Computerized Medical Imaging and Graphics 99 (2022) 102091

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag

Spine-GFlow: A hybrid learning framework for robust multi-tissue segmentation in lumbar MRI without manual annotation

Xihe Kuang [a,1], Jason Pui Yin Cheung [a,1], Kwan-Yee K. Wong [b], Wai Yi Lam [a], Chak Hei Lam [a], Richard W. Choy [a], Christopher P. Cheng [c], Honghan Wu [a], Cao Yang [d], Kun Wang [d], Yang Li [e,*], Teng Zhang [a,*]

[a] Department of Orthopaedics and Traumatology, Li Ka Shing Faculty of Medicine, University of Hong Kong, Hong Kong, China
[b] Department of Computer Science, Faculty of Engineering, University of Hong Kong, Hong Kong, China
[c] Department of Computer Science, Dartmouth College, Hanover, NH 03755, USA
[d] Department of Orthopedics, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430022, China
[e] School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China

ARTICLE INFO

Keywords:
Multi-tissue segmentation
Auto-annotation
Weakly-supervised segmentation
Spine

ABSTRACT

Most learning-based magnetic resonance image (MRI) segmentation methods rely on the manual annotation to provide supervision, which is extremely tedious, especially when multiple anatomical structures are required. In this work, we aim to develop a hybrid framework named Spine-GFlow that combines the image features learned by a CNN model and anatomical priors for multi-tissue segmentation in a sagittal lumbar MRI. Our framework does not require any manual annotation and is robust against image feature variation caused by different image settings and/or underlying pathology. Our contributions include: 1) a rule-based method that automatically generates the weak annotation (initial seed area), 2) a novel proposal generation method that integrates the multi-scale image features and anatomical prior, 3) a comprehensive loss for CNN training that optimizes the pixel classification and feature distribution simultaneously. Our Spine-GFlow has been validated on 2 independent datasets: HKDDC (containing images obtained from 3 different machines) and IVDM3Seg. The segmentation results of vertebral bodies (VB), intervertebral discs (IVD), and spinal canal (SC) are evaluated quantitatively using intersection over union (IoU) and the Dice coefficient. Results show that our method, without requiring manual annotation, has achieved a segmentation performance comparable to a model trained with full supervision (mean Dice 0.914 vs 0.916).

1. Introduction

MRIs are widely used in the clinic for the diagnosis of degenerative lumbar diseases (Benneker et al., 2005; Cheung et al., 2019; Jensen et al., 1994; Lai et al., 2021a, 2021b; Pfirrmann et al., 2001). Since an MRI allows the visualization of the 3D structure of soft tissues including intervertebral discs (IVD) and the spinal canal (SC) (Fig. 1 A and B), it is the gold standard for the assessment of IVD herniation (Benneker et al., 2005; Pfirrmann et al., 2001) and spinal stenosis (Cheung et al., 2019; Lai et al., 2021a, 2021b). Currently, analysis of lumbar MRIs relies heavily on the experience and subjective judgment of specialists, which makes the process laborious and potentially inaccurate with inevitable interrater variations. Thus, automated and objective lumbar MRI assessments are highly desirable. Semantic segmentation is important for auto-analysis of lumbar MRIs as it provides the locations and pixel-wise anatomical information of spinal tissues, which serve as precursors for further pathology and disease progression predictions.

Conventional semantic segmentation methods for lumbar MRI are rule-based and based on graphical and anatomical priors of target tissue (Carballido-Gamio et al., 2004; Egger et al., 2012; He et al., 2017; Michopoulou et al., 2009; Neubert et al., 2012). Pre-determined templates, detectors, and rules are manually designed for the segmentation task. However, these rule-based methods are not robust against the highly variable image features in MRI caused by systematic and/or individual deviations (Cheng and Halchenko, 2020). The systematic deviation is usually caused by different MRI protocols, equipment settings,

* Corresponding authors.
E-mail addresses: liyang@buaa.edu.cn (Y. Li), tgzhang@hku.hk (T. Zhang).
[1] The first two authors contributed equally to this publication

https://doi.org/10.1016/j.compmedimag.2022.102091
Received 14 February 2022; Received in revised form 30 May 2022; Accepted 13 June 2022
Available online 16 June 2022

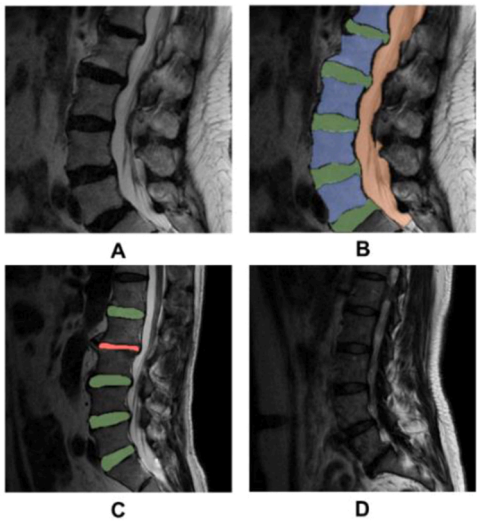

Fig. 1. A and B present an example of a sagittal lumbar MRI that clearly shows multiple spinal tissues including vertebral bodies (blue), intervertebral discs (green), and the spinal canal (orange). C illustrates serious shape distortion of an intervertebral disc (red) due to disc degeneration. D presents an MRI with low image quality including low pixel intensity, low contrast, unclear edges, and noise. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and human operations, which are common when MRIs are obtained from different institutions. The individual deviation is usually caused by underlying pathologies, such as shape and alignment deformity, which are random and vary widely between individuals. Several examples of image feature variation (Fig. 1C and D), including shape distortion, low pixel intensity, low contrast, unclear edges, and noise, can be observed. These rule-based methods can detect approximate tissue locations but often fail to obtain accurate shape information. As a result, they are not suitable to be used directly in clinical practice. Furthermore, these rule-based methods are usually designed based on specific tissue, thus they can only segment a single tissue. Multi-tissue segmentation is important considering that clinical diagnosis often requires a comprehensive analysis of multiple tissues (Lai et al., 2021a, 2021b; Pfirrmann et al., 2001).

Recently, with the rapid development of convolutional neural networks (CNN), learning-based methods have achieved remarkable performance in semantic segmentation. For medical images, a CNN model trained with full pixel-wise annotation (full-supervision) can obtain accuracy comparable to clinical specialists (Lu et al., 2018; Ronneberger et al., 2015; Zhou et al., 2019). However, the required manual annotation is extremely laborious and time-consuming, which makes full-supervision costly and large-scale annotated datasets scarce. To address this limitation, weakly-supervised methods were developed. Weakly-supervised methods train models with weak annotations (Kervadec et al., 2019, 2020; Qu et al., 2020; Rajchl et al., 2016; Valvano et al., 2021; Yoo et al., 2019), which can significantly reduce the cost of full-supervision; priors of tissues such as pixel value, shape, and size are usually utilized to support training. Nevertheless, for 3D images such as MRI and CT, weak annotation is still expensive since each slice needs to be annotated separately. Furthermore, since the CNN model is data-sensitive and vulnerable to the variation of image features, new annotations may be required to fine-tune the model for images acquired under a different setting, and the well-trained model may also fail to the case with underlying pathology.

We combine rule-based and learning-based methods, and propose a hybrid framework for multi-tissue segmentation in lumbar MRI that requires no manual annotation. A rule-based method is designed to automatically generate the incomplete (in a few MRI slices) and inaccurate (missing and location deviation) weak annotation. It first identifies approximate tissue locations and a rough spinal region, and further determines the initial seed areas. We then propose an iterative optimization procedure to train a CNN model with the initial seed areas. The CNN model can generate multi-scale feature maps and pixel classification from MRI. The optimization procedure iterates between two steps: 1) proposal generation and 2) CNN training. In proposal generation, we integrate the multi-level information within the multi-scale feature maps to produce the segmentation proposals based on the seed areas. The rule-based proposal fine-tuning is adopted to explicitly embed the anatomical prior. In CNN training, a comprehensive loss is adopted to optimize the pixel classification and feature distribution of feature maps simultaneously based on the proposals. We hypothesize that with the iterative optimization procedure, our framework can gradually optimize the proposals and CNN model, and the optimized CNN model can produce accurate multi-tissue segmentation in the lumbar MRI. Since no manual annotation is required in our framework, it can automatically fine-tune the CNN model on the target MRI, which can effectively improve the robustness of the model against image feature variation caused by different image settings and/or underlying pathology. Unlike other unsupervised segmentation methods (Harb and Knöbelreiter, 2021; Hwang et al., 2019; Mirsadeghi et al., 2021; Van Gansbeke et al., 2021) that do not use any annotation in the training process, our framework utilizes automatic annotation, which can guide the model to generate more semantic features, rather than focusing on the shallow image features.

We aim at establishing and evaluating a hybrid framework, named Spine-GFlow, for the robust segmentation of multiple tissues including vertebral bodies (VB), IVD, and SC in sagittal lumbar MRI images without relying on any manual annotation or human intervention. The name is derived because, 1) this framework is specifically tuned based on the anatomical knowledge of the spine, which is a complex organ consisting of multiple types of tissues; and 2) "G" stands for "Generative" as we do not require manual annotations but generating masks automatically. Our objectives include: 1) designing a rule-based method that automatically generates the weak annotation (initial seed area), 2) developing a proposal generation method that integrates the multi-scale feature maps and anatomical prior, 3) training the CNN model with a comprehensive loss that optimizes the pixel classification and feature distribution of feature maps simultaneously, 4) validating our framework and comparing with other existing methods.

2. Related work

2.1. Rule-based spine/lumbar MRI segmentation

Rule-based segmentation methods for spine/lumbar MRI are usually developed based on the graphical or anatomical priors of specific tissues. Normalized cut (NCut) was adopted (Carballido-Gamio et al., 2004) to segment VBs from midline sagittal spine MRIs. A multi-feature and adaptive spectral segmentation was proposed (He et al., 2017) to segment spinal neural foramina within preselected ROIs. A statistics-based method was proposed (Neubert et al., 2012) to segment IVD and VB with statistical shape analysis and registration of grey level intensity profiles. An atlas-based segmentation method for IVD that relied on manually-designed templates was proposed (Michopoulou et al., 2009). Shape information was utilized (Egger et al., 2012) to produce the segmentation of VB, which relied on manually selected seed points for initialization. All rule-based methods above could only produce the segmentation of one kind of tissue at a time, and modification was required to transfer these methods to different tissues. Additionally, some methods required human intervention to guide segmentation (Egger et al., 2012; He et al., 2017).

2.2. Weakly-supervised segmentation

Training a model for a segmentation task with weak annotations such as image tag (Pathak et al., 2015), bounding boxes (Dai et al., 2015; Khoreva et al., 2017; Kulharia et al., 2020; Lee et al., 2021; Song et al., 2019), scribbles (Lin et al., 2016; Tang et al., 2018), and points (Bearman et al., 2016) is an attractive problem. A key idea for weakly-supervised segmentation is to integrate the priors about the object (shape, size, relative location, etc.) and image (color, texture, brightness, etc.) in the training process. BoxSup (Dai et al., 2015) proposed an iterative procedure that iterates between proposal generation and model training to gradually improve the proposals and the model. Previous work (Khoreva et al., 2017) demonstrated that with a carefully designed proposal, the model could achieve better performance with much fewer training rounds. Attention mechanism was applied (Kulharia et al., 2020; Song et al., 2019) to guide the model to focus on specific areas of objects in the image. Pixel-embedding learning was adopted (Kulharia et al., 2020) to generate pixel features with high intra-class affinity and inter-class discrimination. Priors of objectness filling rates were adopted (Song et al., 2019) to support training. The BBAM (Lee et al., 2021) utilized higher-level information to identify small informative areas in the image, which served as a pseudo-ground-truth for training the segmentation model. The CCNN (Pathak et al., 2015) adopted a constrained loss to integrate the priors in the training process, which imposed linear constraints on a latent distribution of the model output and trained the model to be close to the latent distribution. A generic objectness prior was directly incorporated in the loss to train a CNN model with point supervision (Bearman et al., 2016). Priors of shallow image features were employed in the loss function (Lin et al., 2016; Tang et al., 2018) to propagate information from scribbles to unmarked pixels.

