Abstract—Accurate segmentation of rectal cancer and rectal wall based on high-resolution T2-weighted magnetic resonance imaging (MRI-HRT2) is the basis of rectal cancer staging. However, complex imaging background, highly characteristics variation and poor contrast hindered the research progress of the automatic rectal cancer segmentation. In this study, a multi-task learning network, namely mask segmentation with boundary constraints (MSBC-Net), is proposed to overcome these limitations and to obtain accurate segmentation results by locating and segmenting rectal cancer and rectal wall automatically. Specifically, at first, a region of interest (RoI)-based segmentation strategy is designed to enable end-to-end multi-task training, where a sparse object detection module is used to automatically localize and classify rectal cancer and rectal wall to mitigate the problem of background interference, and a mask and boundary segmentation block is used to finely segment the RoIs; second, a modulated deformable backbone is introduced to handle the variable features of rectal cancer, which effectively improves the detection performance of small objects and adaptability of the proposed model. Moreover, the boundary head is fused into the mask head to segment the ambiguous boundary of the target and constrain the mask head to obtain more refined segmentation results. In total, 592 annotated rectal cancer patients in MRI-HRT2 are enrolled, and the comprehensive results show that the proposed MSBC-Net outperforms state-of-the-art methods with a dice similarity coefficient (DSC) of 0.801 (95% CI, 0.791-0.811), which can be well extended to other medical image segmentation tasks with high potential clinical applicability.

Index Terms—Medical image segmentation, rectal cancer, MRI, deep learning.

I. INTRODUCTION

According to the global cancer statistics recorded in 2020 [1], China ranks second in colorectal cancer incidence and fifth in mortality. Therefore, the screening and treatment of colorectal cancer are an urgent requirement. Preoperative chemoradiotherapy can reduce the local recurrence of the locally advanced rectal cancer, but unnecessary overtreatment can cause unexpected complications, so accurate T staging of rectal cancer is critical. According to the diagnostic criteria of rectal cancer, radiologists first identify the location and shape of the cancer and the rectum from medical images, and then diagnose the staging of rectal cancer by determining the depth of rectal cancer infiltrating the rectal wall. Therefore, accurate segmentation of rectal cancer and rectal wall is the primary task to guide the staging of rectal cancer and determining a suitable treatment plan. However, manual segmentation of rectal cancer requires multi-sequence images and clinical information to accurately locate the cancer. Fully automatic segmentation can greatly reduce the workload of radiologist and also reduce segmentation error caused by individual bias and clinical experiences, to a certain extent.

The application of deep learning based automatic segmentation technology could considerably facilitate the study of colorectal cancer. However, to the best of our knowledge, only a limited amount of research has been conducted on the automatic localization and segmentation of rectal cancer. Most of the existing algorithms focus on lung segmentation [2]– [4], brain segmentation [5]– [7], pancreas segmentation [8]– [10], retinal vessel segmentation [11]– [13], and so on. Different types of cancer possess different characteristics. Particularly, rectal cancer is difficult to locate and segment accurately compared with other types of cancer.

Magnetic resonance imaging (MRI) can assess the depth of the tumor penetration into rectal wall and can clearly show the soft tissue structure of the surrounding pelvic region and thus is a standard method used to determine the stage of rectal cancer. However, the automation of lower abdomen imaging procedures is hampered by several drawbacks, as follows.

1) Complex backgrounds in images and objects lack positional priors: In rectal MR images, the background relative to the segmentation targets includes the intestinal content, adipose tissue, and other normal organs, etc. Moreover, lesions such as those of the lung or liver have a simple background and, since they are solid viscera, show minimal morphological changes on slice scans. However, the rectum is a hollow viscera, so the background around rectal wall and rectal cancer are highly variable in different slices. Therefore, it is difficult to accurately locate and segment rectal cancer and rectal wall in such a complex background. In addition, the axial sampling direction is scanned along the axis of the diseased intestine, which may change with different patients, resulting in the
variable positions of the rectum and the tumor in the images.

2) Objects lack shape characteristics: The rectum has different shapes at different scanning levels (e.g., round, strip, and tube). Additionally, contraction and dilation of the rectum as well as motion produce different shapes on the scanning slices. Therefore, it is difficult for the model to adapt to the high variation in the scale and shape of the targets.

3) Low contrast between tumors and normal tissue: The boundary of rectal cancer usually has a low contrast, making it difficult to distinguish cancer from normal tissue. The deep learning network structures applied in the medical image segmentation mostly adopt full convolution or encoder-decoder structure, such as FCN [14], SegNet [15], DeepLab [16]–[19] series, U-Net [20] and many variants of U-Net. However, a common limitation of these methods is that the process of learning increasingly abstract feature representations involves consecutive pooling operations or convolution striding, which reduce the feature resolution and lead to the loss of some spatial and shape information, making it more difficult for the networks to segment the boundary of targets.

