Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 8351725, 11 pages
https://doi.org/10.1155/2020/8351725

Research Article

An Improved Mask R-CNN Model for Multiorgan Segmentation

Jian-Hua Shu,[1] **Fu-Dong Nian**,[2,3] **Ming-Hui Yu**,[4] **and Xu Li**[5]

[1]*School of Medical Information Engineering, Anhui University of Chinese Medicine, Hefei 230012, China*
[2]*School of Advanced Manufacturing Engineering, Hefei University, Hefei 230601, China*
[3]*School of Computer Science and Technology, Anhui University, Hefei 230601, China*
[4]*School of Electrical Engineering and Automation, Anhui University, Hefei 230601, China*
[5]*Sino-German Institute of Applied Mathematics, Hefei University, Hefei 230601, China*

Correspondence should be addressed to Fu-Dong Nian; nianfd@hfuu.edu.cn

Received 8 June 2020; Accepted 1 July 2020; Published 24 July 2020

Guest Editor: Shaohui Wang

Medical image segmentation is a key topic in image processing and computer vision. Existing literature mainly focuses on single-organ segmentation. However, since maximizing the concentration of radiotherapy drugs in the target area with protecting the surrounding organs is essential for making effective radiotherapy plan, multiorgan segmentation has won more and more attention. An improved Mask R-CNN (region-based convolutional neural network) model is proposed for multiorgan segmentation to aid esophageal radiation treatment. Due to the fact that organ boundaries may be fuzzy and organ shapes are various, original Mask R-CNN works well on natural image segmentation while leaves something to be desired on the multiorgan segmentation task. Addressing it, the advantages of this method are threefold: (1) a ROI (region of interest) generation method is presented in the RPN (region proposal network) which is able to utilize multiscale semantic features. (2) A prebackground classification subnetwork is integrated to the original mask generation branch to improve the precision of multiorgan segmentation. (3) 4341 CT images of 44 patients are collected and annotated to evaluate the proposed method. Additionally, extensive experiments on the collected dataset demonstrate that the proposed method can segment the heart, right lung, left lung, planning target volume (PTV), and clinical target volume (CTV) accurately and efficiently. Specifically, less than 5% of the cases were missed detection or false detection on the test set, which shows a great potential for real clinical usage.

1. Introduction

Diagnostic imaging plays an important role in modern medicine. Computed tomography (CT), magnetic resonance imaging (MRI), and other imaging modalities provide important assistance for diagnosis and treatment planning. Take esophageal cancer as an example; esophageal cancer is a primary malignant tumor of the esophagus. At least 200,000 people suffer from esophageal cancer every year [1], and radiotherapy is one of the main treatments in China. However, treatment planning of radiotherapy is highly dependent on planning target volume (PTV) and accurate description of the organs at risk. The accuracy of organ countersegmentation determines the quality of dose planning optimization in radiotherapy and thus affects the success or failure of radiotherapy or the incidence of complications [2].

With the increasing scale and quantity of medical images, organ segmentation via manual delineation by the clinical experience of radiologists is inefficient [3]. And it is necessary to use computers for processing and analyzing the medical images automatically. With the development of computer vision technology, many different automatic image segmentation and delineation algorithms have been developed. These algorithms are called medical image segmentation or organ segmentation [4] in the literature.

Conventional medical image segmentation/organ segmentation algorithms can be roughly divided into eight categories [4]: (a) thresholding approaches, (b) region-growing approaches, (c) classifiers, (d) clustering approaches, (e) Markov random field models, (f) deformable models, (g) artificial neural networks, and (h) atlas-guided approaches. Although these methods have made some progress, the accuracy is not sufficient.

Benefit from the continuous progress of deep learning technology, medical image segmentation/organ segmentation is currently dominated by the CNN (convolutional neural network) [5]. Similar to the object detection method, CNN-based organ segmentation can also be divided into two types: (a) one-stage algorithm, which deems the organ segmentation as a one-stage pixel classification task. The typical structure is fully convolutional networks (FCNs) [6]; (b) two-stage algorithm, which decouples the organ segmentation into organ localization and instance segmentation stages. The typical structure is region CNN (R-CNN) [7]. The most well-known one-stage CNN architecture for organ segmentation is U-Net, published by Ronneberger et al. [8]. Most state-of-the-art organ segmentation methods are the invariants of U-Net [9–11]. Although they have achieved encouraging performance, two shortcomings exist. On the one hand, many literature studies focus on single-organ segmentation, while only few works are made effort to address the multiorgan segmentation problem [12, 13]. On the other hand, two-stage segmentation methods work well for multiobject segmentation on the natural image segmentation dataset [14] but worse than the one-stage algorithm on medical image segmentation [15]. Therefore, mining the potential of the two-stage multiorgan segmentation algorithm has great research value.

In this paper, to address the shortcomings mentioned above, we present an improved Mask R-CNN framework for multiorgan segmentation. Original Mask R-CNN [16] is presented to address the multi-instance segmentation problem on the natural image. Although the original Mask R-CNN has achieved state-of-the-art instance segmentation performance on general image datasets, the latest research [15] shows that it is able to accurately find bounding boxes for organs, while its performance on segmentation is worse than U-Net on the medical image segmentation dataset. We think a major reason for this is that the semantic representation obtained from the original Mask R-CNN framework is too rough for organ segmentation because organ boundaries may blur and organ shapes are various. To address it, we have made two improvements to the original Mask R-CNN: (a) a ROI (region of interest) generation method is presented in the RPN which is able to utilize multiscale semantic features; (b) a prebackground classification subnetwork is integrated to improve the precision of multiorgan segmentation. Moreover, CT images of 44 esophageal cancer patients are collected and annotated as benchmark to evaluate the proposed method.

To sum up, our contributions are as follows:

(1) We applied the Mask R-CNN to esophageal cancer medical image processing successfully. Most existing methods focus on single-organ segmentation, while this paper devotes to address the multiorgan segmentation problem.

(2) To provide a better multiorgan segmentation model, we propose two improvements compared with the original Mask R-CNN framework.

(3) We conduct extensive experiments and analysis on the collected real multiorgan dataset and demonstrate the excellent performance of our proposed method on the multiorgan segmentation task.

The rest of this paper is organized as follows. Section 2 reviews and discusses the related works. Section 3 describes the proposed improved Mask R-CNN model in detail. Experimental results and comparisons are discussed in Section 4, and conclusions with the future work are described in Section 5.

