Two-Path Augmented Directional Context Aware Ultrasound Image Segmentation

Yuhan Song
School of Information Science
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa, Japan
songyuhan98@gmail.com

Armagan Elibol and Nak Young Chong
Human Information Science Research Area
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa, Japan
{*aelibol, nakyoung*}*@jaist.ac.jp*

***Abstract*—The segmentation of ultrasound (US) images plays a crucial role in the development of end-to-end smart diagnosis systems. In the diagnostic stage, specific diagnosis programs can be applied to well-cropped sub-regions within a single US image, catering to different medical interests. In this study, we propose and test a neural network model designed to perform end-to-end segmentation on abdominal US images, with a focus on five different anatomical structures: liver, kidney, vessels, gallbladder, and spleen. The main contribution of our work lies in the exploration of multi-organ/tissue segmentation. Unlike previous research, our approach takes into account two inherent features of US images: (1) significant variations in spatial sizes among different organs and tissues, and (2) the relatively consistent spatial relationships among anatomical structures within the human body.**

To address these considerations, we introduce a novel image segmentation model that combines the feature pyramid network (FPN) and the spatial recurrent neural network (SRNN). In our paper, we describe the utilization of FPN for extracting anatomical structures of varying scales, as well as the implementation of SRNN to capture spatial context features within abdominal US images. Our model incorporates both top-down and bottom-up pathways, enhancing both semantic features and spatial context features. We refer to this as the "two-path augmented" approach. Furthermore, we incorporate a directional attention mechanism, which selectively leverages spatial context information from four principal directions. This is the essence of our "directional context aware" component. The performance of our proposed model is evaluated through both quantitative and qualitative measures. The evaluation results demonstrate the competitiveness of our approach, and the inclusion of spatial contextual information has resulted in improved performance compared to using the pure feature pyramid network alone.

Index Terms - Artificial Intelligence, Medical Image Segmentation, Robotic Ultrasonography

I. Introduction

A. Background

Entering the 21st century, population aging is becoming a serious problem for many countries. Taking Japan as an example, according to Japan Statistical Yearbook 2023 released by the Ministry of Internal Affairs and Communications [1], Japan's population aged 65 and over currently stands at just over 36.21 million, accounting for 28.9% of the total population. This number is still on its way to the peak. One of the most challenging tasks brought about by population aging is that senior citizens require regular and continuous health support and/or monitoring. To make sure senior citizens can get medical help timely and easily, besides medical facilities, diagnosing methods in the caring center or personal residence are also in need. Moreover, the popularity of personal medical inspection devices will bring convenience to other citizens as well. With the help of those devices, anyone in need can do simple physical checks to monitor their health condition just within their residence. For example, people with limited mobility or pregnant mothers will not bother traveling far to the hospital to get their regular physical check.

The demand for real-time/convenient health monitoring requires remote or portable inspection devices to be popularized. In clinical practice, US imaging is one of the most commonly implemented imaging modalities. Because it is approachable, effective, informative, and of low cost, US devices are easy to be implemented. Also benefiting from its non-invasive and non-radioactive nature, the operation of US devices is of low threshold (Fig. 1). Medical US imaging requires an accurate delineation or segmentation of different anatomical structures for various purposes. For example, doctors can make a naive health condition preview by assessing the organ size [2]. In some surgeries, precise US image segmentation can provide guidance for the interventions [3]. However, in contrast to US devices' convenience in use, US images are hard to process because of low contrast, acoustic shadows, and speckles, to name a few [4]. Even experienced doctors consider it a challenging task to tell the accurate contour of various organs and tissues. To realize a robust computer-aided diagnosis system, an automated and robust US image segmentation method is expected to help with locating and measuring important clinical information. Along these lines, we are developing a control algorithm for the robot arm to perform automatic US scans (see Fig. 2). As this system is expected to operate without human intervention, an evaluation metric for the robot's movement is necessary. Besides resolution and clarity, the integrity of anatomical structures is important as well. To this end, a segmentation algorithm needs to be incorporated into the robot trajectory control system. US images derived from abdominal scanning are our primary research target. In clinical practice, doctors can leverage abdominal US scanning images to evaluate health conditions of various anatomical structures [2].

Fig. 1. Wireless US Probe

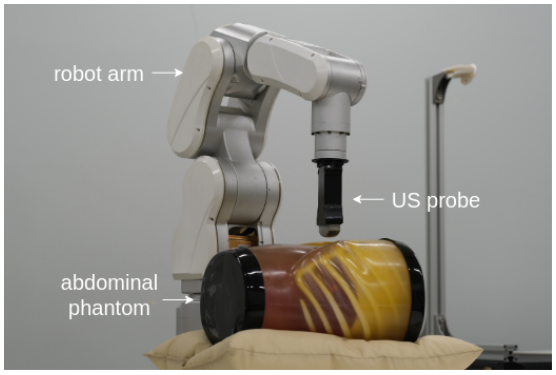


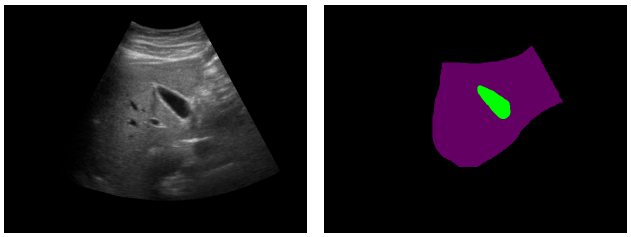
Fig. 2. Our Robot-assisted US Imaging System

B. Research Contributions

We have designed and tested a new network model that can predict semantic masks on convex US images. Compared with previous research on US image segmentation, this network model can perform segmentation on several different anatomical structures simultaneously rather than focusing on a single target tissue. To achieve this goal, two important inherent properties of US imaging are taken into consideration.

