Automatic detection and segmentation of colorectal tumor based on multimodality magnetic resonance imaging fusion and co-predictive learning

Yunhao Ge [a] , Weixin Yan [a], Bin Li [a]

[a] Robotics Institute of Shanghai Jiao Tong University, Shanghai, China

ABSTRACT

Colorectal tumor detection and segmentation by a deep learning algorithm in magnetic resonance imaging (MRI) is a difficult task due to many mutually-affected challenges, including the complicated anatomical environments, limited information extracted from radiology images and inadequate training samples. Regarding the diagnosis algorithm, some end-to-end deep convolutional neural networks (CNNs) have achieved remarkable success in some medical image detection and segmentation tasks. However, due to the limitation of the features extracted from single modality MRI with a single algorithm, which are the basics of diagnosis, the precision of detection and segmentation of colorectal tumor diagnosis are hard to improve. Instead, experienced doctors make their final decision by combining radiological data, pathology data, morphological data, and clinical information. In this paper, a co-predictive learning method, which consists of two parallel prediction algorithms, is proposed to imitate the diagnostic process of doctors by fusing multimodality MRI ((T1 weighted (T1-w), T2 weighted (T2-w) and diffusion-weighted imaging (DWI)) information and combining different detection algorithms (CNNs and machine learning methods). The T2-w MRI, which performs best in single modality detection after the experiment, is used to perform main-stream detection and segmentation. While the other MRI modalities (T1w, DCE), which have different complementary information on the classification task, can extract both high level features and artificial features with different algorithms and predict the possibility whether the target image contains tumors by the eXtreme Gradient Boosting (XGBoost) algorithm. The image-level predictions are used to make main-stream co-prediction to improve the detection precision. Meanwhile, with the help of specific diagnostic characteristics of colorectal tumor and specific skills for the analysis of the medical image, the hardness of segmentation can be reduced to improve the quality of segmentation. Experimental results demonstrate that our co-predictive method can perform better both qualitatively and quantitatively in colorectal tumor automatic detection and segmentation, compared with the state-of-the-art methods.

Keywords: Colorectal tumor segmentation, Automatic detection, Multimodality magnetic resonance imaging fusion, Deep convolutional neural networks, co-predictive learning.

1. Introduction

Colorectal cancer is one of the most common malignant tumors of the digestive system. It is ranked as the third most common cancer worldwide making up about 10% of all cases. Due to the inapparent symptoms, most patients are at an advanced local stage at the time of diagnosis, thus missing the optimal time for treatment (Lambregts, D. M. J. et al., 2015). Therefore, it is critical to diagnose and assess colorectal cancer timely. Magnetic Resonance Imaging (MRI) provides a noninvasive and accurate method for colorectal cancer detection by generating detailed location and geometry information of the organs and tissues (Martens, M. H. et al., 2015; Seierstad et al., 2015). MRI images can also help to predict clinically relevant endpoints, by predicting whether the patients who show a positive response to treatment can achieve significant clinical meaning, as organ-preserving treatment strategies could be considered for them as an alternative to standard surgical resection (Lambregts, D. M. J. et al., 2015). Encouraging results have been achieved for volumetric measurements by multimodality MRI. However, most of results are analyzed based on regions of interest (ROI) of the tumors that are typically obtained after manual segmentation by experienced clinicians. The tumor segmentation procedure is very time-consuming and thus greatly limits their use in clinical practice (Van Heeswijk et al., 2016). In the field of biomedical imaging, deep learning has been largely utilized for automatic detection and segmentation purposes (LeCun, Y et al., 2015; Greenspan, H et al., 2016). To the best of our knowledge, few studies have been conducted on the automatic localization and segmentation of colorectal cancer. Colorectal cancer detection is a difficult task due to many mutually-affected challenges, including the complicated anatomical environments, limited information extracted from radiological images and inadequate training samples. Atlas-based colorectal tumors auto-segmentation could

improve contouring efficiency in the clinical practice setting (Gambacorta et.al., 2013), while the method (shortage) without an automatic segmentation procedure uses paired dynamic contrast-enhanced MRI (DCE-MRI) and achieves a successful result. However, the requirement of image-pair large-scale data analysis greatly limits its usage because in some situations the DCE-MRI pair-data is infeasible to obtain (Trebeschi et al., 2017). A deep learning method has been proposed to automatically segment the colorectal tumor from actual patients (Trebeschi et al., 2017). However, in the actual application for medical image diagnosis, judging if the target patient contains a colorectal tumor or not is significantly more important and difficult than segmenting the tumor in positive patients. Moreover, most deep learning methods in medical imaging diagnosis are end-to-end network, which complete the detection or segmentation using a single algorithm depending only on the information from a single modality image. Compared with the experienced radiologist, who prefers making diagnosis after combining the information from multimodality MRI and other pathological or dynamic information to improve the precision of detection and segmentation. To solve the above problems, in this study, we aimed to develop a co-predictive neural network to automatically localize and segment colorectal rectum tumors.

There were nearly 30 slides MRI images for every patient in each modality. For the patients with colorectal cancer, not all the 30 slides contained the tumors, which increased the difficulty of detection and segmentation. Since T1 weighted (T1-w) MRI images and T2 weighted (T2-w) MRI images could extract more features by deep convolutional neural networks (CNNs) (LeCun, Y et al., 2015) about the shape and texture, which could be useful for the tumor location and segmentation. DCE images provide a much clearer and less noisy signal to make image-level prediction about whether the target image contains a tumor. Different modalities of MRI might be more sensible for machine learning algorithm, especially in limited dataset circumstance. Besides, doctors can also make segmentation by specific characteristics of colorectal cancer, such as by limiting the output regain (only around rectal tissues) and limiting the output of the location, which performs better than a non-maximum suppression algorithm. These are much more complex for end-to-end algorithm.

To combine the advantages of different features extracted by multimodality MRI, the proposed co-predictive neural network includes two parallel prediction algorithms, namely the main-stream and the side-stream algorithms. The main-stream algorithm achieves detection and segmentation on the T2-w MRI images, because T2 performs best among different single modality MRI detection in our CNNs experiments. The other MRI modalities (T1-w, DCE), provide different complementary information on classification tasks that could be used to extract high level features and perform other statistical, texture or gray features analysis by different methods, which can be used to predict the possibility of whether the image contains tumors by the XGBoost (eXtreme Gradient Boosting) algorithm (Chen, Tianqi et al., 2016). The image-level predictions are used to make co-prediction in main-stream CNNs to improve the detection precision. This adjustment is useful to improve the precision of image-level detection. To take advantage of the specific characteristics of colorectal tumor in diagnosis and improve the accuracy of detection and segmentation, crop methods are applied before the detection and segmentation using a simple CNN. In addition, to reduce the computational cost, in the main-stream neural network, the detection and segmentation task share the weight in the CNN part, which is used to extract the crucial high-level features.

