A Two-Stage Network for Segmentation of Vertebrae and Intervertebral Discs: Integration of Efficient Local-Global Fusion Using 3D Transformer and 2D CNN

Zhiqiang Li[1,2], Xiaogen Zhou[1,2], and Tong Tong[1,2,3](✉)

[1] College of Physics and Information Engineering, Fuzhou University, Fuzhou, China
`ttraveltong@gmail.com`

[2] Fujian Key Lab of Medical Instrumentation and Pharmaceutical Technology,
Fuzhou University, Fuzhou, China

[3] Imperial Vision Technology, Fujian, China

Abstract. In the field of computer-aided diagnosis (CAD) for spinal diseases, the fundamental task of multi-label segmentation for vertebrae and intervertebral discs (IVDs) assumes a significant role. However, the distinctive characteristics inherent to the spinal structure pose considerable challenges to the segmentation process, impeding its practical applicability in clinical settings. Convolutional neural networks have been widely used in this task; however, their limited receptive field restricts their capacity to capture extended-range spatial correlations. Consequently, the model's ability to accurately delineate vertebral boundaries is compromised, leading to a notable deterioration in the quality of segmentation outputs. To address this limitation, we propose a novel two-stage convolutional neural network (CNN) framework that incorporates both 3D Transformers and 2D CNNs. By synergistically leveraging the advantages of Transformers in facilitating the integration of long-range dependencies and the ability of CNNs to learn global and local features, our proposed approach exhibits promising potential in enhancing the segmentation performance for vertebrae and intervertebral discs. Moreover, we introduce a graph convolution module into our network architecture to exploit the inherent spatial dependencies present in MRI scans of spinal structures, thereby extracting semantic feature representations and further augmenting the efficacy of segmentation. The evaluation of our proposed method is conducted on the MRSpineSeg Challenge dataset, encompassing T2-weighted MR images.

Keywords: Two-stage · Combined with 3D Transformers and 2D CNN · multi-label · Graph Convolution module · Segmentation of Vertebrae and Intervertebral Discs · deep learning

Supported by National Natural Science Foundation of China under Grant 62171133.

B. Luo et al. (Eds.): ICONIP 2023, CCIS 1964, pp. 467–479, 2024.
https://doi.org/10.1007/978-981-99-8141-0_35

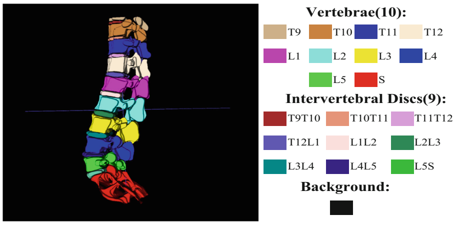

Fig. 1. The task involves the multi-label segmentation of volumetric MR images depicting the vertebrae and intervertebral discs, encompassing 10 distinct labels for vertebrae and 9 for intervertebral discs. It is worth noting that the labels correspond to vertebrae located in the thoracic (T), sacral (S), and lumbar (L) regions.

1 Introduction

The spinal serving as the central axis of the skeletal structure, assumes a vital role in protecting essential organs, blood vessels, and nerves [1]. As the population ages, the incidence of spinal disorders has witnessed a significant increase. In the domain of computer-aided diagnosis and treatment of spine-related diseases, the multi-label segmentation of volumetric magnetic resonance (MR) images pertaining to vertebral bones and intervertebral discs assumes a critical significance. Accurate segmentation of the spinal region, as depicted in Fig. 1, empowers medical practitioners to assess the structural characteristics and overall health of vertebrae and intervertebral discs, thereby facilitating early detection, diagnosis, and surgical planning for various spinal conditions, including deformities, traumas, tumors, and fractures.

Currently, with the progress of artificial intelligence, contemporary medical image spinal segmentation techniques are predominantly built upon two predominant strategies:1) Traditional machine learning based methods. Bao et al. [2] employed a linear iterative clustering algorithm to acquire superpixel MRI images of the spine, enabling the subsequent segmentation of the spinal region. Viji et al. [3] applied a probabilistic boosting tree (PBT) approach in conjunction with fuzzy support vector machine segmentation to achieve automated detection of the spinal canal.2) Deep learning-based methods. In contrast to conventional methodologies, deep learning techniques have demonstrated remarkable efficacy in the domain of spinal segmentation.

