https://dx.doi.org/10.1016/j.bspc.2023.104743

RsALUNet: A reinforcement supervision U-Net-based framework for multi-ROI segmentation of medical images

Yi Huang[a], Jing Jiao[a], Jinhua Yu [a,b], Yongping Zheng[c,d,*], Yuanyuan Wang[a,b,*]
[a] *Biomedical Engineering Center, Fudan University, Shanghai, 200433, China*
[b] *Key Laboratory of Medical Imaging Computing and Computer Assisted Intervention of Shanghai, Fudan University, 200433, China*
[c] *Department of Biomedical Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, China*
[d] *Research Institute for Smart Ageing, The Hong Kong Polytechnic University, Hong Kong SAR, China*

Abstract

A new multiple region of interest (multi-ROI) segmentation framework, RsALUNet, is proposed in this paper, whose backbone was U-Net. Rs represented the reinforcement-supervision strategy by utilizing adversarial learning (AL) between U-Net's decoders and an additional discriminator, which was based on the differences among the segmentation results ($diff_{Segs}$) and labels of multi-ROI ($diff_{Labels}$). As the AL progressed, $diff_{Segs}$ was increasingly similar to $diff_{Labels}$, and this further contributed to a more accurate segmentation of each member in the multi-ROI. In addition to the Rs strategy, three blocks were proposed to enhance RsALUNet, namely, a dilated convolution chain providing diverse and large receptive fields to accommodate different target sizes, a fusion block integrating features of large targets to small ones to optimize the segmentation of the latter, and a location-encoder block extracting multi-scale positional information to enhance the model's attention to the ROI. RsALUNet was evaluated through three multi-ROI segmentation tasks using different imaging modalities, including X-ray, ultrasound, and magnetic-resonance (MR) imaging. The mean Dice coefficient (Dice) increased from 1.1% to 8.5% compared to the other frameworks. The results demonstrate the promising adaptability and extendibility of our strategy and RsALUNet for multi-ROI segmentation in X-ray, ultrasound, and MR images.

Keywords: Multi-ROI segmentation framework, Difference, Reinforcement supervision, Adversarial learning

1 Introduction

Automatic and accurate segmentation plays a basic and important role in medical-image analysis, and is also a critical step in computer-aided diagnosis [1, 2, 3, 4]. Numerous clinical practices require the segmentation of multiple physiological structures to support diagnoses. Consequently, the segmentation of multiple targets (multi-targets) in medical images is necessary and more challenging than the segmentation of a single target. Multitarget segmentation is widely used in a variety of medical-imaging modalities. For example, anteroposterior X-ray images are commonly used to diagnose spinal diseases, and the segmentation of the spine plays a critical role in measuring the Cobb angle [5, 6, 7]. In addition, the progression of scoliosis is often accompanied by vertebral deformities [8, 9, 10, 11]. Vertebral segmentation is also important for detecting them.

Additionally, intravascular ultrasound (IVUS) is an effective imaging method for diagnosing atherosclerosis, a common cardiovascular disease [12, 13]. The media adventitia (MA) and luminal intima (LI) are two membranes in the vascular wall that are reflected in IVUS images; however, atherosclerosis is difficult to diagnose, based solely on MA or LI segmentation. Fortunately, the area difference between the MA and LI provides valuable information for diagnosing atherosclerosis [14, 15]. A similar case exists in the comprehensive diagnosis of brain glioma, based on brain magnetic-resonance (MR) images; whole tumors (WTs), enhancing tumors (ETs), and tumor cores (TCs) are required to be segmented [16]. Thus, it can be concluded that segmenting multiple regions of interest (multi-ROIs) in medical images is extremely necessary and valuable for clinical practices. Medical-image segmentation has been widely studied, and deep-learning methods have demonstrated outstanding capabilities for this. U-Net [17], a prominent deep-learning framework for medical-image segmentation, along with its variant versions, such as UNet++ [18] and nnUNet [19], can extract features at different scales and reuse them, based on the skip connection between the encoder and decoder. Numerous models based on the U-Net architecture have been proven to perform well.

However, their supervision is directly and simply based on corresponding manual labels. If an effective strategy exists to establish the correlation among multi-ROIs and then employ it to reinforce supervision, further improvements are possible. The introduction of prior physiological knowledge can be considered a strategy to strengthen the supervision because the more consistent the segmentation is, with a priori knowledge, the more accurate the result will be. Unfortunately, the prior knowledge of different medical images is quite different. As a result, it is difficult to employ prior knowledge universally. In this situation, effectively achieving reinforcement supervision for the segmenting multi-ROIs in different medical images is a valuable research question.

After a great deal of investigation, we found a common phenomenon in multi-ROI segmentation for medical images; namely, one ROI is embedded into or surrounded by another one, such as the examples shown in Fig. 1. Additionally, more applications are required in clinical practice. To evaluate heart function based on cardiac MR images, the left ventricle and ventricular myocardium must be segmented [20]. Liver cancer is the sixth most frequent cancer and is usually diagnosed, based on computed tomography (CT) or MR images [21]. To comprehensively assess a patient's condition, both the liver and its lesion must be extracted [22]. Additionally, the MA and LI also exist in carotid-artery ultrasound images, and they must be detected simultaneously to assess the cardiovascular risk [23]. For these cases, the difference among multi-ROIs is likely to be an effective instrument for delineating their correlation. We

* Corresponding author.
E-mail address: Yongping Zheng (email: yongping.zheng@polyu.edu.hk) and Yuanyuan Wang (email: yywang@fudan.edu.cn)

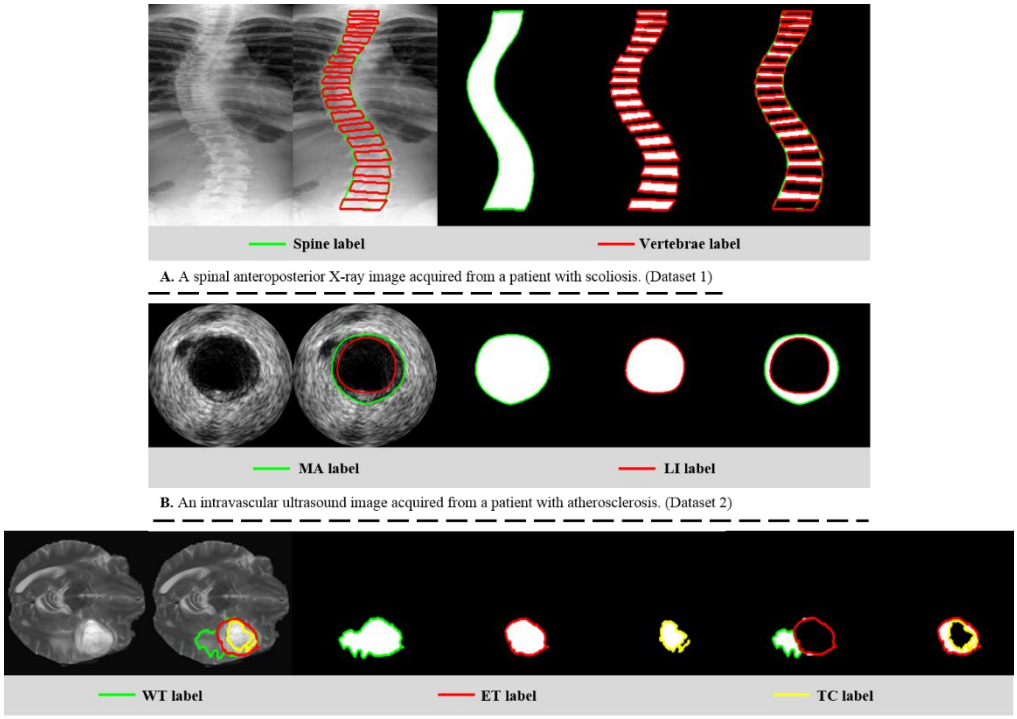

Fig. 1. Examples of multi-ROIs and their differences. Parts A, B, and C are X-ray, ultrasound and MR imaging modalities and come from datasets 1, 2 and 3, respectively. MA and LI mean media adventitia and luminal intima, respectively. WE, ET and TC mean whole tumor, enhancing tumor and tumor core, respectively.

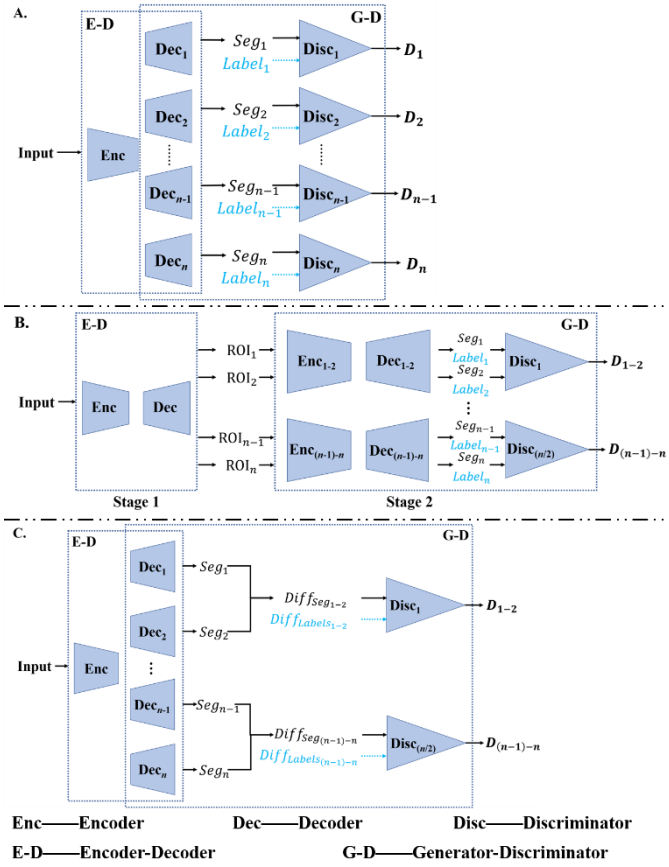

Fig. 2. Backbones of segmentation frameworks with adversarial learning. Parts A and B correspond to existing frameworks; C shows the proposed one.

refer to the difference among multi-ROIs as $diff$ for convenience of description. Their counterparts in the segmented ROIs and corresponding labels are defined as $diff_{Segs}$ and $diff_{Labels}$, respectively. If the differences in the segmentation results and labels become more similar, each member in the multi-ROI will be better segmented. To achieve effective reinforcement supervision based on this consideration, we consider generative adversarial networks (GANs) [24], particularly adversarial learning. A typical GAN-based framework contains a discriminator module after the generator. Under the guidance of adversarial learning, the generator attempts to deceive the discriminator by generating similar output-space distributions, and the discriminator attempts to correctly identify the generated output as real or fake. Thus, the framework can effectively learn the input characteristics without supervision. Moreover, the discriminator promotes the generator; that is, the former urges the latter to generate more real output.

Recently, segmentation studies inspired by adversarial learning have been conducted. Although these models achieved good performance, they either aimed only at a single ROI segmentation, requiring a two-stage implementation with new generators and discriminators, or required an additional discriminator for each member of the multi-ROI, as demonstrated in Figs. 2A and B. To overcome these limitations, in this study, we propose a new reinforcement-supervision strategy based on $diff$. Based on our strategy, a U-Net-based framework with reinforcement supervision achieved via adversarial learning (RsALUNet) was designed.

The backbone of RsALUNet is shown in Fig. 2C. The first module is a segmentation module that segments the multi-ROI. Then, $diff_{Segs}$ is obtained and becomes the input for the discriminator. The second module is considered a GAN module because decoders can be used as generators. Through adversarial learning, $diff_{Segs}$ tends to become more similar to $diff_{Labels}$. During this process, the generator, namely, the decoders, is also optimized to achieve more precise segmentation. The main contributions of our

study are threefold.

(1) A general and concise reinforcement-supervision strategy was designed for segmenting multi-ROIs in medical images. The method exploits $diff$ as the goal of adversarial learning, allowing it to be easily used in medical images of various modalities.

(2) A generative-adversarial structure with a reused decoder was proposed. The simplified network structure is conducive to reducing the training difficulty and improving the inference efficiency.

(3) Three modules, including a dilated convolution chain (DCC), fusion block, and location encoder, were proposed to strengthen the segmentation performance of the framework. They enable larger and more flexible perceptual fields, small-target segmentation, based on the fusion of large-target features, and multiple scale (multi-scale) spatial-attention mechanisms.

The remaining parts of this manuscript is arranged as following: Some relevant works are discussed in section 2. Then, section 3 describes our reinforcement supervision strategy and the RsALUNet in detail. Section 4 gives the description concerning employed datasets, experimental settings and evaluation strategies. Moreover, we show adequate experimental results to demonstrate the effectiveness of our strategy and model in section 5, too. Finally, discussions are proceeded comprehensively in Section 6 and conclusions are given in Section 7.

2 Related work

2.1 U-Net-based frameworks

Owing to their outstanding performance, U-Net-based models have been widely used in numerous medical-image segmentation tasks. To further improve segmentation performance, researchers have focused on enhancing the models' capability for feature extraction. They have also considered the efficient propagation, reuse, and fusion of features. Several outstanding medical-image segmentation frameworks have been proposed, based on these mechanisms.