To address above issues, we propose a novel segmentation network termed mask segmentation with boundary constraints (MSBC-Net). The proposed method integrates classification, regression and segmentation into a single network, which is a sparse, flexible, and versatile multi-task learning network that enable end-to-end multi-task training, and is effectively used to segmentation of rectal cancer and rectal wall. In this method, shallow representations are shared among multiple tasks and the parameter-based sharing mechanism has two effects on the learning of the main task of the final rectal cancer segmentation task: firstly, sharing and complementing the learned domain information through shallow shared representation, so as to promote each other to learn and improve the penetration and acquisition of information; secondly, when multiple tasks are backpropagated at the same time, the shared representation takes into account the feedback from multiple tasks, and compared with single task, the overfitting risk is reduced and the generalization ability is enhanced.

The motivations for the design of MSBC-Net stems from the fact that U-Net and some end-to-end semantic segmentation network improved networks based on U-Net to perform segmentation task on the rectal MRI dataset, resulting in the mis-segmentation of the background as objects. Moreover, small rectal cancers are easily overlooked and positive subjects are identified as negative cases. The detection-based segmentation network can effectively reduce the interference of the background, thus decreasing false positive rate. To simplify the detection pipeline, we abandoned dense-to-sparse method such as R-CNN relying on dense candidate, most of which are based on a large set of proposals [21], [22], anchors [23], or window centers [24], [25]. Their performances are significantly affected by the post-processing steps, due to which a sparse object detection strategy is adopted in this study. A modulated deformable ResNet-FPN (MDRF) is employed as the backbone to perform feature extraction on the pre-processed single-scale images, and then generate features with strong semantic information pyramids at all scales, the extracted features are fully utilized at each stage. This can effectively alleviate the problem of high false negative rate caused by too small cancer area. The mask head and boundary head parallel to the detection head are used to generate the segmentation maps. The contributions of this paper are summarized as follows:

- We propose a novel segmentation network, MSBC-Net, for automatic localization and segmentation of rectal cancer and rectal wall, which provides an effective and accurate automatic tool for the clinical management of rectal cancer. The proposed method enables end-to-end multi-task training for simultaneous classification, regression, and segmentation tasks, and it is experimentally verified that MSBC-Net can be extended to other segmentation tasks and obtain state-of-the-art segmentation accuracy.

- The proposed three parallel branches of classification, regression, and mask generation in the multi-task dynamic head (MTDH) block are integrated, and the proposed boundary head (BH) and mask head (MH) are fused to obtain finer segmentation details. To ensure that the MSBC-Net is more suitable for irregular and variable targets, a modulated deformable backbone (MDRF) is introduced.

- Extensive ablation experiments are conducted to confirm the importance of each proposed component and to demonstrate the efficacy of the proposed approach by comparing it with several advanced methods used in the medical image segmentation.

II. RELATED WORKS

In the field of medical image analysis, medical image segmentation can be used in image-guided interventional diagnosis and treatment, directional radiotherapy and other processes. Currently, various segmentation methods based on deep learning have been applied to medical image segmentation.

Ronneberger et al. [20] designed U-Net for biomedical images. Owing to its excellent performance, U-Net and its variants have been widely used in various subfields of computer vision, these methods lies on a U-shaped structure and integrate multi-scale features and realizes the combination of low-resolution information and high-resolution information. However, the continuous pooling and strided convolution used in these methods results in the loss of some spatial information. Therefore, Gu et al. [30] proposed a context encoder network for 2D medical image segmentation, named CE-Net, to capture more high-level information and retain spatial information and applied it to different segmentation tasks. Nevertheless, the inception module in CE-Net is complex, increasing the difficulty in the modification of the model. And Zhou et al. [28] reported that the skip connection to combine the low-level features of encoder in U-Net with the high-level features of decoder is not appropriate, it suffers from the problem of the semantic gap, U-Net++ redesigned the skip connection, and realized the flexible feature fusion in the decoder, which is an improvement over the restrictive skip connection that only fuses the feature maps of equal size in U-Net. Furthermore, Fan et al. [31] proposed Inf-Net for CONVID-19 lung CT infection segmentation, which employs
reverse attention and edge attention to improve the recognition of the infected areas. However, in complex segmentation tasks such as rectal cancer segmentation, the edge attention of InfNet is not really effective. The details are analyzed in the following section.

In the fields of object detection and segmentation, He et al. [32] proposed Mask R-CNN, which is one of the most widely used object detection and segmentation algorithms, a parallel mask branch is added based on the classification and regression branch of the Faster R-CNN [33]. We employed the idea of parallel branches and attempted to use this mask branch as the mask head of our model, but the results indicated that the proposed novel MBSB block in our study is more effective, which is demonstrated in Section IV. A recent study by Carion et al. [34] introduced the DETR that uses a transformer to detect objects efficiently. This is the first object detection framework that successfully integrates the transformer into the central building block of the detection pipeline. However, the DETR training period is long, 10-20 times slower than Faster R-CNN, and the performance of small objects detection is relatively poor. Sparse R-CNN [35] is similar to the DETR, instead of complex traditional target detection routines, there is no proposal (Faster R-CNN), no anchor (YOLO [36], [37]), no center (CenterNet [25]), and no NMS. It directly predicts the detection boxes and categories. In this study, the Sparse R-CNN object classification and regression strategy is adopted, and some improvements are proposed, the details of which are provided in Section III.