Particularly, convolutional neural networks (CNNs) [4–8] have been widely adopted, yielding significant advancements in spinal MR image segmentation. Noteworthy models such as the fully convolutional neural network (FCNNs) [4,5] and U-Net [6,7] have played a prominent role in these advancements. However, the effectiveness of FCNNs is limited by the restricted spatial range of the convolutional layers, impeding the model's ability to capture long-range spatial correlations. Despite the increasing diversity of models employed in spinal segmentation, they often overlook the distinctive chain structure of the spine and neglect the structural interdependencies among neighboring vertebrae and

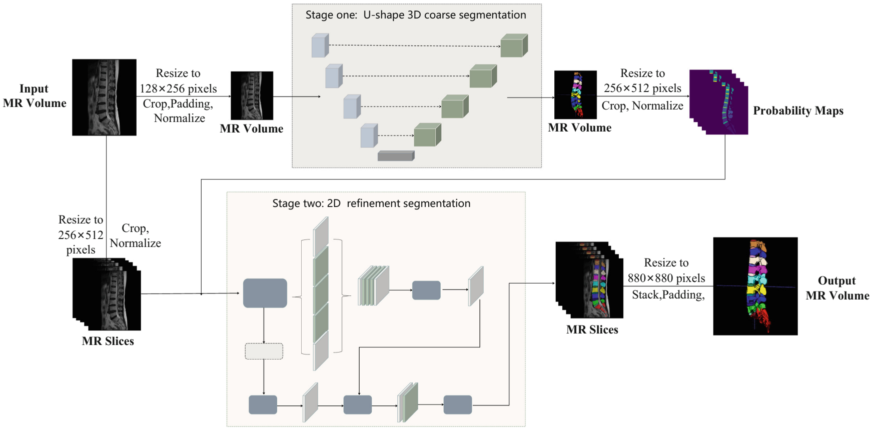
Stage one: U-shape 3D coarse segmentation
Input MR Volume
Resize to 128×256 pixels Crop,Padding, Normalize
MR Volume
MR Volume
Resize to 256×512 pixels Crop, Normalize
Probability Maps
Resize to 256×512 pixels
Crop, Normalize
Stage two: 2D refinement segmentation
MR Slices
MR Slices
Resize to 880×880 pixels Stack,Padding,
Output MR Volume

Fig. 2. Our proposed segmentation network consists of two stages, namely 3D coarse segmentation and 2D refinement segmentation.

lumbar discs. These approaches overlook the holistic architecture of the spine, the persistent long-range dependencies between vertebrae, and the inherent relationships among them. Furthermore, the significant computational and memory requirements associated with these methods impose limitations on their adaptability in diverse spinal segmentation scenarios.

Our work presents the following main contributions:

1. We propose a novel two-stage network architecture designed specifically for the segmentation of biomedical 3D MR images. Our approach involves the integration of a coarsely segmented 3D Transformer to capture long-distance dependencies, along with a finely segmented 2D CNN to capture local high-level features effectively.
2. The incorporation of both 3D and 2D networks enables our model to assimilate a broader range of feature information from images with varying dimensions, thus enhancing its ability to learn diverse representations.
3. To further augment the segmentation performance of our proposed two-stage network, we introduce graph convolution modules within both the 3D and 2D networks. This integration harnesses the power of graph convolution to exploit spatial relationships, leading to improved segmentation outcomes.

2 Methods

2.1 Overall Architecture Design

We presents an innovative methodology for multi-class segmentation, employing a two-stage approach. In particular, we introduce a U-shaped 3D coarse segmentation network, leveraging Transformers as the foundation for the initial segmentation stage, followed by a refinement segmentation network based

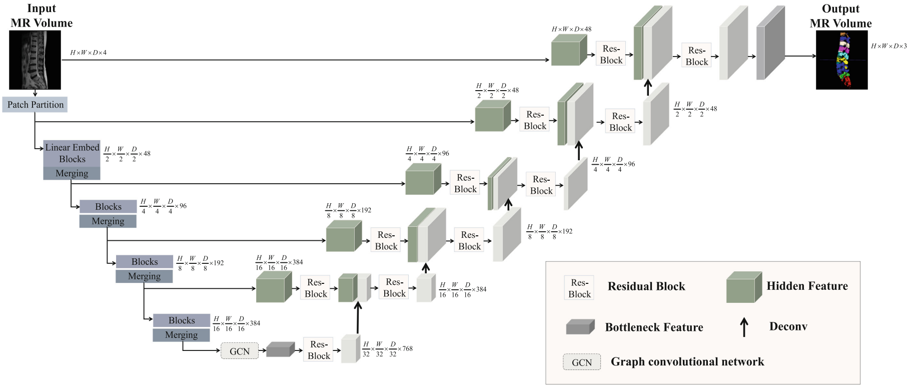
Input
MR Volume
$H \times W \times D \times 4$
Patch Partition
Linear Embed
Blocks
Merging
$\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2} \times 48$
Blocks
Merging
$\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4} \times 96$
Blocks
Merging
$\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8} \times 192$
Blocks
Merging
$\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 384$
GCN
Res-Block
$\frac{H}{32} \times \frac{W}{32} \times \frac{D}{32} \times 768$
$H \times W \times D \times 48$
Res-Block
Res-Block
$\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2} \times 48$
Res-Block
Res-Block
$\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2} \times 48$
$\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4} \times 96$
Res-Block
Res-Block
$\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4} \times 96$
$\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8} \times 192$
Res-Block
Res-Block
$\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8} \times 192$
$\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 384$
Res-Block
Res-Block
$\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 384$
Output
MR Volume
$H \times W \times D \times 3$
Res-Block Residual Block
Hidden Feature
Bottleneck Feature
Deconv
GCN Graph convolutional network