In Dai and Dong's model, a new pair of residual blocks was designed to promote feature extraction at multiple scales, and a module was proposed to optimize the feature fusion of multiple channels. Finally, skin lesions were effectively segmented [25]. Pi et al. introduced an architecture for mass segmentation of mammograms. This architecture was combined with a parallel dilated convolution module to enlarge the receptive field to extract features at multiple scales. In addition, a novel module based on similarity estimation was proposed to strengthen the feature fusion [26]. Mahmud et al. modified a network for polyp segmentation in colonoscopy images. They also enlarged the receptive field through a depth-dilated inception module to assist with feature extraction. Moreover, a deep-fusion skip module and a deep-reconstruction module were employed to promote the fusion of features and aggregation of feature maps on a multiscale, respectively [27]. Zhao et al. proposed a robust retinal-vascular segmentation network. A nested pyramid architecture was established to strengthen the capability of extracting local vessel details. Furthermore, an attention module promoted feature extraction by weighting each channel from the multi-scale architecture and highlighting channels containing vasculature features [28]. He et al. developed a two-stage framework for prostate segmentation using computed tomography (CT) images. Using a new voxel-metric learning strategy, based on online sampling, to assist the extraction of more representative features, their model performed better than existing ones [29]. The model architecture of single encoder and multiple decoders is also classical and effective for the multi-ROI segmentation. Rashed et al. presented a network consisting of an encoder and parallel multiple decoders, connected by dense skip connections to preserve abundant multi-scale features, for the segmentation of different components of a human head in MR images [30]. Qiu et al. also constructed a model containing single encoder and multiple decoders to achieve myocardial pathology segmentation in cardiac MR image sequences. A novel module and an inclusiveness loss further contribute to the outstanding performance of their model, via regularizing myocardium anatomy consistency and localizing the pathologies [31].

2.2 Prior-knowledge-based frameworks

Prior knowledge is commonly used in multiple-organ or whole-brain segmentation of MR and CT images [30,31]. Atlas-based segmentation is widely employed, which can effectively introduce prior knowledge. Huo et al. reported a supervoxel-atlas-based framework for brain segmentation in MR images. They established references of the brain atlas and performed supervoxel segmentation on them and a target image. The target-image result was matched to one of the atlas results. They were then combined and optimized to achieve the final segmentation [32].

Atlas-based strategies have also been successfully employed in other applications. Zhang et al. proposed a framework for pancreatic segmentation. First, the pancreas was located in a bounding box via multiple-atlas (multi-atlas)-based image registration. Then, the pancreas with its bounding box was sequentially fed into two convolutional neural networks (CNNs) to be further segmented. In their work, the pancreas was located using multi-atlas-based image registration, which effectively employed prior knowledge [33]. An esophagus-segmentation method from planning CT images was introduced by Diniz et al. In their method, the atlas was a probabilistic volume used to find the probable location of the esophagus from a specialist's marks; then, a residual U-Net was combined and the final segmentation result was obtained [34]. Dong et al. first proposed a deep atlas-based network for three-dimensional (3D) left-ventricle (LV) segmentation of echocardiography. Two parts were included in this network: a transformer part and a deformable part. The transformer was used to generate parameters, and these parameters were employed to place the atlas close to the target object; thus, the segmentation was optimized [35].

2.3 GAN-based frameworks

Generative adversarial networks (GANs) were first used for image synthesis. With the deepening study of GANs, some researchers have also employed them to improve the segmentation of medical images.

For example, Pachade et al. designed a two-stage framework for joint optic disc and cup segmentation in fundus images [36]. In the first stage, a CNN was employed

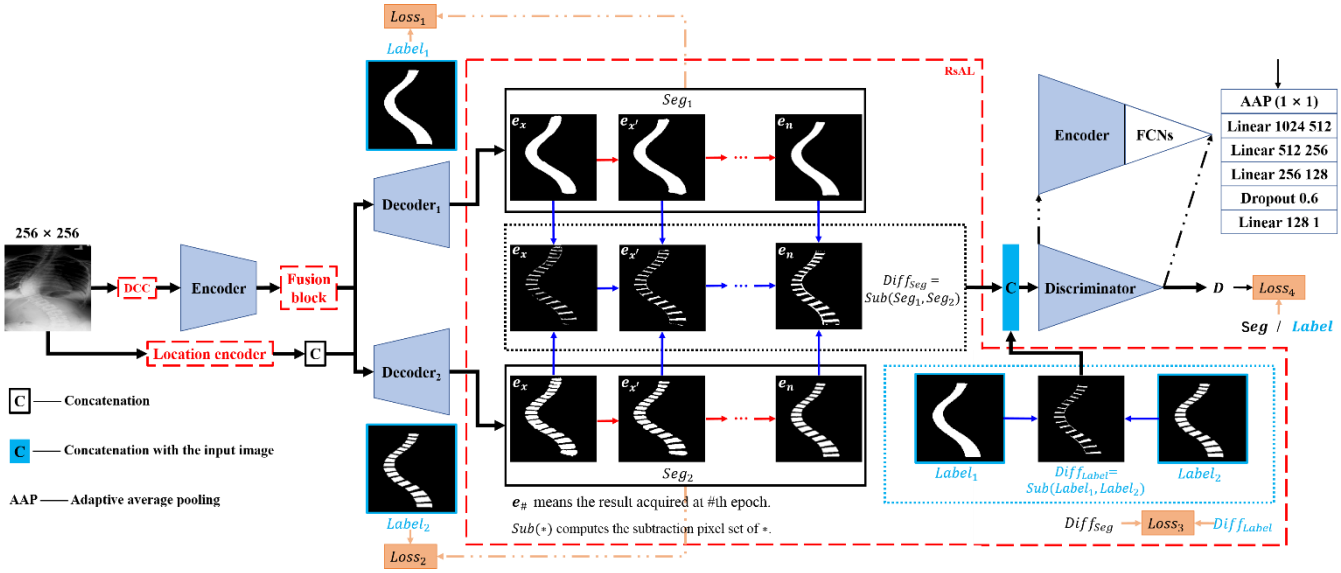

Fig. 3. Illustration of RsALUNet, which is composed of a segmentation module with a GAN and three improved blocks for implementing the RsAL strategy. First, a U-Net-based module, improved by dense convolution, segments the multi-ROI. Then, $diff_{Segs}$ is obtained and fed into the following discriminator; via adversarial learning, it tends to become more similar to $diff_{Labels}$. During this, the generator, namely, the decoders, also are optimized to achieve more precise segmentation.

Table 1. Encoder and decoder architecture.

Layers	Size of output	Details	Encoder	Decoder (Post-processing)
Dense block 1 (Db-1)	128 × 128	$\begin{bmatrix}\text{Conv } (1\times1)\\ \text{Conv } (3\times3)\end{bmatrix}\times 6$		
Transition layer 1 (Tl-1)	64 × 64	$\text{Conv } (1\times1)$ $\text{Averagepooling } (2\times2), \text{stride} = 2$		
Dense block 2 (Db-2)	64 × 64	$\begin{bmatrix}\text{Conv } (1\times1)\\ \text{Conv } (3\times3)\end{bmatrix}\times 12$		
Transition layer 2 (Tl-2)	32 × 32	$\text{Conv } (1\times1)$ $\text{Averagepooling } (2\times2), \text{stride} = 2$		
Dense block 3 (Db-3)	32 × 32	$\begin{bmatrix}\text{Conv } (1\times1)\\ \text{Conv } (3\times3)\end{bmatrix}\times 24$		
Transition layer 3 (Tl-3)	16 × 16	$\text{Conv } (1\times1)$ $\text{Averagepooling } (2\times2), \text{stride} = 2$		
Dense block 4 (Db-4)	16 × 16	$\begin{bmatrix}\text{Conv } (1\times1)\\ \text{Conv } (3\times3)\end{bmatrix}\times 16$		

to locate the ROIs. In the second stage, the ROIs were fed into a generator for segmentation. After combining them with a discriminator to optimize the local segmentation details, precise segmentation results were obtained. Lei et al. proposed a network with two discriminators to segment skin lesions in dermoscopy images [37]. One of the two discriminators focused on the spatial features of the images, and the other concentrated on the details of the segmentation mask. Subsequently, via adversarial learning, they encouraged the segmentation network to provide more precise results. Wu et al. introduced a segmentation framework with two discriminators [38]. It was used to segment the LV in MR images. Similar to Lei's work, segmentation masks, concatenated with images, were the input of one discriminator, and the input of the other discriminators was only masks. The difference between their works was that, in Wu's framework, the output of the dual discriminators was a confidence map, and the loss function was computed based on it. In addition to the above applications, segmenting gross tumors in CT images is a challenging task. Liu et al. designed a framework to achieve this [39]. Adversarial learning was combined with an encoder–decoder segmentation network to balance the distribution differences between small and large targets in a sample. The experimental results indicated that their model could effectively segment targets of different sizes.

As summarized in Figs. 2A and B, the above-mentioned works required additional generators and discriminators or were two-stage implementations. More importantly, they did not attempt to effectively segment multi-ROIs using different medical-imaging modalities. Unlike these methods, our GAN is used to implement our $diff$-based reinforcement supervision strategy. Meanwhile, the discriminator can directly drive decoders to more precisely segment, since the decoders are reused as the generator in our framework due to the introduction of $diff$. The model implementing our strategy is more concise and has great extendibility for different multi-ROI segmentation tasks.

3 RsALUNet

Fig. 3 shows the architecture of RsALUNet, which has two main modules. Fig. 3 shows the architecture of RsALUNet, which has two main modules, an encoder-multiple decoders part and a generator-discriminator one. The input image is firstly fed into the encoder to extract features and multiple decoders segment multi-ROI based on these features. Then, multiple decoders are reused as a generator of $diff_{Segs}$. Concatenating with the input image,

the $diff_{Segs}$ is fed into the following discriminator and the $diff_{Segs}$ tends to be increasingly similar to the $diff_{Labels}$ by adversarial learning, in which the segmentation of each target in multi-ROI become more precise. As can be seen, $diff$ is the core of our study, which is the aim of the adversarial learning. Reused decoders, acting as the generator of $diff_{Segs}$, combine with the added discriminator to construct a GAN module to implement reinforcement supervision based on $diff$. Specifically, reused decoders generate $diff_{Segs}$, which is a type of input for the discriminator. $diff_{Labels}$ is an additional input. With the concatenation of the original input images, the discriminator gives the probability of whether its input is in $diff_{Labels}$. The generator attempts to deceive the discriminator and the discriminator attempts to correctly distinguish the input origin. Their confrontation makes $diff_{Segs}$ more similar to $diff_{Labels}$, which means that the segmentation result of each member in a multi-ROI is closer to its own label.

Additionally, the U-Net-based segmentation module plays an important role in effective multi-ROI segmentation. Dense convolutions and connections improve the backbone of the segmentation module to enhance its basic capability. Moreover, three blocks are proposed to further assist it: a dilated convolution chain (DCC) provides a large and diverse receptive field, a fusion block optimizes the segmentation of a small target via fusing features from a large one, and a location encoder provides multi-scale spatial attention to highlight ROIs.

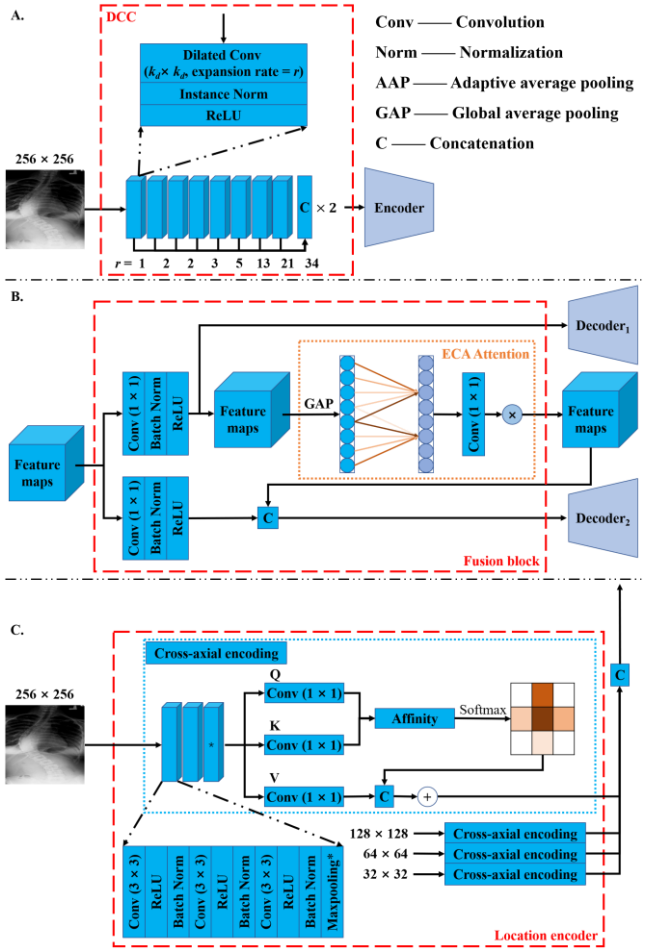

Fig. 4. Block diagrams of the proposed modules. Parts A, B, and C indicate the DCC, fusion block, and location encoder, respectively.

3.1 Segmentation module

The original U-Net is a potential backbone, but lacks targeted improvement for multi-ROI segmentation in different medical images. Its capabilities in three aspects must be enhanced to be competent for this task. First, its multiscale feature-extraction capability should be enhanced to fit ROIs of various sizes in different medical images. Second, an effective strategy should be established to integrate features from different ROIs, thereby assisting their segmentation. Third, it must be able to use contextual information, which has been demonstrated in numerous studies to be helpful for segmentation tasks.