Limited research has been conducted on the automatic localization and segmentation of rectal cancer [38]–[42]. The methods involved in these works still rely on FCN or U-Net, where the input images are encoded and then decodes to recover the per-pixel classification, but continuous down sampling causes the loss of spatial information and up sampling can not recover it well. In addition, it is insensitive to the details in the image, can not achieve more fine segmentation, and is vulnerable to the interference of the background, resulting in the rise of false positives.

III. PROPOSED METHOD

The MSBC-Net consists of a modulated deformable backbone (MDRF), a multi-task dynamic head (MTDH) for classification, box regression and mask generation, and a mask and boundary segmentation block (MBSB) fuse the mask head (MH) with the boundary head (BH), as shown in Fig. 1, the structure of the MTDH is also detailed in Fig. 2. Overall, the proposed MSBC-Net is a simple and versatile multi-task segmentation network, and many manual design components, such as anchor generation or non-maximum suppression procedures, are effectively eliminated.

A. Modulated Deformable Backbone

A modulated deformable ResNet50-FPN (MDRF) is adopted as the backbone network to generate multi-scale feature maps from the input images. The traditional convolution kernels usually have a fixed geometric structure and a fixed size and cannot make adaptive changes when the target is enlarged or rotated. Although data augmentation or the use of transformation-invariant features and algorithms can alleviate the geometric deformation problem to a certain extent, it is assumed to be fixed and known, which is a priori information. It is irrational to use these known deformations to resolve the unknown deformations, and hand-designed features or algorithms cannot handle excessively complex deformations.

The modulated deformable convolution is introduced into our method, as defined as in (1), to improve the adaptive deformation ability of MSBC-Net. Specifically, in order to apply the modulated deformable convolution as a single layer without affecting other layers, in practice, the convolution kernel is not really extended, instead the learned offset is used to recalculate the pixels position of the feature maps before the convolution to realize the expansion of the convolution kernel. In addition, when the image pixels are reconsolidate, the pixels need to be offset. The data type of generated offset is float and must be converted to integer type, if the offset is rounded directly, cannot back propagation, then adopt the bilinear interpolation to obtain the corresponding pixels, and different weights are assigned to the positions after offset correction to achieve more accurate feature extraction.

Given the convolution kernel of K sampling positions, $\omega_k$ and $p_k$ represent the weight and pre-specified offset of the k-th position, respectively. $x(p)$ and $y(p)$ represent the features at position p from input feature map x and output feature maps y, respectively. $\Delta p_k$ and $\Delta m_k$ are the learnable offset and modulation scalar for the k-th location, offset is the position of the region where effective information is to be found, and modulation scalar is to give weight to the position. Both aspects guarantee the accurate extraction of effective information.

$$y(p) = \sum_{k=1}^{K} \omega_k \cdot x(p + p_k + \Delta p_k) \cdot \Delta m_k$$

(1)
Fig. 2. An overview of Multi-task Dynamic Head. P2-P5: multi-scale feature maps extracted from MDRF, P2 as the boundary features and P2 through P5 as the mask features; DH: detection head, used for RoIs classification and box regression; MBSB: mask and boundary segmentation block containing the mutual learning process of MH and BH. \( N \) refers to the number of proposal boxes, \( N' \) is the number of foreground proposal boxes, \( B \) represents the batch size, and \( D \) is the dimension.

Fig. 3. Detection head. The RoIs features and the proposal features after multi-head attention serve as input to the dynamic instance interactive module to extract features for each target. The final classification prediction is calculated from a module containing linear layer, layer normalization (LN) layer and ReLU activation, and the regression prediction from 3-layer perception with linear layer, LN layer and ReLU activation function.

In this study, the Feature Pyramid Network [43] based on ResNet [44] is adopted as the backbone of the proposed model. And the number of bottleneck blocks is 4, 6, and 3 respectively in the res3, res4, and res5 stages of ResNet, and all the 13 conv 3 × 3 layers are replaced by modulated deformable convolutions [45]. A feature pyramid structure with P2-P5 layers is then conducted to deal with multi-scale changes in the task and finally output the feature maps of different sizes with high-level semantic information.

B. Multi-task Dynamic Head

The multi-task dynamic head (MTDH) is an iterative recursive multi-task learning structure. The core of this module involves iteratively learning the proposal boxes \((N \times 4)\) and proposal features \((N \times D)\) that are initialized, where the proposal features is used to encode rich instance features. Specifically, the proposal boxes, the proposed features, and the feature maps extracted by the MDRF module are taken as input to MTDH, with the proposal boxes and features learned from the previous iteration as input to the next iteration. Note that the feature maps extracted by the MDRF module are shared by the DH and MBSB, as shown in Fig. 2.

