Context Guided for Rectal Tumor Segmentation in Transrectal Ultrasound

Yanzhou Su[1]

[1] University of Electronic Science and Technology of China, Chengdu, P.R. China
chengjian@uestc.edu.cn
[2] the Affiliated Cancer Hospital, School of Medicine,
University of Electronic Science and Technology of China, Chengdu, P.R. China
graceof@163.com

Abstract. Colorectal cancer is the third most common malignancy in the world. Automatic rectal tumor segmentation in transrectal ultrasound (TRUS) images is of essential importance for rectum diagnostic and treatment planning. At present, the method based deep learning has been widely used in medical image segmentation. Most of the previous methods concatenated/summed every location of the Encoder and Decoder features equally, which could be disadvantage for the rectal tumor segmentation when only a subset of locations actually matters at a semantic level. In this paper, we leverage the global contextual information as a guidance to learn discriminative feature representation that is shared across relevant locations. Specifically, we further proposed a novel module, named Context Guided Module, to exploit the relevant contextual information between the global contextual information and feature vectors corresponding to every location in the course of feature aggregation, which can reduce the influence of unrelated areas and improve the performance of rectal tumor segmentation. To the best of our knowledge, our work is the first time to use deep learning for rectal tumor segmentation, so we evaluate the proposed method on the challenging transrectal ultrasound dataset we collected. The experimental results demonstrate that our proposed method outperforms current state-of-the-art medical image segmentation methods in rectal tumor segmentation.

1 Introduction

Rectal cancer is one of the common malignant tumors of the digestive tract. The 2019 US cancer data[8] demonstrates that the new incidence and estimated mortality of colorectal cancer ranks third among all tumor types. More than 90% of rectal adenocarcinomas come from benign rectal adenoma[7]. However, in early stages, benign rectal adenoma often has few symptoms and is very difficult to discover, but are easier to treat (Only need local resection or anus-preserving surgery). By the time of diagnosis, the benign rectal adenoma can develop into rectal cancer, which is one of the diseases that seriously affect people's health. It has the worst prognosis of colorectal cancer surgery (e.g., the 5 year survival

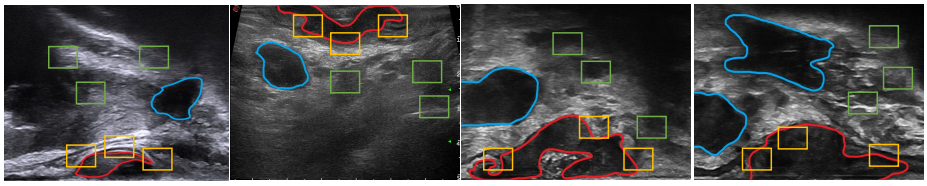


Fig. 1. Example TRUS images. There are large inter-patient variation in terms of shape and size, ambiguous boundary and inhomogeneous intensity distribution of the rectal tumor. Red contour denotes the rectal tumor, and the blue contour are error-prone region. Yellow box indicate the region relevant to tumor segmentation, the main discriminative information for judging rectal tumors, and the green box are noise region that irrelevant to rectal tumor segmentation.

rate of colorectal malignant tumor was 43%[3]) due to its deep location, complex anatomical relationship, and easy to recurrence. Therefore, diagnosis and treatment in time and correctly is very important. Rectal tumors are characterized by localized thickening of the rectal wall, changes in echo, masses growing inside or outside the cavity, and the affected intestinal wall and the surrounding normal intestinal wall structure are discontinuously interrupted, so the depth of tumor invasion can be judged by transrectal ultrasound (TRUS). The advantages of dynamic and real time observation, good reproducibility, cheap and no radiation are important criteria for assessing the T stage of rectal cancer[9]. Contrast-enhanced ultrasound can not only display the morphological characteristics of rectal disease, but also dynamically observe the microcirculation blood flow perfusion of the tumor, which is a reliable indicator for the diagnosis and differential diagnosis of rectal cancer. Therefore, we adopt the transrectal ultrasound (TRUS) as an important basis for our diagnosis.