Our contributions in this work are three fold:

1. Proposed a Co-prediction learning method which combining multimodality Magnetic Resonance (MR) images and using complementary decision algorithms to imitate the diagnostic process of doctors to improve image-level detection accuracy
2. Embedded two parallel prediction algorithms in the Co-prediction method that combined the advantage of the neural network's high-level feature learning ability and the statistical and handcraft feature selection ability of XGBoost
3. Fused multimodality MR images (T1 weighted, T2 weighted and Diffusion Weighted Imaging) by models in a comprehensive view and achieved the colorectal tumor detection accuracy improved to 92% as well as the segmentation AP to 0.7

2. Materials and Methods

2.1 Dataset

This retrospective study was approved by the Institutional Review Board (IRB) of the Sir Run Run Shaw Hospital, College of Medicine, Zhejiang University (Zhejiang, China); and the informed consent requirement was waived from the IRB. The colorectal tumor database (CTD) comprised 18,720 images from 208 patients, who were diagnosed between June 2009 and November 2014, in accordance with the inclusion and exclusion criteria. The inclusion criteria for patients in this study were: (i): rectal cancer confirmed by biopsy; (ii) locally advanced disease determined by pre-treatment MRI images (stage T3 or T4); (iii) pre-treatment MRI scan performed within 4 weeks of treatment. The exclusion criteria included: (i): received complete neoadjuvant chemoradiotherapy before MRI scans; (ii) suffered from synchronous distant metastasis; (iii) poor MRI image

quality due to motion artifacts.

All patients underwent MRI scans including T1-weighted, T2-weighted and dynamic contrast-enhanced (DCE). All the MRI scans were performed at our institution using a 3.0 Tesla MRI scanner (Signa HDxt, GE Medical Systems, Milwaukee, WI, USA) with a phased-array body coil. No special bowel preparation was performed during the scan. A quality assurance check was performed on the MRI machine to ensure the consistency of the image quality. The T2-weighted images were acquired using T2-weighted fast spin echo sequence. The acquisition parameters were as follows: repetition time (TR): 2,840ms; echo time(TE): 131ms; image resolution 0.49×0.49×4 mm; matrix: 512×512. The array spatial sensitivity encoding technique (ASSET) was used with an acceleration factor of 2. The T1-weighted sequences were obtained using a spoiled gradient echo sequence. The T1-weighted sequences parameters were: TR: 4.4 ms; TE: 1.9 ms; flip angle: 12°; bandwidth: 325.5 kHz; image resolution: 0.7×0.7×2 mm; matrix: 512×512. For the DCE scan, contrast injection and data acquisition were performed simultaneously. All patients were injected with 0.1 mmol/kg body-weight gadolinium-diethylenetriamine penta-acetic acid (Gd-DTPA) at a speed of 2.5 mL/s. The data from four phases were acquired, before the injection of the contrast agent (Time-1), at 15 seconds after the injection (Time-2), at 60 seconds after the injection (Time-3), and at 120 seconds after the injection (Time-4). All images were reviewed with the MIM® software (MIM software Inc., Cleveland, OH, USA) by an experienced colorectal MRI radiologist. Then the three-dimensional (3D) tumor volumes, excluding the intestinal lumen, were manually segmented as ground truth. The volumes in the T1-w and T2-w images were segmented separately. The segmentation results were also reviewed by another colorectal clinician.

The CTD was divided in two independent datasets, called the training-validation dataset and test dataset. The training-validation dataset including 188 patients was used to train the co-predictive model. The test dataset comprising 20 patients was used to test the performance of the model developed. The detailed information of the CTD is summarized in Table 1.

Table1. Information of the colorectal tumor database (CTD)

Dataset	Amount	T1-w images	T2-w images	DCE	Ground Truth	Total(images)
Train-validation	188 patients	5640	5640	5640	5640	16920
test	20 patients	600	600	600	600	1800

2.2 Proposed Method

Considering the multiparametric MRI (mp-MRI) data of a patient, our goal is to (1) automatically classify whether each slice of the mp-MRI contains tumor or not, and (2) for slices classified as positive for colorectal tumor, localize the position of the cancerous tissues and segment the area to obtain the pixel level classification information. The framework of our automated co-predictive colorectal tumor detection and segmentation system is depicted in Fig. 1. The framework consists of 2 main steps. First, after recording of the multimodality MRI images, an automated method is applied to detect the rectal region using a squared bounding box (bbox) on every T2-w slice. Then, the bbox is used to align all T1-w and DWI slices and the content within each bbox is cropped and the intensities are normalized. The ground truth in training is extracted from the manually constructed CTD. Second, the cropped T2-w image is input into the main-stream co-predictive algorithm and 2 items are output: (1) the bbox regression as detection output, which contains the rectal tumors (2) the output of the binary mask of the ROI, where each pixel is predicted as a colorectal tumor region. Meanwhile, the T1w and DWI modality images go through the shared structure CNNs in the main-stream algorithm to output the high-level features. Other artificial features are also obtained by the machine learning algorithm, which contain statistical features, gray-level co-occurrence matrices (GLCM) features (Hall-Beyer et al., 2000) and local features. Then, the extracted features are input into a XGBoost classifier and output the possibility of the image-level prediction on whether this cropped rectal square contains a tumor or not, the prediction is used to make the co-prediction in the classification part of the main-stream algorithm. Our co-predictive CNNs has two main advantages: (1) high accuracy of image-level detection with the novel fusion solution by different modalities of MRI. By using different algorithms to extract high-level and artificial features from the aligned T1w and DWI, the XGBoost prediction depending on these combined features can improve the detection accuracy on the image-level (most of slice in each patient may contain no tumor regain), as well as reduce the difficulty of the segmentation of the ROI. (2) The bbox output and binary mask depend on the same CNN structure and the same heatmap feature, which significantly reduce the computational cost in main-stream. Additionally, the main-stream and side-stream also have shared CNNs structures to avoid the repetitive computation.

The framework of a fully automatic co-predictive method for colorectal tumor detection and segmentation includes five key components, as shown

in Fig.1.