Fig. 3. An outline of the architecture of the 3D coarse segmentation network is presented. The input to the initial segmentation stage consists of 3D multi-modal MRI images with 4 channels. The encoded feature representations in the Swin transformer are transmitted to a CNN-decoder via skip connections at multiple resolutions. The final segmentation output comprises 3 output channels.

on DeepLabv3+ in the subsequent stage. The 3D coarse segmentation network utilizes Swin Transformers as the encoder, which is connected to FCNN-based decoders via skip connections. The decoder generates probability maps for the coarse segmentation task. Subsequently, during the refinement segmentation stage, the volumetric MR image and the probability map derived from the 3D coarse segmentation network serve as inputs for the 2D refinement segmentation network, aiming to achieve more precise and intricate segmentation results. Our proposed two-stage network is specifically tailored for multi-category segmentation of vertebrae and intervertebral discs in volumetric MR images. Figure 2 provides a visual depiction of the network architecture, offering an overview of its structural components.

2.2 3D Coarse Segmentation Stage

Inspired by the effectiveness of the "U-shaped" network architecture, we present a U-shaped 3D coarse segmentation network built upon the Swin Transformer. This network is designed for application during the coarse segmentation stage. The structural configuration of the coarse segmentation network is illustrated in Fig. 3 of this study.

Our coarse segmentation network follows a contracting-expanding pattern, incorporating a stack of transformers as the encoder and establishing connections with the decoder through skip connections. The input token $X \epsilon R^{H \times W \times D \times S}$ to the coarse segmentation network exhibits a patch resolution of $(\hat{H}, \hat{W}, \hat{D})$ and a dimension of $\hat{H} \times \hat{W} \times \hat{D} \times S$. To facilitate the projection of a 3D token sequence with a dimensional parameter $[\frac{H}{\hat{H}}] \times [\frac{W}{\hat{W}}] \times [\frac{D}{\hat{D}}]$ onto an embedding space of dimensional parameter C, we employ a patch partition layer. This layer enables the transformation of the input token sequence into an embedded representation.

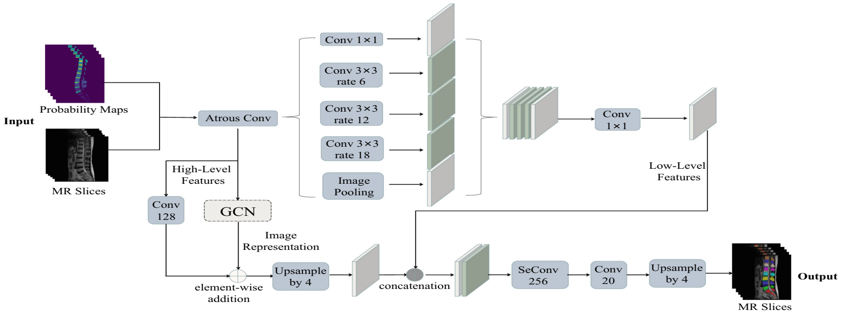

Fig. 4. The architecture of the 2D refinement segmentation network is delineated. The inputs of this network include both the 2D MR sagittal slice and its corresponding coarse probability map, which is produced by the 3D coarse segmentation network.

In order to capture token interactions effectively, we incorporate a self-attention mechanism that operates across non-overlapping windows generated during the partitioning phase. Within the transformer encoder architecture, at a specific layer denoted as l, we employ windows of size M×M×M to evenly divide a 3D token sequence into $[\frac{\hat{H}}{M}] \times [\frac{\hat{D}}{M}] \times [\frac{\hat{W}}{M}]$ regions. These partitioned window segments are subsequently shifted by $([\frac{M}{2}], [\frac{M}{2}], [\frac{M}{2}])$ voxels in layer l+1. Instead of the conventional multi-head self-attention (MSA) module, the Swin Transformer utilizes a shifted windows module, which constrains self-attention calculations to non-overlapping local windows using the shifted windows strategy. This approach not only facilitates efficient computation but also enables the modeling of token dependencies across the entire sequence.