2. Related Work

Pham et al. [4] and Litjens et al. [5] reviewed the conventional and deep learning-based organ segmentation methods, respectively. In this section, we briefly review the previous methods which are most related to our work including the conventional medical image segmentation method, deep learning-based single-organ segmentation method, and deep learning-based multiorgan segmentation method.

2.1. Conventional Medical Image Segmentation Method. Conventional medical image segmentation method can be roughly divided into eight categories: (a) thresholding approaches [17]: thresholding approaches first attempt to determine an intensity value (threshold), then group all pixels with intensity greater than the threshold into one class, and all other pixels into another class. (b) Region-growing approaches [18]: region-growing approaches utilize intensity information and/or edges in the medical image to predefine criteria for extracting a region of the image that is connected. (c) Classifiers [19, 20]: classifier methods convert the medical image from the image space to the feature space first and then train classifiers on the feature space to distinguish which class of the pixel they belong to. (d) Clustering approaches [21]: commonly used clustering approaches for medical image segmentation are K-means, fuzzy c-means, and expectation-maximization. Compared with the classifiers, the clustering approaches are unsupervised approaches. (e) Markov random field models: Markov random field (MRF) is a statistical model which can be used within segmentation methods by modeling model spatial interactions between neighboring or nearby pixels. (f) Deformable models: deformable models use closed parametric curves or surfaces to delineate region boundaries. (g) Artificial neural networks (ANNs) [22]: the most widely applied use of the ANN in conventional medical image processing is as a classifier. (h) Atlas-guided approaches [23, 24]: the atlas is generated by compiling information on the anatomy that requires segmenting. This atlas is then used as a reference frame for segmenting new images. In addition, level set optimization is also utilized for multiorgan segmentation [25]. Though the methods mentioned above have achieved some progress, the accuracy of organ segmentation is not too high because all conventional methods depend on manual feature representation.

2.2. Deep Learning-Based Single-Organ Segmentation Method. Ronneberger et al. [8] first presented a novel CNN architecture (U-Net) and became the most popular structure in medical image analysis. The main novelty in U-Net is the combination of an equal amount of upsampling and downsampling layers. Inspired by U-Net, Zhou et al. [26] presented U-Net++, a more powerful architecture for medical image segmentation. Milletari et al. [27] proposed V-Net (a 3D variant of U-Net architecture) performing 3D image segmentation using 3D convolutional layers with an objective function directly based on the Dice coefficient. Drozdzal et al. [11] investigated the use of short ResNet-like skip connections in addition to the long skip connections in a regular U-Net. Besides CNN, Xie et al. [28], Stollenga et al. [29], Chen et al. [30], and Poudel et al. [31] utilized the recurrent neural network (RNN) for organ segmentation tasks. To combat spurious responses, few papers attempt to combine the CNN/RNN with graphical models like MRFs [32] and conditional random fields (CRFs) [33] to refine the segmentation output. Although these methods have achieved encouraging performance, they were presented to address the single-organ segmentation problem, which may not be suitable/optimal for multiorgan segmentation (It is difficult to segment multiple organs at the same time, which damages the clinical auxiliary effect.).

2.3. Deep Learning-Based Multiorgan Segmentation Method. The research on the deep learning-based multiorgan segmentation method is in its early phase. Tong et al. [34] introduced discriminative dictionary learning for abdominal multiorgan segmentation. Lay et al. [35] used context integration and discriminative models for rapid multiorgan segmentation. Roth et al. [36] and Chen et al. [37] adopted the 3D fully convolutional network. Recently, Dong et al. [38] presented a generative model (U-Net-GAN), and Wang et al. [39] proposed densely connected U-Net for multiorgan segmentation. Lei et al. [40] presented a review of deep learning in multiorgan segmentation. Different from these methods, the proposed method in this paper aims to improve the two-stage instance segmentation algorithm which is widely used in the natural image dataset, making it suitable for the multiorgan segmentation task.

3. Methods

In this section, we introduce the proposed method (which is named improved Mask R-CNN) for multiorgan segmentation. As shown in Figure 1, the proposed method is based on the existing well-known multi-instance segmentation method, Mask R-CNN. Compared with the original Mask R-CNN, we have made two improvements: (a) a ROI (region of interest) generation method is presented in the RPN which is able to utilize multiscale semantic features; (b) a prebackground classification subnetwork is integrated to improve the precision of multiorgan localization. The detailed proposed approach is presented in two sections: (a) the network structure and (b) loss function.

3.1. The Network Structure. The network of the proposed algorithm can be mainly divided into three modules. The first module is called feature extraction and ROI generation, which is mainly composed of ResNet50 + FPN + RPN. In this module, we generate multilayer feature maps first. Then, each point on the feature map is mapped into the original image to acquire the corresponding ROI.

The second module is named region of interest alignment, which pools the ROIs obtained from the first module to a fixed size and avoids quantization error. The third module is mask acquisition. In this module, the fixed-size ROIs obtained from the second module are sent to the organ region segmentation network for generating organ mask. And at the same time, they are also sent to the fully connected layer for organ-position rectangular bounding box regression and organ classification. The above three modules are detailed as follows.

3.1.1. Feature Extraction and ROI Generation. The purpose of this step is to extract the features of the input image and generate the ROI in the corresponding feature layer. First, a medical CT image containing multiple organs is input to the ResNet50 network. Res2, Res3, Res4, and Res5 are the feature output layers of the ResNet [15, 41], respectively. Then, feature pyramid network (FPN) [42] is adopted to fuse these multilayer features to obtain strong semantic information and improve the accuracy of organ detection. As shown in Figure 2, the specific approach is to conduct dimensionality reduction operation on the features above Res4 (that is, to add a layer of $1 * 1$ convolution layer) and upsampling operation on the features above P5 to make them have the same size. Then, addition operation (adding corresponding elements) is performed on the processed P5 and the processed Res4 to output the obtained results to P4, P2, P3, and so on. Then, the RPN network is used to predict in different output layers, P2, P3, P4, and P5, to obtain ROIs.

3.1.2. Region of Interest Alignment. This step aims to pool all ROIs remaining on the feature maps to a fixed size. Since the ROI position is usually obtained by the regression model, it is generally a floating-point number, while the pooled feature map requires a fixed size. In order to avoid quantization errors, the ROI align [15] (illustrated in Figure 3) layer is adopted. In the presented framework, we use the ROI align layer to traverse each ROI first, keeping the floating-point number boundary unquantized. Then, the ROI is divided into $k \times k$ cells with the boundary of each cell not quantized. Then, the fixed four coordinate positions are calculated in each cell, the values of these four positions are calculated by bilinear interpolation, and the max-pooling operation is carried out finally. Through the above operations, the fixed size ROI can be obtained with no quantization error.