To overcome these limitations, we employ DenseNet121 as the backbone; its architecture is listed in Table 1. It effectively improves the feature-extraction capability via efficient feature reuse and propagation with dense connections [40]. To match the feature-encoding capability of DenseNet121, the decoder is also improved by a dense connection. The details of the decoder are listed in Table 1. It is followed by a post-processing part, including three 3×3 convolution layers, a batch-normalization layer, and a sigmoid layer to restore the output to 256×256. In addition, skip connections, which communicate features without information losses and fuse features on a multiscale, are utilized between the transition layers in the encoder and decoder. To further enable our framework to be applicable to different image modalities, we introduced some functional blocks.

Dilated convolution chain (DCC): ROIs of various sizes appear in different medical images. A variety of receptive fields is necessary to effectively extract features for different cases. A dilated convolution can effectively enlarge the receptive field and, hence, extract more multiscale features [41]. Its capability has been demonstrated in numerous segmentation tasks. We further compose a series of dilated convolutions into a chain structure, as shown in Fig. 4A. In this manner, a previous convolution fills the holes of a later one, which reduces the loss of information. Their expansion rates r_i are set according to the Fibonacci sequence [42]:

$$r_i = \frac{1}{\sqrt{5}} \times \left[\left(\frac{1+\sqrt{5}}{2} \right)^i - \left(\frac{1-\sqrt{5}}{2} \right)^i \right], \qquad (1)$$

where i is the ordinal number of dilated convolutions in the DCC. In addition, the corresponding kernel size k_d for the i^{th} dilated convolution unit is set as (2):

$$k_d = k + (k-1) \times (r_i - 1), \qquad (2)$$

where k is the kernel size of the first dilated convolutional block. The previously dilated convolution unit fills the hole caused by the posterior one. Therefore, after a series of dilated convolutions, a large and diverse receptive field is acquired without holes, and more features are effectively extracted. Each dilated convolution layer, followed by instance normalization and a rectified linear unit (ReLU), creates 32 feature maps. Finally, these feature maps are concatenated as the output of the DCC.

Fusion block (FB): Multiple decoders are employed for multi-ROI segmentation. In this case, we designed a feature fusion block to integrate the features of the external ROI with those of the internal ROI. The purpose is to reuse the features of the external ROI and guide the segmentation of the internal ROI. As shown in Fig. 4B, the efficient channel

attention (ECA) mechanism [43] plays a critical role in the fusion block. The feature map is obtained through global average pooling to extract a $1\times1\times C$ (channel dimension) feature, which indicates the weight of each channel. Subsequently, a one-dimensional (1D) convolution causes cross-channel information to interact without reducing dimensionality. The ECA mechanism is defined as follows:

$$\omega = \mathrm{Sigmoid}\left(Conv_{1\times1}^{k_1}(y)\right), \tag{3}$$

where $Conv_{1\times1}$ represents a 1D convolution, k_1 is the kernel size of $Conv_{1\times1}$, and y is the aggregation feature without dimensionality reduction. Furthermore, k_1 is an important parameter because it determines not only the kernel size of the 1D convolution, but also the range of interactive information. k_1 is adaptively set as follows:

$$k_1 = odd\left(\left|\frac{\log_2(C)}{\gamma}+\frac{b}{\gamma}\right|\right), \tag{4}$$

where γ and b are empirically set to 2 and 1 [43], respectively, and $odd(*)$ indicates the odd number nearest to $*$.

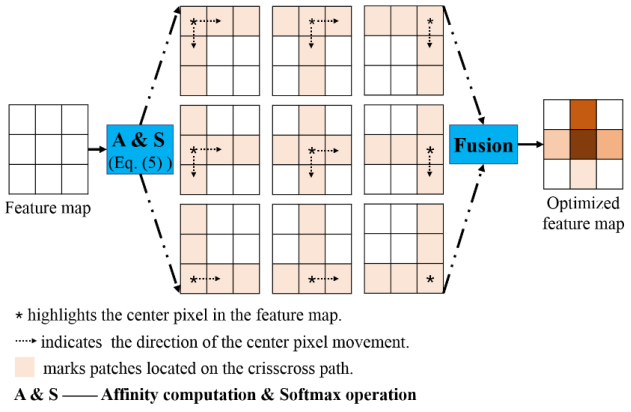

Fig. 5. Encoding example following the crisscross path in LE.

Location encoder (LE): Contextual information is valuable for segmentation tasks. Traditional contextual-information methods, such as pyramid convolution-based methods, can only extract contextual information from the surrounding pixels; therefore, it is difficult to collect dense contextual information for a full feature map. More importantly, these dilated convolution-based methods must set dilated ratios based on experience. These limitations prevent them from effectively assisting in the segmentation of ROIs of different sizes from different modalities. To overcome these problems, a strategy based on crisscross encoding is proposed.

An example given in Fig. 5 briefly describes the implementation of crisscross encoding. Each pixel in the feature map becomes the center pixel, and the affinities between itself and the surrounding pixels located in the top, bottom, left, and right directions, are computed. All the affinity maps are fused into one as the final encoding result. The implementation of this strategy indirectly encodes the location of each pixel in the feature map, and this strategy has been proven to be effective [44, 45]. Consequently, in our framework, a crisscross contextual-information extractor is employed. For convenience, this functional block is called a location encoder (LE).

The structure of the proposed location encoder is shown in Fig. 4C. After the convolution layers, a local $C\times W\times H$ feature map $\boldsymbol{F}$ is obtained, and two convolutional layers, each having a 1×1 kernel, are used to generate two feature maps ($\boldsymbol{Q}$ and $\boldsymbol{K}$). Subsequently, affinity $\boldsymbol{A}$ is calculated:

$$\boldsymbol{A} = \mathrm{Softmax}\left(\boldsymbol{Q}_p\boldsymbol{K}^{*\mathrm{T}}_{i,p}\right), \tag{5}$$

where $\boldsymbol{Q_p}$ is a vector obtained at position p in the spatial dimension of $\boldsymbol{Q}$, and $\boldsymbol{K}^*_{i,p}$ indicates the i^{th} elements of the feature vectors extracted from $\boldsymbol{K}$. Then, the remaining convolutional layer with a 1×1 kernel is employed to generate $\boldsymbol{V}$ for feature extraction; the i^{th} elements of the acquired vector at position p are $\boldsymbol{V}^*_{i,\boldsymbol{p}}$. Finally, $\boldsymbol{A}$ and $\boldsymbol{V}^*_{i,p}$ are aggregated and added to $\boldsymbol{F_p}$ to obtain the location-encoding results, as follows:

$$\boldsymbol{F}'_{\boldsymbol{p}} = \sum_{i=0}^{H+W-1}\boldsymbol{A}_{i,\boldsymbol{p}}\boldsymbol{V}^*_{i,\boldsymbol{p}} + \boldsymbol{F}_{\boldsymbol{p}}. \tag{6}$$

In (6), $\boldsymbol{F}'_p$ is a feature vector in $\boldsymbol{F}_p$ at position p and $\boldsymbol{A}_{i,p}$ denotes a scalar value for channel i and position p in $\boldsymbol{A}$. Except for axial encoding, four branches are established, and their inputs have different scales. The block aims to encode positional information at different scales, and is concatenated with feature maps before each upsampling layer in the decoder. In this manner, more effective assistance can be provided for the decoder.

3.2 Reinforcement supervision via adversarial learning (RsAL)

A difference ($diff$) usually exists, as shown in Fig. 1 and described in Section 1, and it indirectly but effectively describes the relationship between each member in a multi-ROI. It can be considered as shape prior knowledge. It can be imagined that as $diff_{Segs}$ becomes more similar to $diff_{Labels}$, which is more in line with the shape prior knowledge, the segmentation results also improve. Motivated by this, we introduce indirect-shape prior knowledge, based on the difference between ROIs, as reinforcement supervision. Moreover, it is concisely implemented by reusing decoders to establish adversarial learning. The decoders in the segmentation module provide predictions for the multi-ROI, acting as a generator of $diff_{Segs}$. Consequently, our reinforcement supervision can be implemented by adding an additional discriminator and combining it with adversarial learning. The discriminator has a similar architecture to the encoder except for the final layers, which contain an adaptive average pooling layer, a dropout layer, and five fully connected layers, as shown in Fig. 3. The decoders are considered to be a generator, combined with the added discriminator, and then a generator–discriminator (GAN) module is established.

Further, the input of the discriminator is set to concatenate the $diff_{Segs}$ masks and the original image. $diff_{Segs}$ can simultaneously highlight border details belonging to each member of the multi-ROI, and the original image contains spatial information that maps to the segmentation mask, which can also help the discriminator make correct decisions. The design can be summarized as follows

Table 2. Details of the three introduced datasets.

Dataset	Modality	Imaging system	Number of target kinds	Number of images
Dataset 1	Xray	Si5 (Volcano Corporation device)	2 (Spine and vertebrae)	609, 481 for training and 128 for testing
Dataset 2	Ultrasound		2 (MA and LI)	435, 105 for training and 330 for testing
Dataset 3	MR (Four modalities: T1, T1-Gd, T2 and T2-FLAIR)	Scanners vary from 1T to 3T	3 (WT, ET and TC)	472 scan data, via random selecting, 372 for training and 100 for testing (including 4608 slices and 1200 ones, respectively)

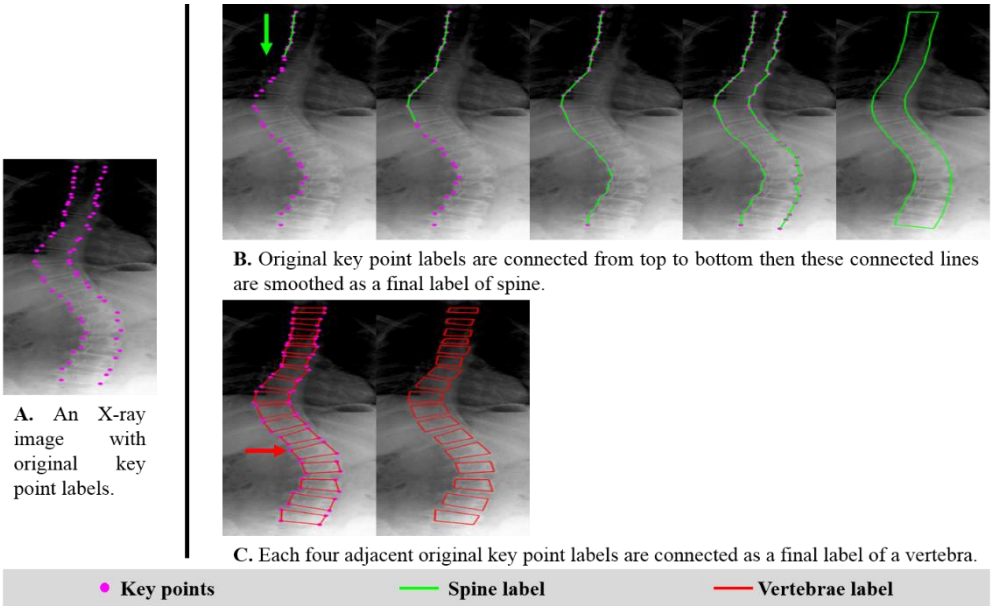

Fig. 6. Examples of spine and vertebrae label generation in dataset 1. Green arrow indicates that lines are connected from top to bottom, and red arrow highlight the adjacent four key points construct a vertebral label.

$$\min_G \max_D \mathrm{E}_{\boldsymbol{x}_i,\boldsymbol{x}_t\sim p_{data}(\boldsymbol{x}_i,\boldsymbol{x}_t)}[\log D(C(\boldsymbol{x}_i,\boldsymbol{x}_t))] + \mathrm{E}_{\boldsymbol{x}_i\sim p_{data}(\boldsymbol{x}_i)}[\log(1 - D(C(\boldsymbol{x}_i, G(\boldsymbol{x}_i))))], \tag{7}$$

$$G(\boldsymbol{x}) = diff_{Segs_{n-1,n}} = Sub(Seg_{n-1}, Seg_n),$$
$$\boldsymbol{x}_t = diff_{Labels_{n-1,n}} = Sub(Label_{n-1}, Label_n), \tag{8}$$

$\boldsymbol{x}_i$ and $\boldsymbol{x}_t$ are the input image and the corresponding $diff$ label, respectively, and $G(*)$ and $D(*)$ represent the output from the generator and discriminator, respectively. The $Sub(*)$ computes the subtraction set of $*$. Seg_{n-1} and Seg_n are binary masks for the segmentation results of the ROI_{n-1} and the ROI_n, respectively. $Label_{n-1}$ and $Label_n$ are defined similarly by the corresponding labels. Moreover, $C(*_1,*_2)$ indicates a concatenation between $*_1$ and $*_2$. G tries to minimize this function by making $diff_{Segs}$ and $diff_{Labels}$ similar, while D attempts to maximize it by distinguishing $diff_{Labels}$ as correctly as possible when training the RsAL. The competition between them gradually facilitates output from the generator, namely, the reused decoders in our model, which is in line with the indirect shape prior knowledge provided by $diff$.

Most published frameworks that used a GAN module constructed a discriminator for each segmentation object and optimized the paired decoder independently; this performed well, but required too many discriminators for multi-ROI segmentation tasks. Moreover, independent optimization makes it difficult to utilize the correlation among targets to enhance the segmentation performance. Relative to these, our strategy not only integrates outputs from each decoder and avoids establishing excessive discriminators, but also jointly optimizes each prediction via their difference. More significantly, the RsAL has strong applicability because it can be effectively utilized for medical images in which $diff$ exists.