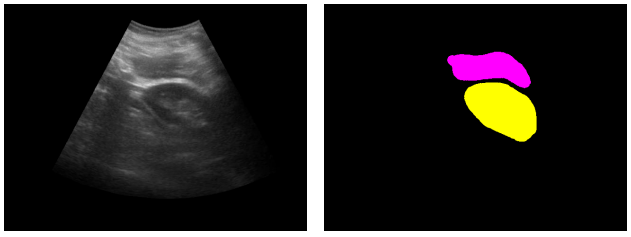
On one hand, different organs/tissues in the abdominal part vary largely in shape and size. For example in Fig. 3, the liver (violet) is much larger than the pancreas (green). To alleviate the influence of such class imbalance problem brought by the difference in scale, the network model is built in an FPN structure. On the other hand, different anatomical structures usually form a constant spatial correlation pattern. For example in Fig. 4, the spleen (pink) and kidney (yellow) are maintaining a similar spatial correlation in different US images.

Utilizing these two properties, the proposed model has managed to predict semantic masks for 5 different organs and tissues (liver, kidney, spleen, vessel, and gallbladder). Compared with previous research, this model realized true end-to-end US image segmentation aiming at multiple target structures using one single model. This work can be generalized to other tasks where spatial context information can help a lot with analysis.



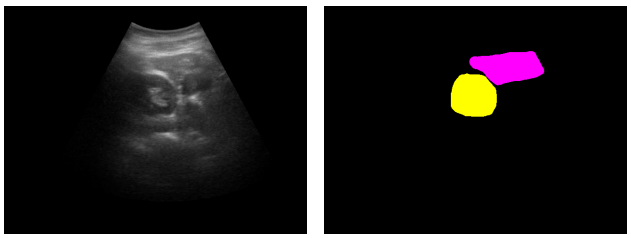
(a) Liver & Gallbladder (b) Mask Label

Fig. 3. Scale Variance Example.



(a) Spleen & Pancreas(1) (b) Segmentation Mask(1)

(c) Spleen & Pancreas(2) (d) Segmentation Mask(2)



(e) Spleen & Pancreas(3) (f) Segmentation Mask(3)

Fig. 4. Spatial Correlation Example

II. Related Work

Traditional US image segmentation methods usually focus on the detection of textures and boundaries based on morphological or statistical methods (Fig. 5). Mishra et al. [5] proposed an active contour solution using low-pass filters and morphological operations to make a prediction of the cardiac contour. Mignotte et al. [6] developed a boundary estimation algorithm based on a Bayesian framework, where the estimation problem was formulated as an optimization algorithm to maximize the posterior possibility of being a boundary. Previously Mignotte's team used statistical external energy in a discrete activate contour for the segmentation of short-axis parasternal images [7], in which a shifted Rayleigh

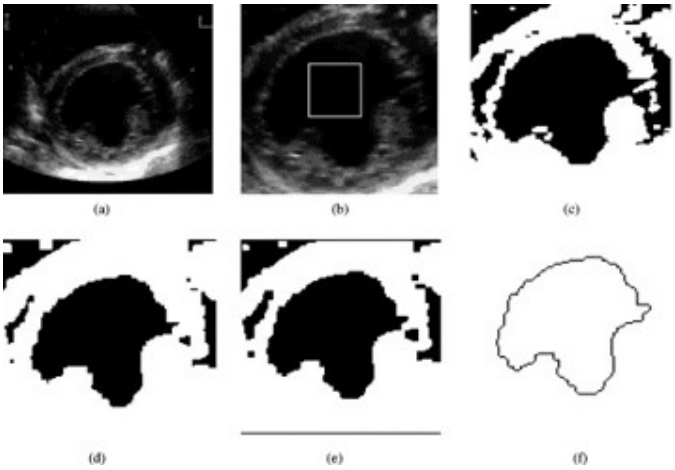


Fig. 5. Traditional Segmentation Methods [5]

distribution was used to model gray-level statistics. Boukerroui et al. [8] also proposed a Bayesian framework to conduct robust and adaptive region segmentation, taking the local class means with a slow spatial variation into consideration to compensate for the non-uniformity of US echo signals. Those methods serve more as a reference for doctors than an independent diagnosis procedure.