In the original Mask R-CNN segmentation algorithm, the ROI obtained by the RPN network is aligned to extract the ROI features. In this step, each ROI is aligned by a single-layer (single-scale) feature. In the presented method, as shown in Figure 4, we replace the single-layer features with multilayer features, that is to say, each ROI needs to do ROI

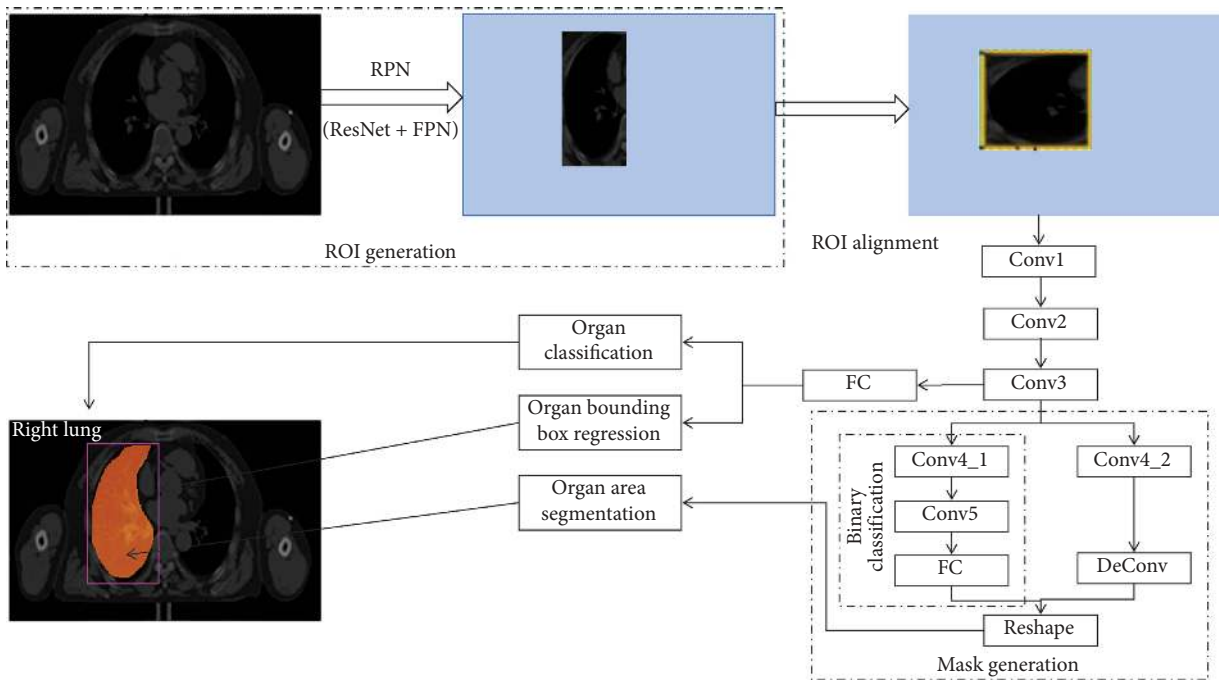


Figure 1: The framework of the proposed method.

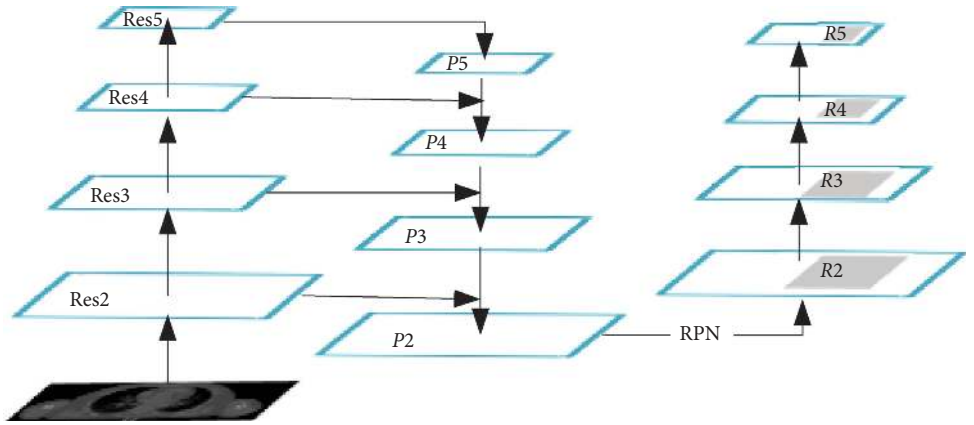


Figure 2: From left to right are the ResNet50 network, the feature pyramid network (FPN), and the region proposal network (RPN).

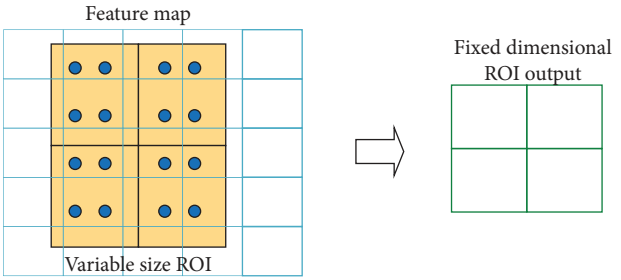


Figure 3: The illustration of ROI alignment operation. The blue-dotted box represents the feature map obtained after convolution, and the black solid box represents the ROI feature.

alignment operation with multilayer features, and then the ROI features of different layers will be fused together so that each ROI feature will have multilayer features.