2.3. Weakly-supervised segmentation in medical image

In the scenario of medical images, since full annotation is expensive and priors about objects are usually well-established, interest in weakly-supervised segmentation is increasing rapidly. DeepCut (Rajchl et al., 2016) adopted an iterative updating procedure to train a CNN model for fetal MRI segmentation based on a bounding box. Prior work (Kervadec et al., 2019) introduced a differentiable penalty in the loss function to enforce inequality constraints, which was applied to the cardiac, vertebral body, and prostate segmentation on MRI images. Kervadec et al. (2020) leveraged the tightness prior via constrained loss for the segmentation of spinal and brain MRI. Edge information was utilized in PseudoEdgeNet (Yoo et al., 2019), which trained the model to segment the nuclei with point annotations. Prior work (Qu et al., 2020) generated two types of coarse labels from point annotations to train a model for the segmentation of histopathology images. In another study (Valvano et al., 2021), the model was trained with an adversarial game for segmentation from scribble annotations in MRI images. Prior work (Ma et al., 2021) proposed a two-stage method containing a coarse segmentation and a refinement for the segmentation of organs at risk in nasopharyngeal carcinoma radiotherapy. Wang et al. (2021) proposed a semi-supervised method that utilized two auxiliary tasks to leverage the task-level consistency of unlabeled data for segmentation. Hu et al. (2022) proposed another semi-supervised method for the segmentation of nasopharyngeal carcinoma, that adopted a semi-supervised mean teacher model to generate the ROI-focused segmentation results. The UG-Net was proposed in (Tang et al., 2022) that consisted of a coarse segmentation module, an uncertainty guided module, and a feature refinement module for accurate segmentation on CT and retinal fundus image. The MRI-SegFlow (Kuang et al., 2020) also adopted the idea of automatic annotation and proposed a two-stage process for VB segmentation without relying on manual annotation. It adopted a rule-based method to automatically generate the suboptimal region of interest (ROI) and trained the CNN model with the suboptimal ROI. However, unlike our proposed method, the suboptimal ROI in MRI-SegFlow was not further optimized with the CNN training process. Besides, the rule-based method of MRI-SegFlow required further modification to transfer to other tissues.

3. Methodology

3.1. Overview of Spine-GFlow

The overall framework of our Spine-GFlow is illustrated in Fig. 2A. A rule-based method is first applied on the MRI image $\boldsymbol{E}_0$, which utilizes anatomical priors of tissue including texture, relative location, and size, to detect the approximate tissue locations and a rough spinal region. For each tissue, its locations are only detected in its midline sagittal slices. The detection result is utilized to initialize the seed areas Ψ, and the initial seed areas are served as the automatic weak annotation of our framework. For each tissue, the initial seed area consists of small 3D neighborhoods around the tissue locations, which are not necessarily in the same slice due to potential scoliosis (Fig. 2 B). The initial seed area of the background is determined according to the rough spinal region. More details about the rule-based seed area initialization will be discussed in Section 3.2.

Further, the MRI image $\boldsymbol{E}_0$ is fed into a CNN model that can generate multiple pixel-wise feature maps, $\boldsymbol{E}_1$, …, $\boldsymbol{E}_M$, with different scales. The proposal generation method integrates the MRI image $\boldsymbol{E}_0$, multi-scale feature maps, $\boldsymbol{E}_1$, …, $\boldsymbol{E}_M$, and seed areas Ψ to generate the segmentation proposals Ω. Each proposal consists of pixels belonging to a specific tissue or background, and pixels that are not in any proposals are defined as ambiguous pixels. The seed areas are also updated for the next iteration of proposal generation, which expand to adjacent slices and get closer to the proposals during the updating (Fig. 2 B). More details about proposal generation will be discussed in Section 3.3.

Based on the proposals, a comprehensive loss is calculated to train the CNN model, which will be discussed in detail in Section 3.4. Only pixels within the proposals are involved in CNN training, and ambiguous pixels are ignored. The proposal generation and CNN training are conducted iteratively. In each iteration, the MRI image is first fed into the CNN model, which will produce multi-scale feature maps. Then, based on the feature maps, the proposals are generated, which are further used to calculate the comprehensive loss for CNN training. The optimized CNN model can produce the feature maps for better proposal generation in turn.

In the following statement, (u, v) represents the 2D coordinates of a pixel in a 2D image, and $p = (x, y, z)$ represents the 3D coordinates of a pixel in a 3D scan, where z is the index of the slice. Let $t = 0, 1, 2, 3$ represent the background, VB, IVD, and SC.

3.2. Rule-based initialization

We adopt a rule-based method to generate the initial seed areas for VB, IVD, SC, and background. The VB area is identified first via gradient thresholding, size selection, and location selection, which determines approximate VB locations as well as a rough spinal region. Then, the IVD and SC are localized based on their relative location to VB. Further, the seed areas are initialized according to the tissue locations and rough spinal region.

In gradient thresholding, the normalized image gradient g_n and amplified image gradient g_a are defined as:

$$g_n(u, v) = g(u, v)/ave(u, v) \tag{1}$$

$$g_a(u, v) = g(u, v) * ave(u, v) \tag{2}$$

where $g(u, v)$ is the image gradient magnitude calculated with a Sobel operator, and $ave(u, v)$ is the average pixel value in the 3×3 neighborhood at $(u, \quad v)$. Considering the whole MRI scan as a 3D volume, we

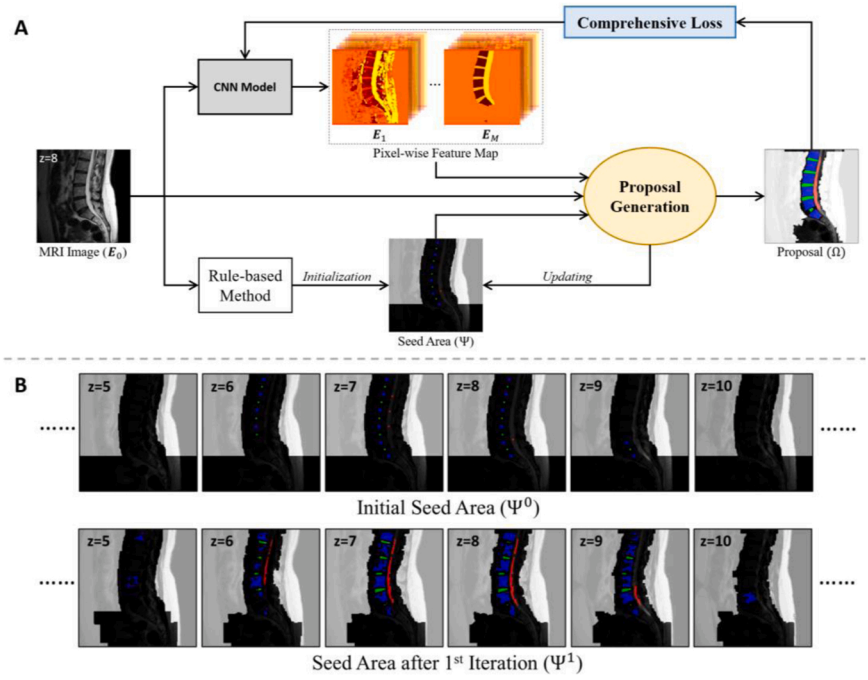

Fig. 2. A presents the overall framework of our proposed Spine-GFlow. Seed areas are first initialized with a rule-based method. In the iterative optimization procedure, the proposals are generated based on the MRI image, pixel-wise feature maps, and seed areas. The seed areas are updated for the next iteration. The generated proposals are further used to calculate a comprehensive loss to train the CNN model. An example of the initial seed areas and seed areas after 1st iteration is shown in B (blue: VB, green: IVD, red: SC, white: background), which shows that the initial seed areas are only in a few slices that are not necessarily the same. The seed areas expand to adjacent slices and get closer to the proposals during the updating. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

calculate the g_n and g_a in transverse, coronal, and sagittal views separately. The normalized and amplified gradients of 3D volume, G_n and G_a, are the pixel-wise maximum of g_n and g_a in three views, respectively (Fig. 3 B and C). The potential VB area is defined as: $V = \{p : G_n(p) < T_n, G_a(p) < T_a\}$, where T_n and T_a are two threshold values (Fig. 3 D).

The potential VB area is further processed via size selection and location selection (Fig. 3 E and F). We first consider the potential VB area as several 2D connected components (CCs) in each slice and find the minimum bounding rectangle (MBR) for each 2D CC. We measure the height, width, and aspect ratio of each MBR and remove CCs whose measurements are out of a certain range. Then, the processed VB area is treated as several 3D CCs in an MRI scan. For each 3D CC, we measure its thickness (i.e., how many slices it spans) and select those with the requested thickness. The midline slice of each selected 3D CC is projected onto one image and morphological closing using a square kernel ker_1 with a size of s_{ker} is applied. The morphological closing can merge VB projections and isolate non-VB projections. We remove all 3D CCs corresponding to isolated projections, and the remaining CCs are

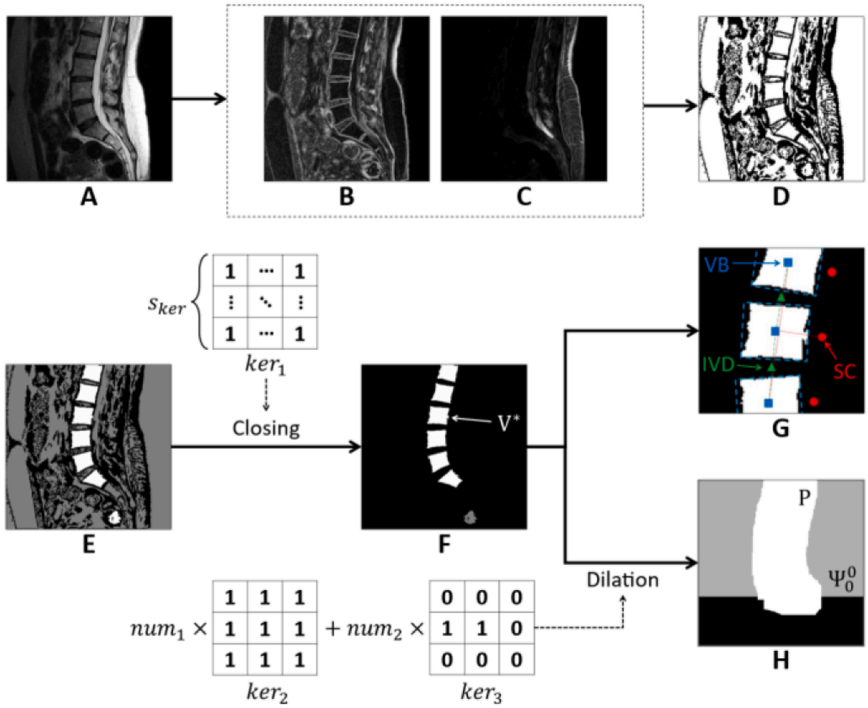

Fig. 3. The rule-based seed area initialization. A presents an MRI, whose G_n and G_a are presented in B and C. D is the potential VB area V. E and F illustrate the size and location selection on V, and the white area in E and F represents the selection result. G presents center locations of tissues. H shows the projection of rough spine area P and the initial seed area of the background Ψ_0^0. ker_1, ker_2, and ker_3 are 3 kernels for location selection and determination of the rough spinal region.

denoted as V^*.