First, layer P2 of the feature maps extracted by the MDRF module is taken as the boundary features, and layers P2-P5 are taken as the mask features and the box features together; second, RoIALign is performed with the proposal boxes according to different pooler resolution to extract the features of each box to obtain the region of interest (RoIs) features, noting that for the MBSB module, the foreground boxes in the proposal boxes are selected as input. Then the RoIs features, foreground-based RoIs mask features and RoIs boundary features are obtained, in which the RoIs features are fed to the DH module and the rest input to MBSB. In summary, this simpler and sparse pipeline avoids many manually designed components compared to two-stage pipelines.

C. Detection Head

The detection head (DH) contains classification and regression of the RoIs, detailed in Fig. 3. In this module, the critical process is the dynamic interaction of the RoIs features and proposal features. The dynamic interaction process follows the dynamic instance interactive head in [35], which aims to extract features for each instance and then predict the category and coordinate offset of each box. Note that the dynamic dimension is set to 16. Specifically, the proposed features interact one-to-one with the RoIs features extracted from the multi-scale feature maps by the proposed boxes, highlighting the bin \((7 \times 7)\) that contribute the most to the foreground, thereby affecting the target location and classification prediction. If it is the background, none of the bins has a high output value. Furthermore, the multi-head attention in DH acts on the proposal features designed to infer the relationship between instances.

D. Mask and Boundary Segmentation Block

The MBSB, shown in Fig. 4, is a newly proposed module, which can be considered as a simple encoder-decoder structure. The spatial domain information is crucial for segmentation, first, for rectal cancer segmentation task, the target lacks
shape characteristics, specifically the heterogeneous texture and ambiguous structural boundary observed in tumors, as shown in Fig. 5; second, in the usual segmentation methods, although the details lost in the down-sampling process can be collected to a limited extent by the multi-level features, the stride convolution or continuous pooling still cause the loss of spatial and shape information and cannot be recovered by upsampling. Therefore, BH is proposed to predict the structural boundary of the target to better extract its boundary and shape information.

In MBSB, the RoIs boundary features and RoIs mask features obtained after RoIAlign of foreground proposal boxes with P2 and P2 through P5 feature maps respectively serve as the input of MBSB, exhibiting fusion between the two branches.

Fig. 4. Overview of the proposed MBSB. MBSB is composed of MH and BH. The RoIs boundary features and RoIs mask features obtained by RoIAlign of foreground proposal boxes with P2 and P2 through P5 feature maps respectively serve as the input of MBSB, exhibiting fusion between the two branches.

Fig. 5. Typical examples of MR slices with rectal cancer. The red, green, and yellow areas represent rectal cancer, rectal wall, and transitional zone, respectively. It shows that the targets area lack shape features and location priori, and cancers boundary is ambiguous.

E. Loss Function

As shown in (2), the multi-task loss for each sampled RoI is defined, allowing the network to generate mask for each class without competition between classes, that is,

$$\mathcal{L} = \lambda_{cls} \cdot \mathcal{L}_{cls} + \mathcal{L}_{box} + \mathcal{L}_{mask} + \mathcal{L}_{boundary},$$  \hspace{1cm} (2)

where $\mathcal{L}_{cls}$ is the focal loss [23] of the ground truth category labels and predicted classifications, and $\lambda_{cls}$ is the coefficient of the classification loss. $\mathcal{L}_{mask}$ denoted the average binary cross-entropy loss. The box regression loss $\mathcal{L}_{box}$ is defined as (3) and the boundary loss $\mathcal{L}_{boundary}$ is defined in (4). For the RoI corresponding to the ground-truth class category k, the $\mathcal{L}_{mask}$ is only defined for the k-th mask.

$$\mathcal{L}_{box} = \lambda_{L_1} \cdot \mathcal{L}_{L_1} + \lambda_{giou} \cdot \mathcal{L}_{giou}$$ \hspace{1cm} (3)

where $\mathcal{L}_{L_1}$ and $\mathcal{L}_{giou}$ are the $L_1$ loss and IoU loss [45], respectively, $\lambda_{L_1}$ and $\lambda_{giou}$ are the coefficients corresponding to the $L_1$ loss and IoU loss, respectively.

$$\mathcal{L}_{boundary} = \mathcal{L}_{BCE} + \mathcal{L}_{Dice}$$ \hspace{1cm} (4)

In medical image segmentation tasks, the anatomy of interest usually occupies only a small scanning area, which often cause the learning process to fall into the local minimum value of the loss function, resulting in a network with a strong prediction bias to background. Consequently, the foreground area is often lost or only partially detected. For MH which generates mask based on proposal box, where the foreground dominates and considering the stability of training, we define $\mathcal{L}_{mask}$ with the average binary cross-entropy loss; while in boundary segmentation, there are few pixels in RoI, resulting in class imbalance issue.