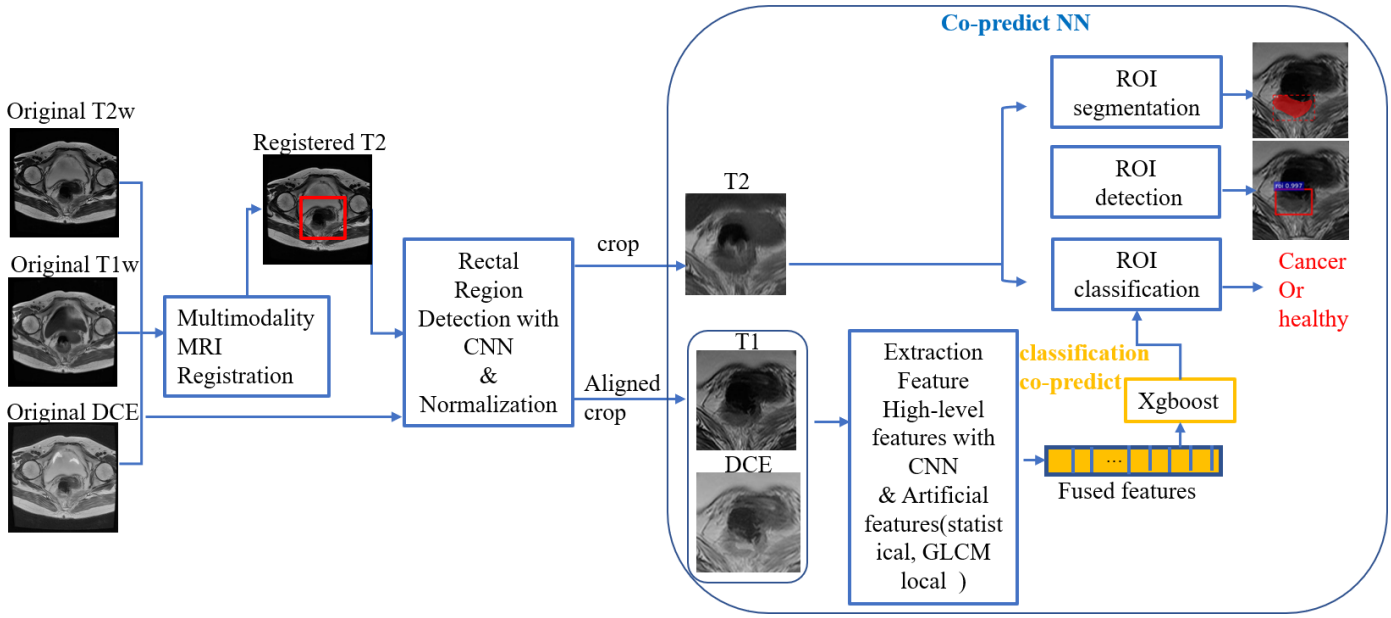


Fig. 1. The framework of the automated rectal localization system, which includes five key components: (1) multimodality (T1-w, T2-w and DWI) image registration; (2) rectal detection with CNN and normalization; (3) main-stream of co-predictive algorithm with ROI detection and segmentation; (4) tumor classification via XGBoost; (5) combination of the main-stream and side-stream to make the co-prediction.

2.2.1 Image preprocessing

The image preprocessing is performed by including the multi-modality image registration and intensity normalization. For the multi-modality MRI images, the DCE images were selected as the reference image, and the T1-weighted and T2-weighted images were deformed to match them using the MIM software (Klein, S et al., 2010). The images were normalized, and the formula is as follows:

$$G_{norm} = (G - G_{\min}) / (G_{\max} - G_{\min}) \quad (1)$$

where G is the grey intensity for each pixel in the image; G_{norm} is normalized grey value; G_{min} is the minimum grey value of the image and G_{max} is the maximum grey value of the image.

2.2.2 Automated rectal detection and normalization

Once the multi-modality image registration and intensity normalization were completed, we used a simple CNN regression to automatically crop a square region including the whole rectal region for every T2-w slice along the transverse direction. The architecture of our CNN model is depicted in Fig. 2. The CNN model is trained by a set of original T2-w slices using the manually labeled squared bbox to note the rectal regions. There are three output parameters from the CNN regression: the *(x, y)* coordinate position of the center of the square bbox, and the length *l* of the side of the square bbox. Note that these three parameters are normalized by the origin T2-w size, the range of the *x, y* coordinate position is [-1, 1] and the range of length *l* is [0,1]. We adopted the activation of the tanh function as the final output layer of the x, y position, and the sigmoid function as the final output layer of the length l. and the corresponding loss function is defined as:

$$loss = \frac{1}{3}(|\tanh(p_1 - x_t)| + |\tanh(p_2 - y_t)| + |sigmoid(p_3 - l_t)|) \quad (2)$$

where *p1*, *p2*, and *p3* are the three final outputs of the CNN as shown in Fig. 2, and (x_t , y_t , l_t)are the normalized center coordinates and the length of the rectal region.

$$x_t = \frac{2x - w}{w}$$
$$y_t = \frac{2y - w}{w} \quad (3)$$
$$l_t = \frac{l}{w}$$

where *(x, y, l)* indicates the non-normalized coordinates and length ranging from 0 to *w*, and *w* is the width and height of the input square image with the size of 256x256. The CNN parameters are updated for the rectal region detection by minimizing the loss function.

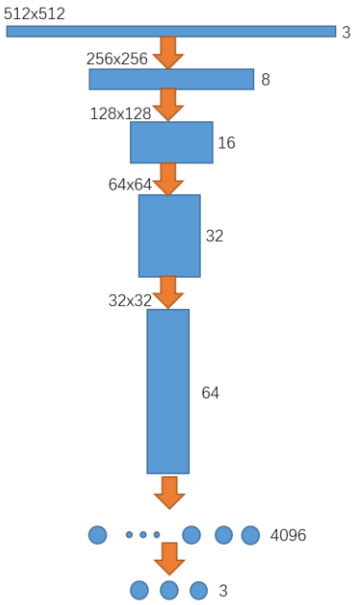


Figure 2. CNN for rectal region detection

2.2.3 Tumor detection and Segmentation with co-predictive CNNs

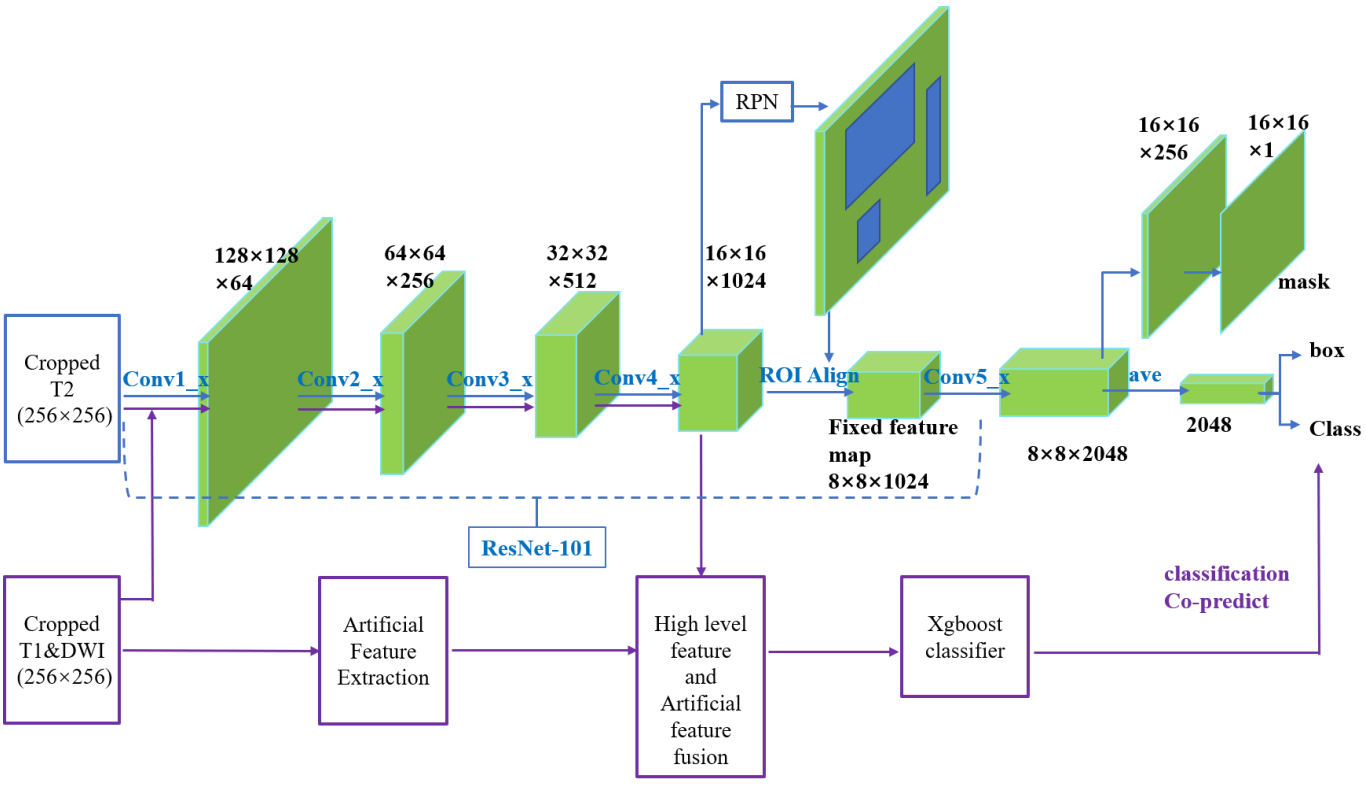


Fig. 3 Framework of the co-predictive algorithm for fusing multimodality (T2-w, T1-w and DWI) information: the blue structure is main-stream, while the purple part is side-stream, the Conv1_x to Conv4_x is the backbone architecture, which is used for feature extraction over a cropped input image. We extended the existing Faster Region-based CNN (Faster R-CNN) algorithm (Ren, S et al., 2015) by adding a mask branch. The

numbers denote spatial resolution and channels. The arrows denote either conv, de-conv, or fc layers, as can be inferred from the context (conv preserves spatial dimension, while de-conv increases it). All convs are 3×3, except the output conv which is 1×1, all de-convs are 2×2 with a stride of 2, and we use ReLU (Nair et al., 2010) in hidden layers.