The tissue locations l_t are determined based on V^*. We find the MBR for the midline slice of each 3D CC in V^*. For each VB, the center location $l_1 = (x_1, y_1, z_1)$ and width w_1 are measured from the MBR. Further, the center locations of IVD l_2 and SC l_3 are determined by:

$$l_2^n = \left((x_1^n + x_1^{n+1})/2, (y_1^n + y_1^{n+1})/2, (z_1^n + z_1^{n+1})/2\right) \tag{3}$$

$$l_3^n = \left(x_1^n - \alpha * \sin\theta^n * w_1^n, y_1^n + \alpha * \cos\theta^n * w_1^n, z_1^n\right) \tag{4}$$

where $\theta^n = \arctan\left((y_1^{n+1} - y_1^{n-1})/(x_1^{n+1} - x_1^{n-1})\right)$, α is a constant which is set to 0.80, and (x_1^n, y_1^n, z_1^n) and $(x_1^{n\pm1}, y_1^{n\pm1}, z_1^{n\pm1})$ are the center locations of three adjacent VBs (Fig. 3G).

The initial seed areas Ψ_t^0 of VB, IVD, and SC are defined as 3D neighborhoods at corresponding center locations. To determine the initial seed area of background Ψ_0^0, we project the midline slice of each 3D CC in V^* onto one image and apply the morphological dilation to generate a rough spinal region P. The morphological dilation takes num_1 and num_2 iterations with the kernel ker_2 and ker_3, respectively. Let the location of the lowest VB in V^* be $l_1^{max} = (x_1^{max}, y_1^{max}, z_1^{max})$, while Ψ_0^0 is defined as: $\Psi_0^0 = \{p|(x,y) \notin P, x < x_1^{max}\}$ and is the same in each MRI slice (Fig. 3 H). The specific configuration of rule-based initialization is described in Section 4.2.

3.3. Proposal generation

Unlike most iterative optimization methods, which generate proposals based on CNN output, our framework combines different levels of information by integrating multi-scale feature maps in proposal generation. Inspired by Segsort (Hwang et al., 2019), a clustering-based method is applied on each feature map first to divide pixels into several clusters, and each pixel cluster is further decomposed into several CCs. Specific CCs are selected according to each seed area and assembled into the corresponding proposal, which is further fine-tuned with several rule-based operations to explicitly embed the anatomical prior. Finally, the seed areas are updated based on the proposals and pixel clustering results for the next iteration of proposal generation.

3.3.1. Clustering-based pixel division

We adopt the k-means algorithm for pixel clustering, which iteratively conducts the assignment and update steps. In the assignment step, each pixel is assigned to the cluster with the most similar mean feature. The assignment step produces a set of pixel clusters C (Fig. 4 C), which is defined as:

$$C_k = \left\{p_i : \|\boldsymbol{e}_i - \boldsymbol{\rho}_k\|_2 \le \|\boldsymbol{e}_i - \boldsymbol{\rho}_j\|_2 \quad \forall j, \quad 1 \le j \le K\right\} \tag{5}$$

where $\boldsymbol{e}_i$ is the feature of pixel p_i, $\boldsymbol{\rho}_k$ is the mean feature of C_k, and K is the number of total pixel clusters. In the update step, the mean feature of each pixel cluster is calculated as:

$$\boldsymbol{\rho}_k = \frac{1}{|C_k|}\sum_{p_i \in C_k} \boldsymbol{e}_i \tag{6}$$

where $|C_k|$ is the number of pixels in C_k. The mean feature is initialized with K randomly selected pixel features from the feature map, and the clustering stops after 10 iterations. Formally, we define the pixel clustering process on feature map $\boldsymbol{E}$ as:

$$\mathrm{Clu}(\boldsymbol{E}) = \{C_k\} \tag{7}$$

We conduct the pixel clustering on the original MRI $\boldsymbol{E}_0$ (Fig. 4 A) and the multi-scale feature maps, $\boldsymbol{E}_1$, …, $\boldsymbol{E}_M$, (Fig. 4 B), generated by the CNN model, individually. For $\boldsymbol{E}_0 = \{v_i\}$, v_i represents the pixel value of p_i, which is the feature with the smallest scale. For $\boldsymbol{E}_m = \{\boldsymbol{e}_i\}_m (m \ge 1)$, the feature of each pixel is normalized as $\widetilde{\boldsymbol{e}}_i = \boldsymbol{e}_i / \|\boldsymbol{e}_i\|$ before clustering. Since the location of each pixel is not involved in the clustering process, each pixel cluster may not be spatially aggregated, which can be represented as $C_k = \bigcup_n cc_{k,n}$, where $cc_{k,n}$ is the 2D CC in C_k. Further, pixels in the MRI scan can be divided into multiple 2D CCs (Fig. 4 D) based on the clustering result of $\boldsymbol{E}$. Formally, we define the pixel division process as:

$$\mathrm{Div}(\boldsymbol{E}) = \{cc_{k,n}\}$$

3.3.2. Pixel selection

The 2D CCs in the pixel division result of each feature map are

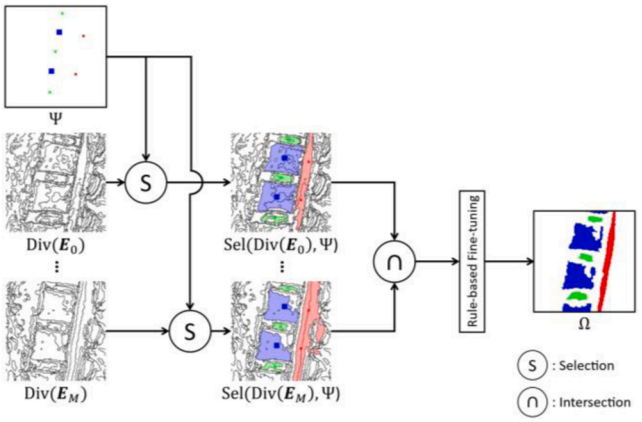

Fig. 5. The pixel selection process. The 2D CCs that overlap with the seed areas are selected and assembled. The intersections of all selection results derived from $\boldsymbol{E}_0$, …, $\boldsymbol{E}_M$ are further processed with the rule-based fine-tuning to generate the final proposals.

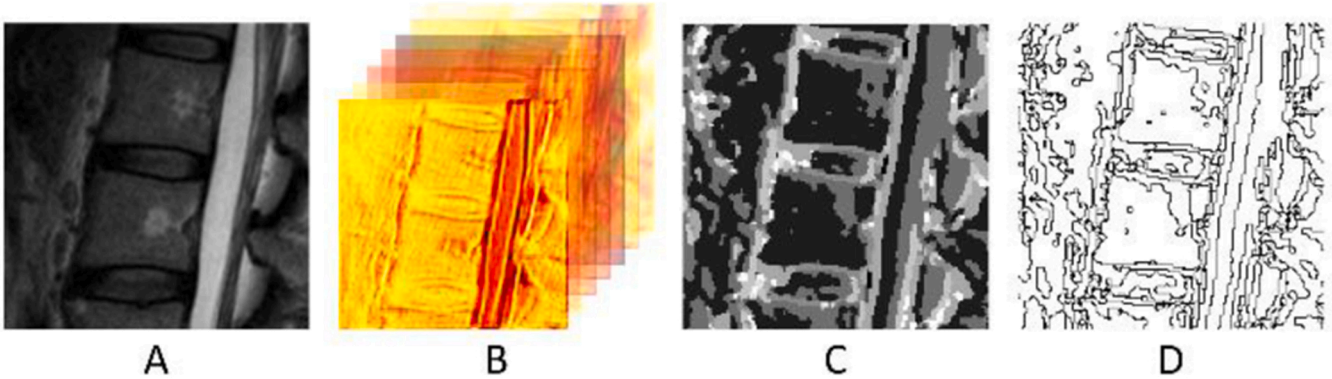

Fig. 4. The clustering-based pixel division. A and B present an MRI patch and one of its feature maps generated by the CNN model. C is the clustering result of the feature map, and D is the pixel division based on the clustering result.

selected according to the seed areas Ψ_t (Fig. 5). Specifically, we select and assemble the CC that overlaps with Ψ_t. The selection process is defined as:

$$\mathrm{Sel}(\{\mathrm{cc}_{k,n}\}, \Psi_t) = \bigcup_{\mathrm{cc}_{k,n} \cap \Psi_t \neq \varnothing} \mathrm{cc}_{k,n} \tag{8}$$

and is conducted on the pixel division result of original MRI $\boldsymbol{E}_0$ and feature maps, $\boldsymbol{E}_1$, …, $\boldsymbol{E}_M$, individually, and the proposals Ω_t are defined as the intersections of all selection results:

$$\Omega_t = \bigcap_{m=0}^{M} \mathrm{Sel}(\mathrm{Div}(\boldsymbol{E}_m), \Psi_t) \tag{9}$$

The proposals are further fine-tuned with three rule-based operations:

(1) ***3D Morphological Closing***: Considering the tissue should be solid in an MRI scan, we apply 3D morphological closing on the proposals to remove any potential small inner cavities.
(2) ***3D Morphological Opening***: Since the tissue has relatively fixed positions and no drastic shape variation in adjacent slices, we apply 3D morphological opening on the proposals to remove structures with insufficient thickness.
(3) ***Exclusivity***: Each pixel can only be in one proposal. If a pixel belongs to more than one proposal, it will be removed from all proposals.

3.3.3. Seed area updating

To update the seed area of each tissue, we first determine the dominant pixel cluster D for each proposal based on the clustering results of feature maps. We select the pixel cluster from the clustering result of each feature map that contains the most pixels in the proposal. The dominant pixel cluster is the intersection of all selected pixel clusters, which is defined as:

$$\mathrm{D}_t = \bigcap_{m=1}^{M} \underset{\mathrm{C}_k \in \mathrm{Clu}(\boldsymbol{E}_m)}{\mathrm{argmax}} |\mathrm{C}_k \cap \Omega_t| \tag{10}$$

Note that only the feature maps generated by CNN model are utilized, and the original image $\boldsymbol{E}_0$ is not involved in the determination of dominant pixel cluster. In each round of updating, the seed area will expand to adjacent slices. Let pixel $p = (x, y, z)$ and its slice neighborhood sn_p be defined as $\mathrm{sn}_p = \{(x, y, z \pm 1)\}$. Let the expanded part of seed area Ψ_t^* be defined as $\Psi_t^* = \{p : p \in \mathrm{D}_t,\ p \notin \Omega_t,\ \mathrm{sn}_p \cap \Omega_t \neq \varnothing\}$. The expanded part covers the pixels from the dominant pixel cluster whose slice neighborhood overlaps with the proposal. The seed area is then updated as (Fig. 6):

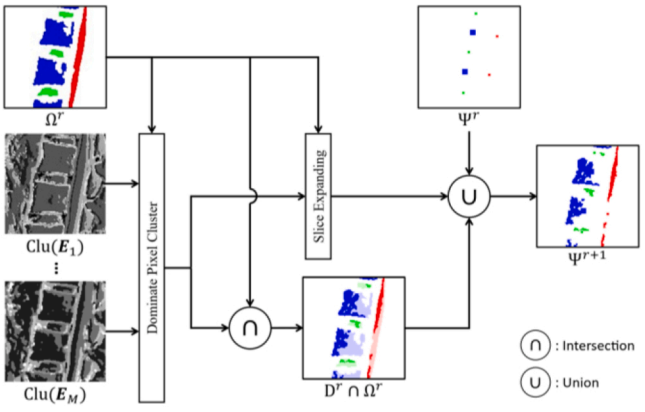

Fig. 6. The seed area updating. The dominant pixel cluster is first determined based on the clustering results of feature maps, $\mathrm{Clu}(\boldsymbol{E}_1)$, …, $\mathrm{Clu}(\boldsymbol{E}_M)$, and proposal Ω^r. The slice expanded part is further determined. The updated seed area is the union of previous seed area, slice expanded part, and intersection of dominant pixel cluster and proposal.

$$\Psi_t^{r+1} = (\mathrm{D}_t^r \cap \Omega_t^r) \cup \Psi_t^{*r} \cup \Psi_t^r \tag{11}$$

where r represents the number of iterations. For the background, the seed area is simply updated as:

$$\Psi_0^{r+1} = \Omega_0^r \tag{12}$$

3.4. CNN model and training

3.4.1. CNN architecture

The CNN model in our framework adopts the U-Net+ + (Zhou et al., 2019) as the backbone, which can generate multi-scale pixel-wise feature maps from input MRI images. As illustrated in Fig. 7, the CNN model can generate M feature maps, $\boldsymbol{E}_1$, …, $\boldsymbol{E}_M$, where M is determined by the number of levels in the U-Net+ +. All feature maps are concatenated and further processed by two convolutional layers (conv-layers) with a kernel size of 1×1 and a softmax layer, which produce the pixel classification $\boldsymbol{Y}$.