The early diagnosis of rectal tumor requires much expertise in reading the scanned TRUS images and making decisions, but the increasing number of cases makes it impossible for a limited number of experienced radiologists to check all TRUS images manually. For radiologists, segmenting rectal tumor can improve diagnostic efficiency, develop treatment plans. Consequently, an automatic tumor segmentation system for improve the diagnosis efficiency is in need. When the radiologists is conducting tumor screening, the system first gives a plausible tumor region, and then manually confirms the result, which can save time and improve the diagnosis efficiency.

However, accurate rectal tumor segmentation in TRUS images remains very challenging due to large inter-patient variation in terms of shape and size, as well as several irrelevant noise region, as illustrate in Fig.1. With the development of convolutional neural network (CNN), near-radiologits level performance can be achieved in automated medical image segmentation[11, 2, 17, 13, 5]. Nonetheless, the problem of automatic rectal tumor segmentation in TRUS images haven't been exploited to my best knowledge. The success of U-Net[6] has significantly promoted widespread applications of segmentation on medical images, but there are still limitations in U-Net structure[6, 15]. Most of the previous methods based

on U-Net concatenated/summed every location of the Encoder and Decoder features equally, which could be disadvantage for the rectal tumor segmentation when only a subset of locations actually matters at a semantic level. For TRUS images, there are several region similar to rectal tumor (See Fig. 1 blue region), which make it difficult to segment real tumor using U-Net, always leading to false predictions because it still lacks enough capacity to extract discriminative contextual information relevant to rectal tumor, such as rectal wall in rectal (see Fig.1 yellow box). Above issue is our motivation, We want to learn discriminative feature representation shared across relevant location and reduce the influence of irrelevant locations in the course of feature aggregation, which can improve the performance of rectal tumor segmentation and decrease false positive situation.

Consequently, we leverage the global contextual information as a guidance to learn discriminative feature representation that is shared across relevant locations. We further proposed a novel module, named Context Guiding Module, to exploit the relevant contextual information between the global contextual information and feature vectors corresponding to every location in the course of feature aggregation, which can reduce the influence of unrelated areas and improve the performance of rectal tumor segmentation.

Our contributions are three folds: 1) We made manual annotations of a rectal tumor transrectal ultrasound dataset, which is the first rectual tumor dataset to the best of our knowledge. 2) Context Guided Module exploits the relevant contextual information between the global contextual information and feature vectors corresponding to every location in the course of feature aggregation to reduce the influence of unrelated areas and improve the performance of rectal tumor segmentation. 3) We evaluate the proposed method on rectal tumor segmentation task. Results demonstrate the proposed method outperforms state-of-the-art performance than before, which shows a promising direction on the screening and diagnosis of rectal cancers.

2 Methodology

Fig 2. demonstrates the architecture of proposed method. The proposed framework is an Encoder-Decoder framework consists of three phases: the Encoder, the Context Guided Module and the Decoder. Like U-shape Net architecture, the Encoder includes four encoder blocks, and the residual network (ResNet) block was employed as our backbone for each block to learn hierarchical representations. The Context Guided Module contains the Attention Block and the Context Guide Operator. Fig.3. shows the architecture of the Context Guide Module, and details will be illustrated as follows. The Decoder is adopted to restore the spatial detail, which comprise a stack of 1×1 convolution, a 3×3 transposed convolution and a 1×1 convolution. The output feature of the Decoder will be upsampled to the size of the original image, and then the probability map is obtained through softmax. Finally, the segmentation map can be obtained through a single argmax operation on this probability map.