Traditional US image segmentation is time-consuming and prone to irregular anatomical structure shapes because of the inherent physical defects of US. In recent years, AI has been showing great success in the field of image processing. Compared with morphological and statistical methods, CNN-based solutions are more powerful and flexible thanks to their strong nonlinear learning ability. Among all the CNN-based methods, U-Net [9] is definitely one of the most popular network models for biological and medical image segmentation. Many researchers have proposed modified U-Net for various semantic segmentation tasks. For example, Oktay et al. [10] proposed a U-Net model with an attention mechanism. They implemented attention gate units to trim features that are not relevant to the ongoing task in order to improve the segmentation performance without adding excessive computational complexity to the model. Matteo et al. [11] designed a Siam-U-Net for knee cartilage tracking. In that work, Matteo et al. extended the encoder of U-Net up to a parallel structure to extract the cross-correlation depth wisely. Although U-Net has achieved great success in biological and medical image segmentation, in the context of abdominal multi-organ segmentation, the performance of U-Net is limited by the class imbalance problem. Despite the fact that the total amount of instances may be almost equal in training, the relatively large organs and tissues occupy much more pixels in the US images. As shown in Fig. 6, the violet part is the liver which occupies most of the pixels, and the green part is the gallbladder. This makes the algorithm classify as many pixels into the liver as possible since a majority class has a much bigger influence on the final score than the minority class. Fig. 7 shows the segmentation result of an US image containing the liver (violet) and kidney (yellow). Compared with the ground truth, the result tends to ignore the kidney to focus on drawing the true mask of the liver. There could be some solutions to this problem, such as adding different loss weights for different classes [12] or enlarging the weight of classification loss in the total loss calculation. However, to better leverage this property, we build our proposed network model upon an FPN structure [13].

In the context of abdominal US image segmentation, most of the existing methods are targeted at specific organs or anomalies. Chen et al. designed a multi-scale and deep-supervised CNN architecture for kidney image segmentation [14]. They implemented a multi-scale input pyramid structure to capture features at different scales and developed a multi-output supervision module to enable the network to predict segmentation results from multi-scales. Huang et al. [15] developed a detection algorithm for pulmonary nodules based on deep three-dimensional CNNs and ensemble learning. However, the importance of multi-organ segmentation is still ignored.

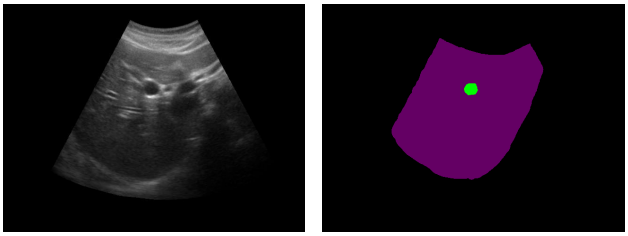


Fig. 6. Example of Class Imbalance Problem

Fig. 7. U-Net Segmentation Result

III. Proposed Method

A. Feature Pyramid Network

In the abdominal section, different anatomical structures vary greatly in shape and size, which inspires us to leverage the FPN network structure. FPN is not an independent object detector by itself. It usually serves as a feature extractor for other detectors. Also, FPN is not the exclusive name for any specific network model. We can build any FPN structure based on our own modified backbone targeting different tasks.

The most important difference between FPN and its competitors is that FPN takes the feature maps from multiple

layers of the encoder backbone as outputs rather than only from the deepest output [13]. Before FPN, there have been other kinds of network models following a pyramid structure. For example, the SSD [16] is one of the first attempts at leveraging the feature pyramid hierarchy. The SSD reuses multi-scale feature maps from different layers in the forward pass. This pyramid network structure is scale-invariant in the sense that an object's scale changes with shifting its level in the feature pyramid. In other words, smaller objects are usually easier to be detected from smaller yet deeper feature maps, and vice versa. Compared with other pyramid network structures like SSD, FPN not only utilizes the relation between scale and layer depth, but also uses a top-down pathway to construct higher-resolution layers from a semantic layer. This solves the problem that features maps composed of low-level structures (closer to the original level) is too naive for accurate object detection. As the reconstructed layers are semantically strong, but the locations of objects are not precise after all the down-sampling and up-sampling, the authors then added lateral connections between reconstructed layers and the corresponding feature maps to help the decoder predict the locations better. Then the detector heads will make predictions on all the output layers.

B. SRNN Structure

One important property of US images is that the anatomical structures form a constant spatial relationship under the same scan pattern. Experienced sonographers rely heavily on such spatial context information to locate the target organs. This prior knowledge inspired us to take spatial context information into consideration.

Many studies have explored the utilization of RNNs to gather contextual information. Traditionally, RNNs are utilized to extract context from a sequence (sentence, speech, or video). For example, Schuster and Paliwal [17] proposed a BRNN that passes both forward and backward across a time map to ensure the information is propagated across the entire timeline. Tang et al. [18] designed a context-aware natural language generation model, which encodes the contexts into a continuous semantic representation and then decodes the semantic representation into text sequences with recurrent neural networks. When it comes to the context of spatial information, Graves and Schmidhuber [19] proposed a multi-dimensional RNN to recognize handwriting. Byeon et al. [20] built a long short-term memory RNN structure for scene labeling.