After the cropping and normalization steps, we obtained a sequence of T2-w, T1-w and DCE groups, and each group is spatially aligned and the intensity is well normalized. The framework of our proposed co-predictive algorithm is depicted in Fig. 3. In the next subsection, we first describe the details of main-stream to detect the rectal tumor and segment the exact region using only the T2-w single-modality MRI as input. Then, the side-stream is described, explaining how to extract the artificial features using different extraction methods. Ultimately, the details about the combination of main-stream and side-stream to construct the co-predictive neural network is described.

2.6 Main-stream Network Architecture

The two main functions of the co-predictive neural network are implemented in main-stream, ROI detection and segmentation. Regarding the object detection task, the framework of the Faster R-CNN, which consists of two stages. The first stage, the Region Proposal Network (RPN) (Ren, S et al., 2015), proposes candidate object bounding boxes. The second stage, which is similar to the Fast R-CNN (Ren, S et al., 2015), extracts features using the ROI Pool operator from each candidate box and performs classification and bbox regression. This structure also shares the features used by two parts for faster inference. For adding the segmentation function as well, Mask R-CNN (He, K et al., 2017) has an excellent CNN framework and a mature pipeline for combining the object detection and segmentation tasks. Mask R-CNN improves the function of the Faster R-CNN, which has two outputs for each candidate object, a class label and a b-box offset, by adding a third branch that outputs the object mask. Thus, the Mask-R-CNN achieves a conceptually simple, flexible, and general framework for object instance segmentation. The design of the main-stream structure is similar to that of the Mask-R-CNN, which consists of three main parts. First, as shown in Table.2, we use ResNet101 (He, K et al., 2016) as the backbone architecture, which extracts the high-level features from the input image. The architecture of Resnet 101 is shown in Table 2. In the main-stream structure, the 91conv layers from conv1 to conv4-x are shared parts of the CNNs, the output of conv4_x is used as the input of the ROI align (He, K et al., 2017) layers. In addition, the output of conv4_x is also used as input of the second part——the RPN (Ren, S et al., 2015), which proposes the candidate object b-boxes that may contain the colorectal tumor. In other words, we share parts of the backbone architecture with the RPN. The proposed region is mapped into the feature map of the ROI align layers in main-stream, the features in the proposed region then go through the conv5_x to output the final high-level feature map of the input image. Ultimately, the third part——head Architectures is used to construct three parallel outputs: classification, detection and segmentation.

Table. 2 Architecture of the ResNet-101

Layer name	Output size	structure
Conv1	128×128	7×7, 64, stride=2
Conv2_x	64×64	3×3 max pool, stride=2
		$\begin{bmatrix} 1\times1, & 64 \\ 3\times3, & 64 \\ 1\times1, & 256 \end{bmatrix}\times3$
Conv3_x	32×32	$\begin{bmatrix} 1\times1, & 128 \\ 3\times3, & 128 \\ 1\times1, & 512 \end{bmatrix}\times4$
Conv4_x	16×16	$\begin{bmatrix} 1\times1, & 256 \\ 3\times3, & 256 \\ 1\times1, & 1024 \end{bmatrix}\times23$
Conv5_x	8×8	$\begin{bmatrix} 1\times1, & 512 \\ 3\times3, & 512 \\ 1\times1, & 2048 \end{bmatrix}\times3$

2.7 Side-stream Network Architecture with Multimodality MRI Fusion

The input of the side-stream includes two modalities of MRI, namely T1-w and DCE, these two modalities do not perform as well as the T2-w in the single modality ROI detection and Segmentation based on our previous experiment. Thus, we utilize the information in T1-w and DCE with the side-stream to improve the classification decision in the main-stream. Two methods are simultaneously used to extract the features in the ROI image. First, the shared part of the CNNs in main-stream (from conv1 to conv4-x in Table.2) was used to extract a set of complex high-level convolutional features. The global average pooling (GAP) operation generates the final output by averaging all scores within the feature maps and the dimension of the final output is 512-d. Second, the artificial features composed of GLCM features, location features, and artificial features are obtained. Before the extraction of artificial features, the ROI image needs to be converted to the gray size. The artificial features consist of the statistical parameters, GLCM features and location features. Below is a list of twelve features extracted from the mask area of the original image; they were selected to be combined with the output of the final convolutional layer to train and test our method. Statistical parameters: four statistical parameters of the ROI region were extracted: mean, variance, skewness and kurtosis. For each unique pixel value $f_k p_f(f_k)$ is the probability of this unique pixel value in the whole ROI. The four parameters are calculated as follows (Eq.4-7):

$$mean: \mu = \sum_{k=1}^{N} f_k p_f(f_k) \tag{4}$$

$$\text{var}iance: \sigma^2 = \sum_{k=1}^{N} (f_k - \mu)^2 p_f(f_k) \tag{5}$$

$$skewness: ske = \sum_{k=1}^{N} [(f_k - \mu)^3 p_f(f_k)] / \sigma^3 \tag{6}$$

$$kurtosis: kur = \sum_{k=1}^{N} [(f_k - \mu)^4 p_f(f_k)] / \sigma^4 \tag{7}$$

The GLCM features: the GLCM features consists of the sum entropy (SE), sum average (SA), difference variance (DV) and difference entropy (DE). The *SE* is a logarithmic function of ROI in consideration; *SA* is calculated from the ROI and the size of the gray scale; *DV* is a variance measure between the ROI intensities calculated as a function of the *SE* previously calculated; *DE* is an entropy measure which provides a measure of no uniformity, while taking into consideration a measure of the difference obtained from the original image. These four parameters are calculated as follows:

$$SE = -\sum_{i=2}^{2N_g} p_{x+y}(i) \log\{p_{x+y}(i)\} \tag{8}$$

$$SA = \sum_{i=2}^{2N_g} i p_{x+y}(i) \tag{9}$$

$$DV = \sum_{i=2}^{2N_g} (i - SE)^2 p_{x-y}(i) \tag{10}$$

$$DE = -\sum_{i=2}^{2N_g} p_{x-y}(i) \log\{p_{x-y}(i)\} \tag{11}$$

Location features: four parameters about the ROI location and shape were extracted: convexity (C(S)), compactness (C), aspect ratio (AR), area ratio (R_area). The four parameters are calculated as follows (Eq.12-15):

$$C(S) = A / area(CH(S)) \tag{12}$$

$$C = P^2 / 4\pi A \tag{13}$$

$$AR = D_y / D_x \tag{14}$$

$$R\_Area = area\_ROI(in\_pixels) / area\_window(in\_pixels) \tag{15}$$

where S is a ROI, *CH(S)* is its convex hull and *A* is the ROI's area, *P* is the ROI's perimeter, and $Area\_window = D_x * D_y$, D_x is the width's ROI and D_y is the height's ROI.