In general boundary segmentation research, a common practice to assign weights to different classes to alleviate the class imbalance problem in the boundary prediction, which give rare labels more importance based on standard cross-entropy loss. Although valid for some unbalanced problems, may encounter bottlenecks for highly unbalanced datasets, assigning a large weight means that it may also amplify noise and cause instability.

The boundary learning is optimized in this study by combining the binary cross-entropy loss and dice loss as the loss of BH, as shown in (4), $\mathcal{L}_{Dice}$ is the dice coefficient loss which has a greater effect on the problem of the category
imbalance in this task. The dice coefficient is used to measure the overlap between the predictions and ground truths when the ground truths are available. The boundary ground truths are generated from the binary mask ground truths using Laplacian operator, which is a two-step operator that can produce thin boundaries. The resulting boundary is converted into binary maps using a threshold of 0 as the final ground truth.

Dice coefficient loss is expressed as:

\[ L_{\text{Dice}}(p_b, g_b) = 1 - \frac{2 \sum_{i=1}^{H \times W} (p_{bi}^2 + \epsilon) + \sum_{i=1}^{H \times W} (p_{bi}^2 + \epsilon)}{\sum_{i=1}^{H \times W} (p_{bi})^2 + \sum_{i=1}^{H \times W} (g_{bi})^2 + \epsilon} \]  

where \( i \) represents the \( i \)-th pixel, and \( \epsilon \) is a smoothing term used to avoid division by zero (we set \( \epsilon = 1e^{-8} \)). \( H \) and \( W \) denote the height and width of the predicted boundary map, respectively, \( p_b \in \mathbb{R}^{H \times W} \) represents the predicted boundary, and \( g_b \in \mathbb{R}^{H \times W} \) represents the corresponding boundary map.

IV. EXPERIMENT

A. Dataset and Metrics

1) Dataset: The experiment was performed using the MR images of rectal cancer patients obtained from Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, from February 2011 to June 2020, and this retrospective study was approved by the ethics committee of Sir Run Run Shaw Hospital, Zhejiang University School of Medicine. A total of 592 patients are enrolled after excluding the preoperative abandoned. A GE Medical Systems - Discovery MR750w. The number of pathological staging information, and primary tumors of other patients, constituting 268 2D images, and can be downloaded for free from the official website 2.

The disease well in the lung CT images can be identified by finding the lungs first, and it is necessary to accurately segment the lungs. Moreover, to demonstrate that MSBC-Net can be extended to more other segmentation tasks, MSBC-Net is also applied to the Kvasir-SEG [51] and ISIC-2017 [52] dataset. A brief description of the two datasets is as follows: Kvasir-SEG contains 1000 gastrointestinal polyp images and corresponding segmentation masks, each containing at least one polyp region, with the image resolutions ranging from 332 × 487 to 1920 × 1072. ISIC-2017 contains a training set of 2000 annotated dermatological images and a total of 600 images for testing, with image sizes from 540 × 722 to 4499 × 6748.

2) Evaluation Metrics: In this study, three widely adopted metrics are used to evaluate the final segmentation results: dice similarity coefficient (DSC), specificity (Spec.) and sensitivity (Sen.).

B. Experimental Setup

1) Training details: MR images of 592 subjects are randomly divided into a training set of 572 cases and a test set of 65 subjects, with a total of 13,396 slices, which are completely separated from the training set. Data augmentation includes random horizontal flip and random vertical flip. The shortest edge is 256 pixels, the longest edge is 512 pixels at most while maintaining the aspect ratio of the original images. The performance of the method is evaluated by performing multi-class segmentation (i.e., segmenting rectal cancer and rectal wall) and calculating the DSC, specification, and sensitivity of the targets in the segmentation map and the objects in the manual annotation results. The annotators and reviewers jointly defined the ground truth about the cancers and rectums in all the 2D MR images.

In the lung segmentation dataset, 90% of all images are used as the training and the remaining 10% were used as testing. The HU values of the original lung CT images are intercepted to [-125, 275] and normalized to [0, 1]. In addition, random contrast and random saturation were added on the basis of the data enhancement strategy of rectal cancer segmentation. For the Kvasir-SEG dataset, 80% of the dataset are used as the training set and 10% as the test set. In the ISIC-2017 dataset, the original training set and test set are retained without repartition. With the above two datasets keeping the original image aspect ratio, the shortest edge is 256 pixels and the longest edge is up to 512 pixels.

AdamW [46] with a weight decay of 0.0001 is adopted as the optimizer, the batch size is set to 8 and the model is trained on an RTX5000 GPU. The backbone of MSBC-Net is initialized with the pretrained weights on ImageNet [48], the learning rate is initialized to \(2.5 \times 10^{-5} \) and the warm up [44] learning rate optimization method is adopted during the train. The maximum number of iterations is set to 270,000, and the learning rate was reduced by 0.1 times and 0.01 times after 210000 and 250000 iterations, respectively. The data augmentation includes random horizontal and random transformations of the image contrast. The default number of proposal boxes is set to 300, and the number of MTDH

1https://www.itksnap.org/  
2https://www.kaggle.com/kmader/finding-lungs-in-ct-data/data/
iterations is 6. The MH and BH are turned on by default. To apply the proposed model for object detection tasks instead of segmentation tasks, these two heads can be conveniently turned off to convert this network into an object detection network.

