A Kidney Dynamic Ultrasound Image Segmentation Method Based on STDC Network

Ziyu Wang
College of Artificial Intelligence
Nankai University
Tianjin, China
mathilda11235@163.com

Yunmei Guan
College of Computer Science
Nankai University
Tianjin, China
gyunmei0531@163.com

Ziang Chen
College of Software
Nankai University
Tianjin, China
sobremesachen@163.com

Wendi Zhang
College of Cyber Science
Nankai University
Tianjin, China
ztrpasp@163.com

Gongping Chen
Tianjin Key Laboratory of Intelligent Robotics, College of Artificial Intelligence
Nankai University
Tianjin, China
cgp110@stu.xhu.edu.cn

Yu Dai
Tianjin Key Laboratory of Intelligent Robotics, College of Artificial Intelligence
Nankai University
Tianjin, China
daiyu@nankai.edu.cn

2024 36th Chinese Control and Decision Conference (CCDC) | 979-8-3503-8778-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/CCDC62350.2024.10588191

***Abstract*—Ultrasound dynamic images of the kidney serve as crucial tools in renal diagnosis, dynamically displaying the anatomical structure and pathological information of the kidney. However, traditional ultrasound image segmentation heavily relies on the experience of ultrasound doctors, leading to inaccuracies due to repetition and redundancy. In this paper, we present a deep learning method based on Short-Term Dense Concatenate network (STDC network) for kidney dynamic ultrasound (KDU) images segmentation. STDC network adopts Short Term Dense Concatenate as the basic module. In the decoder, the learning of spatial information is integrated into the low-level layer through a single stream approach. Finally, fuse the low-level features and deep features to predict the final segmentation results. Experiments based on our self-collected dataset shows that our method achieves 0.966 in term of Jaccard, 0.983 in term of Precision and 0.982 in term of Recall in the segmentation task. Furthermore, we compare our network with other classic real-time segmentation models, showing that our method outperforms in both accuracy and speed.**

Keywords—STDC network, Kidney Dynamic Ultrasound Image, Accuracy, Real-Time

I. INTRODUCTION

Ultrasound images play a significant role in medical diagnosis, especially in the detection, diagnosis, and treatment of kidney diseases. Kidney ultrasound image segmentation is a crucial step in kidney ultrasound diagnosis, as it enables the non-invasive and radiation-free extraction of anatomical structures and lesion information from kidney regions, facilitating more accurate medical decisions[1]. Dynamic images, in comparison to static images, offer continuous observation of organ lesions, making KDU images essential for efficient diagnosis. Therefore, achieving real-time segmentation of KDU images is of paramount importance in assisting medical professionals with their diagnosis.

KDU images segmentation involve separating the kidneys and surrounding tissues to locate the kidney position. Traditional medical image segmentation methods mainly rely on image features for segmentation. One type relies on spatial locality such as texture and grayscale of the image, while the other uses gradient information to determine the boundaries of the object to be segmented[2]. Methods such as edge detection[3] and grayscale segmentation[4]. However, these methods often require significant manual intervention and may yield unsatisfactory results when dealing with complex medical images.

In recent years, deep learning has made significant progress in image segmentation, as it automatically learns high-level features to achieve better segmentation results. Various excellent medical image segmentation models have been proposed based on the fully convolutional network[5] (FCN) introduced by Long et al. in 2015. The Segnet model[6], proposed by Badrinarayanan et al., performs pixel-wise classification to achieve segmentation. The famous U-Net[7] network, presented by Olaf Ronneberger et al., has shown promising results in medical image segmentation. However, many of these models primarily focus on static images, such as CT images, MRI images, and static ultrasound images, with limited research on KDU images.

To achieve real-time segmentation of dynamic images, network models need to strike a balance between speed and accuracy. Researchers have explored this trade-off, with methods like ENet[8], proposed by Adam Paszke et al., based on an asymmetric encoder-decoder structure, significantly increasing processing speed. Yu[9] et al. proposed the BiSeNet model, employing a multi-path structure to capture global context and preserve local details. But due to the adoption of a multi-path structure, BiSeNet has to some extent increased the computational time of the network. Additionally, these methods have been mainly tested on datasets like CamVid and Cityscapes, with minimal research conducted on medical images. KDU images possess unique characteristics:

- Poor image quality, with numerous artifacts and noise[1].
- Kidney shape and position vary during scanning.
- Different individuals may have variations in kidney size and position.