As a network that conducts our strategy, RsALUNet is extensible and can be adapted to segmentation tasks with more targets via simple modifications. For example, the MA and LI in IVUS images are two targets that must be segmented. In this case, the architecture of RsALUNet is the same as that shown in Fig. 3, namely, two decoders for two targets and one fusion block (FB) for integrating their feature maps; however, RsALUNet's architecture should be extended when three targets need to be extracted. WT, ET, and TC are three ROIs existing in brain MR images; three decoders should be used to segment them, and an additional FB should be introduced. An additional discriminator should also be introduced because two differences can be obtained via three ROIs. If more ROIs require segmentation, a similar modification can make RsALUNet competent for the task.

3.3 Loss function

Four loss functions are employed in RsALUNet, as shown in Fig. 3, two of them for the segmentation module and the remaining parts for the RsAL. It should be noted that

Table 3. Mean and standard deviation of evaluation metrics calculated between segmentation results and manual labels.

Model	Dataset 1		Dataset 2		Dataset 3		
	Spine	Vertebrae	MA	LI	WT	ET	TC
				Dice			
RsALUNet	**0.93 ± 0.03**	**0.86 ± 0.06**	**0.95 ± 0.03**	**0.95 ± 0.02**	**0.95 ± 0.06**	**0.93 ± 0.13**	**0.90 ± 0.22**
SegAN	0.91 ± 0.03	0.82 ± 0.05	0.94 ± 0.04	0.93 ± 0.05	0.94 ± 0.03	0.90 ± 0.11	0.82 ± 0.18
MedT	0.90 ± 0.04	0.78 ± 0.07	0.92 ± 0.04	0.92 ± 0.03	0.90 ± 0.08	0.83 ± 0.18	0.75 ± 0.21
TransUNet	0.91 ± 0.04	0.82 ± 0.06	0.94 ± 0.04	0.94 ± 0.06	0.93 ± 0.05	0.91 ± 0.11	0.82 ± 0.21
nnUNet	0.90 ± 0.03	0.84 ± 0.04	0.92 ± 0.05	0.94 ± 0.03	0.91 ± 0.13	0.84 ± 0.22	0.83 ± 0.14
				HD (pixels)			
RsALUNet	**10.55 ± 4.75**	**10.98 ± 4.70**	**7.21 ± 3.61**	**5.80 ± 2.33**	**5.92 ± 4.57**	**5.16 ± 3.76**	**4.90 ± 3.86**
SegAN	13.68 ± 10.73	14.78 ± 10.28	8.11 ± 4.19	7.16 ± 5.09	10.82 ± 10.08	6.01 ± 4.71	5.82 ± 4.83
MedT	12.59 ± 6.28	14.98 ± 9.31	11.64 ± 4.96	8.38 ± 3.53	9.45 ± 6.93	8.19 ± 6.14	7.57 ± 6.54
TransUNet	14.38 ± 12.99	16.18 ± 12.15	8.39 ± 5.23	6.49 ± 5.17	6.45 ± 5.67	6.38 ± 6.40	7.07 ± 7.34
nnUNet	12.12 ± 4.17	14.53 ± 4.38	10.96 ± 5.79	7.06 ± 4.04	8.46 ± 5.90	8.06 ± 4.60	7.49 ± 4.81

Table 4. Mean and standard deviation of evaluation metrics calculated between segmentation results and manual labels in ablation experiments.

Model	Dataset 1		Dataset 2		Dataset 3		
	Spine	Vertebrae	MA	LI	WT	ET	TC
				Dice			
$UNet_{backbone}$	0.89 ± 0.05	0.80 ± 0.06	0.94 ± 0.06	0.95 ± 0.02	0.91 ± 0.10	0.86 ± 0.18	0.72 ± 0.29
$UNet_{backbone}$ + DCC	0.89 ± 0.03	0.83 ± 0.05	0.95 ± 0.04	0.95 ± 0.02	0.94 ± 0.06	0.89 ± 0.12	0.79 ± 0.22
$UNet_{backbone}$ + FB	0.91 ± 0.03	0.84 ± 0.05	0.95 ± 0.03	0.95 ± 0.02	0.91 ± 0.11	0.84 ± 0.21	0.76 ± 0.26
$UNet_{backbone}$ + LE	0.86 ± 0.11	0.80 ± 0.09	0.88 ± 0.12	0.90 ± 0.12	0.90 ± 0.13	0.83 ± 0.22	0.72 ± 0.30
$UNet_{backbone}$ + RsAL	0.91 ± 0.04	0.83 ± 0.05	0.94 ± 0.05	0.95 ± 0.02	0.93 ± 0.05	0.87 ± 0.16	0.76 ± 0.25
$UNet_{backbone}$ + discriminators	0.91 ± 0.03	0.84 ± 0.05	0.95 ± 0.03	0.94 ± 0.02	0.91 ± 0.10	0.84 ± 0.21	0.71 ± 0.29
				HD (pixels)			
$UNet_{backbone}$	13.77±7.84	15.37±11.10	7.02±4.52	5.69±2.07	8.66±6.50	6.30±4.18	6.42±4.52
$UNet_{backbone}$ + DCC	13.06±5.07	12.92±5.64	6.86±3.50	5.74±2.23	6.54±5.07	5.84±4.06	5.63±3.97
$UNet_{backbone}$ + FB	13.14±5.64	12.96±5.51	6.81±3.33	5.75±2.35	7.77±5.82	6.41±4.35	6.19±5.08
$UNet_{backbone}$ + LE	17.87±21.72	17.65±18.80	12.44±6.53	10.50±13.22	8.30±5.94	6.57±4.45	6.46±4.67
$UNet_{backbone}$ + RsAL	12.33±6.49	12.94±6.61	7.68±4.93	5.51±2.05	7.72±5.37	6.43±4.22	6.05±4.27
$UNet_{backbone}$ + discriminators	12.77±7.97	12.55±6.25	7.14±3.25	6.34±2.09	8.66±6.53	6.91±5.28	6.91±5.33

$UNet_{backbone}$ + $*$ indicates the block or strategy of $*$ is further employed based on the U-Net-based segmentation module ($UNet_{backbone}$).

the following description is based on a situation in which two ROIs exist; it can be extended for scenarios in which more ROIs exist. Specifically, in the segmentation module, the combined loss function is defined as follows:

$$\begin{aligned} Loss_{\text{seg}} &= \omega_1 Loss_1 + \omega_2 Loss_2 \\ &= \omega_1 Loss_{BCE}(Seg_1, Label_1) \\ &+ \omega_2 Loss_{BCE}(Seg_2, Label_2), \end{aligned} \tag{9}$$

where ω_1 is the weight of the loss function. Furthermore, $Loss_{\text{BCE}}$ represents the binary cross-entropy (BCE) loss function, which is widely employed for segmentation tasks, and calculates the difference between the probability distributions of the segmentation result and the corresponding label. In RsAL, to highlight the local details of $diff$, the structural-similarity index measurement (SSIM) is modified to become a loss function ($Loss_3$), as follows:

$$\text{SSIM}(a, b) = \frac{(2\mu_a\mu_b + C_1)(2\sigma_a\sigma_b + C_2)}{(\mu_a^2 + \mu_b^2 + C_1)(\sigma_a^2 + \sigma_b^2 + C_2)}, \tag{10}$$

$$Loss_{\text{SSIM}} = 1 - \text{SSIM}\left(diff_{Segs}, diff_{Labels}\right), \tag{11}$$

where μ, σ, and σ^2 represent the mean, variance, and convenience, respectively. Additionally, C_1 and C_2 are constants that ensure division. Furthermore, the objective function of RsAL, namely, (7), is also a loss function ($Loss_4$) belonging to the discriminator, as demonstrated in (12):

$$\begin{aligned} Loss_D = E_{\boldsymbol{x}_i, \boldsymbol{x}_t \sim p_{data}(\boldsymbol{x}_i, \boldsymbol{x}_t)}[\log D(\boldsymbol{x}_i, \boldsymbol{x}_t)] \\ + \mathrm{E}_{\boldsymbol{x}_i \sim p_{data}(\boldsymbol{x}_i)}[\log(1 \\ - D(\boldsymbol{x}_i, G(\boldsymbol{x}_i)))]. \end{aligned} \tag{12}$$

Combining these two parts of formulas (11) and (12), the final loss function belonging to the RsAL strategy is summarized as follows:

$$Loss_{\text{RsAL}} = Loss_{\text{SSIM}} + Loss_D. \tag{13}$$

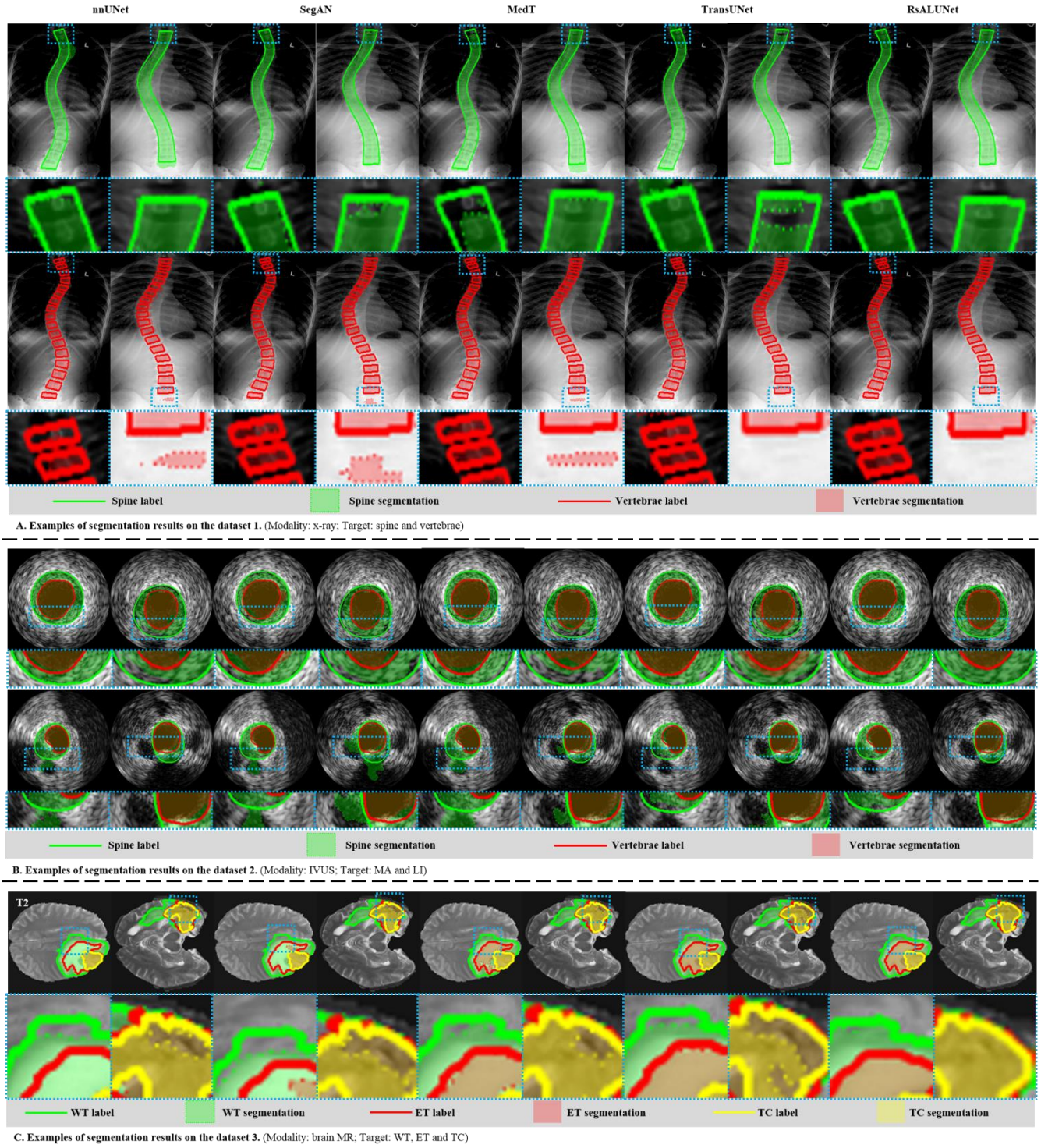

Fig. 7. Examples of segmentation results from RsALUNet and four compared models. Parts A, B, and C refer to datasets 1, 2, and 3, respectively. Blue dashed lines highlight differences between segmentation results and labels. MA and LI mean media adventitia and luminal intima, respectively. WE, ET and TC mean whole tumor, enhancing tumor and tumor core, respectively.

4 Experiment

In this section, we first present three employed datasets, including details on its acquisition, annotation and organization for training and testing (4.1), then report implementation details and evaluation metrics (4.2 and 4.3). Furthermore, the introduction of compared networks and the strategy of ablation studies are illustrated in 4.3, too.