The Swin Transformer module consists of a multi-head self-attention (MSA) module with a shifted window and a two-layer MLP, embedded between Gaussian Error Linear Units (GELU) nonlinearities. Prior to each MSA and MLP module, a LayerNorm (LN) layer is applied. Moreover, residual connections are established between two Swin Transformer modules, enhancing information flow within the network. The introduction of the shifted window division method in the Swin Transformer module optimizes its computational efficiency. The calculation process of the Swin Transformer module, employing this method, can be outlined as follows:

$$\hat{Z}^l = W - MSA(LN(Z^{l-1})) + Z^{l-1} \tag{1}$$

$$Z^l = MLP(LN(\hat{Z}^l)) + \hat{Z}^l \tag{2}$$

$$\hat{Z}^{l+1} = SW - MSA(LN(\hat{Z}^l)) + Z^l \tag{3}$$

$$Z^{l+1} = MLP(LN(\hat{Z}^{l+1})) + \hat{Z}^{l+1} \tag{4}$$

where $\hat{Z}^l$ and Z^l stand for the (S) W-MSA modules and the MLP module's respective block l output characteristics. Similar to other studies [9,10], the following formula is used to calculate self-attention:

$$Attention(Q, K, V) = SoftMax(\frac{QK^T}{\sqrt{d}} + B)V \tag{5}$$

where d is the query/key dimension,M^2 is the number of patches in a window, and Q, K, and V are the queries, key, and value metrics. Since the range of the relative position along each axis is $[-M+1, M-1]$, we parameterize a bias matrix with a smaller size, $B \in R^{(2M-1)\times(2M-1)}$, and values in B are obtained from $\hat{B}$.

2.3 2D Refinement Segmentation Stage

During the 2D segmentation stage, our methodology is primarily guided by the design principles of DeepLabv3+ [?]. In the encoder phase, we employ parallel atrous convolution at multiple rates, commonly referred to as Atrous Spatial Pyramid Pooling (ASPP) [11], to effectively encode multi-scale context information. For the segmentation task, we adopt the Xception architecture and incorporate depthwise separable convolution to enhance both the efficiency and precision of network training. Furthermore, to refine the segmentation outcomes, we introduce a straightforward yet highly efficacious decoder module, which builds upon the aforementioned foundation. The architectural details of the refinement segmentation network are presented in Fig. 4.

The 2D refinement segmentation network takes as input the 2D MR sagittal slice and the coarse probability map corresponding to that slice, which is generated by the 3D coarse segmentation network. Incorporating the coarse probability map enables the 2D refinement segmentation network to leverage the implicit 3D semantic information of the image. By effectively integrating the semantic features of the spinal structure with detailed information, the network achieves accurate segmentation. The high-resolution MR slices contain detailed information, and the 2D refinement segmentation network combines this information with the 3D semantic information to produce fine segmentation.

2.4 Graph Convolution Module

The graph convolution module consists of three consecutive stages: Pooling, Graph Convolutional Network (GCN), and Unpooling. In the Pooling stage, the input image representation is transformed into a graph-based representation to facilitate subsequent processing by the GCN stage. The GCN stage aims to generate graph representations enriched with semantic information through the application of graph convolution operations. In the final Unpooling stage, the obtained semantic graph representation is mapped back to the semantic image representation and passed to the convolution layer for further processing.

Table 1. The mean DSC (%) for the proposed method and other methods on the MRSpineSeg Challenge dataset.

Methods	T9	T10	T11	T12	L1	L2	L3	L4	L5	S
nnUNet	0.00 ± 0.00	3.07 ± 15.02	70.71 ± 28.28	79.86 ± 18.10	80.55 ± 16.81	79.17 ± 18.17	80.59 ± 14.55	84.44 ± 8.43	85.58 ± 8.83	86.71 ± 2.07
VNet	0.00 ± 0.00	28.96 ± 29.87	72.45 ± 27.34	80.64 ± 20.83	81.98 ± 19.47	84.33 ± 13.57	86.63 ± 4.70	86.69 ± 3.31	86.16 ± 4.03	85.54 ± 2.49
UNETR	0.00 ± 0.00	26.17 ± 35.85	66.99 ± 26.52	75.66 ± 18.59	79.23 ± 16.82	79.23 ± 15.32	79.21 ± 15.25	80.80 ± 13.49	82.80 ± 7.99	83.43 ± 4.59
3D Graphonomy	20.78 ± 30.97	44.32 ± 38.50	75.67 ± 23.83	82.14 ± 14.62	83.56 ± 14.73	82.22 ± 13.82	82.65 ± 12.51	82.80 ± 12.88	84.48 ± 10.82	82.52 ± 4.28
3D Deeplabv3+2D ResUNet	24.18 ± 25.39	49.91 ± 37.25	78.86 ± 23.43	86.59 ± 13.35	**88.20 ± 9.62**	87.67 ± 7.88	87.27 ± 6.78	86.76 ± 7.04	86.93 ± 6.11	**87.58 ± 3.45**
3D Graphonomy+2D Deeplabv3	23.59 ± 23.12	44.77 ± 35.39	77.09 ± 23.52	84.78 ± 13.90	86.27 ± 13.09	86.07 ± 12.69	86.35 ± 11.52	85.92 ± 11.57	85.87 ± 9.96	85.82 ± 3.33
Ours	**31.12 ± 21.99**	**56.90 ± 34.25**	**80.75 ± 20.34**	**87.34 ± 9.74**	88.19 ± 9.76	**87.68 ± 9.64**	**88.54 ± 3.83**	**88.31 ± 3.15**	**87.83 ± 3.08**	87.53 ± 2.54

Table 2. The mean DSC (%) for the proposed method and other methods on the MRSpineSeg Challenge dataset.