3.4.2. Comprehensive loss

A comprehensive loss is calculated based on the proposals and consists of the pixel classification loss (PCL) and feature distribution loss (FDL) (Fig. 7) to optimize the final output, that is the pixel classification, and pixel-wise feature maps generated by the CNN model simultaneously. The PCL introduces the penalization in the pixel classification $\boldsymbol{Y} = \{\boldsymbol{y}_i\}$ generated by the CNN model to optimize the model output. For proposal Ω_t, the PCL is defined as:

$$\mathrm{PCL}(\Omega_t) = \frac{1}{|\Omega_t|} \sum_{p_i \in \Omega_t} \mathrm{ce}(\boldsymbol{y}_i, \widehat{\boldsymbol{y}}_i) \tag{13}$$

where $\boldsymbol{y}_i$ and $\widehat{\boldsymbol{y}}_i$ are the CNN classification and ground truth of p_i and $\mathrm{ce}()$ is the cross entropy.

To optimize the pixel feature distribution of the feature maps, beyond the conventional cross entropy PCL, we also introduce the FDL, which encourages the CNN model to generate homogeneous features for pixels from the same proposal, and inhomogeneous features for pixels from different proposals. For the feature map $\boldsymbol{E}_m = \{\boldsymbol{e}_i\}_m (m \geq 1)$, the

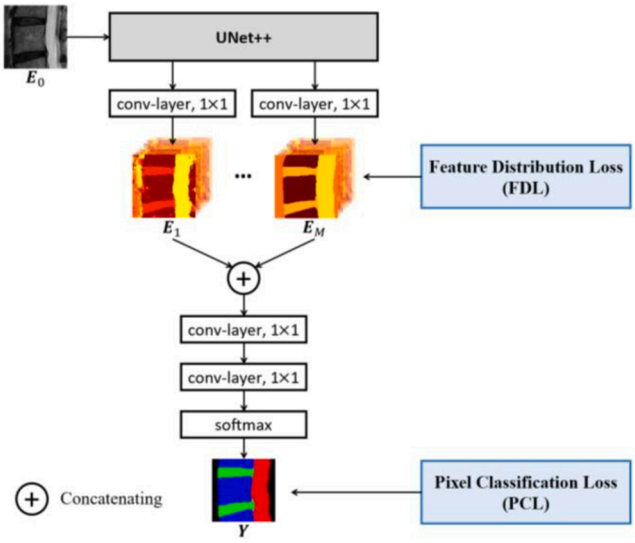

Fig. 7. The CNN model and comprehensive loss. Our CNN model adopts the U-Net+ + (Zhou et al., 2019) as the backbone to generate multi-scale feature maps, which are further concatenated and processed by two convolutional layers and a softmax layer to generate pixel classification. The comprehensive loss consists of the pixel classification loss (PCL) and feature distribution loss (FDL), which optimize the pixel classification and feature distribution of feature maps, respectively.

FDL of tissue proposal $\Omega_t (t \in [1,3])$ is defined as:

$$\mathrm{FDL}(\boldsymbol{E}_m, \Omega_t) = -\frac{1}{|\Omega_t|}\sum_{p_i \in \Omega_t} \log \frac{\exp(\boldsymbol{\varphi}_t{}^{\mathrm{T}} \widetilde{\boldsymbol{e}}_i)}{\sum_{s\in[1,3],\ s\neq t} \exp(\boldsymbol{\varphi}_s{}^{\mathrm{T}} \widetilde{\boldsymbol{e}}_i)} \tag{14}$$

where $\widetilde{\boldsymbol{e}}_i = \boldsymbol{e}_i / \|\boldsymbol{e}_i\|$ is the normalized pixel feature and $\boldsymbol{\varphi}_t = (\sum_{p_i \in \Omega_t} \widetilde{\boldsymbol{e}}_i) / \left\| \sum_{p_i \in \Omega_t} \widetilde{\boldsymbol{e}}_i \right\|$ is the mean feature of Ω_t. The numerator encourages each pixel feature to be close to the mean feature of its own proposal, and the denominator pushes each pixel feature away from the mean feature of other proposals. The FDL of background proposal Ω_0 is defined as:

$$\mathrm{FDL}(\boldsymbol{E}_m, \Omega_0) = -\frac{1}{|\Omega_0|}\sum_{p_i \in \Omega_0} \log \frac{1}{\sum_{s\in[1,3]} \exp(\boldsymbol{\varphi}_s{}^{\mathrm{T}} \widetilde{\boldsymbol{e}}_i)} \tag{15}$$

Due to the diversity of pixel features in the background, we do not calculate the mean feature for Ω_0 and the FDL only encourages the pixel feature to be far away from the mean feature of all tissue proposals.

Only pixels within the proposals are involved in CNN training and ambiguous pixels are ignored. Furthermore, the average loss is calculated over pixels in each proposal separately, which prevents the weight of the small-size tissue from being diluted. The final loss is calculated as:

$$Loss = \sum_{t=0}^{3} \sum_{m=1}^{M} a_{t,m} \times \mathrm{FDL}(\boldsymbol{E}_m, \Omega_t) + \sum_{t=0}^{3} b_t \times \mathrm{PCL}(\Omega_t) \tag{16}$$

where $a_{t,m}$ and b_t are the weights for different losses.

3.4.3. Training protocol

We train the CNN model with small image patches instead of the whole MRI to make the model focus on the area covering the tissue proposals. The patches are randomly selected from the MRI slices, where the proposals of all tissues appear. Overlapping or repetition of selected patches is acceptable. During proposal generation, to provide the feature map of the whole MRI scan, we uniformly select patches with a constant stride from the input MRI and merge the feature map of each patch generated by the CNN model.

In our framework, the CNN model can be trained with different protocols. First, when a set of unlabeled MRI scans are available, the CNN model can be trained with patches selected from different MRI scans. The FDL enforces the model to extract similar features for pixels of the same tissue in different MRI scans, which helps the model learn general features. After each iteration of training, proposals of all MRI scans are updated simultaneously based on the trained model. We call this training protocol *holistic training*. The unlabeled MRI scans can be simply collected as a clinical routine. Since no manual annotation is required, our framework can provide another CNN training protocol called *individual training*, where the CNN model is trained on the target MRI directly. In individual training, patches are selected from the target MRI only, which makes the model adapt to potential feature variations in each MRI scan and allows our framework to boot up with only one MRI scan. To obtain better performance, our framework can take advantage of both holistic and individual training. The CNN model is first trained with a set of prepared MRI scans and further fine-tuned on the target MRI. Much fewer patches are used in the fine-tuning process compared with only individual training.

4. Dataset and implementation details

4.1. Dataset

4.1.1. HKDDC

The expert anatomically annotated Hong Kong Disc Degeneration Cohort (HKDDC) dataset (Samartzis et al., 2012) included 40 T2-weighted MRI scans collected from 40 different subjects. This was a population-based dataset with subject recruitment from open advertisement. The MRI scans were obtained via 3 different MRI machines with resolutions from 448 × 448 to 512 × 512. The detailed composition of MRI scans in the dataset is presented in Table 1. Each MRI scan contained at least 5 lumbar vertebrae from L1 to L5, and there are at least 7 slices in each scan containing annotated spinal structures. For each scan, the pixel-wise manual annotations of VB, IVD, and SC were provided (from L1 to S1). All annotation work was completed by three readers who are medically trained, with a fourth reader (a spine surgeon with more than 20 years of clinical experience) to compare the outcomes and confirm precision as well as consistency (the pixel-wise agreement of annotation is 98%). The MRI scans are split into 20:10:10 as the training, validation, and testing set.

4.1.2. IVDM3Seg

The MICCAI 2018 Challenge on Intervertebral Disc Localization and Segmentation (IVDM3Seg) dataset contains 16 MRI cases collected from 8 subjects in two stages. Each case consists of four aligned high-resolution 3D MRI scans with different modalities, including in-phase, opposed-phase, fat, and water, as well as the manually labeled binary mask for IVD. The MRI was scanned with a 1.5-Tesla MRI scanner of Siemens using Dixon protocol. Each MRI scan has a size of 256 × 256 × 36. More detailed information about the IVDM3Seg dataset could be found on the official website (https://ivdm3seg.weebly.com). For each MRI scan, we only focus on the area lower than the T11 vertebra (lumbar region).

4.2. Implementation details

4.2.1. Rule-based initialization configuration

The rule-based initialization was experimentally configured according to the training set. For the HKDDC dataset, in the gradient threshold, we first calculate the normalized and amplified image gradients on 5 cases randomly selected from the training set and experimentally determine the threshold values that can the distinguish potential VB area. The threshold values for the normalized and amplified image gradients T_n and T_a were set as 2.5 and 0.2. In size selection, we calculated the minimum (min) and maximum (max) for dimensions of 10 VBs randomly selected from the training set and determined the requested range as [0.7 × min, 1.3 × max]. Thus, the requested range for height, width, aspect ratio, and thickness were [20,70], [20,70], [0.5,2], and [5,15], respectively. In the morphological closing of location selection, the kernel size w_{ker} was 25. The sizes of 3D neighborhoods in initial seed areas of VB, IVD, and SC were set as 7 × 7 × 3, 3 × 3 × 3, and 3 × 3 × 1. The iteration numbers of the morphological dilation for the rough spinal region, num_1 and num_2, were set as 35 and 25.

For the IVDM3Seg dataset, the image intensity and size were significantly different from the clinical T2-weighted MRI of the HKDDC dataset, thus minor adjustments were required in the gradient threshold and size selection. The normalized image gradient and the amplified image gradient were calculated on fat modality and on opposed-phase modality. The threshold values T_n and T_a were set as 4.0 and 0.1. In size selection, the requested range for height, width, and thickness were set as [10,50], [15,50], and [10,30]. The iteration numbers num_1 and num_2 were 20 and 5.

Table 1
Composition of MRI Scans in the Dataset.

Institution	MRI Machine	Scan Number	Image Number	Gender	Age
Hong Kong Sanatorium Hospital	GE Healthcare Signa	10	110	60 % F	47.3 ± 7.4
	Siemens Trio	12	180	50 % F	50.4 ± 8.8
St. Teresa's Hospital	Siemens Prisma	18	306	50 % F	52.7 ± 8.3

4.2.2. Proposal generation configuration

For the k-means algorithm, the number of pixel clusters, K, was set as 10. The kernel for the 3D morphological closing and opening in proposal fine-tuning was a cuboid with a size of $5 \times 5 \times 2$ and $1 \times 1 \times 3$. The 3D morphological opening was not applied in the first iteration.