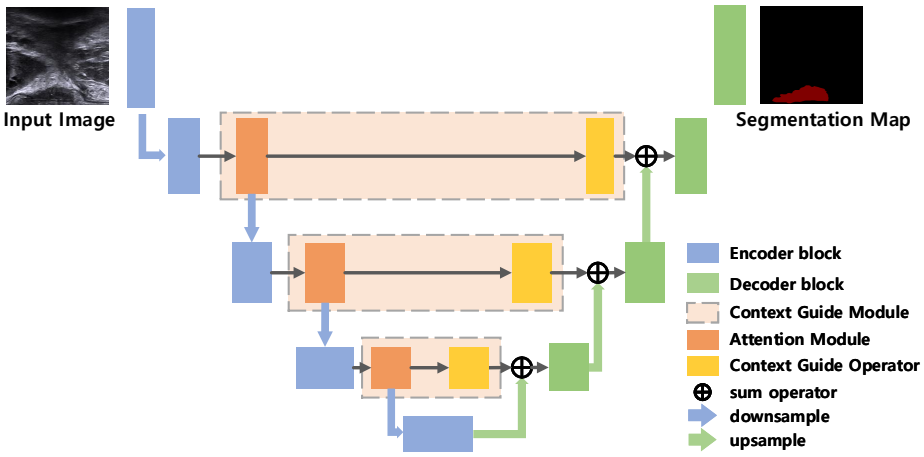


Fig. 2. Illustration of our proposed architecture, which are an Encoder-Decoder framework consists of three phases: the Encoder, the Context Guided Module and the decoder module. The model takes the TRUS images as the input, then outputs the segmentation map in an end-to-end manner.

2.1 Context Guided Module

The Context Guided Module contains the Attention Block and the Context Guide Operator.

Context Guided Operator. Low-level features from Encoder blocks are rich in spatial details but lack semantic information, but high-level features from Decoder blocks are the opposite.[14]. Previous methods which simply concatenate/sum low-level and high-level features from equally resolution of encoder and half resolution of the last decoder respectively, which will be biased toward learning features that are unrelated, because irrelevant information, namely noise, always represent a larger quantity of information in TRUS images which cause misclassification. Inspired by [4], the global context can augment the features at each location, but different locations require different global contexts in the course of aggregate the multi-scale features[1]. So, it is essential to achieve different global context for each location, namely, some relevant locations need more global context for segmentation, but some irrelevant are not. From this motivation, we introduce Context Guided Operator, one critical step in Context Guided Module, leverages global contextual information from high-level feature as a guidance to exploit the relevant contextual information between the global contextual information and feature vectors corresponding to every location of low-level feature in the course of feature aggregation. Context Guided Operator can help reduce the influence of unrelated areas and improve the performance of rectal tumor segmentation. Formally, the low-level feature $E \in \mathbb{R}^{C\times H\times W}$ from Encoder block firstly go through an Attention block to extract the discriminative contextual information, followed by a convolution layer to generator the global

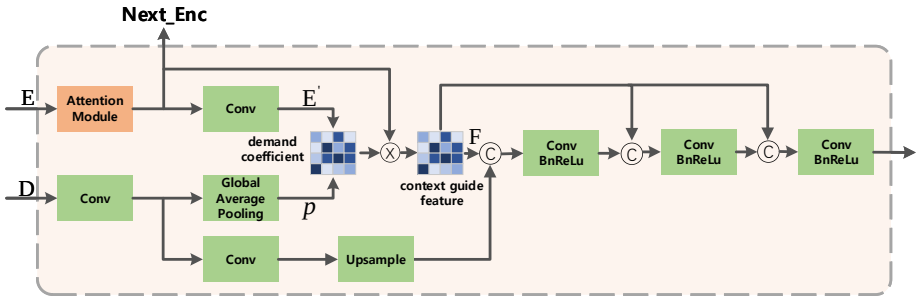


Fig. 3. Architecture of Context Guided Module, which inputs the high-level feature from decoder block and low-level feature from Encoder block, then output the feature aggregated to the next Decoder block.

context $E' \in \mathbb{R}^{C' \times H \times W}$. Then, we use a convolution layer followed by a global average pooling to generate a global context vector $p \in \mathbb{R}^{C' \times 1 \times 1}$ of high-level feature map $D \in \mathbb{R}^{C' \times H' \times W'}$, where $H' = \frac{1}{2}H$ and $C' = 2C$. We utilize the global context vector p to guide the low-level feature context E' to aggregate the feature maps efficiently. We apply the L_2 norm to learn the contextual coefficient which represents the amount of global contextual information required for each location in different location before aggregating multi-scale features. Formally, let $w_i \in \mathbb{R}^{H \times W}$ be the contextual coefficient of each pixel location, $i \in [1, 2, ..., H \times W]$ is i^{th} location in E', which denoted as $w_i = ||E'_i - \alpha * p||_2$, where α is a scaling parameter, which counteract excessive global context vector p. The smaller w_i, the greater the demands for global context. And sigmoid also be utilized to restrict the contextual coefficient range to [0,1]. After that, we multiply the global context E_i by w_i to generate the context guided feature $F \in \mathbb{R}^{C' \times H \times W}$, which can be denoted as $F_i = \sum_{i=1}^{H \times W} w_i E_i$.