Taking these into consideration, we propose a novel and efficient structure named STDC network[10] by removing structure redundancy to balance segmentation accuracy and inference speed of KDU images. We will introduce in detail to the structure in the next section.

II. METHOD

STDC network utilizes the STDC module as its fundamental building block and integrates the learning of spatial information into low-level layers in a single-stream

Research supported by the National Key R&D Program of China (Grant No. 2022YFB4700203).

manner in the decoder. Finally, it combines the features from low-level and deep layers to predict the ultimate segmentation results. An overview of the STDC network is shown in Fig. 1, and in the following sections, we will discuss each detail thoroughly.

A. Short-Term Dense Concatenate Module

The STDC module is a critical component of the network. The structure of the STDC module is shown in Fig. 2. Each module is separated into several blocks and we use $ConvX_i$ to denote the operation of the i-th block. As a result, the output of the i-th block can be calculated as follows:

$$x_i = ConvX_i(x_{i-1}, k_i) \tag{1}$$

where x_{i-1} and x_i are the input and output of i-th block, separately. *ConvX* includes one convolutional layer, one batch normalization layer and ReLU activation layer, and k_i is the kernel size of convolutional layer. We cascade the output features of each block as the final output of the STDC module：

$$x_{output} = F(x_1, x_2, ..., x_n) \tag{2}$$

where x_{output} denotes the STDC module output, F is the fusion operation and $x_1, x_2, \ldots, x_n$ are feature maps from all n blocks. We adopt concatenation as our fusion operation.

The number of output channels for each STDC module is N, and the size depends on the step size of the second *ConvX*, as shown in Fig. 2(b) and Fig. 2(c). When the step size is 1, the size remains unchanged, and when the step size is 2, the size is reduced by 1/4.

The STDC module has two advantages:

1) As the network deepens, gradually decrease the number of feature channels to reduce computational complexity;
2) The output of STDC integrates the output feature maps of multiple blocks, containing multi-scale information.

B. Segmentation Architecture

We use the pretrained STDC network as the backbone of our encoder and adopt the context path of BiSeNet to encode the context information. As shown in Fig. 1(a), Stage 3, 4, 5 in the network down-sample feature maps to produce feature maps with down-sample ratios of 1/8, 1/16, and 1/32, respectively. Then, global average pooling is used to extract global context information. We use a U-shaped structure to up-sample the features stem from global feature and combine them with the counterparts from Stage 4 and Stage5.

We adopt Attention Refine Module (ARM) to refine the combination features of every two stages. Fig. 1(b) shows the ARM structure. For the final semantic segmentation prediction, we adopt Feature Fusion Module (FFM) to fuse the 1/8 down-sampled feature from Stage 3 in the encoder and the counterpart from the decoder, then the scale is balanced by batch normalization. These two types of features are at different levels of feature representation. The features from the encoding backbone retain rich detail information, while the features from the decoder contain contextual information due to inputs from the global pool layer. FFM structure shown as Fig .1(c).

According to the different number of STDC modules in each stage, the network can be divided into STDC813 and STDC1446. The General STDC network architecture is shown in Fig. 2(a). *ConvX* operation refers to the Conv-BN-ReLU. Stage3, 4, 5 are composed of several STDC modules.