4.2.3. CNN architecture and training configuration

A UNet+ + with 4 levels was adopted in the CNN model that could generate 3 pixel-wise feature maps with different scales. All conv-layers in the model had 64 filters except the output layer, which had 4 filters. For the HKDDC dataset, the input of the CNN model was the patch of raw clinical MRI. For the IVDM3Seg dataset, the input of the CNN model was the patch of the concatenation for 4 modalities of MRIs.

We evaluated the framework with 3 different CNN training protocols: (1) only holistic training (HT), (2) only individual training (IT), (3) both holistic and individual training (HT+IT). For HT, the CNN model was trained with only the training set and applied on the target MRI directly without any fine-tuning. The training took 15 iterations. For each iteration, 5120 patches were selected from each MRI. For IT, the CNN model was trained from scratch based on only the target MRI, which also took 15 iterations, and 5120 patches were selected from the target MRI for each iteration. For HT+IT, the CNN model was first pretrained on the training set with 15 iterations, and for each iteration, 3840 patches were selected from each MRI. The CNN model was further fine-tuned on each target MRI with 8 iterations. For the first 7 iterations, 512 patches were selected from the target MRI, and for the last iteration, 3072 patches were selected. For all training protocols, all weights of the loss were set as 1. The mini-batch strategy with a batch size of 16 was adopted. Adam was used as the optimizer with an initial learning rate of 0.0006.

4.3. Evaluation metrics

Two different metrics were adopted to quantitatively evaluate the segmentation performance of our Spine-GFlow: Intersection over Union (IoU) and the Dice coefficient (Dice), which were calculated as follows:

$$IoU = \frac{TP}{TP + FP + FN} \tag{17}$$

$$Dice = \frac{2TP}{2TP + FP + FN} \tag{18}$$

where TP, FP and FN denoted the number of true positive, false positive, and false negative pixels in the segmentation results respectively. The mean IoU and mean Dice were defined as the average IoU and Dice of all tissues.

5. Results

5.1. Multi-tissue segmentation

The quantitative evaluation results of the multi-tissue segmentation performance achieved by our Spine-GFlow on the HKDDC dataset were shown in Table 2, and Table 3 showed the IVD segmentation result on the IVDM3Seg dataset. We compared the framework with 3 different CNN training protocols, including HT, IT, and HT+IT. Furthermore, we also compared our method with the model trained with the constrained losses in (Kervadec et al., 2019), the automatic annotation of MRI-SegFlow (Kuang et al., 2020), and the full supervision. The constrained losses (Kervadec et al., 2019) trained the model using small regions within the ground-truth mask, which were similar to the initial seed areas used in our framework. We generated the weak annotation for constrained losses according to (Kervadec et al., 2019). The MRI-SegFlow (Kuang et al., 2020) provided a rule-based method to generate automatic annotation of VB, and we modified the parameters and transferred it to IVD and SC. For IVDM3Seg dataset, we also compared with the best result reported in (Zheng et al., 2017) (achieved by team UNILJU), and the result achieved by IVD-Net (Dolz et al., 2018) that was a state-of-the-art method for IVD segmentation. For all training strategies, the CNN model adopted the same network architecture.

Table 3
Evaluation of IVD Segmentation Performance on IVDM3Seg Dataset.

Method	IVD IoU	IVD Dice
Constrained Losses	0.773 ± 0.015	0.872 ± 0.013
MRI-SegFlow	0.783 ± 0.021	0.878 ± 0.017
Spine-SegLoop (HT)	0.792 ± 0.017	0.884 ± 0.014
Spine-SegLoop (IT)	0.812 ± 0.018	0.895 ± 0.015
Spine-SegLoop (HT+IT)	0.820 ± 0.015	0.901 ± 0.013
Full Supervision	0.845 ± 0.015	0.916 ± 0.013
UNILJU	–	0.918 ± 0.021
IVD-Net	–	0.919 ± 0.018

The results showed that our Spine-GFlow consistently outperformed the model trained with constrained losses (Kervadec et al., 2019) and MRI-SegFlow (Kuang et al., 2020) for all tissues. We conducted the t-test on the mean Dice of HKDDC dataset and IVD Dice of IVDM3Seg dataset achieved by our method and constrained losses as well as MRI-SegFlow. On HKDDC dataset, the p-values were 0.00009 for constrained losses and 0.006 for MRI-SegFlow. On IVDM3Seg dataset, the p-values were 0.0007 for constrained losses and 0.01 for MRI-SegFlow. HT and IT achieved similar overall performance. For VB, HT produced 1.7 % higher IoU and 1.1 % higher Dice than IT on the HKDDC dataset. For IVD, IT achieved 1.6 % higher IoU and 0.9 % higher Dice than HT on the HKDDC dataset, as well as 2.0 % higher IoU and 1.1 % higher Dice on the IVDM3Seg dataset. For SC, there was no significant difference between these two protocols. By combining HT and IT, our framework could further improve segmentation accuracy for all tissues. Moreover, the Spine-GFlow with HT+IT obtained performance comparable to the model trained with full supervision. We achieved only 2 % lower Dice for SC, and even 1 % and 0.4 % higher Dice for IVD and VB on the HKDDC dataset. On the IVDM3Seg, the Dice of our Spine-GFlow was only 0.017 and 0.018 lower than two state-of-the-art fully supervised methods, UNILJU (Zheng et al., 2017) and IVD-Net (Dolz et al., 2018).

Fig. 8 visually presented several multi-tissue segmentation results on the HKDDC dataset. All segmentation results in Fig. 8 were the original outputs of the CNN models trained with different methods and no further post-processing was applied. Fig. 8 A and B illustrated the initial seed areas and multi-tissue segmentations, respectively, produced by

Table 2
Evaluation of Multi-tissue Segmentation Performance on HKDDC Dataset.

Method	IVD		VB		SC	
	IoU	Dice	IoU	Dice	IoU	Dice
Constrained Losses	0.745 ± 0.036	0.854 ± 0.026	0.801 ± 0.036	0.889 ± 0.022	0.794 ± 0.035	0.885 ± 0.022
MRI-SegFlow	0.806 ± 0.036	0.892 ± 0.025	0.829 ± 0.021	0.907 ± 0.012	0.782 ± 0.034	0.877 ± 0.022
Spine-GFlow (HT)	0.830 ± 0.035	0.907 ± 0.021	0.860 ± 0.029	0.925 ± 0.017	0.809 ± 0.045	0.894 ± 0.029
Spine-GFlow (IT)	0.846 ± 0.029	0.916 ± 0.017	0.843 ± 0.041	0.914 ± 0.025	0.807 ± 0.051	0.893 ± 0.034
Spine-GFlow (HT+IT)	**0.847 ± 0.028**	**0.917 ± 0.016**	**0.866 ± 0.022**	**0.928 ± 0.012**	0.811 ± 0.040	0.896 ± 0.026
Full Supervision	0.830 ± 0.039	0.907 ± 0.024	0.859 ± 0.041	0.924 ± 0.024	0.846 ± 0.026	0.916 ± 0.015

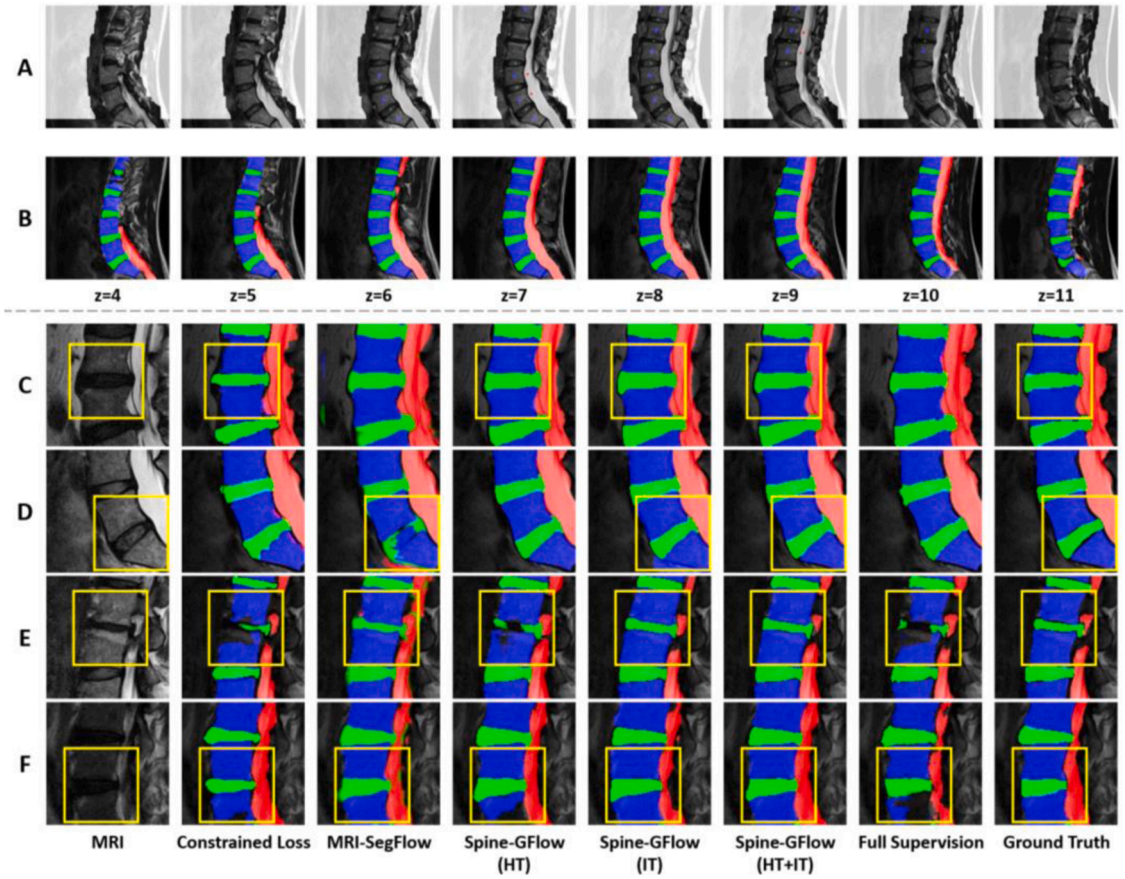

Fig. 8. A and B present the initial seed areas and multi-tissue segmentation (blue: VB, green: IVD, red: SC) produced by Spine-GFlow on an MRI scan with alignment deformity. C-F are the visual comparisons of multi-tissue segmentation produced by different methods on MRI patches. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

our method on an MRI scan displaying alignment deformity. The result showed that initial seed areas were presented in different slices since the midline sagittal slices of each tissue were different. Our method could adapt the alignment deformity and produce accurate segmentation on different slices. Fig. 8 C-F visually compared the segmentation on MRI patches produced by different methods. It showed that our framework could identify the shape detail, such as corners and potential deformity, better than the model trained with constrained losses (Fig. 8 C) and reduce the noise and shape distortion compared with MRI-SegFlow (Fig. 8 D). Compared with the framework using only IT, the framework with both HT and IT could reduce noise in the result (Fig. 8 D). For some extreme variations in image features caused by pathologies, such as the Marrow change (Fig. 8 E), our framework with the IT could adapt better than other methods and produce a more accurate result. Moreover, for the image with low contrast (Fig. 8 F), our framework with the IT also showed high robustness.

5.2. Ablation study

To further investigate the effect of different components in our framework, the ablation study was conducted on the HKDDC dataset.