3.1.3. Mask Acquisition. The goal of this step is to get the multiorgan segmentation result. ROI of pooling to a fixed size was sent to the fully connected layer for organ

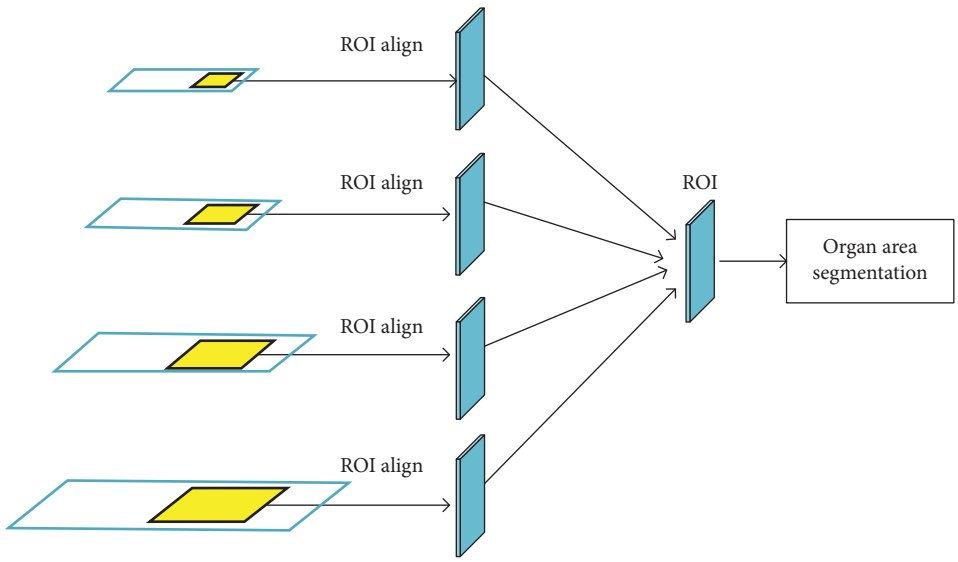


Figure 4: The multilayer ROI shall be merged after the operation of ROI alignment and then sent to the organ area segmentation network.

classification (6 categories including background) and organ-position rectangular bounding box regression. Meanwhile, ROI of pooling to a fixed size was also sent to a mask generation branch (i.e., fully convolution neural network operation in each ROI). Organ area segmentation is a parallel branch to organ classification and organ-position rectangular bounding box regression. As shown in Figure 5(a), the branch consists of four consecutive convolution layers and a deconvolution layer (with 2 times of upsampling). The kernel size and channels of each convolution layer are $3 * 3$ and 25, respectively. A binary classification branch is added to distinguish foreground and background before the original mask branch (illustrated in Figure 5(b)). The new branch contains two $3 * 3$ convolution layers and a fully connected layer. The dimension of the output of the new branch is the same as the original branch via a reshape operation. The output mask of these two branches was fused to get the final multiorgan segmentation result.

3.2. Loss Function. In terms of loss function, a third loss function, which is used to generate mask, is added on the basis of Fast R-CNN [43] so that the total loss function of our improved Mask R-CNN framework is

$$\mathscr{L} = \mathscr{L}_{\text{cls}} + \mathscr{L}_{\text{box}} + \mathscr{L}_{\text{mask}}. \tag{1}$$

Here, the classification and regression losses are defined as $\mathscr{L}_{\text{cls}}$ and $\mathscr{L}_{\text{box}}$, respectively:

$$\mathscr{L}_{\text{cls}} = -\log \quad P_u, \tag{2}$$

$$\mathscr{L}_{\text{box}} = \sum_{i=1}^{4} \text{Smooth}_{\text{L1}}\left(t_{\text{i}}^{\text{u}} - v_{\text{i}}\right), \tag{3}$$

$$\text{Smooth}_{\text{L1}}(X) = \begin{cases} 0.5x^2, & |x| < 1, \\ |x| - 0.5, & \text{otherwise.} \end{cases} \tag{4}$$

P is a $(k+1)$-dimensional vector representing the probability of a pixel belonging to the k class or background. For each ROI, $P = (P_0, P_1, \ldots, P_k)$, and P_u represents the probability corresponding to class u. $t_u = (t_x^u, t_y^u, t_w^u, t_h^u)$ represents the predicted translation scaling parameter of class u, t_x^u, t_y^u refer to the translation with the same scale as the object proposal, and t_w^u, t_h^u refer to the height and width of the logarithmic space relative to the object proposal. t_1, t_2, t_3, and t_4 in equation (3) represent t_x, t_y, t_w, and t_h, respectively. Moreover, v_i represents the corresponding parameter of the ground-truth bounding box.

Note that the smooth $L1$ loss is utilized in equation (3); the reasons are twofold: (a) compared with the widely used $L2$ loss, smooth $L1$ loss is robust for outlier points. (2) Many famous object detection frameworks use smooth $L1$ loss, e.g., Faster-RCNN and Mask R-CNN. We utilize the same bounding loss function which can guarantee the fairness of algorithm comparison. Of course, some box regression loss functions which have been proposed recently (e.g., GIoU, DIoU, and CIoU) are also compatible with the proposed framework.

$\mathscr{L}_{\text{mask}}$ in equation (1) is the mask loss of the newly added background segmentation branch (as described in Section 3.1.3). In our improved Mask R-CNN framework, the output dimension of each ROI is $K * m * m$ for the newly added mask branch, where $m * m$ represents the size of the mask and K represents categories, so a total of K-binary masks were generated in here. After the predicted mask was obtained, the value of the sigmoid function was calculated for each pixel of the mask, and the obtained result was taken as one of the inputs of $\mathscr{L}_{\text{mask}}$ (cross-entropy loss function). It should be noted that only positive sample ROI is used to calculate $\mathscr{L}_{\text{mask}}$. The definition of the positive sample is the same as that of general object detection algorithms, and IOU greater than 0.5 is defined as the positive sample. In fact, $\mathscr{L}_{\text{mask}}$ is very similar to $\mathscr{L}_{\text{cls}}$ except that the former is calculated on the basis of pixels and the latter on the basis of