Methods	T9T10	T10T11	T11T12	T12L1	L1L2	L2L3	L3L4	L4L5	L5S
nnUNet	0.00 ± 0.00	0.00 ± 0.00	74.78 ± 26.38	81.19±19.41	80.44 ± 19.75	81.02 ± 19.13	85.42 ± 13.41	85.37 ± 9.59	85.07 ± 10.09
VNet	0.00 ± 0.00	44.97 ± 33.05	78.68 ± 24.01	83.21 ± 21.37	86.17 ± 14.78	87.17 ± 13.47	89.11 ± 4.05	86.52 ± 7.10	84.83 ± 8.12
UNETR	0.00 ± 0.00	42.97 ± 38.16	73.31 ± 28.29	76.48 ± 21.08	78.53 ± 22.19	80.74 ± 18.55	81.39 ± 16.30	80.69 ± 15.22	82.44 ± 7.76
3D Graphonomy	22.55 ± 39.85	61.39 ± 30.74	80.01 ± 20.09	83.07 ± 15.86	83.84 ± 17.79	83.54 ± 15.16	83.77 ± 15.31	82.42 ± 12.32	82.33 ± 12.38
3D Deeplabv3 + 2D ResUNet	27.15 ± 36.07	74.05 ± 28.64	84.22 ± 20.74	87.78 ± 13.74	89.09 ± 10.75	88.07 ± 12.83	88.34 ± 8.22	85.86 ± 7.39	85.65 ± 11.95
3D Graphonomy + 2D Deeplabv3	26.42 ± 35.10	73.52 ± 25.93	84.33 ± 18.65	87.07 ± 14.31	87.30 ± 15.28	87.11 ± 14.12	87.37 ± 13.96	85.81 ± 11.43	85.72 ± 10.57
Ours	**28.35 ± 34.11**	**76.92 ± 22.48**	**86.03 ± 16.54**	**88.91 ± 10.06**	**89.37 ± 9.94**	**88.83 ± 10.12**	**89.99 ± 4.02**	**87.47 ± 5.45**	**86.73 ± 7.62**

3 Experiments

3.1 Dataset

Our proposed method was evaluated on the MRSpineSeg Challenge dataset, which comprises a total of 215 T2-weighted MR volumetric images. During the experiment, 172 images were utilized, and they were partitioned into training, validation, and testing sets in a ratio of 7:2:1. The volumetric images encompassed 10 vertebrae, 9 intervertebral discs (IVDs), and backgrounds, resulting in a total of 20 distinct categories. The original images exhibited varying dimensions, with widths and heights ranging from 512 to 1024, while the number of slices along the coronal axis ranged from 12 to 20.

3.2 Implementation Details

For 3D networks, the pre-processing stage comprises a series of steps aimed at preparing the input data. These steps include cropping, resizing, padding, and normalization. To begin with, the cropping step involves center-cropping the images along the depth direction to eliminate the non-spine portion, as half of the image does not contain spinal information. Subsequently, the cropped image is resized to a dimension of $18 \times 256 \times 128$ pixels, with zero filling applied in the depth direction to ensure uniformity. Lastly, the normalization process involves computing the mean and variance values across all the images. These values are then utilized to subtract the mean from each pixel and divide by their standard deviation, resulting in a normalized representation of the data.

Our methodology was implemented using the Python programming language based on the PyTorch deep learning framework. The model was trained on an Nvidia RTX 3090 GPU with 24 GB of RAM. During the 3D segmentation stage, a preliminary probability map of dimensions $20 \times 18 \times 256 \times 128$ was generated,

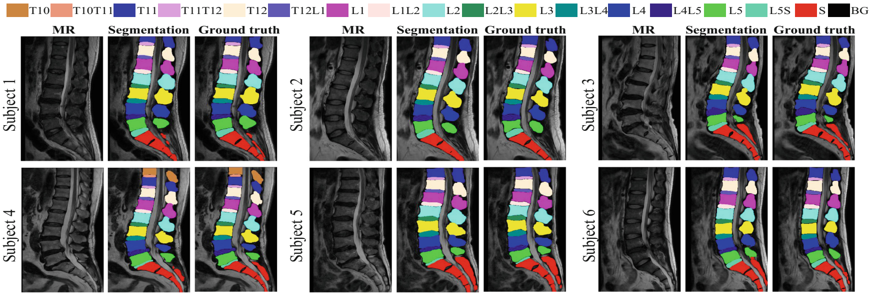

Fig. 5. Using our segmentation method, sagittal slices depicting vertebrae and intervertebral discs from six subjects were acquired. The label BG means the background in these slices.

with an MR volume of size 16 × 256 × 128 serving as input. The Adam [13] optimizer was employed for optimization, with a weight decay of 0.0001. We initiated the learning rate at 0.001, and reduced it by a factor of 5 every 33 epochs. The batch size was set to 2, which was limited by the available GPU memory.