Moreover, for more effective integration of multi-scale features, we concatenate the context guided feature and the high-level fature map upsampled, and then reuse the context guided feature by a concatenate operation followed by a convolution layer for two times. Such a recurrent learning process complements more spatial details for each position, and achieve a coarse-to-fine performance improvement[1].

Attention Block is very essential in Context Guided Module, which can further enhance discriminative ability of feature between real rectal tumor and the tumor-like region. We adopt the self-attention block[12, 10] as the attention block in Context Guided Module. Let F represent the input to the Attention Block and O represent the output. Specifically, it first computes the response at a position as a weighted sum of feature F at all spatial location S in the input feature maps:

$$O_i = \sum_{\forall j \in S} A(F_i, F_j) \cdot \mathcal{G}(F_j) \tag{1}$$

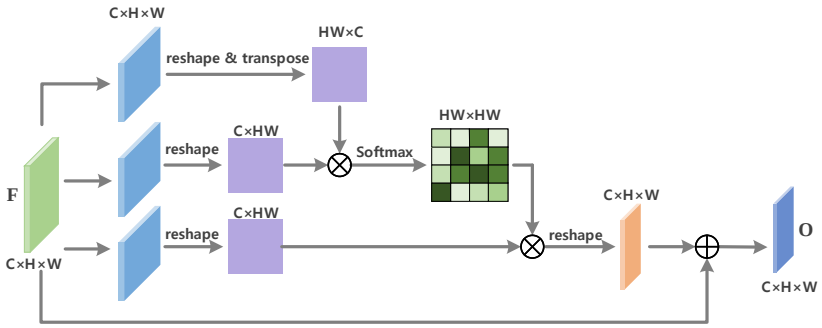


Fig. 4. self-attention block.

where i is the index of an output position whose response is to be computed and j is index that enumerates all possible positions. $A(F_i, F_j)$ is used to compute the spatial attention map between j^{th} position's impact on the i^{th} position. And O is the output with the same size to F. $\mathcal{G}$ computes a representation of the input signal at the position j, here, it's a 1×1 Convolutional layer. The attention are applied as the weighting of the input feature to better prune out irrelevant background features and thereby distinguish the confusing target region. We adopted the dot product version[10] by setting $A(F_i, F_j) = F_i^T F_j$ to compute the attention. After that, the attention map is processed by a 1×1 convolutional layer, and add back the input feature map to obtain the final output O. An illustration of our attention module can be found in Fig.4.

3 Experiments

The proposed method are evaluated on rectal tumor dataset we collected in Sichuang Cancer Hospital. In the next subsections, we first introduce the dataset and implementation details, then we make detail comparisons to evaluate our proposed approaches on rectal tumor dataset. Finally, we present our results compared with state-of-the-art methods on rectal tumor dataset.

3.1 Datasets

Our experimental data collected from rectal tumor patients who performed transrectal ultrasound examination at Sichuan Cancer Hospital from January 2016 to December 31, 2019. (Note that all patients signed the informed consent.) The screening instruments are BK Pro focus 2202 and esaote mylabtwice HD. In total, 615 TRUS images were collected from their electronic records, of which 365 images were randomly selected for training, 100 for validation and 150 for testing. The size of each TRUS image are resized to 448×448, and all TRUS images were manually delineated by a doctor with more than 5 years of experience in rectal ultrasound.