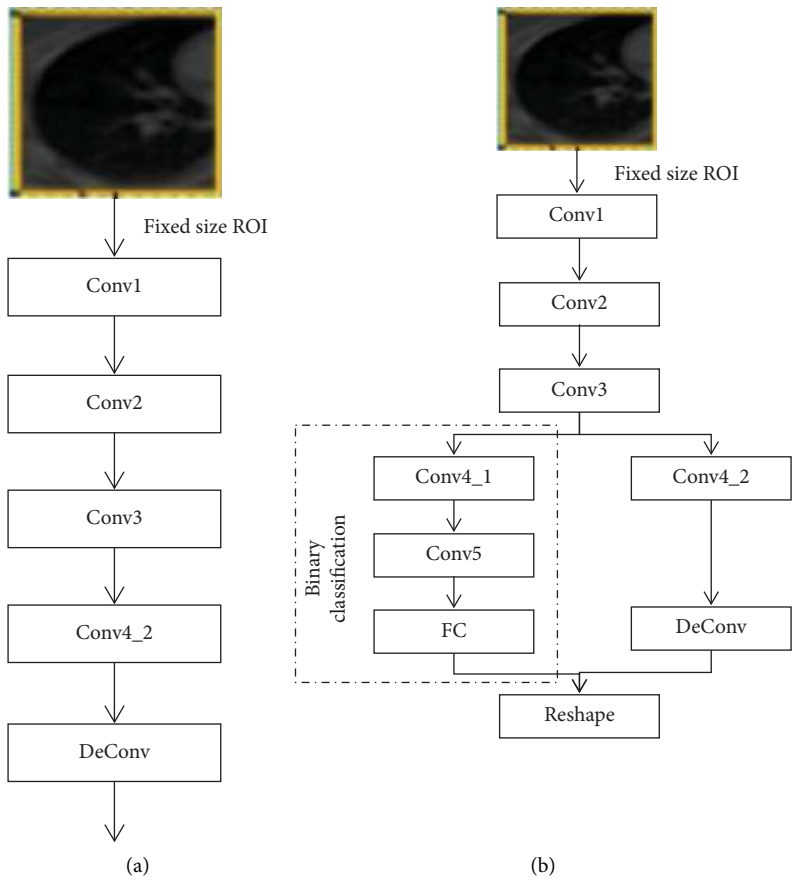


Figure 5: A new binary classification branch is added to the original mask generation branch. (a) The original mask generation branch. (b) The proposed mask generation branch.

images, so it is similar to $\mathscr{L}_{\text{cls}}$ in that although K masks are given here, only the one corresponding to the ground truth is valid in calculating the cross-entropy loss function. A mask contains multiple pixels, so here, $\mathscr{L}_{\text{mask}}$ is the average of the cross-entropy loss of each pixel:

$$\mathscr{L}_{\text{mask}} = -\frac{1}{K}\sum_{i=1}^{K}\sum_{j=1}^{m*m}\left(\log P_{i,j}^{M}\right). \tag{5}$$

Here, $P_{i,j}^{M}$ is the j-th pixel of the i-th generated mask.

4. Experiments

In this section, we conduct extensive experiments to evaluate the proposed improved Mask R-CNN multiorgan segmentation framework. We first introduce the collected and annotated dataset in Section 4.1 followed by the evaluation criteria in Section 4.2. Then, Section 4.3 describes the implementation details. Finally, we discuss the comparison with state-of-the-art methods in Section 4.4.

4.1. Dataset. The utilized multiorgan segmentation dataset consists of all the slice information of 44 esophageal cancer patients, with a total of 4341 CT images. Each image was labeled with five areas (heart, right lung, left lung, PTV, and CTV) by the doctor. We use 80% of these CT images as the training set, 5% as the validation set, and the remaining 15% as the test set.

4.2. Evaluation Criteria. There are many evaluation criteria which are proposed to evaluate the image segmentation results, e.g., region overlap and boundary similarity [44]. Here, we select Dice coefficient (DICE) [45] and Jaccard index (JAC) [46] as criteria to evaluate the overlap between the prediction and the ground-truth organ regions. Suppose that x and y are the organ regions of the prediction and the ground truth, respectively; JAC and DICE are calculated as follows:

$$\begin{aligned}\text{JAC} &= \frac{|X \cap Y|}{|X \cup Y|},\\ \text{DICE} &= \frac{2|X \cap Y|}{|X| + |Y|}.\end{aligned} \tag{6}$$

4.3. Implementation Details. We implement our improved Mask R-CNN model based on the framework of PyTorch. The backbone is the adjusted ResNet50 which is detailed in Section 3.1.1. We use the stochastic gradient descent (SGD) optimizer with the learning rate set to 0.01 initially, and the batch size is set to 64. The maximum number of iterations is

Table 1: Comparisons of the proposed and state-of-the-art multiorgan segmentation methods on the presented dataset.

Methods	Organ	JAC (%)	DICE (%)
Linguraru et al. [47]	Heart	74.9 ± 4.7	71.3 ± 3.1
	Right lung	86.5 ± 2.2	86.3 ± 1.2
	Left lung	85.1 ± 1.3	84.6 ± 0.9
	PTV	82.5 ± 1.6	81.3 ± 2.8
	CTV	80.5 ± 1.2	77.5 ± 1.9
He et al. [48]	Heart	87.5 ± 1.2	86.3 ± 0.7
	Right lung	89.3 ± 1.6	87.3 ± 2.4
	Left lung	90.2 ± 1.9	88.3 ± 1.1
	PTV	86.3 ± 1.7	84.1 ± 2.4
	CTV	85.6 ± 2.3	83.7 ± 3.1
Gauriau et al. [49]	Heart	88.8 ± 0.6	87.5 ± 1.2
	Right lung	91.5 ± 0.9	91.3 ± 0.8
	Left lung	90.3 ± 0.9	90.8 ± 1.5
	PTV	89.1 ± 1.4	86.2 ± 1.7
	CTV	88.7 ± 1.9	87.3 ± 1.5
Original mask R-CNN	Heart	95.1 ± 0.5	94.2 ± 0.7
	Right lung	97.0 ± 0.9	96.2 ± 1.2
	Left lung	96.3 ± 0.4	95.1 ± 0.6
	PTV	94.7 ± 1.1	93.2 ± 0.9
	CTV	94.3 ± 0.5	93.7 ± 0.8
Improved mask R-CNN (ours)	Heart	96.6 ± 1.3	95.1 ± 1.2
	Right lung	98.1 ± 0.5	97.8 ± 0.3
	Left lung	97.6 ± 0.4	96.2 ± 1.4
	PTV	95.3 ± 1.5	95.2 ± 0.7
	CTV	95.8 ± 0.7	94.4 ± 1.2