3.3 Evaluation Metrics

To assess the segmentation performance, several metrics were employed in our experiment, including the Dice similarity coefficient, precision(DSC), and recall. These metrics are computed as follows:

$$Dice = \frac{2TP}{FP + 2TP + FN}. \tag{6}$$

Table 3. Ablation experiments were conducted on the MRSpineSeg Challenge dataset to assess the effectiveness of each component in segmentation of the ten classes of vertebrae T9-S. The mean DSC (%) was used to validate the components.

Coarse Segmentation		Refinement Segmentation		T9	T10	T11	T12	L1	L2	L3	L4	L5	S
3D Swin Transform	3D GCM	2D Deeplabv3+	2D GCM										
✓				21.70 ± 29.97	47.68 ± 34.94	77.39 ± 23.17	84.89 ± 16.34	85.04 ± 15.56	85.48 ± 13.24	86.22 ± 4.78	86.12 ± 3.40	85.53 ± 4.16	85.27 ± 2.82
✓	✓			25.41 ± 24.71	50.19 ± 36.35	78.87 ± 23.12	85.11 ± 16.35	86.14 ± 15.57	86.53 ± 13.26	87.28 ± 4.77	87.15 ± 3.41	86.62 ± 4.13	86.36 ± 2.82
✓		✓		27.11 ± 21.65	53.99 ± 34.63	79.25 ± 22.11	86.49 ± 11.21	87.64 ± 10.37	87.23 ± 10.17	87.81 ± 6.91	87.55 ± 7.26	87.34 ± 4.53	87.27 ± 2.75
✓		✓	✓	28.23 ± 21.47	54.78 ± 33.63	80.37 ± 22.14	86.23 ± 11.20	87.73 ± 10.4	87.29 ± 10.16	87.86 ± 6.90	87.59 ± 7.25	87.41 ± 4.50	87.33 ± 2.72
✓	✓	✓		28.59 ± 21.39	54.39 ± 33.25	80.28 ± 20.44	87.12 ± 9.76	88.10 ± 9.77	87.63 ± 9.61	88.49 ± 3.85	88.28 ± 3.15	87.74 ± 3.13	87.44 ± 2.54
✓	✓	✓	✓	**31.12 ± 21.99**	**56.90 ± 34.25**	**80.75 ± 20.34**	**87.34 ± 9.74**	**88.19 ± 9.76**	**87.68 ± 9.64**	**88.54 ± 3.83**	**88.31 ± 3.15**	**87.83 ± 3.08**	**87.53 ± 2.54**

Table 4. Ablation experiments were conducted on the MRSpineSeg Challenge dataset to assess the effectiveness of each component in segmentation of the nine classes of IVDs T9T10-L5S. The mean DSC (%) was used to validate the components.

Coarse Segmentation		Refinement Segmentation		T9T10	T10T11	T11T12	T12L1	L1L2	L2L3	L3L4	L4L5	L5S
Coarse Segmentation	3D GCM	2D Deeplabv3+	2D GCM									
✓				20.58 ± 25.94	67.33 ± 29.91	81.44 ± 21.70	85.70 ± 16.57	86.96 ± 14.31	86.71 ± 13.70	87.86 ± 4.11	85.37 ± 5.50	84.29 ± 8.15
✓	✓			24.81 ± 35.02	69.92 ± 29.24	82.88 ± 21.65	86.68 ± 16.56	87.92 ± 14.31	87.78 ± 13.73	88.88 ± 4.13	86.37 ± 5.49	85.36 ± 8.11
✓		✓		25.18 ± 26.91	71.08 ± 30.00	83.73 ± 24.41	87.59 ± 10.52	88.23 ± 10.07	88.03 ± 11.61	89.57 ± 6.35	86.88 ± 8.78	86.57 ± 7.66
✓		✓	✓	26.35 ± 26.91	72.92 ± 27.34	84.27 ± 21.00	88.60 ± 10.51	89.26 ± 10.07	88.21 ± 11.61	89.26 ± 10.07	86.95 ± 8.78	86.62 ± 7.66
✓	✓	✓		27.11 ± 34.66	74.33 ± 24.10	85.59 ± 16.69	**88.94 ± 10.07**	**89.41 ± 9.92**	88.75 ± 10.09	89.97 ± 4.00	87.44 ± 5.47	86.65 ± 7.67
✓	✓	✓	✓	**28.35 ± 34.11**	**76.92 ± 22.48**	**86.03 ± 16.54**	88.91 ± 10.06	89.37 ± 9.94	**88.83 ± 10.12**	**89.99 ± 4.02**	**87.47 ± 5.45**	**86.73 ± 7.62**

$$Pre = \frac{TP}{TP + FP}. \tag{7}$$

$$Recall = \frac{TP}{TP + FN}. \tag{8}$$

where TP, FP, FN, and TN denote the number of true positives, false positives, false negatives, and true negatives, respectively.