Ultimately, the output of the fourth layer—including 16 channels of 1x1 array as well as the 12 features extracted from the original image—is fully connected to XGBoost and outputs the category of the input clinical image. The image-level decision of the XGBoost can adjust the classification prediction in main-stream, the predicted result will combine both the main-stream and side-stream classification results. As shown in Fig. 4, the classification result q in the side-stream based on T1w and DWI contains the weight *Wq* in the final prediction, while the main-stream classification results contain the weight *Wp*. If we have a high probability, which means that the image may contain colorectal tumor, the final prediction will be improve by the weight *Wp*. In contrast, if we have low probability, which means that the image may be that of a healthy rectum, the final prediction will be reduced to reluctant the output of the detection box.

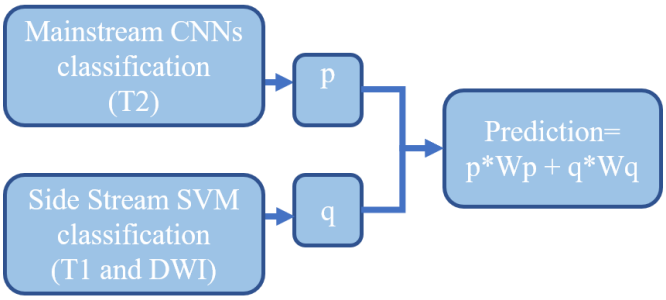


Fig. 4 Framework of the co-predictive network in the ROI classification part

Co-predictive Network Training

Loss function

Formally, during training, we define a multi-task loss function on each sampled ROI as L = Lcls + Lbox + Lmask. Different from the Mask-R-CNN, which contains a multi-class task, our co-predictive network only needs a two-task class. Thus, the classification loss Lcls and b-box loss Lbox are as follows (Eq.16-18):

$$L(\{p_i\},\{t_i\}) = 1/N_{cls}\sum_i L_{cls}(p_i, p_i^*) + \lambda/N_{reg}\sum_i p_i^* L_{cls}(t_i, t_i^*) \tag{16}$$

$$L_{cls}(p_i, p_i^*) = -\log[p_i^* p_i + (1-p_i^*)(1-p_i)] \tag{17}$$

$$L_{reg}(t_i, t_i^*) = \begin{cases} 0.5(t_i - t_i^*)^2 & if \left|t_i - t_i^*\right| < 1 \\ \left|t_i - t_i^*\right| - 0.5 & otherwise \end{cases} \tag{18}$$

where N_{cls} and N_{reg} are the mini-batch size and the number of anchor locations respectively, *i* is the index of an anchor in a mini-batch and p_i is the predicted probability of the anchor i being colorectal tumor. The ground-truth label p_i^* is 1 if the anchor is positive for colorectal tumor, and it is 0 if the anchor is negative; t_i is a vector representing the four parameterized coordinates of the predicted bounding box(Ren, S et al., 2015); t_i^* is that of the ground-truth box associated with a positive anchor.

The classification loss Lcls is a log loss over two classes (object vs. not object) as in Eq.17 (Ren, S et al., 2015). For the regression loss, we use the Eq.18 robust loss function (smooth L1) as previously defined (Ren, S et al., 2015). The outputs of the cls and reg layers consist of $\{p_i\}$ and $\{t_i\}$, respectively. Different from Mask-R-CNN, our definition of Lmask allows the network to generate masks for only 1 class, we typically used a per-pixel softmax and defined Lmask as the average binary cross-entropy loss.

Implement details

The ground-truth definition is like the Faster-R-CNN, where the ROI is considered positive if it has an intersection of union (IOU) with a ground-truth box of at least 0.5 and negative otherwise. The mask target is the intersection between the ROI and its associated ground-truth mask. We trained the main-stream and side-stream separately. The network was trained following the "image-centric" sampling strategy (Ren, S et al., 2015). Each mini-batch arises from a single image that contains many positive and negative example anchors. The main-stream is trained first. Then the transfer learning method is applied due to the size of our training dataset: we randomly initialized all head layers by drawing weights from a zero-mean Gaussian distribution with standard deviation of 0.01. All other layers in the backbone are ResNet101 initialized by pretraining a model for ImageNet classification, as it is standard practice (Ren, S et al., 2015). The training skills are used as the training pipeline in Faster R-CNN to share convolutional layers between the RPN and the backbone. We first trained the head part only. To avoid breaking the function for extracting high-level features, we froze all layers in the backbone: *conv1_x* to *conv4_x*. Then, we fine-tuned all the other layers. The side-stream is trained after the main-stream, as we used the shared structure to acquire the features of T1w and DWI, the image level ground truth and extracted features are used to train the XGBoost. The shared structure in side-stream do not need training again. When predicting, the co-predictive network is implemented with the fusion of the multimodalities MRI.

3. Experiments and Results

Each mini-batch has 4 images per graphics processing unit (GPU) and each image has 32 sampled ROIs, with a ratio of positive to negatives of 1:3 (Ren, S et al., 2015). We trained on 2 GPUs (Titan X with 16G memory per GPU), thus the batch size is 8. There were 100 steps for per training epoch, and we trained the 'head' part with 200 epochs and 'all layers' with another 200 epochs. For the training we used a learning rate of 0.001. We also used a weight decay of 0.0001 and a momentum of 0.9. In the testing dataset, there were 30 image slices for each patient (each slice has 3 modalities of MRI, T1-w, T2-w and DCE), and 30 ground-truth images, which contain the information healthy. if the image has an ROI, the ground-truth will be positive 1 and contain the contour line of the ROI, otherwise if the image is healthy, without the ROI of colorectal tumor, the ground-truth will be negative 0. There were 122 positive image slices and 478 negative image slices in the test dataset.

3.1 Colorectal tumor detection and crop

In the part of the automated rectal detection and crop, our proposed method can achieve a accuracy of 100% with our dataset. Some of the results are shown in Fig. 5. Then we enlarged the region of view and fixed it to 256*256 to encompass some of the surrounding information which can improve the performance in later detection and segmentation.

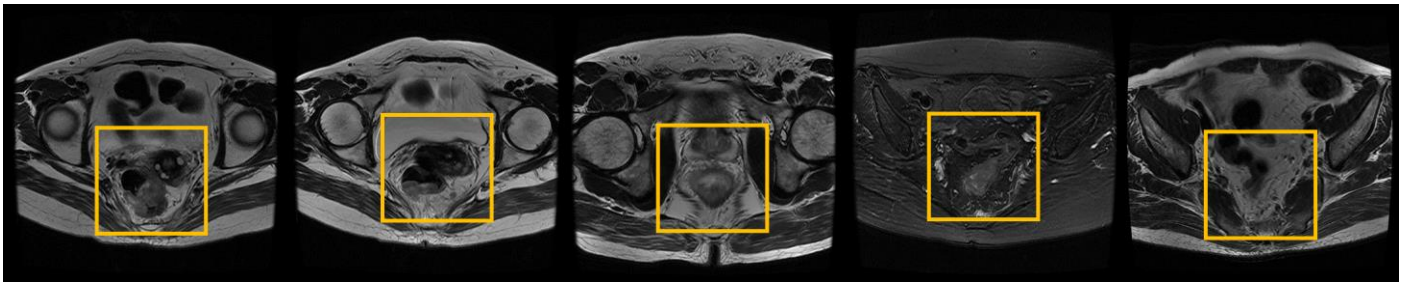


Fig.5 Colorectal tumor region detection and crop based on CNN

3.2 Image-level Tumor Detection

The image-level sensibility (Recall) was calculated and the average precision (AP) was used to evaluate the performance on the image-level colorectal tumor detection, In the actual application for medical image diagnosis, judging if the target MRI contains a rectal tumor or not is significantly more important and difficult than the other tasks. Before the decision of using the T2-w image as input in mainstream, we try different single-modalities as mainstream input to compare the performance on the image-level detection of colorectal tumor, the performance on the validation dataset is as follows:

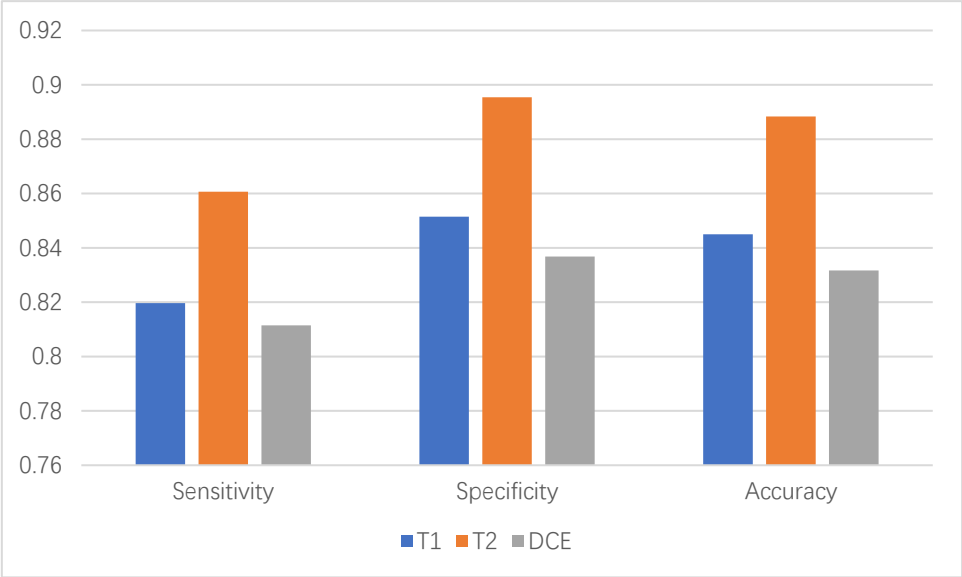


Fig.6 Image level detection performance using single modality in main-stream

Thus, we chose the T2-w image to be the input of the main-stream.

To choose the best machine learning algorithm in the final prediction step in side-stream, we used the validation dataset to evaluate the performance of different algorithms at the image-level detection. We use three machine learning algorithms, XGBoost, SVM and MLP, for the classification model. For XGBoost, the depth of tree is 8 and the number of tree is 110. For SVM, we use the Gaussian RBF kernel and the sigma value of RBF kernel is 0.1. For MLP, we use three hidden layers MLP with 50, 25 and 18 hidden units in each hidden layer. The experimental results are as follows.

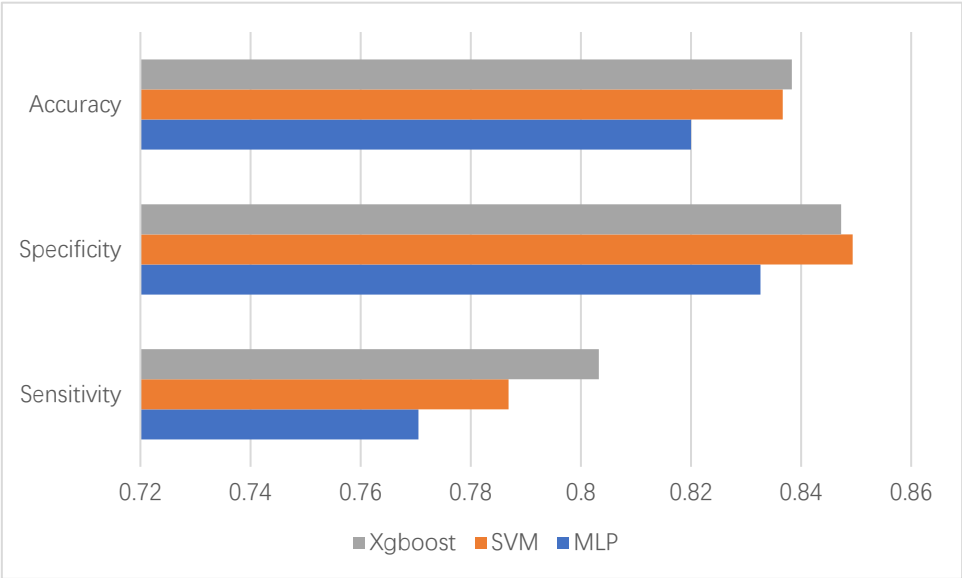


Fig.7 Image level detection performance using single modality in main-stream

Accordingly, we chose the XGBoost to make the final prediction step in side-stream. To optimize the output threshold in main-stream and the combined weight in the co-prediction (ratio between main-stream prediction and side-stream prediction), we manually adjusted the parameters and evaluated the performance at the image-level detection using the validation dataset. Some of the details are listed in Table. 3.

Table. 3 Comparison of the image-level detection performance with different combined weight in the CTD test-dev.

Weight ratio (main-stream : side-stream)	1:0.3	1:0.5	1:0.7	1:1	1:1.2
AP	90.7	91.9	90.1	89.3	88.5

We also performed an experiment to choose the best backbone in the main-stream neural network. Some of the details are shown in Table. 4:

Table. 4 Comparison of the image-level detection performance with different backbone in main-stream.

Base architecture	ZF	Vgg-16	Resnet-50	Resnet-101	Resnet-152
AP	88.4	89.0	89.6	91.8	91.5

We used the training dataset to train both the Mask-R-CNN network and our proposed co-predictive network, and used the test dataset to evaluate them. The evaluation metrics and results are shown in Table. 5.

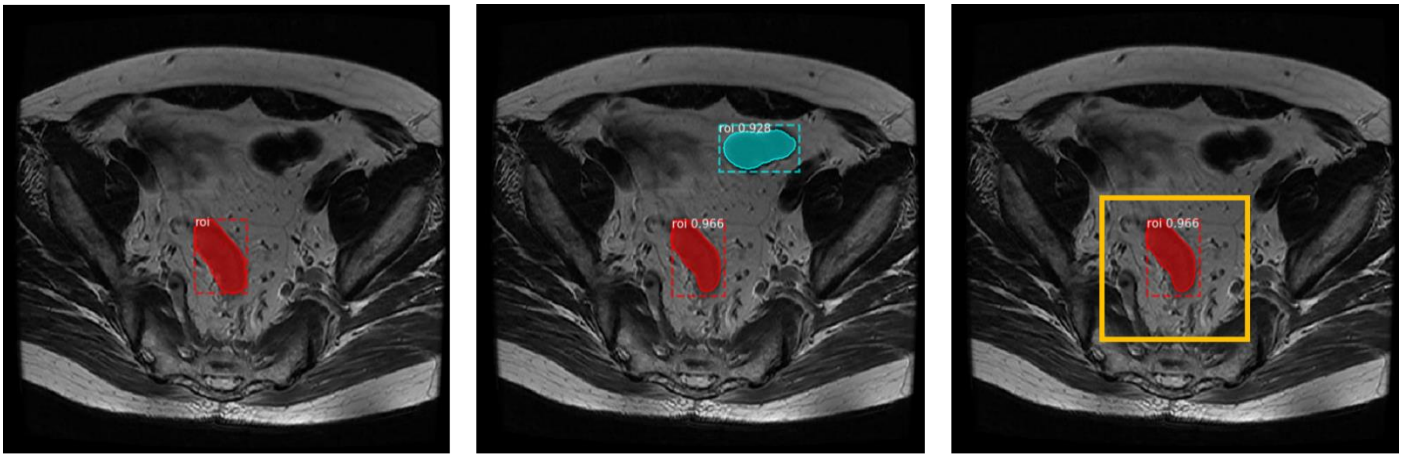
Table. 5 Comparison of the image-level detection performance with different methods