2) Inference details: During testing, for a given input image, the MSBC-Net can directly predict $N$ boxes associated with their score, and the detection boxes larger than the score threshold 0.05 are applied to MBSB. Although this is different from the parallel computation used in training, it can expedite the inference and improve the accuracy owing to the use of less and more accurate RoIs. MSBC-Net is a simple and efficient model that operates at 15 FPS during inference.

C. Main Results

Extensive experiments are conducted to evaluate the performance of the MSBC-Net. First, we apply U-Net , Inf-Net and several other state-of-the-art methods to our rectal cancer MRI dataset. Table I, shows the comparison of the MSBC-Net with the aforementioned methods. This comparison clearly shows that the proposed method achieves superior results with DSC, sensitivity, and specificity are 0.801 (95% CI, 0.791-0.811), 0.811 (95% CI, 0.793-0.828), and 0.998 (95% CI, 0.998-0.998), respectively. Fig. 6 shows the prediction effect of MSBC-Net. Finally, we also evaluated the model complexity of the competitive methods and MSBC-Net, and the results are shown in Table I.

As shown in the first row, by visual predictions, all other models except our method have difficulty in separating them for adherent targets, incorrectly identified the rectal wall as rectal cancer. As for the third row, most of the comparison methods show poor segmentation details, and identify the background as foreground. And in the second and last row, it can be seen that MSBC-Net is able to identify the boundary of the target well and avoids only partial segmentation of the target. Note that TransUNet [49] and Swin-Unet [50] segmentation are not satisfactory, the main reason is that both methods follow the technical route of dividing images into multiple fixed size patches and then embedding, but this may disrupt the semantics in images. On the one hand, the fixed method of dividing the image patch ignores the geometric changes of the same object in different images; on the other hand, it is also common for local structure of objects in images to be split, while a fixed patch is difficult to capture the complete local structure related to the object. Furthermore, Swin-Unet completely removes the CNN, CNN-based method, although it is more difficult than Transformer to learn explicit global and long-range semantic information interaction due to the limitation of its local convolutional operation, it is better than Transformer to extract detailed features of targets, so the visualization results show that the segmentation of Swin-Unet details are poor.

Second, to evaluate the generality and effectiveness of our method, we apply MSBC-Net to three public datasets for different tasks: the lung segmentation dataset, Kvasir-SEG and ISIC-2017. The segmentation results of the proposed model and comparison methods on these three datasets are shown in Table II, Table III and Table IV respectively. The segmentation visualization results are shown in Figure 7, which demonstrates the effectiveness and competitive segmentation performance of the proposed model, with the best segmentation results shown on the ISIC-2017 dataset. In Figure 7, it is
Fig. 7. Segmentation results on public datasets. The bottom two rows are predictions on the ISIC-2017 dataset, and the third to fourth rows show prediction results of the polyp on the Kvasir-SEG. The top two rows represent the results on the lung segmentation dataset.

TABLE I

| Methods     | DSC (95% CI) | Sen. (95% CI) | Spec. (95% CI) | Params./M | Flops/GMac |
|-------------|--------------|---------------|----------------|-----------|------------|
| U-Net       | 0.711 (0.692-0.729) | 0.723 (0.716-0.730) | 0.998 (0.998-0.998) | 17 | 40.7 |
| Inf-Net     | 0.727 (0.712-0.740) | 0.739 (0.730-0.748) | 0.998 (0.998-0.999) | 33 | 7.4 |
| CE-Net      | 0.759 (0.727-0.794) | 0.781 (0.764-0.799) | 0.994 (0.993-0.994) | 29 | 8.9 |
| TransUNet   | 0.656 (0.607-0.696) | 0.629 (0.599-0.655) | 0.995 (0.995-0.996) | 105 | 32.2 |
| Swin-Unet   | 0.616 (0.562-0.663) | 0.625 (0.593-0.654) | 0.994 (0.993-0.995) | 27 | 7.6 |
| MSBC-Net    | 0.801 (0.791-0.811) | 0.811 (0.793-0.828) | 0.998 (0.998-0.998) | 81 | 30.1 |

verified that the MSBC-Net exhibits outstanding segmentation details and detection performance for small targets on these three datasets.

In addition, we conduct a detailed analysis of Inf-Net. Inf-Net adds an edge attention module to significantly improve the representation of the object region boundary, and the module contains four convolutional layers, with one $1 \times 1$ convolution and three $3 \times 3$ convolutions. However, we find that the output of the edge branch does not extract the edge features of objects. Then we perform ablation experiments, and the extraction results of feature maps obtained in each layer of this module are exported. The results are shown in Fig. 8. In a nutshell, the branch used to extract the object edge in Inf-Net is invalid for the task required in this study. However, the proposed boundary head effectively extract object boundary and outputs the effective boundary feature maps.