Bell et al. [21] proposed an object detection network structure called Inside-Outside Net (ION). Besides taking the information near an object's region of interest, the introduction of contextual information has improved the performance, for which a module of four directional RNNs is implemented. Fig. 8 shows the propagation of the RNNs. The structures are placed laterally across the feature maps and move independently in four cardinal directions: right, left, down, and up. The outputs from the RNNs are then concatenated and computed as a feature map containing both local and global

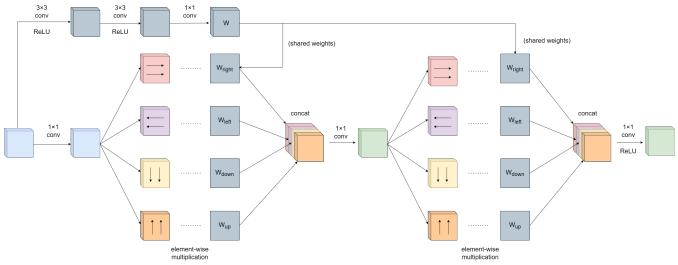


Fig. 8. Spatial RNN Module

contextual information. Upon the base of ION, we put an additional direction-aware attention mechanism from the work of Hu et al. [22]. This attention mechanism aims to making the model selectively leverage the spatial context information propagated through 4 directions.

In this work, SRNN is proposed. The SRNN module follows the idea of the ION network structure. Fig. 9 shows how the RNNs extract the contextual information. A convolution operation is settled at the beginning of the procedure to replace the input-to-hidden translation. Then, four RNNs are propagated through the different directions mentioned above. The outputs from the RNNs are fused into an intermediate feature map. Until this step, each pixel contains the context information aiming at its four principal directions: right, left, up, and down. Then the model will conduct the same process again to extract global-level spatial context information. Finally, a feature map containing the overall context information is generated. For comparison, in the feature map on the left in Fig. 9, each pixel only contains information about itself and its neighbors (depending on the perspective field). After the first round of RNN propagation, the pixels get the context information from its 4 directions. Finally, RNNs propagate through the context-rich pixels to extract the full-directional context information. Therefore, the last feature map is globally context-rich.

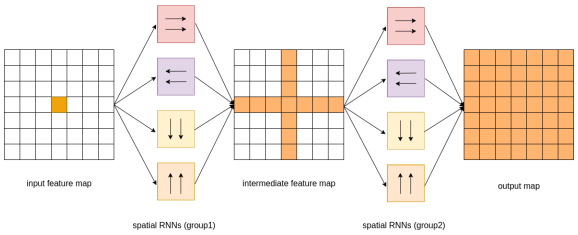


Fig. 9. Illustration of the IRNN propagation

C. Network Structure Overview

Fig. 10 shows the overall structure of the proposed model. On the left side, a ResNet-101 backbone is used as the semantic feature extractor. The input image is propagated from bottom to top, with the network generating feature maps of lower resolution and richer semantic information. We define the layers producing feature maps of the same size as one stage. Then the output of the last layer of each stage represents

the output of the entire stage except the shallowest stage, as this high-resolution layer is computationally demanding due to the low semantic feature. Each of the blue cubes represents an output of the stage called {res2, res3, res4, res5}, respectively. The feature maps go separately through a 1x1 convolution layer and the SRNN module.

The green cubes represent the feature maps after the convolution operation, and the red cubes represent the context feature maps. The deep feature map is concatenated with the context feature map and compressed to reduce depth channels. The feature map from the upper layer, spatially coarser but semantically stronger, is upsampled by a scale factor of 2. (using interpolate function with nearest neighbor upsampling). The upsampled feature maps from the upper pyramid level and the feature map from the current pyramid level are added together(green links) as the new feature map to be concatenated with the spatial feature map.

Similarly, the spatial context information is coarser at the deep level and more precise from those high-resolution levels. Thus, we also build a bottom-up pathway(red links) to deliver precise spatial context information to the higher levels. That is why we call our network "Two-Path Augmented". One thing worth noting is that using adding outperforms the solution of concatenating and dimensionality reducing.

The yellow cubes are the final outputs of the entire feature extractor. After extracting semantic and spatial features, these pyramid feature maps are then sent to RPN [23] and region-based detectors (Fast R-CNN [23], Mask R-CNN [24]). Unlike the classic object detectors, the FPN attaches RPN and Fast R-CNN to each of the output layers. The parameters of the heads are shared across all feature pyramid levels for simplicity, but the accuracy is actually very close with or without sharing parameters (refer to [13]). This is indirect proof that all the levels of the pyramid share similar semantic levels. After that, a DeepMask framework is used to generate masks. The structure of proposers and anchor/mask generators are omitted in the graph, since it is not our main interest.

IV. Experiments and Results

A. Dataset

A dataset of high quality is one of the key factors to train a neural network. However, there are few open-source abdominal US image datasets. Most of the datasets have not been made public for the protection of patients' privacy. In this work, we use the dataset provided by Vitale et al. [25]. This dataset is released on the Kaggle website and contains both artificial US images translated from CT images and images from real US scans (Fig. 11). There are 926 artificial US scans and 61 labeled real US scans, in which we can have the annotations of the liver, kidney, gallbladder, spleen, and vessels. Different organs are assigned segmentation masks of different colors. Table I shows the name of the anatomical structures and the corresponding instance number. We mixed and separated the dataset into 3 subsets: 787 images for training, 100 for testing, and 100 for validation.