Algorithm	AP(Test Dataset)	AP(Train Dataset)
Faster-rcnn(ZF)	85.66%	89.29%
Faster-rcnn(Vgg-16)	84.34%	89.21%
Mask-rcnn (ResNet50-FPN)	85.98%	92.60%
Mask-rcnn (ResNet101-FPN)	88.71%	93.46%
Mask-rcnn (ResNet152-FPN)	89.59%	93.38%
Co-predict	**92.03%**	96.40%

Shortage and future task: After the analysis of the experiment, the result shows most of the false positive instances in the test dataset are start or end parts of the colorectal tumor, where the tumor tissues are fragmented and mixed with healthy tissues. Thus, the 3D convolution (Chen, H et al., 2016) may solve this problem by including the information from across the adjacent tumor slice, while requiring a larger computing space device.

3.3 Colorectal tumor Segmentation

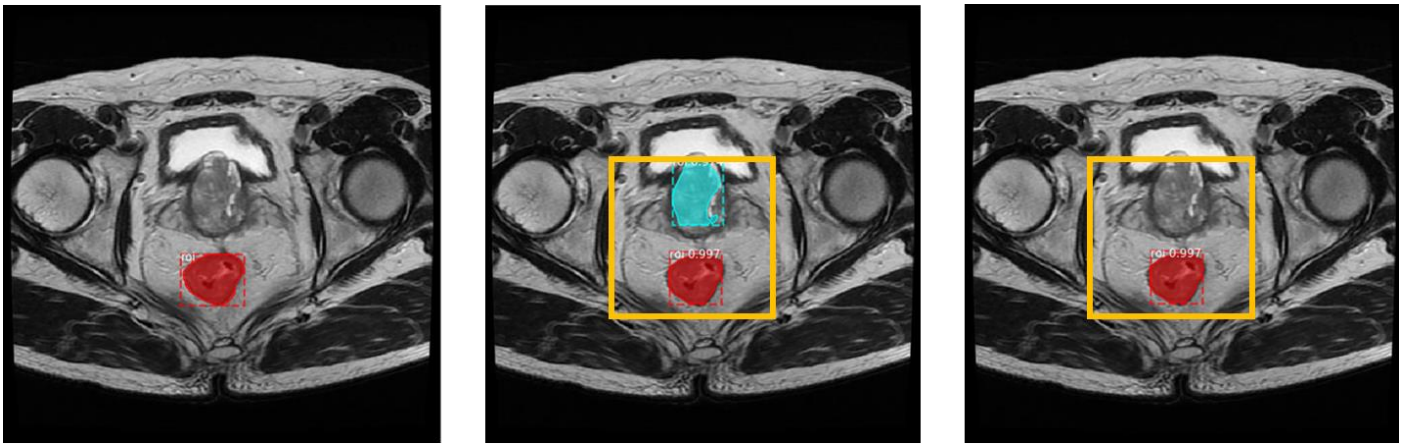
Impact of the automated colorectal detection and crop: The contributions of using the automated colorectal detection and crop to reduce false positives and improve the segmentation results are shown in Fig. 8. Specifically, Fig. 8(a) shows the original MRI image with ground truth, which is highlighted in red, and Fig. 8(b) shows the detection and segmentation results without colorectal detection and crop, which uses the whole axial view original MRI image as input. It can be observed that the cyan output is a false positive. The results obtained by using the colorectal detection and crop, shown in Fig.8(c), reveal that after colorectal detection and crop, the input image change from the original whole image as yellow cropped image, which can focus the detection and segmentation mostly on the correct field of view (FOV), and reduce the false positive output.



(a) (b) (c)

Fig.8 Advantages of using automated colorectal detection and crop to detect and segment the tumor ROI

Effect of applying the diagnostic characteristics of colorectal tumor: The contributions of using the diagnostic characteristics of the colorectal tumor as *a priori* knowledge into co-predictive method are shown in Fig. 9. Different from a normal detection task, there is only one ROI at most in one-axial colorectal slice. Thus, we only output the single ROI which has the highest predicted value. Specifically, Fig. 9(a) shows the original MRI image with ground truth, which is highlighted in red, and Fig. 9(b) shows the detection and segmentation results by applying colorectal detection and crop, which reveals that the red and cyan ROI output are both in the field of the cropped colorectal tissue, which means that using only the colorectal crop does not exclude the false positive. The results obtained by adding the diagnostic characteristics of the colorectal tumor as *a priori* knowledge into the co-predictive method, shown in Fig. 9(c), reveal that by restricting the number of output ROI with predicted values, the false positives in the cropped colorectal FOV can be reduced.



(a) (b) (c)

Fig.9 Advantages of using the diagnostic characteristics of the colorectal tumor to detect and segment the tumor ROI

The results are measured in terms of pixel IOU to quantitatively evaluate the performance in pixel-level segmentation. The quantitative evaluation results of the comparison of the segmentation pixel-level IOU in the CTD testing dataset with different methods, including Mask-rcnn, Unet (Olaf Ronneberger et al., 2015) Vnet (Fausto Milletari et al., 2016) and our co-predicted learning method are shown in Table 6. These results reveal that automated colorectal crop and co-predictive learning can be sued to further improve the results.

Table. 6 Comparison of the segmentation pixel-level IOU in the CTD testing dataset with different methods.

Algorithm	IOU
Mask-rcnn (ResNet101-FPN)	62.7%
U-net	53.5%
V-net	60.2%
Co-predict	70.1%