D. Ablation Study

In this section, a number of experiments are performed to verify the performance of each key component of MSBC-Net, including MBSB, BH, MDRF, proposal box initialization method, and number of proposal boxes.
TABLE II

PERFORMANCE COMPARISON WITH DIFFERENT METHODS ON LUNG SEGMENTATION DATASET (MEAN (95% CI))

| Methods   | DSC    | Sen.    | Spec.    |
|-----------|--------|---------|----------|
| U-Net     | 0.924  | 0.937   | 0.977    |
| Inf-Net   | 0.976  | 0.991   | 0.988    |
| CE-Net    | 0.986  | 0.990   | 0.996    |
| TransUNet | 0.984  | 0.981   | 0.996    |
| Swin-Unet | 0.964  | 0.962   | 0.987    |
| MSBC-Net  | 0.976  | 0.974   | 0.993    |

TABLE III

PERFORMANCE COMPARISON WITH DIFFERENT METHODS ON KVASIR-SEG DATASET (MEAN (95% CI))

| Methods   | DSC    | Sen.    | Spec.    |
|-----------|--------|---------|----------|
| U-Net     | 0.746  | 0.782   | 0.967    |
| Inf-Net   | 0.904  | 0.917   | 0.986    |
| CE-Net    | 0.871  | 0.889   | 0.980    |
| TransUNet | 0.880  | 0.899   | 0.978    |
| Swin-Unet | 0.879  | 0.898   | 0.982    |
| MSBC-Net  | 0.886  | 0.892   | 0.985    |

Fig. 8. Comparison of Inf-Net and MSBC-Net to extract edge features. The first column on the left contains four consecutive MR images of a subject, the second to the fifth columns represent the output of each convolutional layer of the Inf-Net edge attention module, the first column on the right is boundary ground truth of RoI, and the second column on the right shows the boundary maps of RoIs extracted by MSBC-Net.

1) Effectiveness of MBSB: To verify the effectiveness of the proposed MBSB, we replace the MBSB in MSBC-Net with the mask head of Mask R-CNN; training and testing is conducted based on rectal cancer MRI data. The results clearly show in Table V, that the proposed novel MBSB is more effective and obtains a gain of 0.04 DSC. We verify the fusion effect of the mask head in Mask R-CNN and our BH, and also evaluate the fusion performance of MH and BH, demonstrating the effectiveness of BH.

2) Effectiveness of MDRF: To explore the contribution of the MDRF module, we replace the backbone network with the general ResNet50-FPN and conducted a comparative experiment. As shown in Table VI, the MDRF brings in a gain of 0.04 DSC by dynamically adjusting to the geometric deformation of the targets. This shows that the performance of the segmented tasks can be further improved by introducing the MDRF into the MSBC-Net.

3) Initialization of proposal box: We explore the effect of the initialization method of the proposal boxes on the model segmentation. As shown in Table VII, "Image" indicates that all the proposed boxes are initialized to the size of the entire image. "Random" indicates that the center, height, and width of the proposed boxes are randomly initialized with Gaussian distribution. "Zero" indicates that the center, height, and width of the proposed boxes are initialized to zero. The result shows that the performance of the model is optimal when all the proposed boxes are initialized in a "Zero" manner.

4) Effectiveness of number of proposal boxes: The effect of the number of proposal boxes on the MSBC-Net is analyzed in Table VIII, and the results show that increasing the number of proposal boxes can improve the performance of the model, but it is also limited by the performance of the GPU. Therefore, we only evaluate 100 proposal boxes and 300 proposal boxes for the model.

E. Future Work

Based on the excellent segmentation performance of the proposed framework, in the future, we will try to lightweight the MSBC-Net. Although the data and the analysis presented in this paper are preliminary and require further investigation with more samples is required, we will collect more rectal cancer data from multi-center and apply them to a 3D CNN-based segmentation model to further improve the segmentation performance and generalization performance of the model. Finally, our ultimate goal is to build a fully automated system to distinguish all the T stages of rectal cancer, including T1, T2, T3, and T4.
TABLE V
DISPLAY THE EFFECTIVENESS EVALUATION RESULTS OF THE MBSB (MEAN (95% CI))

| Mask R-CNN mask head | MBSB | DSC |
|----------------------|------|-----|
|                      | BH   | MH  | Total | Rectal Cancer | Rectal Wall |
| ✓                    | 0.758 (0.746, 0.770) | 0.794 (0.778, 0.811) | 0.722 (0.708, 0.736) |
| ✓                    | 0.781 (0.769, 0.793) | 0.825 (0.809, 0.840) | 0.737 (0.723, 0.750) |
| ✓                    | 0.772 (0.760, 0.784) | 0.813 (0.799, 0.827) | 0.731 (0.718, 0.744) |
| ✓                    | 0.801 (0.791, 0.811) | 0.847 (0.832, 0.860) | 0.756 (0.742, 0.768) |