B. Detectron2

We built our project on the top of *detectron2*, which is an open-source platform containing many network architectures and training tools [26]. This complete work allows users to train the given networks on their specific tasks or the users can build their own structure efficiently using the encapsulated models. We build the backbone framework based on the implementation of FPN in *detectron2*. We then develop our SRNN structure inserted into the FPN framework as a new context feature extractor. The standardized RPN, ROI, Fast R-CNN, and Mask R-CNN heads are attached after the feature extractor as the proposal generators. Specifically, the output feature maps are from {res2, res3, res4, res5} of the ResNet layers. The size of the anchor generators is set to $32 \times 32, 64 \times 64, 128 \times 128$, and 256×256. For each feature map, FPN gives 1000 proposals. The ROI box head follows the structure of Fast R-CNN with 2 fully convolutional layers and 7×7 pooler resolution. The Mask R-CNN head has 4 convolutional layers and a pooler resolution of 14×14. The ROI heads score threshold is set to 0.5 for both box and mask heads. We reduce the ROI head batch size from 512 to 128, which is computationally efficient while the accuracy is nearly the same. Some modifications for compatibility has been made to the model, enabling it to run under the *detectron2* framework.

C. Loss Functions

Multiple loss functions are included in our training procedure, some of which are listed here:

1) Anchor and bounding box loss: Both RPN and ROI (Box) heads use a smooth l1 loss for the proposed anchors and bounding boxes. The anchors and bounding boxes are represented as a tensor of length 4: (x, y, w, h), namely, the x, y coordinates and width/height of the anchor or bounding box. Then with the ground truth information, 4 deltas (d_x, d_y, d_w, d_h) are calculated by

$$\begin{aligned} d_x &= (g_x - p_x) \\ d_y &= (g_y - p_y) \\ d_w &= \log(g_w / p_w) \\ d_h &= \log(g_h / p_g) \end{aligned} \tag{1}$$

where g represents the ground truth and p stand for the predicted anchor or bounding box. The deltas will be stacked together to compute the smooth l1 loss, given by

$$L_1^{smooth}(x) = \begin{cases} 0.5x^2 & if \quad |x| < \beta \\ |x| - 0.5 * \beta & otherwise \end{cases} \tag{2}$$

where β is a pre-defined smooth parameter.

TABLE I
Dataset

Name	Liver	Kidney	Gallbladder	Vessels	Spleen
Number	591	377	219	289	172
Color	violet	yellow	green	red	pink

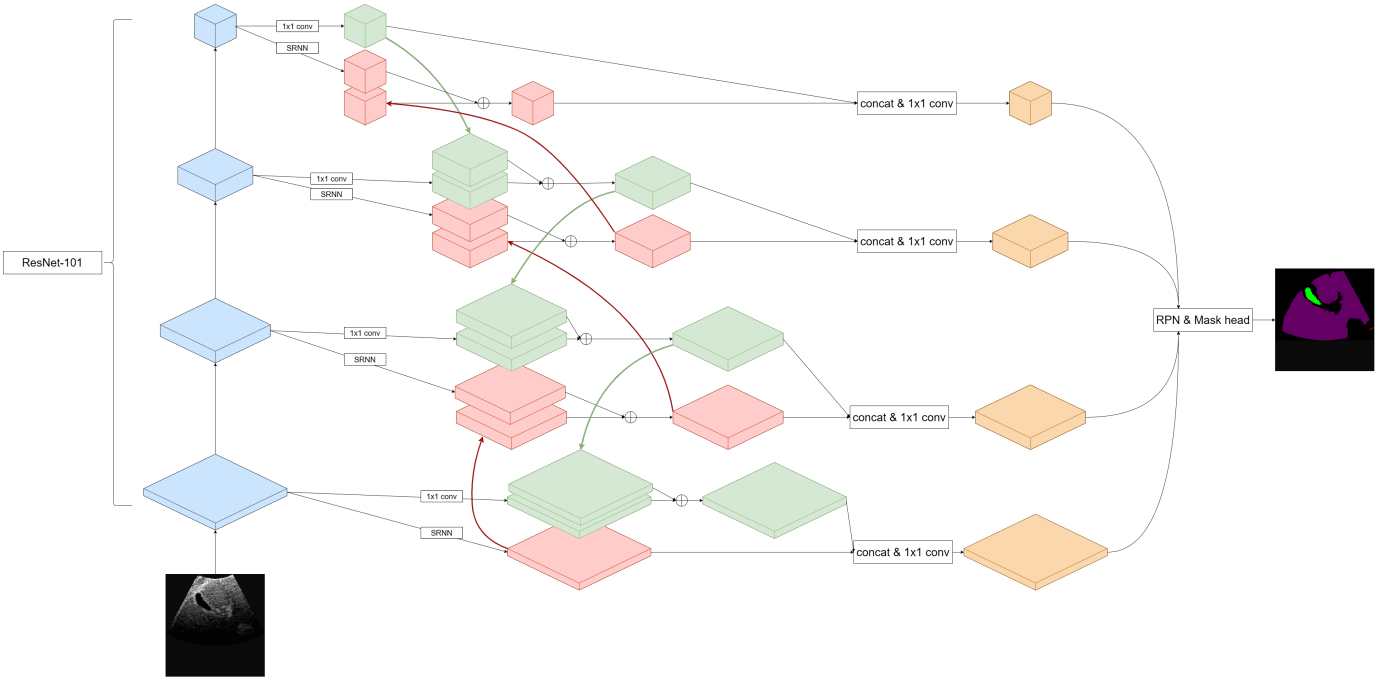
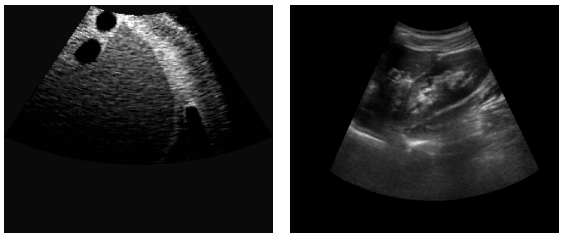