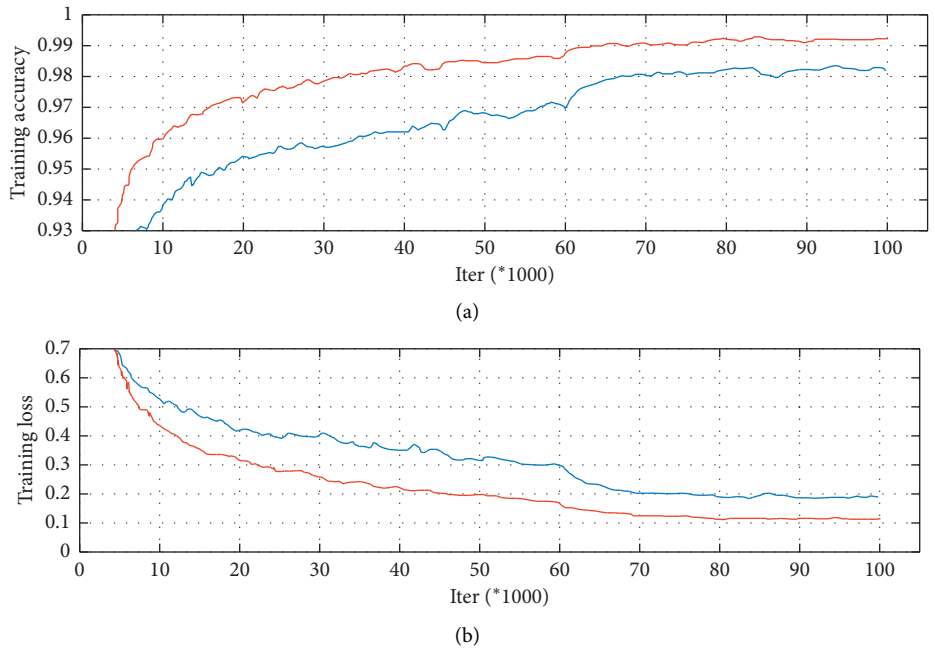


Figure 6: (a) The accuracy curves in the training stage. (b) The loss curves in the training stage. Red curve represents the proposed improved Mask R-CNN model, and blue represents the original Mask R-CNN model.

set to 100,000. When the number of iterations reached 50,000 and 80,000, the learning rate is reduced 10 times. All images are resized to 800 × 1000. The weight decay is set to 0.0001, and the momentum is set to 0.9 for all convolution and fully connected layers. It should be noted that all parameters in the proposed model are trained from scratch.

4.4. Results and Discussion

4.4.1. Quantitative Evaluation with State-of-the-Art Methods. We compare our proposed methods against the current widely used multiorgan segmentation models (Linguraru et al. [47], He et al. [48], and Gauriau et al. [49]), and the

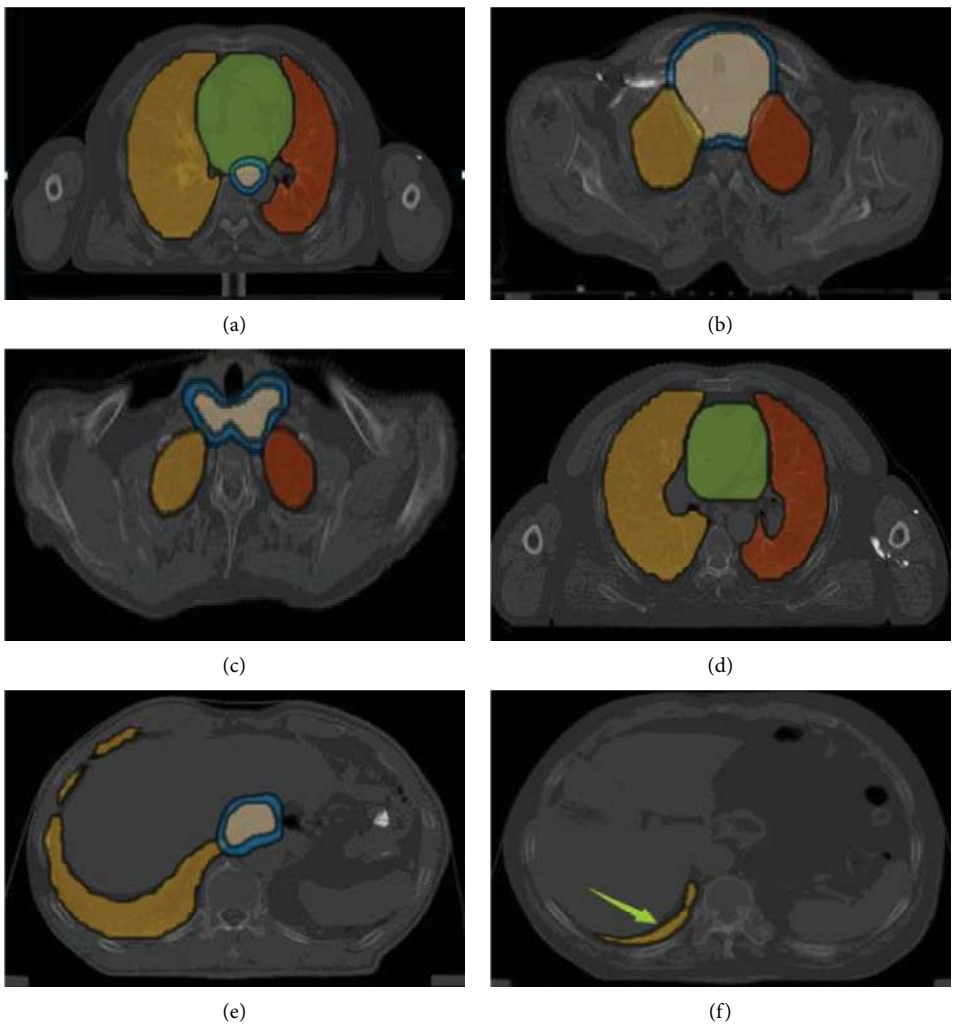
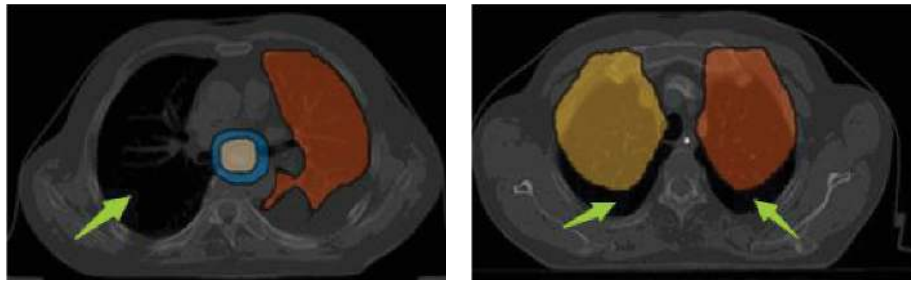


Figure 7: The visualization results. These six images are slices of the 38th, 60th, 66th, 71st, 89th, and 103rd layers of a patient with esophageal cancer (yellow represents the right lung, brown represents the left lung, cyan represents the heart, blue represents PTV, and gray represents CTV). (a) The slice of the 38th layer of a patient. (b) The slice of the 60th layer of a patient. (c) The slice of the 66th layer of a patient. (d) The slice of the 71st layer of a patient. (e) The slice of the 89th layer of a patient. (f) The slice of the 103rd layer of a patient.

comparison results are shown in Table 1. In general, we can observe that the proposed improved Mask R-CNN framework achieved the best performance. Moreover, Figure 6 shows the accuracy (JAC) and loss curves of the improved Mask R-CNN and original Mask R-CNN framework in the training stage. From Table 1 and Figure 6, we can conclude that the presented technique is able to improve the multiorgan segmentation performance of the original Mask R-CNN significantly and steadily.