In our framework, the proposals were generated based on multi-scale feature maps produced by the CNN model. To investigate the effect of integrating the multi-level information, we compared a different proposal generation strategy adopted by (Rajchl et al., 2016), which produced proposals by applying conditional random field (CRF) on the CNN output. For a fair comparison, the generated proposals were further fine-tuned with the same rule-based operations in Section 3.3, and the CNN model was trained with the same protocol and loss function. We denoted this variant as *Spine-GFlow (P-)*.

Our framework introduced the FDL in the CNN training process in addition to the conventional cross entropy PCL to encourage the CNN model to extract more discriminative pixel features. To validate the effect of the comprehensive loss, we compared it to the framework where the CNN model was trained with only PCL. We kept the rest of the framework unchanged and denoted this variant as *Spine-GFlow (L-)*.

Fig. 9 presented the evolution of segmentation performance during

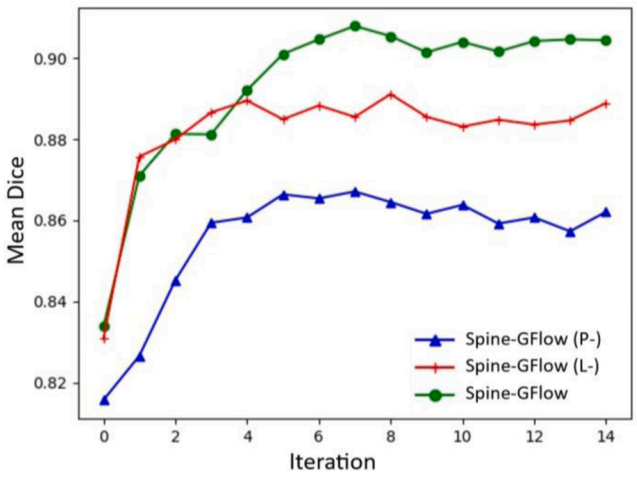

Fig. 9. The evolution of segmentation performance (mean Dice) during the HT process.

the HT process, showing that with multi-scale feature maps and FDL, the CNN model of the standard Spine-GFlow was trained more efficiently. Without FDL, the CNN model of Spine-GFlow (L-) achieved the same learning speed as the standard framework at the beginning of HT; however, its performance did not further improve after the 5th iteration, and after HT its mean Dice were 1% lower than the standard Spine-GFlow. In the Spine-GFlow (P-), the proposals generated based on the only model output more significantly reduced the training efficiency of the CNN model, and after HT its mean Dice were 4% lower than the standard framework. Moreover, as presented in Table 4, after HT and IT, the standard Spine-GFlow ultimately obtained better performance than the other 2 variants.

Fig. 10 presented the proposals generated by Spine-GFlow (P-) and standard Spine-GFlow in different iterations, which showed that the proposals generated with multi-scale feature maps could provide more shape details. Furthermore, when the tissue boundary was not clear, integrating the multi-scale feature maps could avoid the proposals invading the wrong area.

Fig. 11 presented the pixel clustering result of feature maps generated by the CNN model of Spine-GFlow (L-) and standard Spine-GFlow and the corresponding proposals. It demonstrated that the model trained with comprehensive loss could generate feature maps whose pixel clustering results could better reflect the true spatial distribution of different tissues with less noise, especially for feature maps with small scales. More specifically, based on the feature maps of the model trained with comprehensive loss, most pixels of the same tissue would be divided into the same pixel cluster, which could help generate more accurate proposals and in turn improve the CNN training process.

Several rule-based fine-tuning operations were adopted in the proposal generation to explicitly embed the anatomical prior, which could effectively reduce the error in the proposals and have high robustness against the geometry variation caused by tissue deformity. Fig. 12 presented two examples of rule-based fine-tuning for the case with the disc bulge (Fig. 12 A) and Schmorl's node (Fig. 12 B) that showed fine-tuning could fill the cavity (Fig. 12 A) and remove the wrong bulge (Fig. 12 B) in the proposals.

Fig. 13 presented several examples of proposals and segmentation results produced with the defective initial seed areas. The initial seed areas were manipulated with translation and deletion to simulate potential defects. The result demonstrated that location deviation did not significantly affect the final proposals and segmentation result. The partial absence led to missing corresponding tissue in the final proposals but had no obvious influence on the segmentation result.

Fig. 14 presented several examples of seed area initialization on the cases with different feature variations. Fig. 14 A and B had the Schmorl's node, and the image contrast was low. Fig. 14 C had the Marrow change, and Fig. 14 D had the disc bulging. The result demonstrated that these feature variations caused by pathologies and/or image quality did not significantly affect the seed area initialization. There might be the location deviation (SC in Fig. 14 C) and/or partial absence (SC in Fig. 14 B) in the initial seed areas, but no mistake (seed area in wrong tissue).

6. Discussion

In this paper, we have established a hybrid framework, known as Spine-GFlow, for robust multi-tissue segmentation in lumbar MRI without requiring any manual annotations. A rule-based method is first adopted to automatically generate the weak annotation. It detects the approximate tissue locations and rough spinal region, and further determines the initial seed areas. For each tissue, the locations are only detected in its midline sagittal MRI slices, thus the initial seed area is not necessarily in the same slice due to potential alignment deformity, which helps the framework adapt to the case with scoliosis. A CNN model is developed to generate multi-scale feature maps and pixel classifications from the MRI image. A clustering-based method is adopted to generate the segmentation proposals based on multi-scale feature maps and the seed areas. The proposals are further fine-tuned with several rule-based operations to explicitly embed the anatomical prior, and the seed areas are updated according to the fine-tuned proposals. Next, the CNN model is trained with a comprehensive loss, which simultaneously optimizes the pixel classification and feature distribution of feature maps based on the proposals. By iteratively conducting the proposal generation and CNN training, we can obtain a CNN model for accurate multi-tissue segmentation. Since no manual annotation is required in our framework, it can automatically fine-tune the CNN model on the target MRI to improve the robustness of the model against image feature variation.

Our Spine-GFlow is an explainable segmentation framework, that adopts the attention mechanism in CNN training process, and the inference process can be easily interpreted via latent feature space. (Yang et al., 2022) The proposal is generated based on the clustering of image features and the fine-tuning guided by anatomical prior. It serves as a hard attention map for the CNN training process that is meaningful both graphically and anatomically. To interpret the segmentation generated by our CNN model, we can simply visualize the distribution of each pixel feature in both latent feature space and spatial space via clustering. As presented in Fig. 11, pixels belonging to the same tissue have homogeneous features, while pixels from different tissues have inhomogeneous features.

The CNN training process of our Spine-GFlow enables the data harmonisation of our framework. We train the CNN model with small image patches with unified sizes instead of the whole MRI, which can make the model focus on the area covering the tissue proposals, and also serve as an image-processing-based data harmonisation to standardize different shapes of MRIs from different sources. (Nan et al., 2022) In the HT process, the mean feature of FDL is calculated on the patches from different MRIs, which enforces the model to extract similar features for the same tissue of different MRIs. It helps the model learn the source-invariant features, then apply these features for the segmentation task. (Nan et al., 2022).

We validate our method on 2 independent datasets: HKDDC (containing the MRI scans obtained from 3 different machines) and IVDM3Seg. Our framework was quantitatively evaluated with 3 different CNN training protocols, and compared with a CNN model trained with constrained loss (Kervadec et al., 2019), MRI-SegFlow (Kuang et al., 2020), and full supervision. The results showed that our framework consistently outperforms the constrained loss (Kervadec et al., 2019) and MRI-SegFlow (Kuang et al., 2020) for all tissues. The significance test demonstrated the performance improvement was significant. Compared with the constrained loss, our method could produce the result with more shape details, which is important for detecting potential deformity. Compared with the MRI-SegFlow, since our framework can iteratively optimize the proposals for CNN training, our CNN model generates more accurate results with less noise. HT obtains higher segmentation accuracy on VB than IT, while for the IVD, IT performs better. Since the features of VB such as shape and pixel intensity are more consistent than those of IVD, the model trained with HT performed better given it can learn more general features. Otherwise, for IVD, the model trained with IT can adapt to large individual variations better than with only HT. By combining HT and IT, our framework can further improve accuracy on all tissues and achieve a performance comparable with a model trained with full supervision. IT enables the CNN model to be further fine-tuned on the target MRI, which can improve the robustness of our framework against the drastic feature

Table 4
The Ultimate Performance of Three Versions of Spine-GFlow after HT and IT.

Framework	mean IoU	mean Dice
Spine-GFlow (P-)	0.773 ± 0.053	0.870 ± 0.037
Spine-GFlow (L-)	0.809 ± 0.044	0.894 ± 0.029
Spine-GFlow	0.841 ± 0.026	0.914 ± 0.015

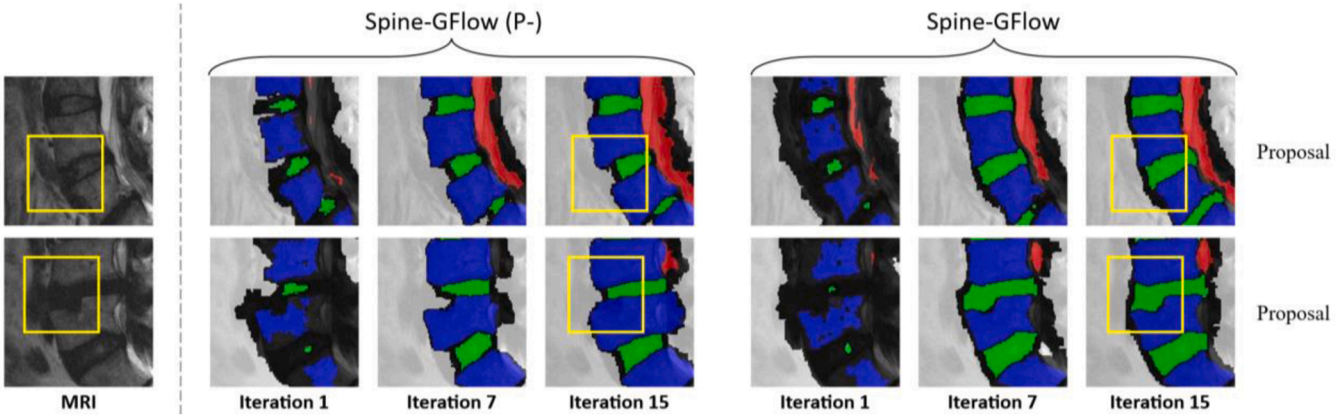

Fig. 10. The proposals generated by Spine-GFlow (P-) and standard Spine-GFlow in different iterations.