4.1 Materials

To comprehensively investigate the performance of RsALUNet, it was evaluated on three different medical-image datasets, including X-ray, ultrasound, and MR images. Detailed information on these datasets is provided in Table 2. The first dataset is an X-ray dataset [46] (http://spineweb.digitalimaginggroup.ca/), which contains 609 spinal anteroposterior X-ray images and the corresponding ground truths of the spine and vertebrae. A total of 481 and 128 images were organized as the training and testing sets, respectively. The key points of 17 vertebrae were delineated by two clinical experts, and the spine and

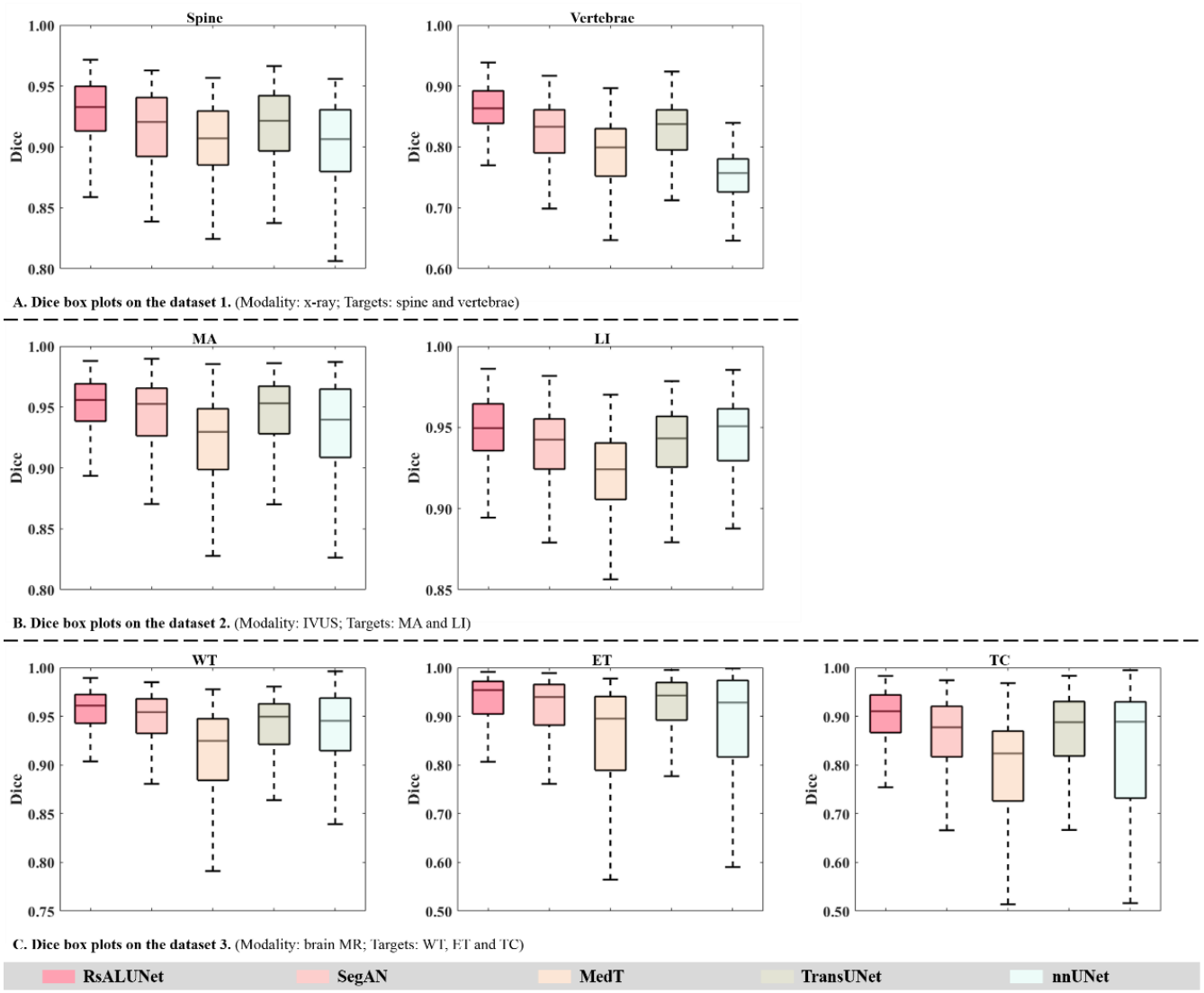

Fig. 8. Dice box plots between RsALUNet and four compared networks. Parts A, B, and C refer to datasets 1, 2, and 3, respectively. MA and LI mean media adventitia and luminal intima, respectively. WE, ET and TC mean whole tumor, enhancing tumor and tumor core, respectively.

vertebrae were manually annotated, based on them. As shown in Fig. 6, these points are linked in order from top to bottom as the label of the entire spine, and every four local points are linked as the label of one vertebra.

The second dataset is an IVUS dataset [47] (http://www.cvc.uab.es/IVUSchallenge2011/dataset.html/),in which 435 IVUS images with manually labeled vessel membranes were included. Of these, 105 and 330 IVUS images were organized as training and testing sets, respectively. Clinical experts provided the borders of the media adventitia and luminal intima as gold standards, as shown in Fig. 1B. Further, it should be noted that the division of the training and test sets in datasets 1 and 2 was determined by the dataset organizer, rather than by us.

The third is a brain MR-image dataset [48] (http://medicaldecathlon.com/) and four different modalities are collected in it; namely, native T1-weight (T1), post-gadolinium (Gd) contrast T1-weight (T1-Gd), native T2-weight (T2), and T2 fluid-attenuated inversion recovery (T2-FLAIR). We selected 472 scans as our dataset, which simultaneously contained WT, ET, and TC. The corresponding labels were confirmed by experienced neuroradiologists. Of these, 372 and 100 were divided into training and testing sets through our random selection; they contained 4608 and 1200 slices, respectively.

The size of all images and corresponding labels are uniformed to 256 × 256 via the bilinear interpolation then used by the RsALUNet. The original size of images in the dataset 2 and 3 are 256 × 256 and 240 × 240 respectively, so the information loss of them is little after resizing. For the dataset 1, although the original sizes of images are various due to different imaging system parameter and individual differences, some preprocessings have been conducted by the organizers of the dataset 1, such as removing cervical and pelvis parts and only retaining the spine ones, which ensures that the whole spine is always retained in images even after resizing. Thus, the resize images and labels in the dataset 1 still are effectively.

4.2 Implementation details

The RsALUNet implementation was based on PyTorch running on an NVIDIA® Tesla V100-SXM2. The Adam optimizer was employed for the network parameters with an initial learning rate of $3e^{-4}$ and a decay rate of $3e^{-7}$. The number of weights in L_{seg} is determined by the number of ROIs in the datasets, and their values should be as balanced as possible to avoid the model becoming biased towards one of the ROIs.

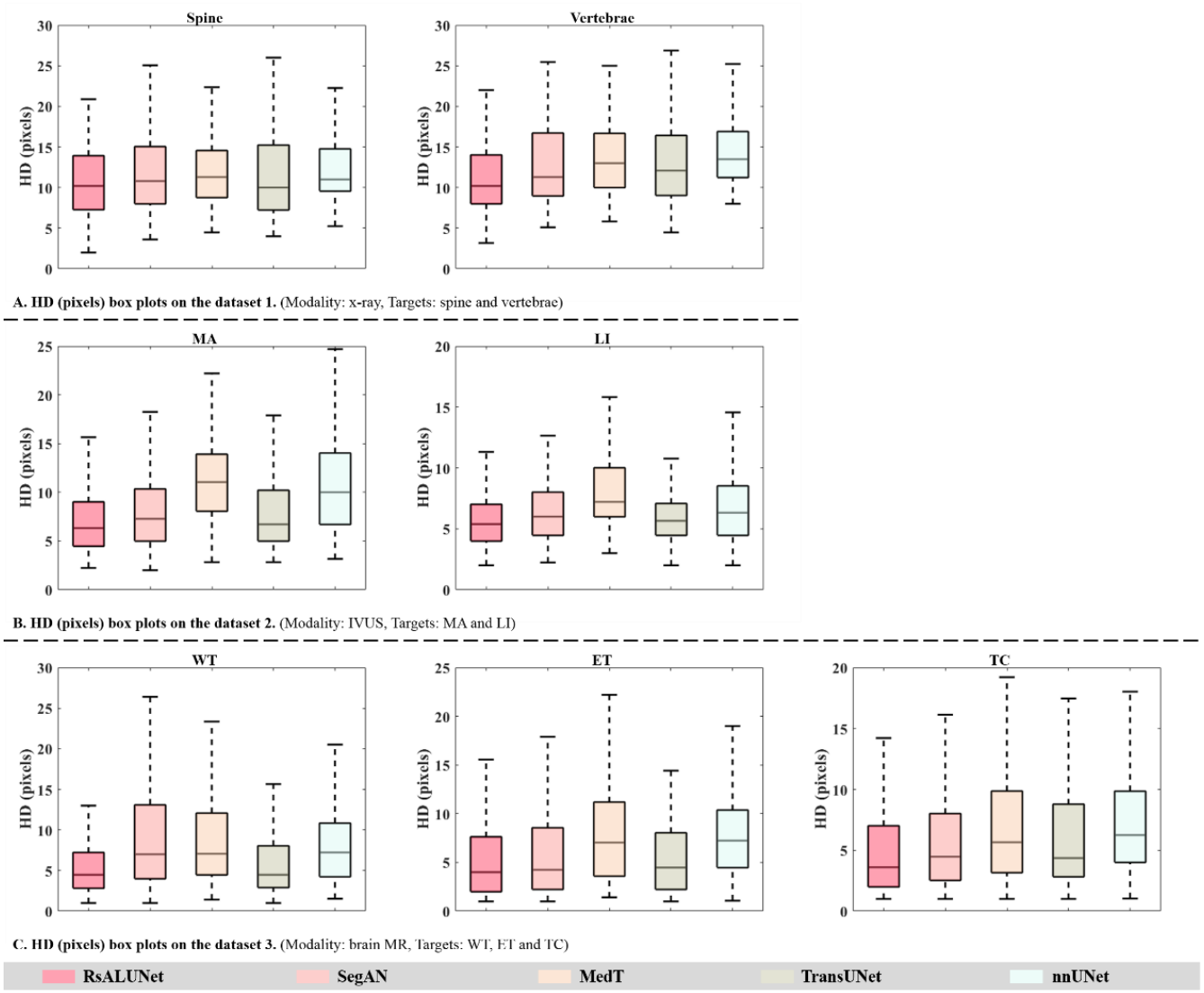

Fig. 9. HD (pixels) box plots between RsALUNet and four compared networks. Parts A, B, and C refer to datasets 1, 2, and 3, respectively. MA and LI mean media adventitia and luminal intima, respectively. WE, ET and TC mean whole tumor, enhancing tumor and tumor core, respectively.

To avoid the model has a serious bias for one target in the multi-ROI, for datasets 1 and 2, two ROIs must be segmented, so two weights (w_1 and w_2) were defined and setto 0.5 and 0.5, respectively. Similarly, for dataset 3, the weights of L_{seg} were extended to w_1, w_2, and w_3 for WT, ET, and TC, respectively, and they were set to 0.3, 0.3, and 0.4, respectively. The model training consisted of two alternate optimization stages. The segmentation module, which includes the encoder, decoders, and three functional blocks, is the first stage. The RsAL, which includes reused decoders and the added discriminator, is the second one. The discriminator parameters were frozen during the first stage, and those belonging to the encoder and functional blocks were frozen during the second stage. Each stage was optimized with 10 epochs. Five-fold cross-validation was performed, and the average performance of the five folds was employed for further comparisons.

4.3 Evaluation metrics and strategies

The Dice coefficient (Dice) and the Hausdoff distance (HD) were employed as quantitative metrics to evaluate the performance of RsALUNet and the compared networks, using the following equations. Dice, a regional indicator, indicates the degree of overlap between the segmentation results and manual annotations, and a larger Dice means that the segmentation results are more precise. HD, a distance-based metric, measures the maximum distance between a point on the border of the segmentation result and the nearest point belonging to the ground truth.

$$\text{Dice} = \frac{2|Seg \cap Label|}{|Seg| + |Label|}, \tag{12}$$

$$\text{HD} = \max_{a \in Con(Seg)} \left\{ \min_{b \in Con(Label)} Dist(a,b) \right\}, \tag{13}$$

where Seg and $Label$ represent the regions of the segmentation result and the manual label, respectively. Furthermore, $Dist(*)$ calculates the Euclidean distance of $*$, and a and b are points belonging to the contours of Seg and $Label$, marked by $Con(Seg)$ and $Con(Label)$, respectively. Box plots were used to visualize these quantitative metrics.