3.2 Implementation Details

In order to accelerate the convergence of training, we employ the ResNet architecture pretrained in imagenet as our backbone network, During training phase, we employ a poly learning rate policy where the initial learning rate is multiplied by $(1 - \frac{iter}{total_iter})^{0.9}$ after each iteration. The base learning rate is set to 0.001. Stoachastic gradient descent (SGD) is adopted to train the segmentation network, which momentum and weight decay coefficients are set to 0.9 and 0.0001 respectively. Moreover, we apply random crop, random rotate and horizontal flip for data augmentation in train phase in order to reduce the risk of overfitting. This experiments are implemented using Pytorch on a GeForce GTX 1080Ti GPU with a batch size of 4. Accuarcy (Acc), Dice Coefficients (Dice), Mean intersection-over-union (mIoU), Precision and Recall are adopted to evaluate the performance of our method.

Loss Function. We need to train the proposed method to predict each pixel to be rectal tumor or not, which is a pixel-wise classification. We achieve this by minimizing the cross-entropy loss of each training sample. Assume y_i is the true label, the $\hat{y}_i$ represents the prediction probability of the model (after softmax), N is the total number of training samples, then the loss function is defined as shown in Equation 2. It encourages the segmentation model to predict the right class label at each pixel location independently. In our experiments, there are only two classes of each pixels in TRUS images, whether rectal tumors or not.

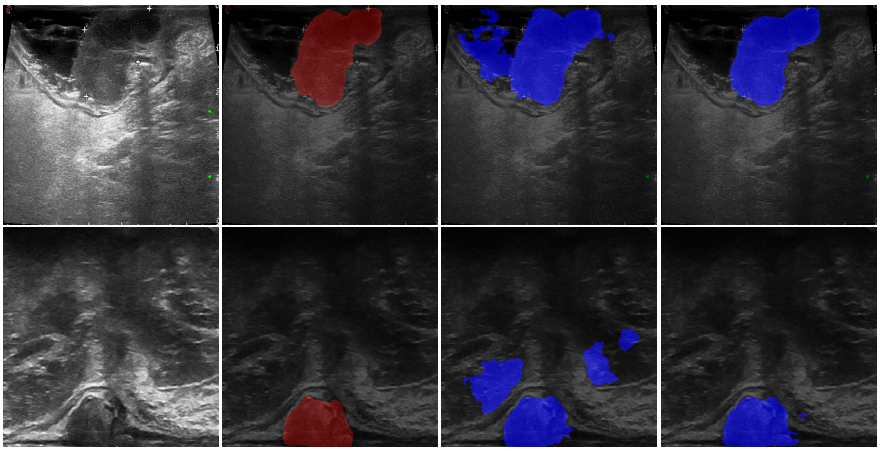
$$\mathcal{L}(y, \hat{y}) = -\frac{1}{N} \sum_{i=0}^{N} y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i) \tag{2}$$

3.3 Segmentation performance

We compared results of our method with several state-of-the-art methods, including U-Net[6], UNet++[16], LadderNet[18], CSNet[5], CENet[2]. For a fair comparison, we obtain the result of our competitions and adjusting training parameters to obtain best segmentation results. The baseline method is similar to the Conventional U-Net, but take the resnet as the Encoder and without the Context Guided Module. This baseline method allows us to analyze the effectiveness of the Context Guided Module. Table 1 lists the metric results of different methods on rectal tumor dataset. It can be observed that our proposed method signicantly better than others. Fig.5 provides visual results of rectal tumor segmentation. Obviously, our segmentation performance is closest to the ground truth facing inhomogeneous intensity distribution and large interpatient variation in terms of shape and size, especially for those TRUS shows several plausible tumor-like region misclassified to rectal tumor. Our methods can successfully segment the real rectal tumor region due to make full use of the contextual information, which demonstrate the effectiveness of our proposed method.

Table 1. Metric results of different methods, including UNet, UNet++, LadderNet, CSNet, CENet and our proposed method (best results are highlighted in bold)

Method	Acc	Dice	mIoU	Precision	Recall
UNet[6]	0.9581	0.8966	0.8220	0.9761	0.9766
UNet++[16]	0.9533	0.8850	0.8051	0.9738	0.9734
LadderNet[18]	0.9606	0.9005	0.8280	0.9744	**0.9813**
CSNet[5]	0.9579	0.8989	0.8254	0.9800	0.9722
CENet[2]	0.9661	0.9168	0.8528	0.9814	0.9803
baseline	0.9649	0.9130	0.8469	0.9794	0.9810
Prop	**0.9675**	**0.9206**	**0.8587**	**0.9828**	0.9805



(a) Image (b) Ground Truth (c) UNet (d) proposed

Fig. 5. Visualization results in rectal tumor dataset. From left to right: original images, ground truth, and predictions from UNet and proposed method.