3.4 Experiment Results

The Table 5 displays the precise values of Mean Recall, Mean Precision, and Dice Similarity Coefficient (DSC) achieved by the two-stage segmentation network for vertebrae, intervertebral discs (IVD), and all 19 spinal structures. We have presented some exemplary images with well-performing segmentation results in Fig. 5.

We conducted a comparative analysis of our proposed spinal segmentation method with several other methods, including nnUNet [14], VNet [15], UNETR [16], 3D Graphonomy [12], 3D Deeplabv3 [11] + 2D ResidualUNet [17], and 3D Graphonomy [12] + 2D Deeplabv3 [11]. The evaluation of the segmentation performance across these methods was based on three crucial metrics: the Dice similarity coefficient (DSC), Precision, and Recall. Tables 1 and 2 present the DSC evaluation indexes specifically for the segmentation of each vertebra and intervertebral disc (IVD). Our proposed segmentation network demonstrated superior performance compared to the other methods, achieving excellent segmentation results for the seven categories of vertebrae T12-S (DSC $>$ 87.34%) and the seven categories of IVDs T11-S (DSC $>$ 86.03%). These quantitative comparison results highlight the notable superiority of our proposed methodology. Furthermore, Fig. 6 showcases specific segmentation results obtained by applying different algorithms to the aforementioned dataset, providing visual evidence of the superior segmentation outcomes achieved by our proposed method.

Table 5. The average values of Recall, Precision, and Recall were computed for the segmentation of vertebrae, intervertebral discs (IVDs), and all 19 spinal structures using our proposed two-stage segmentation network.

	Mean Recall	Mean Precision	Mean Dice
Background	98.96 ± 0.30	98.99 ± 0.50	98.97 ± 0.21
Vertebrae	85.87 ± 8.46	86.16 ± 6.54	85.61 ± 6.68
IVDs	88.76 ± 8.31	85.95 ± 6.71	86.96 ± 7.40
Overall	87.19 ± 7.84	86.04 ± 5.85	86.21 ± 6.57

The performance of our network on the segmentation of T9-T11 vertebrae (DSC $\leq$ 80.75%) and T9-T12 IVDs (DSC $\leq$ 76.92%) is unsatisfactory due to several factors. Firstly, the dataset contains very limited samples of T9-T11 vertebrae and T9-T12 IVDs, with most of them being incompletely shaped. Secondly, the top of the image contains three types of vertebrae (T9-T11) and two types of IVDs (T9-T12), making segmentation difficult due to the limited receptive field at the top. These factors contribute to the suboptimal segmentation results of our network in these regions.

3.5 Ablation Study

To assess the efficacy of each constituent element within our network architecture, a series of ablation experiments were conducted, yielding results that have been presented in Tables 3 and 4. The evaluation process encompassed six distinct configurations involving the integration of the 3D Swin Transform and 3D GCM during the 3D Coarse Segmentation stage, as well as the utilization of the 2D Deeplabv3+ and 2D GCM during the 2D Refinement Segmentation stage. These meticulous experiments effectively demonstrated the augmentation of segmentation performance for both Vertebrae and IVDs through the inclusion of the graph convolutional module and the employment of a dual network strategy.

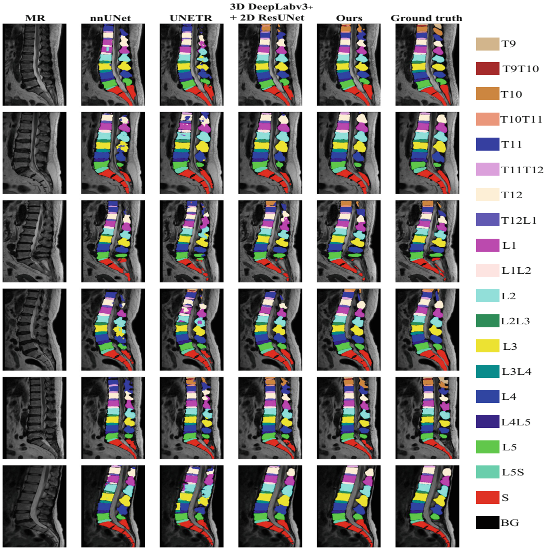

Fig. 6. Visualized comparison of results using different segmentation networks.