The quantitative assessment of RsALUNet and the comparison with other networks were conducted for all three datasets. The compared networks were SegAN, MedT, TransUNet, and nnUNet. SegAN is a representative segmentation framework that employs adversarial learning [49]. MedT has demonstrated its capability to segment preterm neonatal brain ventricles in brain ultrasound images [45]. TransUNet is useful for extracting multiple organs from

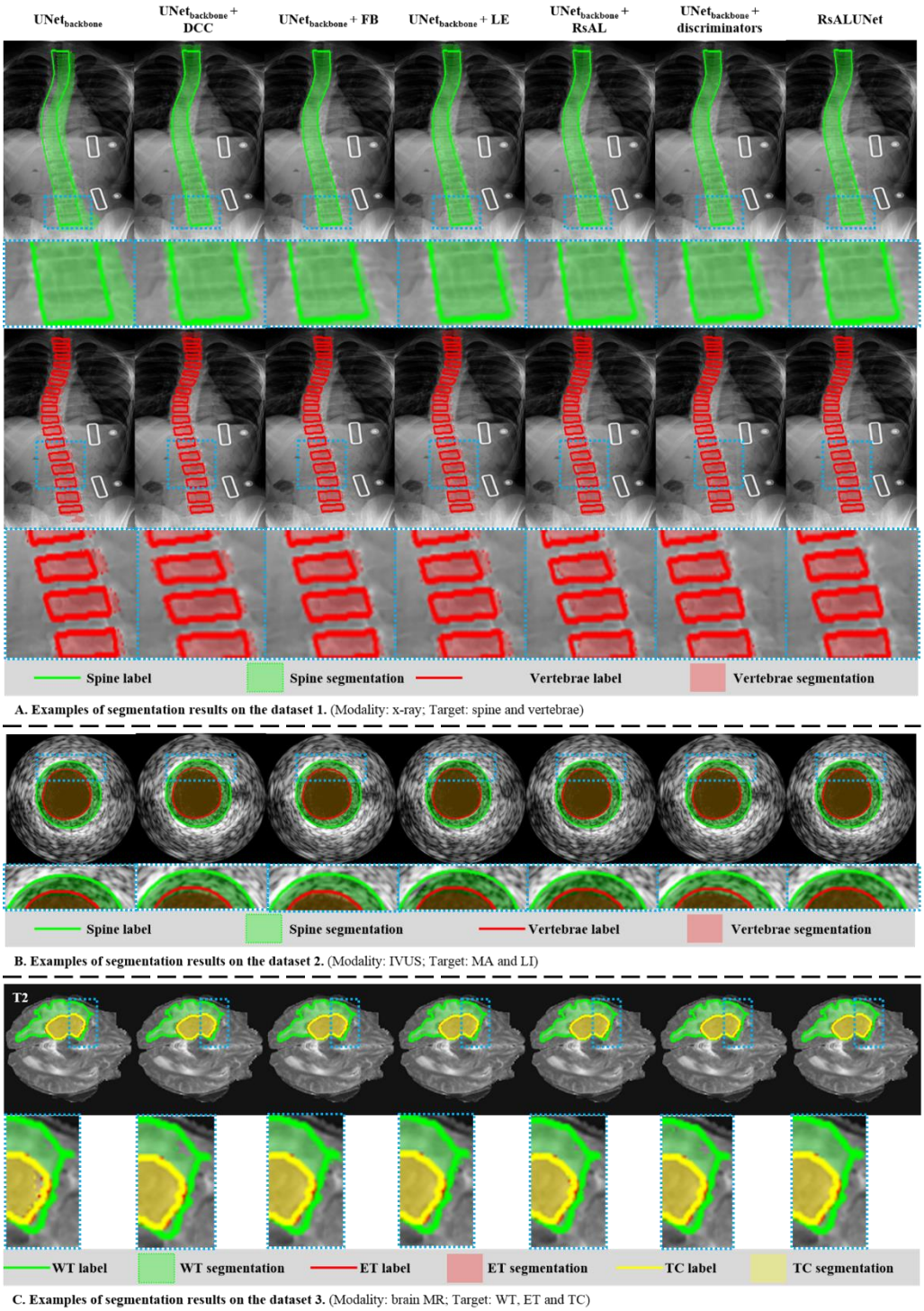

Fig. 10. Examples of segmentation results from ablation experiments. Parts A, B, and C refer to datasets 1, 2, and 3, respectively. Blue dashed lines highlight differences between segmentation results and labels. MA and LI mean media adventitia and luminal intima, respectively. WE, ET and TC mean whole tumor, enhancing tumor and tumor core, respectively.

CT images [50]. nnUNet is an improved version of UNet and performs well for multi-ROI segmentation in brain MR images. All of these are outstanding, and have been proposed in recent years. Furthermore, to highlight the contribution of each function block in the segmentation module, including the DCC, fusion block, and location encoder, ablation experiments were performed on the three datasets. The loss function used in these experiments followed $Loss_{\text{seg}}$.

In addition, to further investigate the effectiveness of RsAL, two more ablation experiments were conducted. In the first one, the backbone of the segmentation module cooperated with the RsAL, and the employed loss function was a combination of $Loss_{\text{seg}}$ and $Loss_{\text{RsAL}}$. In the second, adversarial learning was introduced to strengthen the supervision; however, it was achieved via two decoders working with two discriminators, rather than sharing one discriminator, as in our strategy. Furthermore, the loss function employed was the integration of $Loss_{\text{seg}}$ and $Loss_D$.

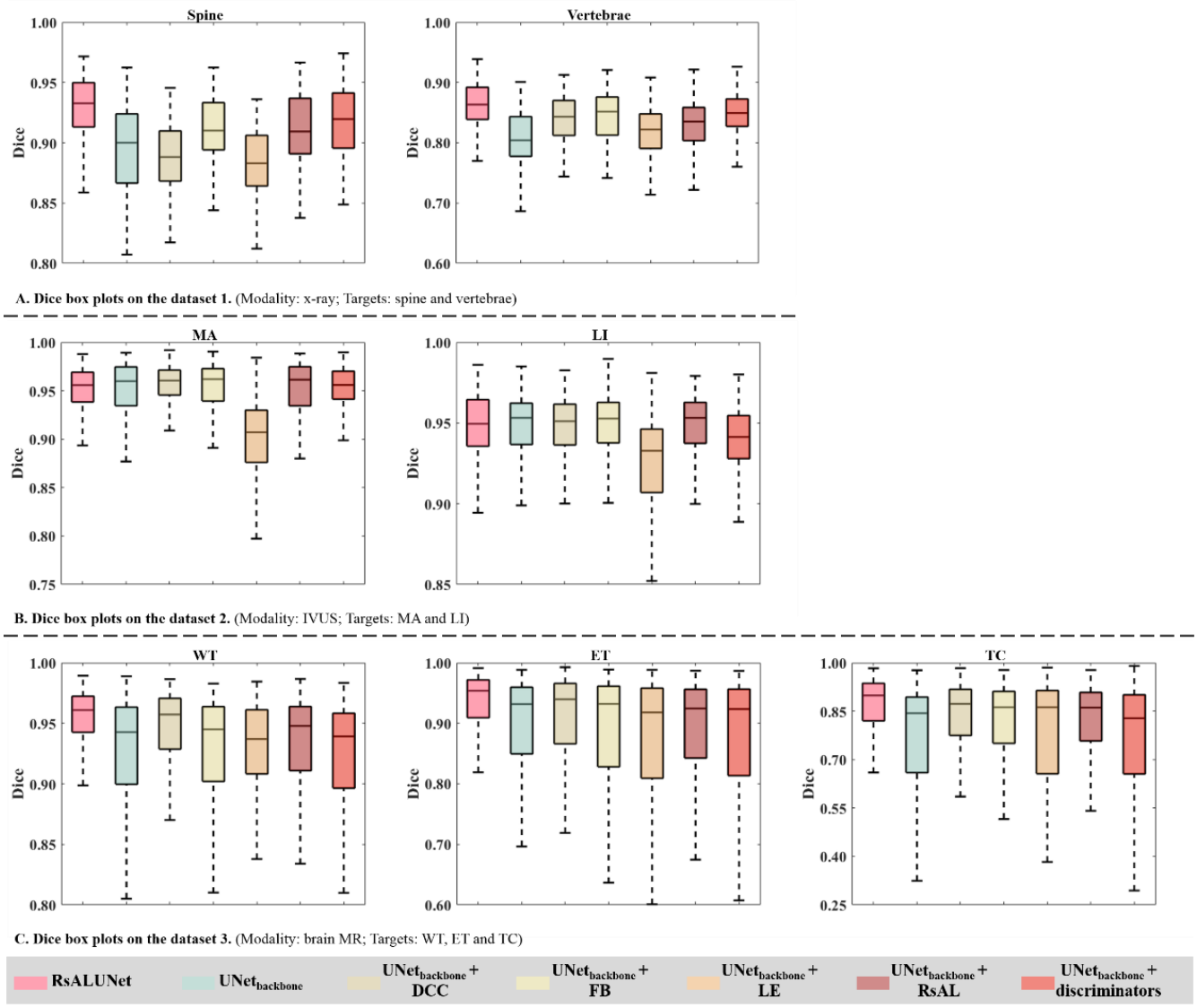

Fig. 11. Dice box plots in ablation experiments. Parts A, B, and C refer to datasets 1, 2, and 3, respectively. MA and LI mean media adventitia and luminal intima, respectively. WE, ET and TC mean whole tumor, enhancing tumor and tumor core, respectively.

5 Results

In this section, adequate experimental results, including comparison with existing models and ablation experiments, are given to comprehensively investigate and effectivelyreveal the effectiveness and superiority of RsALUNet. These results are displayed through three perspectives, including quantitative evaluation metrics, box plots and predication examples.

5.1 Comparison with other frameworks

RsALUNet was trained and tested on three different medical-image datasets. Furthermore, its competitiveness was evaluated by comparisons with four current, outstanding models: SegAN, MedT, TransUNet, and nnUNet.

Segmentation of spine and vertebrae in anteroposterior X-ray images (dataset 1): As quantitative metrics of dataset 1 are given in Table 3, our model outperforms the best of the compared models for the entire spine segmentation, achieving an increase of 2.2% in mean Dice and a decrease of 16.2% in mean HD (pixels). For vertebrae segmentation, RsALUNet further achieved a Dice / HD (pixels) optimization of approximately 4.9% / -25.7%, in comparison with the published networks. These results demonstrate that RsALUNet is outstanding in both spine and vertebral segmentation. Fig. 8A displays the Dice evaluation results via box plots, which further highlights the advantages of our model.

For the spinal segmentation, the mean performances of all networks were close, including RsALUNet and the compared ones; however, the distribution of RsALUNet is more concentrated. For vertebral segmentation, the mean performance of RsALUNet was better, and the distribution was obviously more concentrated. Similarly, those of HD (pixels) in Fig. 9A further highlight the advantages of the proposed model. In addition to the metrics evaluation, some examples of the prediction results from our model and the compared models are shown in Fig. 7A. Compared with the three existing frameworks, the spine border given by RsALUNet is more precise, and the border details of each vertebra are closer to the manual annotations from experts.

Segmentation of MA and LI in IVUS images (dataset 2): Table 3 also shows the quantitative evaluation results for dataset 2. Because of the regularity of the shapes belonging

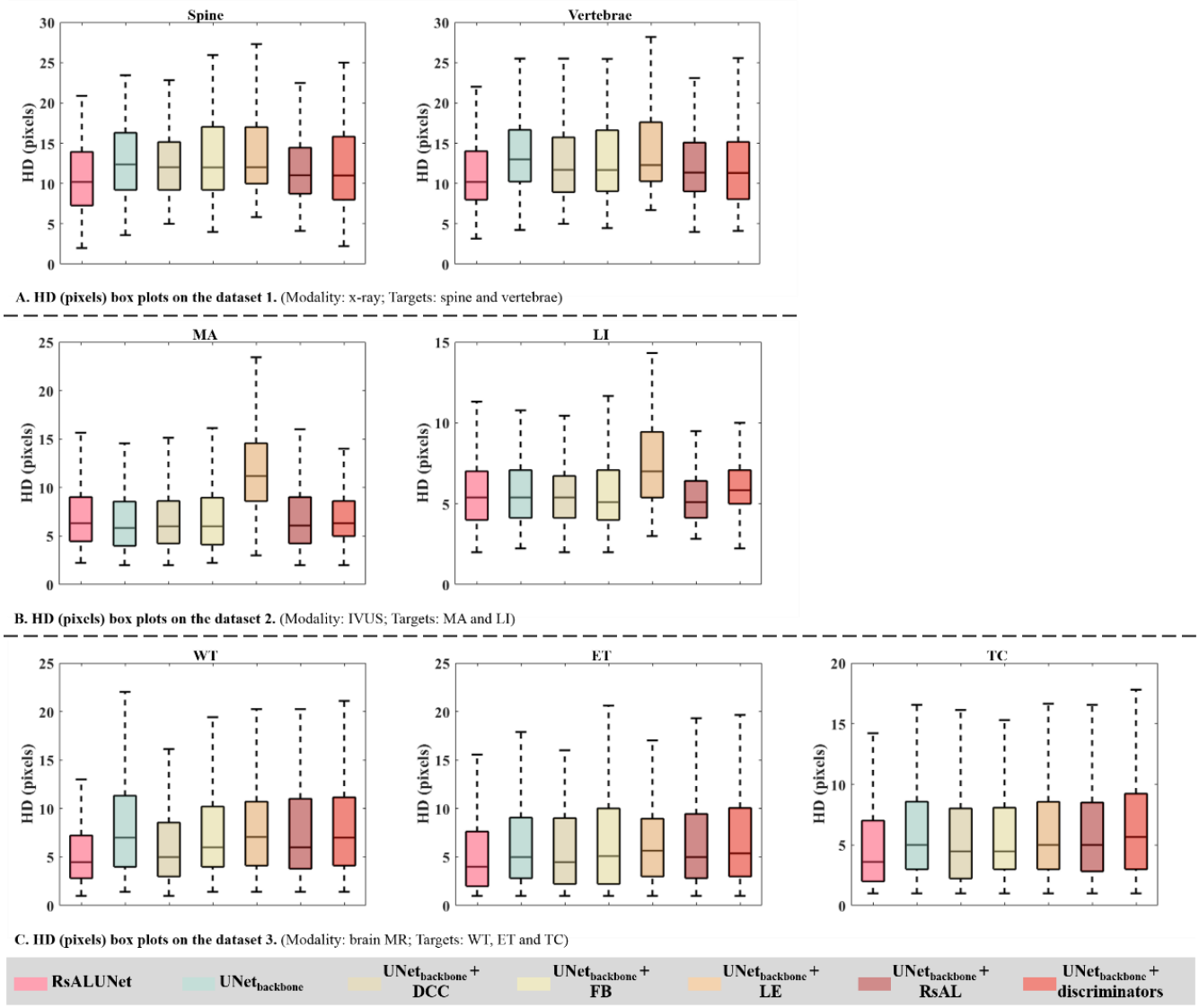

Fig. 12. HD (pixel) box plots in ablation experiments. Parts A, B, and C refer to datasets 1, 2, and 3, respectively. MA and LI mean media adventitia and luminal intima, respectively. WE, ET and TC mean whole tumor, enhancing tumor and tumor core, respectively.

to the MA and LI, this task was relatively easy. It can be seen that both RsALUNet and the compared models achieved promising and close performance for both MA and LI segmentation in IVUS images. The increases in mean Dice, obtained by the RsALUNet, were 1.1% and 1.1% for MA and LI border extraction, respectively. In addition, 11.1% and 10.6% reductions in mean HD (pixels) were achieved by RsALUNet for MA and LI segmentation, respectively. Similar conclusions can be obtained by analyzing the box plots of dataset 2, as displayed in Figs. 8B and 9B.