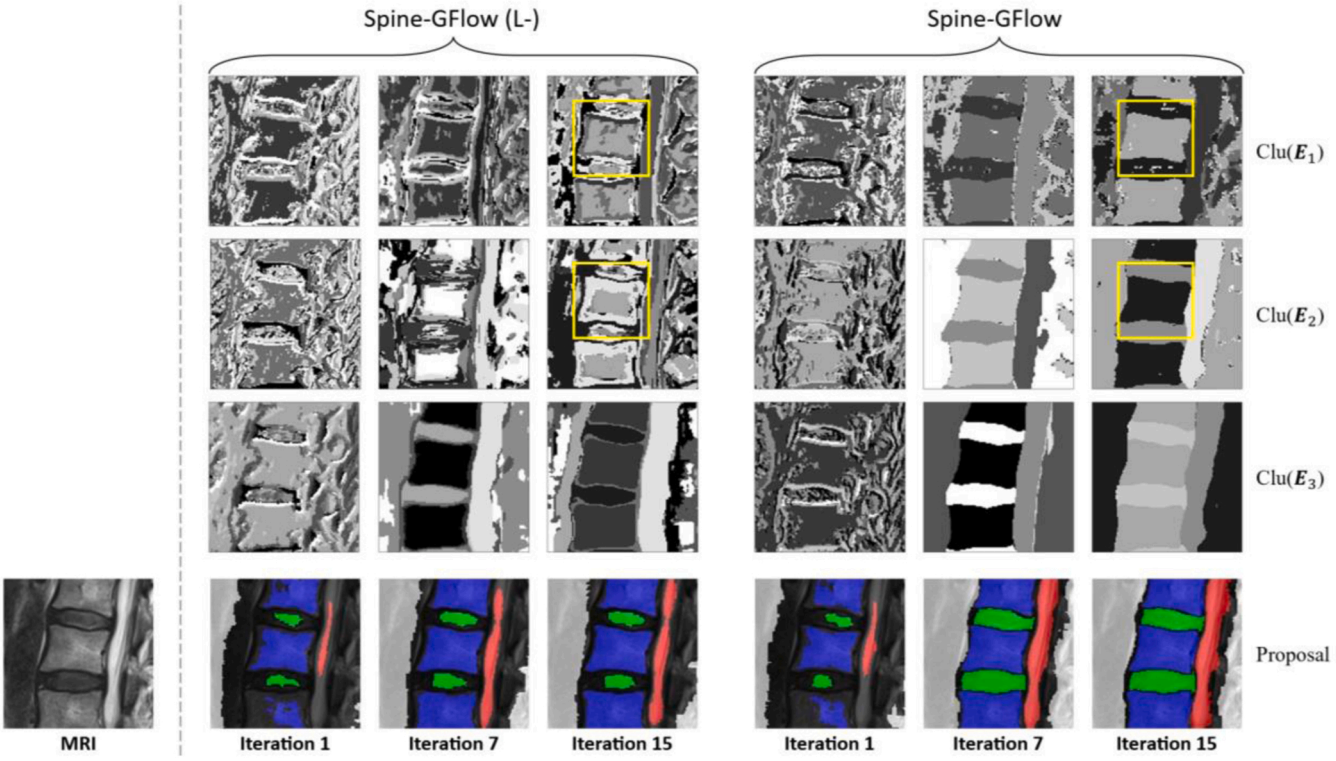

Fig. 11. The pixel clustering results of feature maps generated by the CNN model of Spine-GFlow (L-) and standard Spine-GFlow and the corresponding proposals.

variations caused by pathology or low image quality, such as contrast, which helps our method obtain more accurate results than the weakly-supervised and supervised methods on cases with these feature variations. Since the image feature variations are more likely to appear in VB and IVD regions due to the underlying pathologies, such as Morrow change and disc degeneration, thus our method can achieve better performance in these regions than other methods. The SC has more consistent image features across different cases, thus our pipeline has less advantage in SC. Besides, the multi-source dataset also demonstrated the generalizability of our framework.

Unlike most iterative optimization methods, which generate proposals based on the CNN output, our framework integrates the multi-scale feature maps generated by the CNN model for proposal generation. The output of a CNN model trained with incomplete annotation usually tends to have smooth contours, and the proposals generated with CNN output will lose shape details, especially for tissues with shape deformities. Furthermore, since the tissue boundaries are sometimes fuzzy, such as the edge between IVD and the background, refining methods using low level information such as CRF cannot effectively avoid errors, which will significantly reduce the training efficiency of the CNN model and its ultimate performance.

In addition to the conventional cross entropy PCL, we introduce FDL for the training of the CNN model. Compared with the model trained with only PCL, the model trained with PCL+FDL can generate more discriminative feature maps, where features have large similarities and differences for pixels belonging to the same and different tissues. For feature maps with small scales, this effect of feature aggregation brought by FDL is more significant, which can help the clustering-based method generate more accurate proposals and in turn improve CNN training.

We explicitly embed the anatomical prior in the proposal generation by applying several rule-based fine-tuning operations that utilize the 3D geometry information of adjacent slices. The results show that rule-based fine-tuning can significantly improve the accuracy of the proposals by reducing the potential cavities and wrong bulges (Fig. 12). As

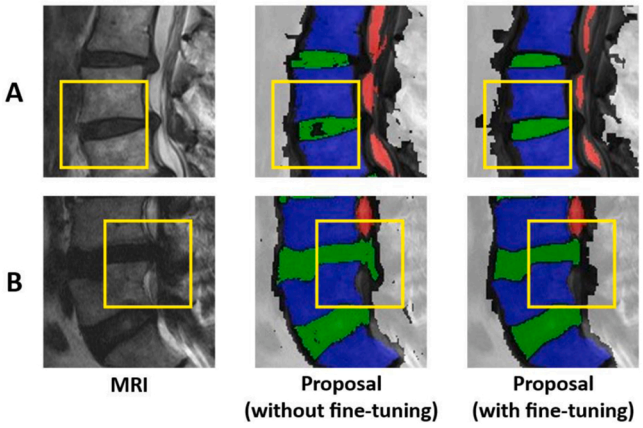

Fig. 12. Two examples of the proposals with and without rule-based fine-tuning.

our rule-based fine-tuning operations are based on the general geometric prior of the tissue, which have high robustness against the geometry variation in different cases due to the potential deformity.

Since our rule-based seed area initialization method only detects the approximate tissue locations and a rough spinal region, it is robust to the image feature variations and can generate satisfactory results for the cases with pathologies and/or low image contrast. There might be a location deviation and/or partial absence in the initial seed areas, but no mistake (Fig. 14). Besides, our framework shows high robustness against suboptimal initial seed areas. The results show that neither location deviation nor partial absence has a significant effect on the final segmentation result. Updating the seed areas can effectively correct location deviation, and the CNN model can be trained with incomplete proposals caused by partial absence.

In future work, our framework will be extended to handle the segmentation of axial lumbar MRI for other spinal tissues such as muscles. Moreover, a prospective clinical study at an independent center will be conducted to further evaluate the performance and stability of our pipeline.

7. Conclusion

In this paper, we have introduced a hybrid framework, named SpineGFlow, for robust multi-tissue segmentation in sagittal lumbar MRIs, which does not rely on any human intervention and manual annotation. We adopted a rule-based method to automatically generate the weak annotation for CNN training. We propose a clustering-based method to generate the proposals by integrating multi-scale feature maps produced by the CNN model, which can produce proposals with shape details. The anatomical prior is explicitly embedded via several rule-based proposal fine-tuning operations. A comprehensive loss is introduced to simultaneously optimize the pixel classifications and feature distribution of feature maps generated by the CNN model, which significantly improves the efficiency of training. Segmentation performance was quantitatively validated and compared with other state-of-the-art methods on the HKDDC dataset that contains the MRI obtained from 3 different machines, and the IVDM3Seg dataset. The results demonstrate that our framework is comparable to a model trained with full supervision. Our framework has significant implications for many MRI analysis tasks, including pathology detection, 3D reconstruction for further auto-diagnosis, and 3D printing.

CRediT authorship contribution statement

Xihe Kuang: Conceptualization, Methodology, Software, Validation, Writing – original draft. **Jason Pui Yin Cheung:** Resources, Investigation, Writing – review & editing. **Kwan-Yee K. Wong:** Methodology,

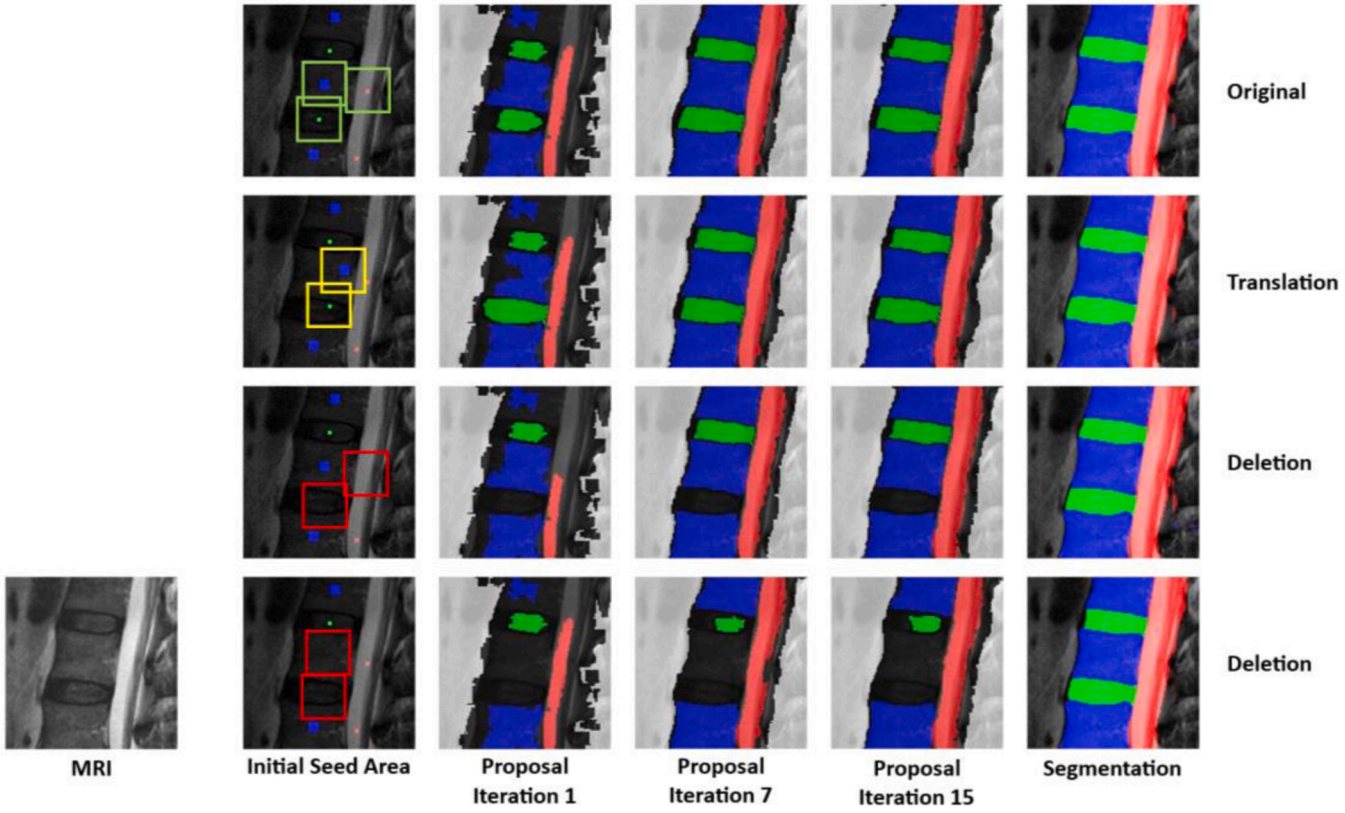

Fig. 13. The proposals and segmentation results produced with the defective initial seed areas.