Fig. 10. Proposed network structure



(a) Artificial US Image (b) Real US Image

Fig. 11. Dataset Content Example

2) Classification loss: Softmax cross entropy loss is calculated for all the foreground and background prediction scores:

$$L_{CE} = -\sum_{i=1}^{n} y_i log(p_i) \tag{3}$$

where y_i is the true label and p_i is the softmax probability for the i^{th} class.

3) Mask loss: The mask loss is defined as the average binary cross-entropy loss. Eq. 4 computes the mask loss for the k^{th} class:

$$L_{mask} = -\frac{1}{m^2} \sum_{1 \le i,j \le m} [y_{i,j} \log \hat{y}^k_{i,j} + (1 - y_{i,j}) \log(1 - \hat{y}^k_{i,j})] \tag{4}$$

where $y_{i,j}$ is the label of a cell(i,j) in the true mask for the region of size $m \times m$, and $\hat{y}^k_{i,j}$ represents the value of the same cell in the predicted mask.

D. Experiment Setup

Our experiment builds upon *detectron2* framework in the PyTorch environment. We modified the original FPN in *detectron2* by adding the SRNN module. We train the model on a single GPU (NVIDIA-A400). The batch size is set to 1, since this GPU has relatively limited memory and the dataset is again relatively small. The model is trained for 300k epochs, taking around 30 hours to converge. The initial learning rate is set to 0.0025. We also tried to explore deeper layers in the backbone, adding 2 extra pyramid layers on the top of the backbone, but the accuracy fails to increase. It takes around 30 hours for the model to converge based on a pre-trained Resent backbone.

V. Results

A. Quantitative Result

The performance of the trained model is evaluated by the dice coefficient. The dice coefficient is twice the number of elements common to two sets X and Y, divided by the sum of the number of elements in each set. In our work, X and Y are the predicted classification map and the ground truth. Therefore, the numerator is regarded as the intersection pixels of the predicted mask and the ground truth, and the denominator is the sum of mask pixels in both. Considering we have 5 object classes (background not included), the coefficient score is computed separately and then averaged as the final score. As there might be no appearance of certain classes, we added a smoothing parameter ϵ to avoid zeros in the denominator. The modified equation is given in (5),

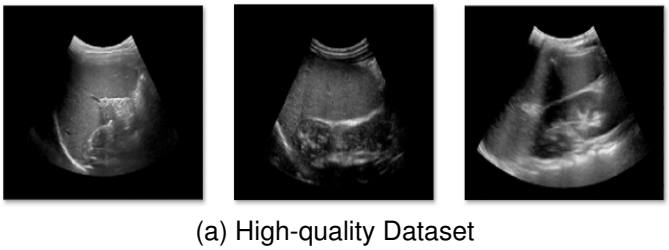
where n is the number of classes.

$$D = \frac{\sum_{i=1}^{n} \frac{2|X_i \cap Y_i|}{|X_i|+|Y_i|+\epsilon}}{n} \tag{5}$$

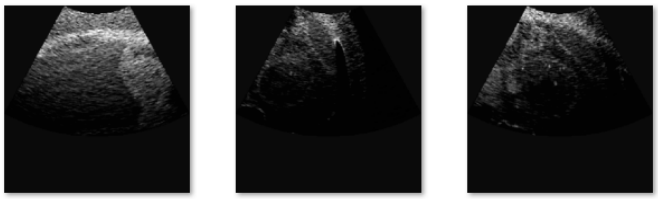
There are a few similar pieces of research surrounding abdominal multi-organ segmentation. To our knowledge, neither any relevant benchmark nor competitionexists. Therefore, we separately pick some comparable results from different research aiming at single-organ segmentation. Respectively, the segmentation result of the liver is compared with the work [27] kidney [28], and their result is taken into comparison as well. The segmentation performance of the gallbladder and spleen are compared with [29] and [30]. The numeric result may not seem encouraging compared with those well-aimed studies. On one hand, the segmentation performance is limited by our lack of high-quality data. For example, in the work of Yuan et al. [30], they trained their model on 420 good quality 2D spleen US images. In our research, we have only 172 instances of training spleen training samples, not to mention that most of the US images are pseudo-US images interpreted from CT images. The difference can be seen in Fig. 12. On the other hand, the specifically targeted studies usually introduced some prior knowledge into their segmentation algorithm like the detection of boundaries. Meanwhile, we trained a pure FPN model for comparison to demonstrate the improvement brought by SRNN. Table II shows the dice score of each class, where we can see that the improvement of performance by SRNN is significant. The proposed model outperformed the pure FPN model.