4 Conclusion

This paper proposed a novel deep neural network for rectal tumor segmentation in TRUS images. Our key idea is to introduce the Context Guided Module which extracts the discrminative feature representation shared across relevant locations in the course of feature aggregations. It reduces the influence of unrelevation locations in TRUS images. In addition, our proposed framework gets a statisfied result in rectal tumor segmentation, which can be practical in assisting radiologists for clinical applications, since the annotation in rectal tumor requires massive labor from radiologists. In future work, we will extend our approach to segment each layer of rectal wall, which can further help stage the tumor and diagnose rectal cancer.

Acknowledgements. This research has been supported Sichuan Province Key Research and Development Plan (2019YFS0427), and in part supported by the National Natural Science Fundation of China under Grant (61671125).

References

1. Fu, J., Liu, J., Wang, Y., Li, Y., Bao, Y., Tang, J., Lu, H.: Adaptive context network for scene parsing. In: Proceedings of the IEEE international conference on computer vision. pp. 6748–6757 (2019)
2. Gu, Z., Cheng, J., Fu, H., Zhou, K., Hao, H., Zhao, Y., Zhang, T., Gao, S., Liu, J.: Ce-net: Context encoder network for 2d medical image segmentation. IEEE transactions on medical imaging (2019)
3. Jemal, A., Siegel, R., Ward, E., Murray, T., Xu, J., Thun, M.J.: Cancer statistics, 2007. CA: a cancer journal for clinicians **57**(1), 43–66 (2007)
4. Liu, W., Rabinovich, A., Berg, A.C.: Parsenet: Looking wider to see better. arXiv preprint arXiv:1506.04579 (2015)
5. Mou, L., Zhao, Y., Chen, L., Cheng, J., Gu, Z., Hao, H., Qi, H., Zheng, Y., Frangi, A., Liu, J.: Cs-net: Channel and spatial attention network for curvilinear structure segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 721–730. Springer (2019)
6. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
7. Savides, T.J., Master, S.S.: Eus in rectal cancer. Gastrointestinal endoscopy **56**(4), S12–S18 (2002)
8. Siegel, R.L., Miller, K.D., Jemal, A.: Cancer statistics, 2019. CA: a cancer journal for clinicians **69**(1), 7–34 (2019)
9. Tsai, C., Hague, C., Xiong, W., Raval, M., Karimuddin, A., Brown, C., Phang, P.T.: Evaluation of endorectal ultrasound (erus) and mri for prediction of circumferential resection margin (crm) for rectal cancer. The American Journal of Surgery **213**(5), 936–942 (2017)
10. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7794–7803 (2018)
11. Wang, Y., Deng, Z., Hu, X., Zhu, L., Yang, X., Xu, X., Heng, P.A., Ni, D.: Deep attentional features for prostate segmentation in ultrasound. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 523–530. Springer (2018)
12. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. In: International Conference on Machine Learning. pp. 7354–7363 (2019)
13. Zhang, S., Fu, H., Yan, Y., Zhang, Y., Wu, Q., Yang, M., Tan, M., Xu, Y.: Attention guided network for retinal image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 797–805. Springer (2019)
14. Zhang, Z., Zhang, X., Peng, C., Xue, X., Sun, J.: Exfuse: Enhancing feature fusion for semantic segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 269–284 (2018)

15. Zhang, Z., Fu, H., Dai, H., Shen, J., Pang, Y., Shao, L.: Et-net: A generic edge-attention guidance network for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 442–450. Springer (2019)
16. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: 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 (2018)
17. Zhu, Z., Xia, Y., Xie, L., Fishman, E.K., Yuille, A.L.: Multi-scale coarse-to-fine segmentation for screening pancreatic ductal adenocarcinoma. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 3–12. Springer (2019)
18. Zhuang, J.: Laddernet: Multi-path networks based on u-net for medical image segmentation. arXiv preprint arXiv:1810.07810 (2018)