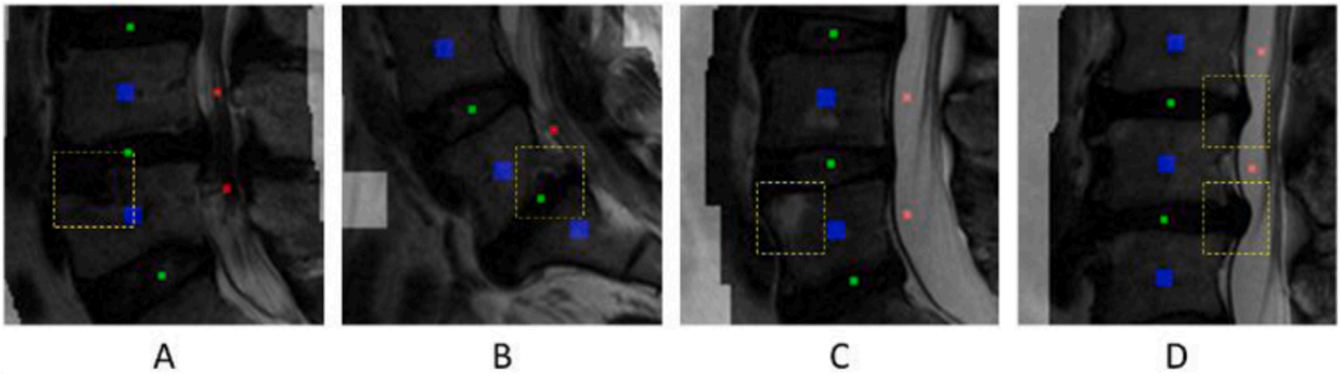

Fig. 14. The initial seed areas in the cases with different feature variations caused by pathologies, including Schmorl's node (A and B), Marrow change (C), disc bulging (D), and/or low image contrast (A and B). The pathology areas were marked by dotted boxes.

Writing – review & editing. **Wai Yi Lam:** Investigation. **Chak Hei Lam:** Investigation. **Richard W. Choy:** Investigation. **Christopher P. Cheng:** Methodology, Software. **Honghan Wu:** Writing – review & editing. **Cao Yang:** Writing – review & editing. **Kun Wang:** Writing – review & editing. **Yang Li:** Supervision, Writing – review & editing. **Teng Zhang:** Conceptualization, Resources, Supervision, Project administration, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank all the participants who joined the study for establishing the Hong Kong Disc Degeneration Cohort (HKDDC) and contributing their MRIs to the dataset. This study was supported by the Innovation and Technology Fund (ITS/404/18, Hong Kong).

References

Bearman, A., Russakovsky, O., Ferrari, V., Fei-Fei, L., 2016. What's the point: Semantic segmentation with point supervision, European Conference on Computer Vision (ECCV). Springer, pp. 549–565.

Benneker, L.M., Heini, P.F., Anderson, S.E., Alini, M., Ito, K., 2005. Correlation of radiographic and MRI parameters to morphological and biochemical assessment of intervertebral disc degeneration. Eur. Spine J. 14, 27–35.

Carballido-Gamio, J., Belongie, S.J., Majumdar, S., 2004. Normalized cuts in 3-D for spinal MRI segmentation. IEEE Trans. Med. Imaging 23, 36–44.

Cheng, C.P., Halchenko, Y.O., 2020. A new virtue of phantom MRI data: explaining variance in human participant data. F1000Research 9.

Cheung, P.W.H., Fong, H.K., Wong, C.S., Cheung, J.P.Y., 2019. The influence of developmental spinal stenosis on the risk of re-operation on an adjacent segment after decompression-only surgery for lumbar spinal stenosis. Bone Jt. J. 101, 154–161.

Dai, J., He, K., Sun, J., 2015. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation, IEEE International Conference on Computer Vision (ICCV), pp. 1635–1643.

Dolz, J., Desrosiers, C., Ben Ayed, I., 2018. IVD-Net: Intervertebral Disc Localization and Segmentation in MRI with a multi-modal UNet, International Workshop and Challenge on Computational Methods and Clinical Applications for Spine Imaging. Springer, pp. 130–143.

Egger, J., Kapur, T., Dukatz, T., Kolodziej, M., Zukić, D., Freisleben, B., Nimsky, C., 2012. Square-cut: a segmentation algorithm on the basis of a rectangle shape. PLoS One 7, e31064.

Harb, R., Knöbelreiter, P., 2021. InfoSeg: Unsupervised Semantic Image Segmentation with Mutual Information Maximization. arXiv preprint arXiv:2110.03477.

He, X., Zhang, H., Landis, M., Sharma, M., Warrington, J., Li, S., 2017. Unsupervised boundary delineation of spinal neural foramina using a multi-feature and adaptive spectral segmentation. Med. Image Anal. 36, 22–40.

Hu, L., Li, J., Peng, X., Xiao, J., Zhan, B., Zu, C., Wu, X., Zhou, J., Wang, Y., 2022. Semi-supervised NPC segmentation with uncertainty and attention guided consistency. Knowl. -Based Syst. 239, 108021.

Hwang, J.-J., Yu, S.X., Shi, J., Collins, M.D., Yang, T.-J., Zhang, X., Chen, L.-C., 2019. Segsort: Segmentation by discriminative sorting of segments, IEEE International Conference on Computer Vision (ICCV), pp. 7334–7344.

Jensen, M.C., Brant-Zawadzki, M.N., Obuchowski, N., Modic, M.T., Malkasian, D., Ross, J.S., 1994. Magnetic resonance imaging of the lumbar spine in people without back pain. New Engl. J. Med. 331, 69–73.

Kervadec, H., Dolz, J., Tang, M., Granger, E., Boykov, Y., Ayed, I.B., 2019. Constrained-CNN losses for weakly supervised segmentation. Med. Image Anal. 54, 88–99.

Kervadec, H., Dolz, J., Wang, S., Granger, E., Ayed, I.B., 2020. Bounding boxes for weakly supervised segmentation: Global constraints get close to full supervision. Med. Imaging Deep Learn. PMLR 365–381.

Khoreva, A., Benenson, R., Hosang, J., Hein, M., Schiele, B., 2017. Simple does it: Weakly supervised instance and semantic segmentation, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 876–885.

Kuang, X., Cheung, J.P.Y., Wu, H., Dokos, S., Zhang, T., 2020. MRI-SegFlow: a novel unsupervised deep learning pipeline enabling accurate vertebral segmentation of MRI images, 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, pp. 1633–1636.

Kulharia, V., Chandra, S., Agrawal, A., Torr, P., Tyagi, A., 2020. Box2seg: Attention weighted loss and discriminative feature learning for weakly supervised segmentation, European Conference on Computer Vision (ECCV). Springer, pp. 290–308.

Lai, M.K.L., Cheung, P.W.H., Samartzis, D., Karppinen, J., Cheung, K.M.C., Cheung, J.P.Y., 2021a. Clinical implications of lumbar developmental spinal stenosis on back pain, radicular leg pain, and disability. Bone Jt. J. 103, 131–140.

Lai, M.K.L., Cheung, P.W.H., Samartzis, D., Karppinen, J., Cheung, K.M.C., Cheung, J.P.Y., 2021b. The profile of the spinal column in subjects with lumbar developmental spinal stenosis. Bone Jt. J. 103, 725–733.

Lee, J., Yi, J., Shin, C., Yoon, S., 2021. BBAM: Bounding box attribution map for weakly supervised semantic and instance segmentation, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2643–2652.

Lin, D., Dai, J., Jia, J., He, K., Sun, J., 2016. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3159–3167.

Lu, J.-T., Pedemonte, S., Bizzo, B., Doyle, S., Andriole, K.P., Michalski, M.H., Gonzalez, R.G., Pomerantz, S.R., 2018. Deep Spine: Automated lumbar vertebral segmentation, disc-level designation, and spinal stenosis grading using deep learning, Machine Learning for Healthcare Conference. PMLR, pp. 403–419.

Ma, Q., Zu, C., Wu, X., Zhou, J., Wang, Y., 2021. Coarse-To-Fine Segmentation of Organs at Risk in Nasopharyngeal Carcinoma Radiotherapy, International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 358–368.

Michopoulou, S.K., Costaridou, L., Panagiotopoulos, E., Speller, R., Panayiotakis, G., Todd-Pokropek, A., 2009. Atlas-based segmentation of degenerated lumbar intervertebral discs from MR images of the spine. IEEE Trans. Biomed. Eng. 56, 2225–2231.

Mirsadeghi, S.E., Royat, A., Rezatofighi, H., 2021. Unsupervised image segmentation by mutual information maximization and adversarial regularization. IEEE Robot. Autom. Lett. 6, 6931–6938.

Nan, Y., Del Ser, J., Walsh, S., Schönlieb, C., Roberts, M., Selby, I., Howard, K., Owen, J., Neville, J., Guiot, J, 2022. Data Harmonisation for Information Fusion in Digital Healthcare: A State-of-the-Art Systematic Review, Meta-Analysis and Future Research Directions. Information Fusion.

Neubert, A., Fripp, J., Engstrom, C., Schwarz, R., Lauer, L., Salvado, O., Crozier, S., 2012. Automated detection, 3D segmentation and analysis of high resolution spine MR images using statistical shape models. Phys. Med. Biol. 57, 8357.

Pathak, D., Krahenbuhl, P., Darrell, T., 2015. Constrained convolutional neural networks for weakly supervised segmentation, IEEE International Conference on Computer Vision (ICCV), pp. 1796–1804.

Pfirrmann, C.W.A., Metzdorf, A., Zanetti, M., Hodler, J., Boos, N., 2001. Magnetic resonance classification of lumbar intervertebral disc degeneration. Spine 26, 1873–1878.

Qu, H., Wu, P., Huang, Q., Yi, J., Yan, Z., Li, K., Riedlinger, G.M., De, S., Zhang, S., Metaxas, D.N., 2020. Weakly supervised deep nuclei segmentation using partial points annotation in histopathology images. IEEE Trans. Med. Imaging 39, 3655–3666.

Rajchl, M., Lee, M.C.H., Oktay, O., Kamnitsas, K., Passerat-Palmbach, J., Bai, W., Damodaram, M., Rutherford, M.A., Hajnal, J.V., Kainz, B., 2016. Deepcut: object segmentation from bounding box annotations using convolutional neural networks. IEEE Trans. Med. Imaging 36, 674–683.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 234–241.

Samartzis, D., Karppinen, J., Chan, D., Luk, K.D.K., Cheung, K.M.C., 2012. The association of lumbar intervertebral disc degeneration on magnetic resonance imaging with body mass index in overweight and obese adults: a population-based study. Arthritis Rheumatol. 64, 1488–1496.

Song, C., Huang, Y., Ouyang, W., Wang, L., 2019. Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3136–3145.

Tang, M., Djelouah, A., Perazzi, F., Boykov, Y., Schroers, C., 2018. Normalized cut loss for weakly-supervised cnn segmentation, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1818–1827.

Tang, P., Yang, P., Nie, D., Wu, X., Zhou, J., Wang, Y., 2022. Unified medical image segmentation by learning from uncertainty in an end-to-end manner. Knowl. -Based Syst., 108215

Valvano, G., Leo, A., Tsaftaris, S.A., 2021. Learning to segment from scribbles using multi-scale adversarial attention gates. IEEE Trans. Med. Imaging.

Van Gansbeke, W., Vandenhende, S., Georgoulis, S., Van Gool, L., 2021. Unsupervised semantic segmentation by contrasting object mask proposals. arXiv preprint arXiv: 2102.06191.

Wang, K., Zhan, B., Zu, C., Wu, X., Zhou, J., Zhou, L., Wang, Y., 2021. Tripled-Uncertainty Guided Mean Teacher Model for Semi-supervised Medical Image Segmentation, International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 450–460.

Yang, G., Ye, Q., Xia, J., 2022. Unbox the black-box for the medical explainable ai via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. Inf. Fusion 77, 29–52.

Yoo, I., Yoo, D., Paeng, K., 2019. Pseudoedgenet: Nuclei segmentation only with point annotations, International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 731–739.

Zheng, G., Chu, C., Belavý, D.L., Ibragimov, B., Korez, R., Vrtovec, T., Hutt, H., Everson, R., Meakin, J., Andrade, I.L., 2017. Evaluation and comparison of 3D intervertebral disc localization and segmentation methods for 3D T2 MR data: a grand challenge. Med. Image Anal. 35, 327–344.

Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J., 2019. Unet++: redesigning skip connections to exploit multiscale features in image segmentation. IEEE Trans. Med. Imaging 39, 1856–1867.