However, one thing is worth noting. Two typical interference shapes, caused by the anatomical structure of the coronary artery, bifurcation, and side vessel, often appear in IVUS images and disrupt the border of the vessel membranes. In this case, the MA and LI are difficult to segment effectively. Facing this problem, RsALUNet still obtained an effective segmentation and did not include the bifurcation region as a part of the LI region, as did the compared models, as demonstrated by examples given in Fig. 7B.

Segmentation of WT, ET, and TC in brain MR images (dataset 3): Unlike spine and vessel membranes, which have regular shapes, the shape of the brain glioma is very irregular. Moreover, there are three ROIs, WT, ET, and TC, existing in brain MR images that need to be extracted; hence, this task is more challenging. Fortunately, the performance of RsALUNet is satisfactory. The evaluation metrics for dataset3 are summarized in Table 3. For the WT and ET segmentation, the mean Dice metrics of 0.95 and 0.93 obtained by RsALUNet were greater than those of the three compared models by approximately 1.1% and 2.2%, respectively. The mean HDs (pixels) of 5.92 pixels and 5.16 pixels were also smaller than the corresponding parts by approximately 8.2% and 14.1%, respectively. The TC-segmentation improvement of our model was more effective, with an 8.5% increase in mean Dice and a 15.8% reduction in mean HD (pixels).

Moreover, by observing the box plots of dataset 3 in Figs. 8C and 9C, it can be found that the concentration levels of the distributions belonging to the published frameworks are inferior to that of our framework, which again indicates the superiority of RsALUNet. Further, as shown in Fig. 7C,

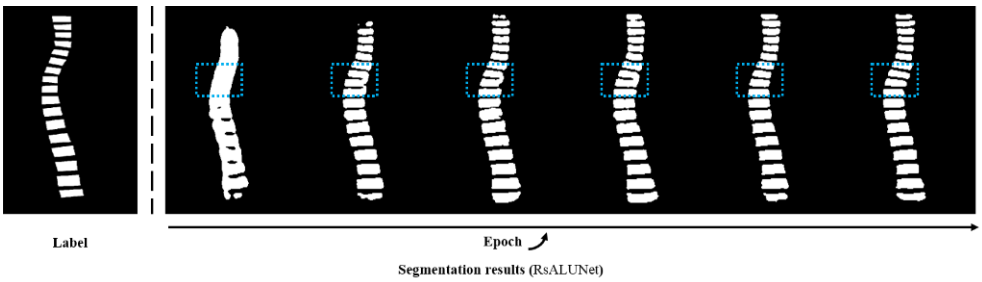

Fig. 13. Iterative segmentation results of RsALUNet for vertebrae. Blue dashed lines highlight the gradual separation of mixed vertebrae.

the delineation details provided by RsALUNet, especially for the TC region, are closer to the ground truth.

5.2 Ablation experiments

The evaluation metrics of the ablation experiments are summarized in Table 4, and their Dice and HD (pixels) box plots and prediction examples are displayed in Figs. 10, 11, and 12, respectively.

Effect of DCC ($UNet_{backbone}$ + DCC): Table 4 shows that the mean Dice of ($UNet_{backbone}$ + DCC) is greater than that of $UNet_{backbone}$ for all three datasets, and the increase was between 1.1% and 11.1%. Further, its mean HD (pixels) also had a decrease between 2.3% and 24.5%, compared with $UNet_{backbone}$. This suggests that a diverse and enlarged receptive field is beneficial, especially for brain-glioma segmentation with diverse shapes. In addition, as shown by the box plots in Figs. 11 and 12, the distribution of ($UNet_{backbone}$ + DCC) is moreconcentrated, which visually highlights the contribution of DCC. The examples in Fig. 10 also indicate that the border details were optimized to some extent.

Effect of fusion block ($UNet_{backbone}$ + FB): Combined with the fusion block, greater mean Dice / HD (pixels) metrics were obtained by ($UNet_{backbone}$ + FB), as listed in Table 4. The box plots in Figs. 11 and 12 also demonstrate that the performance of ($UNet_{backbone}$ + FB) is better than that of $UNet_{backbone}$. The prediction examples in Fig. 10 more effectively highlight the role of the fusion block; namely, integrating the feature maps of the outer ROI to the inner ROI. Consequently, the feasible inner-ROI segmentation region is further bounded. As can be seen in Fig. 10A, the vertebrae segmentation result acquired by ($UNet_{backbone}$ + FB) is better than that of $UNet_{backbone}$, and the significant improvement is that the vertebrae segmentation result does not obviously exceed the border of the spine; however, the prediction obtained by $UNet_{backbone}$ did exceed the border.

Effect of location encoder ($UNet_{backbone}$ + LE): Location information is valuable for a segmentation task; therefore, we also employed a location encoder. Observing the quantitative evaluation metrics given in Table 4, it can be observed that according to their averages, the improvement brought by the location encoder is limited, and larger standard deviations are achieved as well. This phenomenon was more obvious for dataset 2. This can be attributed to location encoding being directly conducted for the entire image. In this case, some interference artifacts, which have a similar intensity distribution and shape as the ROIs, were incorrectly highlighted by the location encoder. The bifurcations or side vessels in IVUS images are a typical example; their intensity distributions and shapes are very close to the LI region. Consequently, it can be concluded that, although the location encoder is powerful, it requires help from other functional modules, such as the fusion block.

Effect of RsAL ($UNet_{backbone}$ + RsAL): RsAL plays a critical role in promoting multi-ROI segmentation in different medical images. With the help of RsAL, $diff_{Segs}$ becomes more similar to $diff_{Labels}$. Equally, the prediction of each member in the multi-ROI becomes closer to the ground truth. The effectiveness of RsAL is reflected by the metrics listed in Table 4, and it can be seen that the improvement based on RsAL is superior to most of the other functional modules. The improvement in vertebrae and TC segmentation are particularly noticeable; approximately 3.8% / 18.1% increases and 15.8% / 7.3% decreases were achieved for the mean Dice and HD (pixels), respectively. The box plots in Figs. 11 and 12 show a similar conclusion. Furthermore, as can be seen in Fig. 10A, the details of the vertebrae borders provided by ($UNet_{backbone}$ + RsAL) are more precise, and the problem of mixing adjacent vertebrae is effectively mitigated by our RsAL strategy.

Effect of discriminators ($UNet_{backbone}$ + discriminators): Reinforcement supervision fulfilled via ($UNet_{backbone}$ + discriminators) was separately performed for each member in a multi-ROI, rather than integrated by $diff$, as in our RsAL strategy. As summarized in Table 4, for spine and vertebrae segmentation, the Dice is optimized by 2.1% and 5.0%, respectively, compared with $UNet_{backbone}$. The delineation of vertebral borders shown in Fig. 10A is better than that of $UNet_{backbone}$, but inferior to that of ($UNet_{backbone}$ + RsAL). Furthermore, the performance of ($UNet_{backbone}$ + discriminators) was unsatisfactory, particularly for TC with diverse shapes. These are caused by separately strengthening the supervision, which not only incapacitates the correlation between multi-ROIs, but also requires more discriminators.

6 Discussion

SegAN is a network employing adversarial learning, but it only mines the correlation between prediction results and corresponding image patches. However, SegAN underperforms on multi-ROI segmentation tasks, since the correlation between different targets is hard to be utilized for the improvement of multi-ROI segmentation. MedT extracts abundant global and local information via a gated axial-axis attention mechanism, this makes it more susceptible to interferences around targets. For those cases with ambiguous borders due to interferences, such as side vessels and bifurcation in IVUS images, it is unable to provide effective segmentations. TransUNet explores long-range dependencies and the global context via combining Transformer layers and convolutional ones, but it requires a

large number of data to train. For small datasets which only contain hundreds of cases, like the dataset 1 and 2, its performance is unsatisfactory. The performance of nnUNet is relatively balanced on three datasets, but there is still room for improvement for targets with irregular shapes or ambiguous borders.

Existing strategies for strengthening supervision are generally unsuitable for certain medical images, such as those based on prior knowledge. Facing this challenge, we selected a phenomenon existing in many medical images, that is, one ROI is embedded into or surrounded by another one, and proposed a new and flexible reinforcement-supervision strategy based on $diff$. An extensible framework (RsALUNet) via adversarial learning was designed to execute our strategy. Thus, we improved the multi-ROI segmentation in spinal X-ray, IVUS and brain MR images. If $diff_{Segs}$ becomes more similar to its intermediate labels, the segmentation of multi-ROI will be more precise. This is the core of our consideration. Another benefit brought by the diff is that it embeds an implicit anatomy prior knowledge into the segmentation network. This can mitigate the negative effect for the segmentation due to image interferences, and promote the multi-ROI segmentation tend to be consistent with the prior knowledge, like mixed vertebrae in spinal x-ray images, exceedingly ambiguous borders of LI caused by side vessels and bifurcation in IVUS images.

As demonstrated by ablation experiments with dataset 1, the spine is an unbroken ROI; however, the vertebrae are separated. Therefore, it can be imagined that $diff$ should also be separated. This characteristic is key to separating the mixed vertebrae highlighted by the cyan boxes in Fig. 13. Coincident with the training progress, the generator, namely, the decoders, becomes more experienced in deceiving the discriminator. In other words, $diff_{Segs}$ becomes more real. Thus, as shown in Fig. 13, the segmentation of the vertebrae gradually became unmixed. Furthermore, from an anatomical perspective, $diff$ maps to cartilage and fibrous connective tissues existing in the vertebral gaps, and these tissues closely adhere to the vertebrae. Thus, $diff$ can reasonably serve as an effective reference. Similarly, in IVUS, the $diff$ between MA and LI should be similar to a ring. In this situation, decoders attempt to generate real outputs that conform to this, with the help of adversarial learning. Consequently, the segmentation of MA and LI becomes more precise. Moreover, the ring is reasonable because it follows the anatomical properties of the vessel. The accumulation of plaque and lipids causes the LI to become locally narrowed to the inner side of the cavity; however, the MA usually maintains a shape similar to an ellipse, which leads to the formation of a ring. In clinical practice, the difference in area between the MA and LI regions, namely, the area change of the ring, is used as a metric to measure the degree of coronary stenosis.

Additionally, for brain MR, our RsAL played a beneficial role in optimizing the segmentation of brain gliomas. More importantly, the comparison with ($UNet_{backbone}$ + RsAL) and ($UNet_{backbone}$ + discriminators) indicates that our reinforcement supervision strategy is very competitive, because the number of discriminators in ($UNet_{backbone}$ + discriminators) must increase as the number of members in the multi-ROI increases; however, our RsAL can effectively avoid this by sharing discriminators via $diff$.

In addition to RsAL, other functional blocks also contribute to the feasibility of RsALUNet for spinal X-ray, IVUS and brain MR images. The DCC block provides a diverse and large receptive field for spinal X-ray, IVUS and brain MR images. For example, it can be found in Table 4 that the brain-glioma segmentation was greatly improved. This is because a convolution kernel with a diverse and large receptive field is more likely to cover a more complete lesion that may appear anywhere in the image and have a different size and shape.

The fusion block is also necessary to improve the segmentation of the inside ROI, as it can bound its feasible region by integrating feature maps belonging to the outside ROI. For instance, the spine is outside and the vertebrae are inside, the MA is outside and the LI is inside, and the WT is outside and the ET and TC are inside. In these cases, segmenting the inside ROI will be more effective, if its achievement is bounded into a smaller feasible region by the border of the outside ROI. Hence, we introduced a fusion block.

Although our proposed framework presents promising multi-ROI segmentation results in X-ray, IVUS and brain MR images, some limitation still exists. When targets have exceedingly irregular shape and ambiguous borders, an accurate segmentation is hard to be obtained. Moreover, a combination between our model and weakly-supervision strategies is required, since our model is fully-supervised but a large amount of high-quality manual annotations is hard to be collected in clinical practices. Additionally, to deal with medical video data, our model should become more light-weight.

7 Conclusion

The segmentation of multi-ROI in medical images is beneficial and helpful for medical-image analysis. In this study, we presented a concise reinforcement-supervision strategy, based on the differences among multi-ROI, and designed a new framework called RsALUNet to implement it. The introduction of $diff$ among multi-ROI effectively reduces the required number of additional discriminators, and our framework also become more concise than existing ones. Meanwhile, implicit anatomy prior knowledge is embedded in the $diff$, so it could promote the model predictions become more consistent with the anatomy. Meanwhile, three functional modules were introduced to further improve the model performance via diverse receptive fields and high-efficient attention mechanisms. Extensive experiments were performed on three different medical image datasets, including X-ray, IVUS and brain MR imaging modalities. The mean Dice obtained by RsALUNet were all above 0.85 for the multi-ROI segmentation of the three datasets. Compared with recent other frameworks, the mean Dice also increased from 1.1% to 8.5% on the three datasets, which further revealed the robustness and competitiveness of RsALUNet.