4.4.2. Qualitative Evaluation. To illustrate the effectiveness of our method more visually, some multiorgan segmentation results are shown in Figure 7. The image we selected is distributed between 35 and 100 slices basically because in this range, each slice contains five organ regions that we need basically, and the information of each organ region is relatively rich. We found that the area of some organs from the 60th to 80th layers of patients is very small, which is difficult to be observed by the naked eye due to the perspective. However, our improved mask R-CNN algorithm can also achieve good results (as shown in Figure 4, especially the area indicated by the arrow in the figure may be difficult for doctors to annotate).

Although the proposed method can achieve encouraging performance, there are still some shortcomings. Examples of false detection and missed detection segmentation are shown in Figure 8. After analyzing all failure results, we find that that the missed detection was mainly concentrated in the slices from the 1st to the 35th layer of the patient, while the missed detection was mainly concentrated in the slices from the 110th to the



(a) (b)

Figure 8: Failure cases. False detection and missed detection segmentation results. Left image is the segmentation result of the 10th slice of a patient, missed detection occurred on the right lung. Right image is the segmentation result of the 115th slice of a patient. The left lung and the right lung areas may be segmented partially (yellow represents the right lung, brown represents the left lung, cyan represents the heart, blue represents PTV, and gray represents CTV). (a) Missed detection. (b) False detection.

130th layer. By observing the constructed dataset, we find that the amount of data of the slice near the front and the slice near the back is relatively small, that is, the slice near the front layer contains relatively less target organ area, so the doctor's label information in these parts is less. Therefore, we believe the major reason for these failure cases is due to the fact that training data are insufficient and unbalanced.

5. Conclusion

In this paper, we present the improved Mask R-CNN segmentation framework for the medical domain that is able to work well on the multiorgan segmentation task. The proposed improved Mask R-CNN framework builds around the original Mask R-CNN framework [15]. Compared with the original Mask R-CNN framework, there are two major improvements on the improved Mask R-CNN: (a) a ROI (region of interest) generation method is presented in the RPN (region proposal network) which is able to utilize multiscale semantic features; (b) a prebackground classification subnetwork is integrated to the original mask generation branch to improve the precision of multiorgan segmentation. Additionally, extensive experiments on the collected and annotated esophageal cancer dataset demonstrate the effectiveness of the proposed framework, i.e., the improved Mask R-CNN framework can segment the heart, right lung, left lung, PTV, and CTV accurately and simultaneously. Since it is time consuming and laborious to label medical images, we will investigate semi-supervised and weakly supervised multiorgan segmentation techniques in the future.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation (NSF) of China (nos. 61702001 and 61902104), the Key Project of the Natural Science Foundation of Anhui University of Traditional Chinese Medicine (2019zrzd10), the Scientific Research Development Foundation of Hefei University (no. 19ZR15ZDA), the Talent Research Foundation of Hefei University (no. 18-19RC54) and Hefei University Annual Academy Research Development Fund Project (Natural Science) (No. 18ZR12ZDA).

References

[1] R. S. Holmes and T. L. Vaughan, "Epidemiology and pathogenesis of esophageal cancer," in *Seminars in Radiation Oncology*, vol. 17, pp. 2–9, Elsevier, Amsterdam, Netherlands, 2007.

[2] Y.-l. Peng, L. Chen, G.-z. Shen et al., "Interobserver variations in the delineation of target volumes and organs at risk and their impact on dose distribution in intensity-modulated radiation therapy for nasopharyngeal carcinoma," *Oral Oncology*, vol. 82, pp. 1–7, 2018.

[3] M. Feng, C. Demiroz, K. A. Vineberg, A. Eisbruch, and J. M. Balter, "Normal tissue anatomy for oropharyngeal cancer: contouring variability and its impact on optimization," *International Journal of Radiation Oncology ∗ Biology ∗ Physics*, vol. 84, no. 2, pp. e245–e249, 2012.

[4] D. Pham, C. Xu, and J. Price, "A survey of current methods in medical image segmentation," *Annual Review of Biomedical Engineering*, vol. 2, 2000.

[5] G. Litjens, T. Kooi, B. E. Bejnordi et al., "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[6] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, Boston, MA, USA, June 2015.

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, Columbus, OH, USA, June 2014.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in

Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, Munich, Germany, pp. 234–241, October 2015.

[9] F. Isensee, J. Petersen, A. Klein et al., "nnu-net: selfadapting framework for u-net-based medical image segmentation," http://arxiv.org/abs/1809.10486.

[10] O. Cicek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: learning dense volumetric segmentation from sparse annotation," in *Ternational Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 424–432, Springer, Berlin, Germany, 2016.

[11] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The importance of skip connections in biomedical image segmentation," in *Deep Learning and Data Labeling for Medical Applications*, pp. 179–187, Springer, Berlin, Germany, 2016.

[12] E. Gibson, F. Giganti, Y. Hu et al., "Automatic multi-organ segmentation on abdominal ct with dense v-networks," *IEEE Transactions on Medical Imaging*, vol. 37, no. 8, pp. 1822–1834, 2018.

[13] M. P. Heinrich, O. Oktay, and N. Bouteldja, "Obelisk-net: fewer layers to solve 3d multi-organ segmentation with sparse deformable convolutions," *Medical Image Analysis*, vol. 54, pp. 1–9, 2019.

[14] X. Chen, H. Fang, T.-Y. Lin et al., "Microsoft coco captions: data collection and evaluation server," http://arxiv.org/abs/1504.00325.

[15] A. O. Vuola, S. U. Akram, and J. Kannala, "Mask-rcnn and u-net ensemble for nuclei segmentation," in *Proceedings of the IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pp. 208–212, Venice, Italy, April 2019.

[16] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2961–2969, Venice, Italy, October 2017.

[17] P. K. Sahoo, S. Soltani, and A. K. C. Wong, "A survey of thresholding techniques," *Computer Vision, Graphics, and Image Processing*, vol. 41, no. 2, pp. 233–260, 1988.

[18] P. Gibbs, D. L. Buckley, S. J. Blackband, and A. Horsman, "Tumour volume determination from mr images by morphological segmentation," *Physics in Medicine and Biology*, vol. 41, no. 11, pp. 2437–2446, 1996.