TABLE VI
ABLATION ANALYSIS OF THE MDRF (MEAN(95% CI))

| Backbone | DSC |
|----------|-----|
|          | Total Rectal Cancer Rectal Wall |
| ResNet50-FPN | 0.769 (0.758, 0.781) | 0.821 (0.806, 0.837) | 0.717 (0.703, 0.731) |
| MDRF     | 0.801 (0.791, 0.811) | 0.847 (0.832, 0.860) | 0.756 (0.742, 0.768) |

TABLE VII
EFFECT OF INITIALIZATION OF PROPOSAL BOXES (MEAN(95% CI))

| Init. | DSC |
|-------|-----|
|       | Total Rectal Cancer Rectal Wall |
| Image | 0.792 (0.780, 0.802) | 0.839 (0.824, 0.854) | 0.744 (0.730, 0.757) |
| Random| 0.783 (0.771, 0.794) | 0.840 (0.825, 0.854) | 0.726 (0.711, 0.740) |
| Zero  | 0.801 (0.791, 0.811) | 0.847 (0.832, 0.860) | 0.756 (0.742, 0.768) |

V. CONCLUSION

In this paper, a novel multi-task learning network for medical image segmentation is proposed, which combines classification, detection and segmentation, and effectively improves the segmentation performance by joint training of multiple branches. The newly proposed BH is fused as the auxiliary information for mask segmentation. Additionally, extensive experiments on rectal cancer dataset demonstrate that the proposed MSBC-Net outperforms advanced segmentation models in the multi-class segmentation task of rectal cancer, and overcomes the limitations of the conventional end-to-end semantic segmentation models for rectal cancer segmentation. The proposed model presents significant potential in medical image segmentation, and it can be easily extended to other detection or segmentation tasks.

ACKNOWLEDGMENT

This work was supported by grants from the National Natural Science Foundation of China (81871403, 61702321, and U1936213), Key Research and Development Program of Zhejiang Province (2019C03014).

REFERENCES

[1] H. Sung, J. Ferlay, R. L. Siegel et al., “Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” CA: A Cancer Journal for Clinicians, vol. n/a, no. n/a.
[2] S. Wang, M. Zhou, Z. Liu et al., “Central Focused Convolutional Neural Networks: Developing a Data-driven Model for Lung Nodule Segmentation,” Medical Image Analysis, vol. 40, 06/01, 2017.
[3] J. Song, C. Yang, L. Fan et al., “Lung Lesion Extraction Using a Toboggan Based Growing Automatic Segmentation Approach,” IEEE Transactions on Medical Imaging, vol. 35, no. 1, pp. 337-353, 2016.
[4] S. P. Pawar, and S. N. Talbar, “LungSeg-Net: Lung field segmentation using generative adversarial network,” Biomedical Signal Processing and Control, vol. 64, pp. 102296, 2021/02/01, 2021.
[5] A. Makropoulos, I. S. Gousias, C. Lediger et al., “Automatic Whole Brain MRI Segmentation of the Developing Neonatal Brain,” IEEE Transactions on Medical Imaging, vol. 33, no. 9, pp. 1818-1831, 2014.
[6] S. Kumar, S. Conjeti, A. G. Roy et al., “InfNet: Fully convolutional networks for infant brain MRI segmentation.” pp. 145-148.
[7] B. H. Menze, A. Jakab, S. Bauer et al., “The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS),” IEEE Transactions on Medical Imaging, vol. 34, no. 10, pp. 1993-2024, 2015.
[8] M. Hammon, A. Cavallaro, M. Erdt et al., “Model-based pancreas segmentation in portal venous phase contrast-enhanced CT images,” Journal of digital imaging, vol. 26, no. 6, pp. 1082-1090, 2013.
[9] H. Roth, A. Farag, L. Lu et al., “Deep convolutional networks for pancreas segmentation in CT imaging.”
[10] J. Cai, L. Lu, Y. Xie et al., “Improving Deep Pancreas Segmentation in CT and MRI Images via Recurrent Neural Contextual Learning and Direct Loss Function,” ArXiv, vol. abs/1707.04912, 2017.
[11] B. Saha Tchinda, D. Tchiotsop, M. Noubom et al., “Retinal blood vessels segmentation using classical edge detection filters and the neural network,” Informatics in Medicine Unlocked, vol. 23, pp. 100521, 2021/01/01, 2021.
[12] Y. Yu, and H. Zhu, “Retinal vessel segmentation with constrained-based nonnegative matrix factorization and 3D modified attention U-Net,” EURASIP Journal on Image and Video Processing, vol. 2021, no. 1, pp. 6, 2021/01/28, 2021.
[13] Y. Ma, X. Li, X. Du et al., “Retinal Vessel Segmentation by Deep Residual Learning with Wide Activation,” Computational Intelligence and Neuroscience, vol. 2020, pp. 8822407, 2020/10/10, 2020.
