L. Zhao, and J. Zhao, Vertebrae identification and localization utilizing fully convolutional networks and a hidden Markov model, IEEE Transactions on Medical Imaging 39, 387–399 (2019).

[10] M. Kolařík, R. Burget, V. Uher, K. Říha, and M. K. Dutta, Optimized high resolution 3d dense-u-net network for brain and spine segmentation, Applied Sciences 9, 404 (2019).

[11] X. Wang, S. Wang, Z. Zhang, X. Yin, T. Wang, and N. Li, CPAD-Net: Contextual parallel attention and dilated network for liver tumor segmentation, Biomedical Signal Processing and Control 79, 104258 (2023).

[12] Y. Zhang, L. Yuan, Y. Wang, and J. Zhang, SAU-Net: efficient 3D spine MRI segmenta- tion using inter-slice attention, in Medical Imaging with Deep Learning, pages 903–913, PMLR, 2020.

[13] S. Pang, C. Pang, Z. Su, L. Lin, L. Zhao, Y. Chen, Y. Zhou, H. Lu, and Q. Feng, DGMSNet: Spine segmentation for MR image by a detection-guided mixed-supervised segmentation network, Medical Image Analysis 75, 102261 (2022).

[14] X. Ji, G. Zheng, D. Belavy, and D. Ni, DSMS-FCN: A Deeply Supervised Multi-Scale Fully Convolutional Network for Automatic Segmentation of Intervertebral Disc in 3D MR Images, in Computational Methods

and Clinical Applications for Spine Imaging, edited by J. Yao, T. Vrtovec, G. Zheng, A. F. Frangi, B. Glocker, and S. Li, volume 10182 of Lecture Notes in Computer Science, pages 38–48, Springer, 2016.

[15] G. Zeng and G. Zheng, DSMS-FCN: a deeply supervised multi-scale fully convolu- tional network for automatic segmentation of intervertebral disc in 3D MR images, in Computational Methods and Clinical Applications in Musculoskeletal Imaging: 5th Inter- national Workshop, MSKI 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 10, 2017, Revised Selected Papers 5, pages 148–159, Springer,2018.

[16] M. Aygün, Y. H. Şahin, and G. Ünal, Multi modal convolutional neural networks for brain tumor segmentation, arXiv preprint arXiv:1809.06191 (2018).

[17] J. Dolz, C. Desrosiers, and I. Ben Ayed, IVD-Net: Intervertebral disc localization and segmentation in MRI with a multi-modal UNet, in International workshop and challenge on computational methods and clinical applications for spine imaging, pages 130–143, Springer, 2018.

[18] B. De Leener, J. Cohen-Adad, and S. Kadoury, Automatic segmentation of the spinal cord and spinal canal coupled with vertebral labeling, IEEE transactions on medical imaging 34, 1705–1718 (2015).

[19] S. Hong, H. Noh, and B. Han, Decoupled deep neural network for semi-supervised semantic segmentation, Advances in neural information processing systems 28 (2015).

[20] W. Luo and M. Yang, Semi-supervised semantic segmentation via strong-weak dual- branch network, in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16, pages 784–800, Springer, 2020.

[21] C. Gros et al., Automatic segmentation of the spinal cord and intramedullary multiple sclerosis lesions with convolutional neural networks, Neuroimage 184, 901–915 (2019).

[22] A. Dosovitskiy et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

[23] R. Tao and G. Zheng, Spine-transformers: Vertebra detection and localization in arbi- trary field-of-view spine ct with transformers, in Medical Image Computing and Com- puter Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24, pages 93–103, Springer,2021.

[24] Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich, Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks, in International conference on machine learning, pages 794–803, PMLR, 2018.

[25] P. A. Yushkevich, J. Piven, H. C. Hazlett, R. G. Smith, S. Ho, J. C. Gee, and G. Gerig, User-guided 3D active contour segmentation of anatomical structures: significantly im- proved efficiency and reliability, Neuroimage 31, 1116–1128 (2006).

[26] A. Paszke et al., Pytorch: An imperative style, high-performance deep learning library, Advances in neural information processing systems 32 (2019).

[27] R. Zhang, X. Xiao, Z. Liu, Y. Li, and S. Li, MRLN: Multi-task relational learning network for mri vertebral localization, identification, and segmentation, IEEE journal of biomedical and health informatics 24, 2902–2911 (2020).

[28] M. Huang, S. Zhou, X. Chen, H. Lai, and Q. Feng, Semi-supervised hybrid spine network for segmentation of spine MR images, Computerized Medical Imaging and Graphics 107, 102245 (2023).

[29] I. Castro-Mateos, J. M. Pozo, M. Pereanez, K. Lekadir, A. Lazary, and A. F. Frangi, Statistical interspace models (SIMs): application to robust 3D spine segmentation, IEEE transactions on medical imaging 34, 1663–1675 (2015).

[30] M. A. Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz, and Y. Zhang, Segment anything model for medical image analysis: an experimental study, Medical Image Analysis 89, 102918 (2023).

[31] Z. Li, J. Fang, R. Qiu, H. Gong, W. Zhang, L. Li, and J. Jiang, CDA-Net: A contrastive deep adversarial model for prostate cancer segmentation in MRI images, Biomedical Signal Processing and Control 83, 104622 (2023).