[19] T. Kapur, W. Eric, L. Grimson, R. Kikinis, and W. M. Wells, "Enhanced spatial priors for segmentation of magnetic resonance imagery," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, Cambridge, MA, USA, pp. 457–468, October 1998.

[20] W. M. Wells, W. E. L. Grimson, R. Kikinis, and F. A. Jolesz, "Adaptive segmentation of mri data," *IEEE Transactions on Medical Imaging*, vol. 15, no. 4, pp. 429–442, 1996.

[21] K.-S. Chuang, H.-L. Tzeng, S. Chen, J. Wu, and T.-J. Chen, "Fuzzy c-means clustering with spatial information for image segmentation," *Computerized Medical Imaging and Graphics*, vol. 30, no. 1, pp. 9–15, 2006.

[22] L. O. Hall, A. M. Bensaid, L. P. Clarke, R. P. Velthuizen, M. S. Silbiger, and J. C. Bezdek, "A comparison of neural network and fuzzy clustering techniques in segmenting magnetic resonance images of the brain," *IEEE Transactions on Neural Networks*, vol. 3, no. 5, pp. 672–682, 1992.

[23] T. Okada, K. Yokota, M. Hori, M. Nakamoto, H. Nakamura, and Y. Sato, "Construction of hierarchical multi-organ statistical atlases and their application to multi-organ segmentation from ct images," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, New York, NY, USA, pp. 502–509, September 2008.

[24] R. Wolz, C. Chu, K. Misawa, M. Fujiwara, K. Mori, and D. Rueckert, "Automated abdominal multi-organ segmentation with subject-specific atlas generation," *IEEE Transactions on Medical Imaging*, vol. 32, no. 9, pp. 1723–1730, 2013.

[25] T. Kohlberger, M. Sofka, J. Zhang et al., "Automatic multi-organ segmentation using learning based segmentation and level set optimization," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, Toronto, Canada, pp. 338–345, September 2011.

[26] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: a nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 3–11, Springer, Berlin, Germany, 2018.

[27] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV)*, IEEE, Stanford, CA, USA, pp. 565–571, October 2016.

[28] Y. Xie, Z. Zhang, M. Sapkota, and L. Yang, "Spatial clockwork recurrent neural network for muscle perimysium segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, Athens, Greece, pp. 185–193, October 2016.

[29] M. F. Stollenga, W. Byeon, M. Liwicki, and J. Schmidhuber, "Parallel multidimensional lstm, with application to fast biomedical volumetric image segmentation," in *Advances in Neural Information Processing Systems*, pp. 2998–3006, MIT Press, Cambridge, MA, USA, 2015.

[30] J. Chen, L. Yang, Y. Zhang, M. Alber, and D. Z. Chen, "Combining fully convolutional and recurrent neural networks for 3d biomedical image segmentation," in *Advances in Neural Information Processing Systems*, pp. 3036–3044, MIT Press, Cambridge, MA, USA, 2016.

[31] R. P. Poudel, P. Lamata, and G. Montana, "Recurrent fully convolutional neural networks for multi-slice mri cardiac segmentation," in *Reconstruction, Segmentation, and Analysis of Medical Images*, pp. 83–94, Springer, Berlin, Germany, 2016.

[32] M. Shakeri, S. Tsogkas, E. Ferrante et al., "Sub-cortical brain structure segmentation using f-cnn's," in *Proceedings of the IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, IEEE, Prague, Czech Republic, pp. 269–272, April 2016.

[33] P. F. Christ, M. E. A. Elshaer, F. Ettlinger et al., "Automatic liver and lesion segmentation in ct using cascaded fully convolutional neural networks and 3d conditional random fields," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, Athens, Greece, pp. 415–423, October 2016.

[34] T. Tong, R. Wolz, Z. Wang et al., "Discriminative dictionary learning for abdominal multi-organ segmentation," *Medical Image Analysis*, vol. 23, no. 1, pp. 92–104, 2015.

[35] N. Lay, N. Birkbeck, J. Zhang, and S. K. Zhou, "Rapid multi-organ segmentation using context integration and discriminative models," in *Proceedings of the International Conference on Information Processing in Medical Imaging*, Springer, Asilomar, CA, USA, pp. 450–462, June 2013.

[36] H. R. Roth, H. Oda, Y. Hayashi et al., "Hierarchical 3D fully convolutional networks for multi-organ segmentation," http://arxiv.org/abs/1704.06382.

[37] S. Chen, H. Roth, S. Dorn et al., "Towards automatic abdominal multi-organ segmentation in dual energy ct using cascaded 3D fully convolutional network," http://arxiv.org/abs/1710.05379.
[38] X. Dong, Y. Lei, T. Wang et al., "Automatic multiorgan segmentation in thoraxCTimages using U-net-GAN," *Medical Physics*, vol. 46, no. 5, pp. 2157–2168, 2019.
[39] Z.-H. Wang, Z. Liu, Y.-Q. Song, and Y. Zhu, "Densely connected deep u-net for abdominal multi-organ segmentation," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, IEEE, Taipei, Taiwan, pp. 1415–1419, September 2019.
[40] Y. Lei, Y. Fu, T. Wang et al., "Deep learning in multi-organ segmentation," http://arxiv.org/abs/2001.10619.
[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, NV, USA, June 2016.
[42] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2117–2125, Honolulu, HI, USA, July 2017.
[43] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, pp. 91–99, MIT Press, Cambridge, MA, USA, 2015.
[44] A. A. Taha and A. Hanbury, "Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool," *BMC Medical Imaging*, vol. 15, no. 1, p. 29, 2015.
[45] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.
[46] P. Jaccard, "The distribution of the flora in the alpine Zone.1," *New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912.
[47] M. G. Linguraru, J. A. Pura, V. Pamulapati, and R. M. Summers, "Statistical 4d graphs for multi-organ abdominal segmentation from multiphase ct," *Medical Image Analysis*, vol. 16, no. 4, pp. 904–914, 2012.
[48] B. He, C. Huang, and F. Jia, "Fully automatic multi-organ segmentation based on multi-boost learning and statistical shape model search," in *VISCERAL Challenge,*, pp. 18–21, ISBI, New York, NY, USA, 2015.
[49] R. Gauriau, R. Cuingnet, D. Lesage, and I. Bloch, "Multi-organ localization with cascaded global-to-local regression and shape prior," *Medical Image Analysis*, vol. 23, no. 1, pp. 70–83, 2015.