TABLE II
EVALUATION RESULT

Organ/Tissue	Related Work	FPN	FPN+SRNN
Liver	0.821 [27]	0.907	0.940
Kidney	0.5 [28]	0.806	0.865
Gallbladder	0.893 [29]	0.799	0.926
Vessels	–	0.801	0.907
Spleen	0.93 [30]	0.810	0.840
Average	–	0.840	0.905

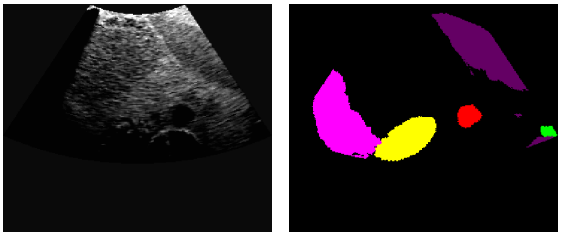


(a) High-quality Dataset

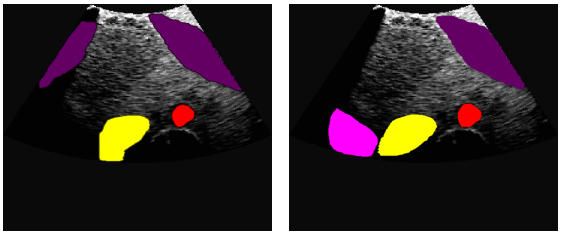


(b) Our Dataset

Fig. 12. Dataset Difference



(a) Original US Image (b) Ground-truth



(c) FPN Result (d) Our Model Result

Fig. 13. Test Results

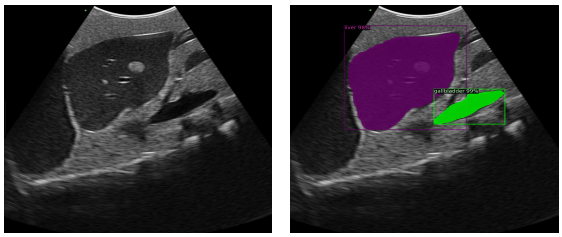
B. Qualitative Result

We have tested the proposed model on artificial and real US images from the evaluation data. Fig. 13 shows an example of semantic segmentation on the US image. (a) is the original US image, and (b) is the corresponding ground truth. (c) and (d) are the segmentation result generated by the pure FPN and our proposed model. We can see that the proposed model outperforms the pure FPN.

Additionally, our proposed model underwent testing using US images manually collected from an abdominal phantom in our laboratory. The exceptional performance of our model is illustrated in Fig. 14, where (a) represents a US image collected from the phantom, and (b) displays the corresponding generated semantic masks. Each bounding box is accompanied by a trust score, which is determined by a hyperparameter. Essentially, any prediction with a confidence score above the threshold value is retained, while those below are discarded. Through multiple experiments, we observed that altering this hyperparameter did not lead to a numerical increase in accuracy score, but it did cause the network to exhibit different prediction trends.

VI. CONCLUSIONS AND FUTURE WORK

In this research, we proposed an FPN based multi-organ/tissue segmentation method combined with the utilization of SRNN. From the experimental results, we can see that the introduction of spatial context information has improved the performance of the original FPN model both in quantitative and qualitative comparison. Notably, our model is competitive even compared with those well-targeted studies. The success of this work lays a solid foundation for feature extensions like



(a) In Vitro US Image (b) Our Model Result

Fig. 14. Robot-assisted US Image Capture/Segmentation

the development of a fully automated US scan system and end-to-end abdominal US diagnosis solution. The findings of this work would also benefit from further research including different scan patterns, since prior knowledge of the US scan pattern would help add more precise spatial context information.

Our future work is to build a control algorithm for a robotic arm to perform an automatic US scan. Along those lines, the improved segmentation algorithm can serve as the evaluation metric of the control system performance.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number JP23K03756 and the Asian Office of Aerospace Research and Development under Grant/Cooperative Agreement Award No. FA2386-22-1-4042.

REFERENCES

[1] "Japan statistical yearbook 2023," Ministry of Internal Affairs and Communications, Nov. 2022.

[2] R. Bisset and A. Khan, *Differential Diagnosis in Abdominal Ultrasound*. Elsevier India, 2012. [Online]. Available: https://books.google.co.jp/books?id=UaY22mAJJnMC

[3] A. Lasso, T. Heffter, A. Rankin, C. Pinter, T. Ungi, and G. Fichtinger, "Plus: Open-source toolkit for ultrasound-guided intervention systems," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 10, pp. 2527–2537, 2014.