3.6 Effect of the Two-Stage Framework

The incorporation of 2D refinement stages into 3D segmentation tasks has demonstrated considerable effectiveness in enhancing the performance of segmentation algorithms. This enhancement is substantiated by the findings presented in Tables 1 and 2, which elucidate the improvements attained through the integration of 2D refinement in the 3D Graphonomy and 3D Graphonomy+2D Deeplabv3 frameworks, respectively. In comparison to the sole utilization of 3D Graphonomy, the inclusion of a 2D refinement stage within the 3D Graphonomy+2D Deeplabv3 approach yielded a notable increase in the average Dice similarity coefficient (DSC) across the eight classes of vertebrae T11-S and T10T11-L5S, as well as the eight classes of intervertebral discs (IVDs). Similarly encouraging results were obtained from the ablation experiments conducted, as evidenced by the outcomes presented in Tables 3 and 4. The incorporation of high-resolution images within the 2D networks contributed to a more comprehensive representation of detailed information, enabling the model to acquire a richer understanding of the underlying features. Consequently, the synergistic combination of 3D and 2D networks facilitated the assimilation of more contextual information from images with varying dimensions, culminating in a discernible enhancement in segmentation performance.

3.7 Effect of the GCM

The quantitative findings elucidated in Tables 3 and 4 provide compelling evidence that the incorporation of the graph convolution module into either the 3D or 2D network yields notable advantages in enhancing the segmentation performance of the model. Notably, it should be acknowledged that during the training phase of both the 3D and 2D networks, the segmentation results may not strictly adhere to the spatial order of the spinal structure. Consequently, the inclusion of the graph convolution module in both the 3D and 2D networks emerges as a more favorable approach for boundary position segmentation.

4 Conclusion

This paper presents a novel two-stage framework designed for achieving precise multi-label segmentation of vertebrae and intervertebral discs. The proposed framework integrates 3D transformers and 2D convolutional neural networks (CNN) to attain accurate and reliable segmentation outcomes. In the initial stage of the framework, 3D transformers are employed to generate preliminary probability graphs, thereby establishing a foundation for subsequent processing. Subsequently, in the second stage, the 2D MR sagittal slice and the corresponding rough probability graph derived from the 3D rough segmentation network are jointly inputted into the 2D network to achieve refined segmentation results with heightened precision. Notably, the integration of graph convolution modules within both the 3D and 2D networks plays a crucial role in addressing pertinent challenges associated with pixel labeling isolation, as well as rectifying errors pertaining to shape and positional segmentation outcomes. These modules contribute to the enhancement of segmentation accuracy by effectively resolving issues related to isolation and correction within the segmentation process. Through comprehensive comparisons with state-of-the-art spinal segmentation methodologies utilizing publicly available datasets, the proposed framework has exhibited superior performance, underscoring its efficacy and potential for advancing the field of spinal segmentation.

References

1. Lopez, I.B., Benzakour, A., Mavrogenis, A., Benzakour, T., Ahmad, A., Lemee, J.-M.: Robotics in spine surgery: systematic review of literature. Int. Orthop. **47**(2), 447–456 (2023)
2. Bao, X.-X., et al.: Recognition of necrotic regions in MRI images of chronic spinal cord injury based on Superpixel. Comput. Methods Programs Biomed. **228**, 107252 (2023)
3. Viji, C., Rajkumar, N., Suganthi, S., Venkatachalam, K., Kumar, T.R., Pandiyan, S.: An improved approach for automatic spine canal segmentation using probabilistic boosting tree (PBT) with fuzzy support vector machine. J. Ambient. Intell. Humaniz. Comput. **12**, 6527–6536 (2021)
4. Pang, S., et al.: SpineParseNet: spine parsing for volumetric MR image by a two-stage segmentation framework with semantic image representation. IEEE Trans. Med. Imaging **40**(1), 262–273 (2020)
5. Pang, S., et al.: DGMSNet: spine segmentation for MR image by a detection-guided mixed-supervised segmentation network. Med. Image Anal. **75**, 102261 (2022)
6. Yang, Z., Wang, Q., Zeng, J., Qin, P., Chai, R., Sun, D.: RAU-Net: u-net network based on residual multi-scale fusion and attention skip layer for overall spine segmentation. Mach. Vis. Appl. **34**(1), 10 (2023)
7. Wang, B., Qin, J., Lv, L., Cheng, M., Li, L., Xia, D., Wang, S.: MLKCA-Unet: multiscale large-kernel convolution and attention in UNet for spine MRI segmentation. Optik **272**, 170277 (2023)
8. Tao, R., Zheng, G.: Spine-transformers: vertebra detection and localization in arbitrary field-of-view spine CT with transformers. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12903, pp. 93–103. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87199-4_9
9. Hu, H., Gu, J., Zhang, Z., Dai, J., Wei, Y.: Relation networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3588–3597 (2018)
10. Hu, H., Zhang, Z., Xie, Z., Lin, S.: Local relation networks for image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3464–3473 (2019)
11. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: Semantic image segmentation with deep convolutional nets, Atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2017)
12. Gong, K., Gao, Y., Liang, X., Shen, X., Wang, M., Lin, L.: Graphonomy: universal human parsing via graph transfer learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7450–7459 (2019)
13. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
14. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)

15. Milletari, F., Navab, N., Ahmadi, S.-A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571. IEEE (2016)
16. Hatamizadeh, A., et al.: UNETR: transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 574–584 (2022)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)