In future work, we will consider a combination between our framework and weakly-supervision strategies, to improve the model applicability in clinical practices. Further, the existence of noisy label is inevitable in medical image datasets. In this case, an effective approach is also required to mitigate the interference caused by them and achieve a robust multi-ROI segmentation.

Acknowledgment

This work is partially supported by National Natural Science Foundation of China (91959127), Hong Kong Research Grant Council Impact Research Fund (RIF) (R5017-18F).

References

[1] H.B. Sezer, A. Sezer, Automatic segmentation and classification of neonatal hips according to Graf's sonographic method: a computer-aided diagnosis system. Appl Soft Comput. 82 (2019) 105516.

[2] Y. Gu, J. Chi, J. Liu, L. Yang, B. Zhang, D. Yu, Y. Zhao, X. Liu. A survey of computer-aided diagnosis of lung nodules from CT scans using deep learning. Comput. Biol. Med. 137 (2021) 104806.

[3] Y. Xu, L.F.F. Souza, I.C.L. Silva, A.G. Marques, F.H.S. Silva, V.X. Nunes, T. Han, C. Jia, V.H.C. Albuquerque, P.P.R. Filho. A soft computing automatic based in deep learning with use of fine-tuning for pulmonary segmentation in computed tomography images. Appl Soft Comput. 112 (2021) 107810.

[4] K.L. Nisha, G. Sreelekha, P.S. Sathidevi, P. Mohanachandran, A. Vinekar. A computer-aided diagnosis system for plus disease in retinopathy of prematurity with structure adaptive segmentation and vessel based features. Comput. Med. Imag. Graph. 74 (2019) 72-94.

[5] J. Cobb. Outline for the study of scoliosis. Instr Course Lect. 5 (1948) 261-275.

[6] L. Wang, Q. Xu, S. Leung, J. Chung, B. Chen, S. Li. Accurate automated Cobb angles estimation using multi-view extrapolation net. Med Image Anal. 58 (2019) 101542.

[7] K. Zhang, N. Xu, C. Guo, J. Wu. MPF-net: an effective network for automated Cobb angle estimation. Med Image Anal. 75 (2022) 102277.

[8] D. Fardon. Nomenclature and classification of lumbar disc pathology. Spine. 26(5) (2001) 461-462.

[9] A.L. Williams, F. Murtagh, S. Rothman, G. Sze. Lumbar disc nomenclature: version 2.0. AM J Neuroradiol. 35(11) (2014) 2029.

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

[11] D. Zhang, B. Chen, S. Li. Sequential conditional reinforcement learning for simultaneous vertebral body detection and segmentation with modeling the spine anatomy. Med Image Anal. 67 (2021) 101861.

[12] G.S. Mintz, S.E. Nissen, W.D. Anderson, S.R. Bailey, R. Erbel, P.J. Fitzgerald, F.J. Pinto, K. Rosenfield, R.J. Siegel, E.M. Tuzcu. American college of cardiology clinical expert consensus document on standards for acquisition, measurement and reporting of intravascular ultrasound studies (IVUS): a report of the American college of cardiology task force on clinical expert consensus documents developed in collaboration with the European society of cardiology endorsed by the society of cardiac angiography and interventions, J Am Coll Cardiol. 37 (2001) 1478-1492.

[13] M. McDaniel, P. Eshtehardi, F. Sawaya, J. Douglas, H. Samady, Contemporary clinical applications of coronary intravascular ultrasound. JACC Cardiovasc inte. 4(11) (2011) 1155-1167.

[14] X. Zhang, C. Mckay, M. Sonka, Tissue characterization in intravascular ultrasound images. IEEE Trans Med Iamg. 17(6) (1998) 889-899.

[15] J. Klingensmith, R. Shekhar, D.G. Vince, Evaluation of three-dimensional segmentation algorithms for the identification of luminal and medial-adventitial borders in intravascular ultrasound images. IEEE Trans Med Iamg. 19(10) (2000) 996-1011.

[16] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J.S. Kirby, J.B. Freymann, K. Farahani, C. Davatzikos, Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci Data. 4 (2017) 170117.

[17] O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention. (2015) 234-241.

[18] Z. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, J. Liang, Unet++: redesigning skip connections to exploit multiscale features in image segmentation. IEEE Trans Med Iamg. 39(6) (2020) 1856-1867.

[19] F. Isensee, P.F. Jager, S.A.A. Kohl, J. Petersen, K.H. Maier-Hein, Automated design of deep learning methods for biomedical image segmentation. (2019) arXiv:1904.08128.

[20] S.E. Petersen, N. Aung, M.M. Sanghvi, F. Zemark, K. Fung, J.M. Paiva, J.M. Francis, M.Y. Khanji, E. Lukaschuk, A.M. Lee, Reference ranges for cardiac structure and function using cardiovascular magnetic resonance (CMR) in caucasians from the UK biobank population cohort. J Cardiovascular Magn Reson. 18 (2017) 19.

[21] B.W. Stewart, C.P. Wild, World cancer report 2014, International Agency for Research on Cancer. (2014).

[22] T. Heimann, B.V. Ginneken, M. Styner, Y. Arzhaeva, V. Aurich, C. Bauer, A. Beck, C. Becker, R. Beichel, G. Bekes, F. Bello, G. Binnig, H. Bischof, A. Bornik, P. Cashman, Y. Chi, A. Crdova, B. Dawant, M. Fidrich, J. Furst, D. Furukawa, L. Grenacher, J. Hornegger, D. Kainmller, R. Kitney, H. Kobatake, H. Lamecker, T. Lange, J. Lee, B. Lennon, R. Li, S. Li, H. Meinzer, G. Nmeth, D. Raicu, A. Rau, E.V. Rikxoort, M. Rousson, L. Rusk, K. Saddi, G. Schmidt, D. Seghers, A. Shimizu, P. Slagmolen, E. Sorantin, G. Soza, R. Susomboon, J. Waite, A. Wimmer, I. Wolf, Comparison and evaluation of methods for liver segmentation from CT datasets. IEEE Trans Med Iamg. 28 (8) (2009) 1251–1265.

[23] T.Z. Naqvi, M. Lee, Intima–media thickness: A tool for atherosclerosis imaging and event prediction. JACC Cardiovasc Imag. 7 (10) (2014) 1025-1038.

[24] I.J. Goodfellow, J.P. Abadie, M. Mirze, B. Xu, D.W. Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets. International Conference on Neural Information Processing Systems. (2014) 2672-2680

[25] D. Dai, C. Dong, S. Xu, Q. Yan, Z. Li, C. Zhang, N. Luo, Ms RED: a novel multi-scale residual encoding and decoding network for skin lesion segmentation. Med Image Anal. 75 (2022) 102293.

[26] J. Pi, Y. Qi, M. Lou, X. Li, Y. Wang, C. Xu, Y. Ma, FS-UNet: mass segmentation in mammograms using an encoder-decoder architecture with feature strengthening. Med Image Anal. 137 (2021) 104800.

[27] T. Mahmud, B. Paul, S.A. Fattch, PolypSegNet: a modified encoder-decoder architecture for automated polyp segmentation from colonoscopy images. Comput. Biol. Med. 128 (2021) 104119.

[28] R. Zhao, Q. Li, J. Wu, J. You, A nested U-shape network with multi-scale upsample attention for robust retinal vascular segmentation. Pattern Recognit. 120 (2021) 107998.

[29] K. He, C. Lian, E. Adeli, J. Huo, Y. Gao, B. Zhang, J. Zhang, D. Shen, MetricUNet: synergistic image- and voxel-level learning for precise prostate segmentation via online sampling. Med Image Anal. 71 (2021) 102039.

[30] E.A. Rashed, J Gomez-Tames, A Hirata. Development of accurate human head models for personalized electromagnetic dosimetry using deep learning. Neuroimage. 202 (2019) 116132.

[31] J. Qiu, L. Li, S. Wang, K. Zhang, Y. Chen, S. Yang, X. Zhuang. MyoPS-Net: Myocardial pathology segmentation with flexible combination of multi-sequence CMR images. Med Image Anal. 84 (2023) 102694.

[32] J. Huo, J. Wu, J. Cao, G. Wang, Supervoxel based method for multi-atlas segmentation of brain MR images. NeuroImage. 175 (15) (2018) 201-214.

[33] Y. Zhang, J. Wu, Y. Liu, Y. Chen, W. Chen, Ed.X. Wu, C. Li, X. Tang, A deep learning framework for pancreas segmentation with multi-atlas registration and 3D level-set. Med Image Anal. 68 (2021) 101884.

[34] J.O.B. Diniz, J.L. Ferreira, P.H.B. Diniz, A.C. Silva, A.C. Paiva, Esophagus segmentation from planning CT images using an atlas-based deep learning approach. Comput Methods Programs Biomed. 197 (2020) 105685.

[35] S. Dong, G. Luo, C. Tam, W. Wang, K. Wang, S. Cao, B. Chen, H. Zhang, S. Li, Deep atlas network for efficient 3D left ventricle segmentation on echocardiography. Med Image Anal. 61 (2020) 101638.

[36] S. Pachade, P. Porwal, M. Kokare, L. Giancardo, F. Meriaudeau, NENet: nested efficientNet and adversarial learning for joint optic disc and cup segmentation. Med Image Anal. 74 (2021) 102253.

[37] B. Lei, Z. Xia, F. Jiang, X. Jiang, Z. Ge, Y. Xu, J. Qin, S. Chen, T. Wang, S. Wang, Skin lesion segmentation via generative adversarial networks with dual discriminators. Med Image Anal. 64 (2020) 101716.

[38] H. Wu, X. Lu, B. Lei, Z. Wen, Automated left ventricular segmentation from cardiac magnetic resonance images via adversarial learning with multi-stage pose estimation network and co-discriminator. Med Image Anal. 68 (2021) 101891.

[39] Y. Liu, X. Yuan, X. Jiang, P. Wang, J. Kou, H. Wang, M. Liu, Dilated adversarial U-Net framework for automatic gross tumor volume segmentation of nasopharyngeal carcinoma. Appl Soft Comput. 111 (2021) 107722.

[40] G. Huang, Z. Liu, L. Maaten, K.Q. Weinberger, Densely connected convolutional networks. (2016) arXiv:1608.06993.

[41] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, G. Cottrell,

Understanding convolution for semantic segmentation. (2017) arXiv:1702.08502.
[42] M. Anthimopoulos, S. Christodoulidis, L. Ebner, T. Geiser, A. Christe, S. Mougiakakou, Semantic segmentation of pathological lung tissue with dilated fully convolutional networks. IEEE J Biomed Health Inform. 23(2) (2018) 714-722.
[43] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, ECA-Net: efficient channel attention for deep convolutional neural networks. (2019) arXiv:1910.03151.
[44] Z. Huang, X. Wang, Y. Wei, L. Huang, H. Shi, W. Liu, T.S. Huang, CCNet: criss-cross attention for semantic segmentation. (2018) arXiv:1811.11721.
[45] J.M.J. Valanarasu, P. Oza, I. Hacihaliloglu, V.M. Patel, Medical transformer: gated axial-attention for medical image segmentation. (2021) arXiv:2102.10662.
[46] L. Wang, C. Xie, Y. Lin, H. Zhou, K. Chen, D. Cheng, F. Dubost, B. Collery, B. Khanal, B. Khanal, R. Tao, S. Xu, U.U. Bharadwaj, Z. Zhong, J. Li, S. Wang, S. Li, Evaluation and comparison of accurate automated spinal curvature estimation algorithms with spinal anterior-posterior x-ray images: the AASCE2019 challenge. Med Image Anal. 72 (2021) 102115.
[47] S. Balocco, C. Gatta, F. Ciompi, A. Wahle, P. Radeva, S. Carlier, G. Unal, E. Sanidas, J. Mauri, X. Carillo, T. Kovarnik, C.W. Wang, H.C. Chen, T.P. Exarchos, D.I. Fotiadis, F. Destrempes, G. Cloutier, O. Pujol, M. Alberti, E.G.M. Ruiz, M. Rivera, T. Aksoy, R.W. Downe, I.A. Kakadiaris, Standardized evaluation methodology and reference database for evaluating IVUS image segmentation, Comput. Med. Imag. Graph. 38 (2014) 70-90.
[48] A.L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. Ginneken, A.K. Schneider, B.A. Landman, G. Litjens, B. Menze, O. Ronneberger, R.M. Summers, P. Bilic, P.F. Christ, R.K.G. Do, M. Gollub, J.G. Pernicka, S.H. Heckers, W.R. Jarnagin, M.K. McHugo, S. Napel, E. Vorontsov, L.M. Hein, M.J. Cardoso, A large annotated medical image dataset for the development and evaluation of segmentation algorithms. (2019) arXiv: 1902.09063.
[49] Y. Xue, T. Xu, X. Huang, Adversarial learning with multi-scale loss for skin lesion segmentation. International Symposium on Biomedical Imaging. (2018) 859-863.
[50] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A.L. Yuille, Y. Zhou, TransUNet: transformer make strong encoders for medical images segmentation. (2021) arXiv:2102.04306.