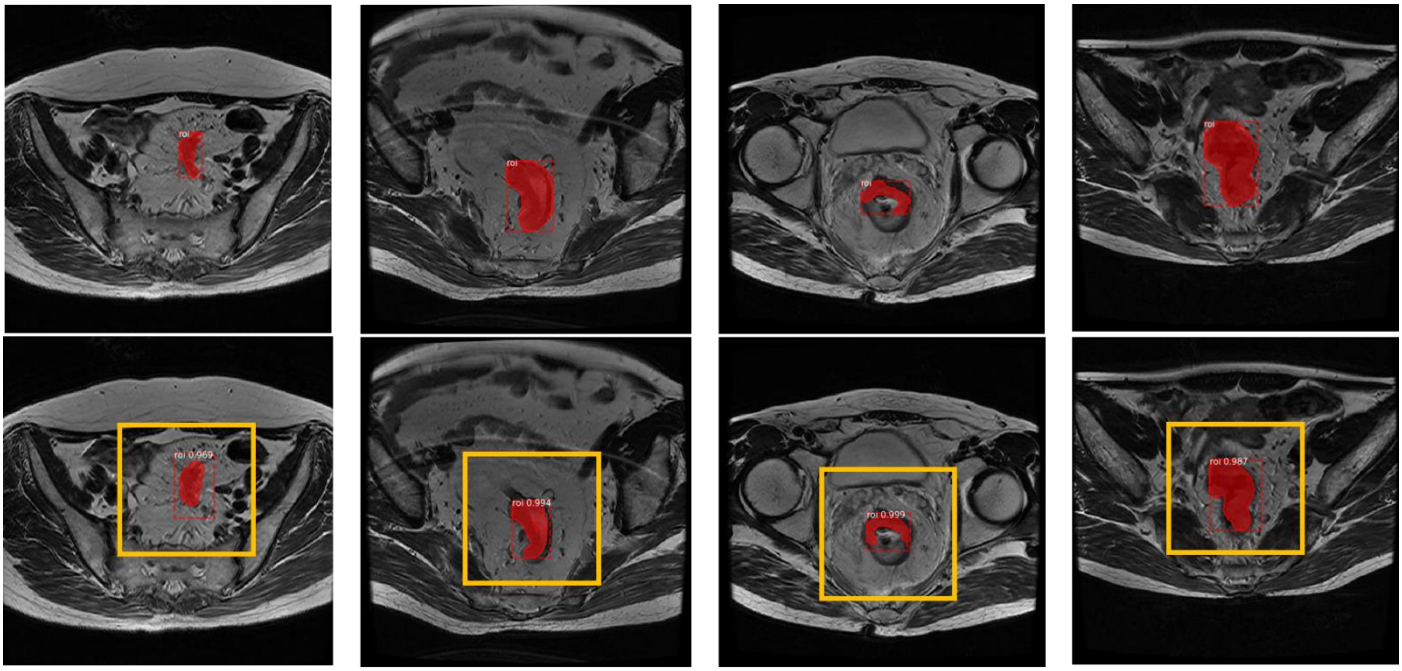


Fig. 10. Samples of the segmentation results obtained with the co-predictive learning method in the CTD testing dataset. (a)-(d) are four pairs of segmentation results where the top figures are original MRI with ground truth which is highlighted with red and the bottom figures are output ROI in red with predicted value. The yellow boxes are FOV after colorectal detection and crop.

The samples of the segmentation results obtained with the co-predictive learning method in the CTD testing dataset are shown in Fig. 10. These results show that after colorectal detection and crop, the FOV is reduced, which can focus the detection on the correct position and eliminate the influence of irrelevant region. The performance of the segmentation illustrated the advantages of using the co-predictive learning method.

Acknowledgment

This work is supported by the National Natural Science Foundation of China under Grant No. 51475305.

4 Conclusions

To accurately detect and segment a colorectal tumor, a fully automatic co-predictive learning method is proposed, which aims to imitate the diagnostic process as performed by physicians by using multimodal fusion of multiple sequences of MRI and making the method more appropriate by combining the diagnostic characteristics of colorectal tumor. First, after the registration of the multimodality MRI and preprocessing, a specific CNN structure is designed to localize the position of the colorectal tumor, (x, y) and l for the center and length of the cropped square, respectively. Then the proposed co-predictive learning method is used to detect the ROI of the colorectal tumor and perform pixel-to-pixel segmentation. The co-predictive learning method consists of two streams. The main-stream neural network detects and segments the ROI of the tumor based on the T2-w modality, while the side-stream, uses both the shared deep network and the combined machine learning algorithm to extract features from T1-w and DWI, to improve the accuracy on the image-level detection in main-stream. In addition, the difficulty of the segmentation process is reduced by the fusion of multimodality MRI and applying the diagnostic characteristics of the colorectal tumor as *a priori* knowledge into the co-predictive method. Moreover, the computing cost is low due to the shared structure in main-stream, as well as between both the side-stream and main-stream. The colorectal tumor dataset (CTD) is used to quantitatively evaluate the efficiency and generalization capability of our co-predictive learning method. The training of the main-stream and XGBoost in side-stream are separated by using different loss functions and data structure. The detection results show that the colorectal detection neural network is able to accurately crop all the colorectal area. The tumor detection and segmentation performance with 92.03% image-level detection accuracy and 70.1% pixel-level IOU indicate that the proposed system is superior compared with the state-of-the-art methods.

References

Chen, H., Dou, Q., Wang, X., Qin, J., Cheng, J. C., & Heng, P. A. 3D fully convolutional networks for intervertebral disc localization and segmentation. In International Conference on Medical Imaging and Virtual Reality. (2016) pp. 375-382.

Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. ACM, (2016).

Fausto Milletari, Nassir Navab, Seyed-Ahmad Ahmadi. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. (2016) arXiv preprint arXiv:1606.04797

Greenspan, H., Ginneken, B. van & Summers, R. M. Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique. IEEE Trans. Med. Imaging (2016)35, pp. 1153–1159.

Gambacorta, Maria Antonietta, et al. "Clinical validation of atlas-based auto-segmentation of pelvic volumes and normal tissue in rectal tumors using auto-segmentation computed system." Acta Oncologica (2013) 52.8, pp. 1676-1681.

Hall-Beyer, Mryka. "GLCM texture: a tutorial." National Council on Geographic Information and Analysis Remote Sensing Core Curriculum (2000).

He, K., Gkioxari, G., Dollár, P., & Girshick, R. Mask r-cnn. In Computer Vision (ICCV), IEEE International Conference on (2017) pp. 2980-2988.

He, K., Zhang, X., Ren, S., & Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (2016) pp. 770-778.

Klein, S., Staring, M., Murphy, K., Viergever, M. A. & Pluim, J. P. W. Elastix: a toolbox for intensity-based medical image registration. IEEE Trans. Med. Imaging (2010)29, pp. 196–205.

Lambregts, D. M. J. et al. MRI and Diffusion-weighted MRI Volumetry for Identification of Complete Tumor Responders After Preoperative Chemoradiotherapy in Patients With Rectal Cancer: A Bi-institutional Validation Study. Ann. Surg. (2015) 262, 1034–9.

LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. Nature (2015)521, pp. 436–444

Martens, M. H. et al. Prospective, multicenter validation study of magnetic resonance volumetry for response assessment after preoperative chemoradiation in rectal cancer: Can the results in the literature be reproduced? Int. J. Radiat. Oncol. Biol. Phys. (2015) 93, pp. 1005–1014.

Nair, Vinod and Hinton, Geoffrey E. Rectified linear units improve restricted Boltzmann machines. In ICML (2010) pp. 807–814.

Olaf Ronneberger, Philipp Fischer, Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. (2015) arXiv preprint arXiv:1505.04597

Ren, S., He, K., Girshick, R., & Sun, J.. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems. (2015) pp. 91-99.

Seierstad, Therese, et al. "MRI volumetry for prediction of tumour response to neoadjuvant chemotherapy followed by chemoradiotherapy in locally advanced rectal cancer." The British journal of radiology 88.1051 (2015): 20150097.

Trebeschi, Stefano, et al. "Deep learning for fully-automated localization and segmentation of rectal cancer on multiparametric MR." Scientific reports (2017)7.1, 5301.

Van Heeswijk, M. M. et al. Automated and semiautomated segmentation of rectal tumor volumes on diffusion-weighted MRI: Can it replace manual volumetry? Int. J. Radiat. Oncol. Biol. Phys. (2016) 94, pp. 824–831.