[4] R. Almajalid, J. Shan, Y. Du, and M. Zhang, "Development of a deep-learning-based method for breast ultrasound image segmentation," in *IEEE International Conference on Machine Learning and Applications*, 2018, pp. 1103–1108.

[5] A. Mishra, P. Dutta, and M. Ghosh, "A ga based approach for boundary detection of left ventricle with echocardiographic image sequences," *Image and Vision Computing*, vol. 21, no. 11, pp. 967–976, 2003.

[6] M. Mignotte, J. Meunier, and J.-C. Tardif, "Endocardial boundary e timation and tracking in echocardiographic images using deformable template and markov random fields," *Pattern Anal. Appl.*, vol. 4, pp. 256–271, 11 2001.

[7] M. Mignotte and J. Meunier, "A multiscale optimization approach for the dynamic contour-based boundary detection issue," *Computerized Medical Imaging and Graphics*, vol. 25, no. 3, pp. 265–275, 2001.

[8] D. Boukerroui, A. Baskurt, J. Noble, and O. Basset, "Segmentation of ultrasound images—multiresolution 2d and 3d algorithm based on global and local statistics," *Pattern Recognition Letters*, vol. 24, no. 4, pp. 779–790, 2003.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *ArXiv*, vol. abs/1505.04597, 2015.

[10] O. Oktay, J. Schlemper, L. L. Folgoc, M. J. Lee, M. P. Heinrich, K. Misawa, K. Mori, S. G. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention u-net: Learning where to look for the pancreas," *ArXiv*, vol. abs/1804.03999, 2018.

[11] M. Dunnhofer, M. Antico, F. Sasazawa, Y. Takeda, S. Camps, N. Martinel, C. Micheloni, G. Carneiro, and D. Fontanarosa, "Siam-u-net: encoder-decoder siamese network for knee cartilage tracking in ultrasound images," *Medical Image Analysis*, vol. 60, p. 101631, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841519301677

[12] T.-Y. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2999–3007, 2017.

[13] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 936–944, 2017.

[14] G. Chen, J. Yin, Y. Dai, J. Zhang, X. Yin, and L. Cui, "A novel convolutional neural network for kidney ultrasound images segmentation," *Computer Methods and Programs in Biomedicine*, vol. 218, p. 106712, 2022.

[15] W. Huang, Y. Xue, and Y. Wu, "A cad system for pulmonary nodule prediction based on deep three-dimensional convolutional neural networks and ensemble learning," *PLOS ONE*, vol. 14, no. 7, pp. 1–17, 07 2019.

[16] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European Conference on Computer Vision*, 2015.

[17] M. Schuster and K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[18] J. Tang, Y. Yang, S. Carton, M. Zhang, and Q. Mei, "Context-aware natural language generation with recurrent neural networks," *arXiv preprint arXiv:1611.09900*, 2016.

[19] A. Graves and J. Schmidhuber, "Offline handwriting recognition with multidimensional recurrent neural networks," in *Advances in Neural Information Processing Systems*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds., vol. 21. Curran Associates, Inc., 2008.

[20] W. Byeon, T. M. Breuel, F. Raue, and M. Liwicki, "Scene labeling with lstm recurrent neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3547–3555.

[21] S. Bell, C. L. Zitnick, K. Bala, and R. B. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2874–2883, 2016.

[22] X. Hu, C.-W. Fu, L. Zhu, J. Qin, and P.-A. Heng, "Direction-aware spatial context features for shadow detection and removal," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 11, pp. 2795–2808, 2019.

[23] R. B. Girshick, "Fast r-cnn," *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1440–1448, 2015.

[24] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask r-cnn," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.

[25] S. Vitale, J. Orlando, E. Iarussi, and I. Larrabide, "Improving realism in patient-specific abdominal ultrasound simulation using cyclegans," *International Journal of Computer Assisted Radiology and Surgery*, 07 2019.

[26] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[27] L. Man, H. Wu, J. Man, X. Shi, H. Wang, and Q. Liang, "Machine learning for liver and tumor segmentation in ultrasound based on labeled ct and mri images," in *2022 IEEE International Ultrasonics Symposium (IUS)*, 2022, pp. 1–4.

[28] M. Marsousi, K. N. Plataniotis, and S. Stergiopoulos, "Atlas-based segmentation of abdominal organs in 3d ultrasound, and its application in automated kidney segmentation," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015, pp. 2001–2005.

[29] J. Lian, Y. Ma, Y. ma, B. Shi, J. Liu, Z. Yang, and Y. Guo, "Automatic gallbladder and gallstone regions segmentation in ultrasound image," *International Journal of Computer Assisted Radiology and Surgery*, vol. 12, 01 2017.

[30] Z. Yuan, E. Puyol-Antón, H. Jogeesvaran, N. Smith, B. Inusa, and A. P. King, "Deep learning-based quality-controlled spleen assessment from ultrasound images," *Biomedical Signal Processing and Control*, vol. 76, p. 103724, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809422002464