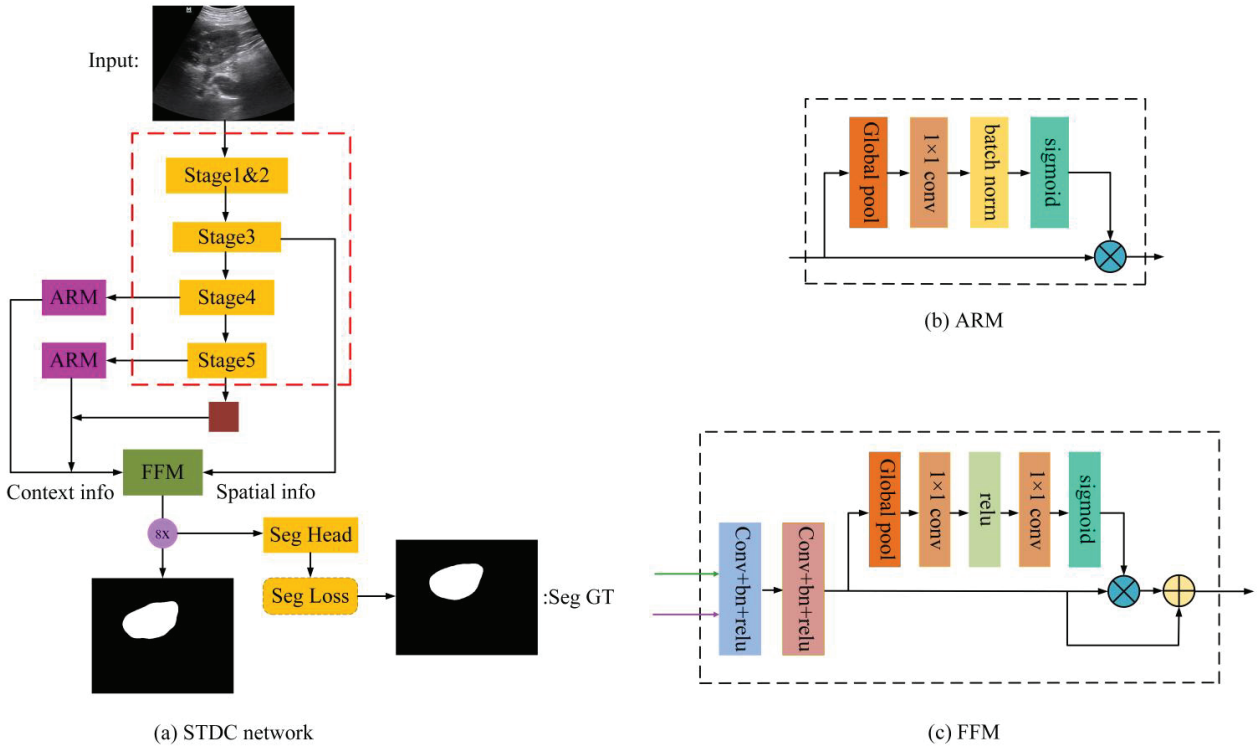

Fig. 1. An overview of STDC network. (a) Network architecture. (b) Components of the attention refinement module (ARM). (c) Components of the feature fusion module (FFM).

C. Loss function

As shown in Fig. 1(a), The Seg Head includes a 3×3 Conv-BN-ReLU operator followed with a 1×1 convolution to get the output dimension N, which is set as the number of classes. We adopt cross-entry loss to optimize segmentation of learning tasks, as (3) shows:

$$\text{loss} = -\frac{1}{N}\sum_i \log\frac{\exp(p_i)}{\sum_j \exp(p_j)} \tag{3}$$

Specifically, during the training phase, Seg Loss consists of three Loss function:

$$l(X; W) = l_1(X_{stage4}; W) + l_2(X_{stage5}; W) + l_p(X_F; W) \tag{4}$$

where l_1, l_2 and l_p are the loss function at stage4, 5 and FFM. X_{stage4} and X_{stage5} are the output feature of stage4 and stage5, respectively. X_F is the FFM fusion feature, W is the weight of network.

III. Experiments

The dataset used in this experiment consists of KDU images collected from 12 patients using laboratory ultrasound detection equipment. After a selection process, we opted for three segments of kidney displays from videos that exhibited relatively complete visualization, with durations of 10 seconds, 15 seconds and 25 seconds, respectively. The shape of each KDU image varies depending on the scanning angle, which contributes to enhancing the generalizability of the network. We first separate the high-quality dynamic images, and then divide them into a training set and a testing set in an 8:2 ratio. The video is further converted into images frame by frame at a standard of 26FPS. Ultimately, we had 1068 images for training and 138 for testing. All images are properly center cropped and resized to 512 × 416 to remove unnecessary background and patient information, and normalize before feeding to the network.

All experiments are conducted on a computer with 6 x Xeon Gold 6142 CPU, NVIDIA RTX 3080 GPU, and 26G of RAM, using PyTorch deep learning framework. In the training phase, the network is optimized using the STDC network's custom optimizer, with lr=0.1, momentum=0.9, and weight_ Decay=5e-4. The batch size is determined to be 16. The other components in STDC network are randomly initialized using PyTorch's default configuration.

To evaluate our method, Jaccard, Precision, and Recall are calculated to quantitatively evaluate network segmentation results, as given in (5) - (7)

$$Jaccard = \frac{TP}{TP+FP+FN} \tag{5}$$

$$Precision = \frac{TP}{TP+FP} \tag{6}$$

$$Recall = \frac{TP}{TP+FN} \tag{7}$$

where TP, FP and FN are the number of true positive, false positive, and false negative pixels, respectively.

Fig. 3 shows the segmentation index results of STDC813 network on some test sets. From the figure, we can see that our method reach obtained 96.62% average Jaccard, 98.29% average Precision and 98.24% average Recall.

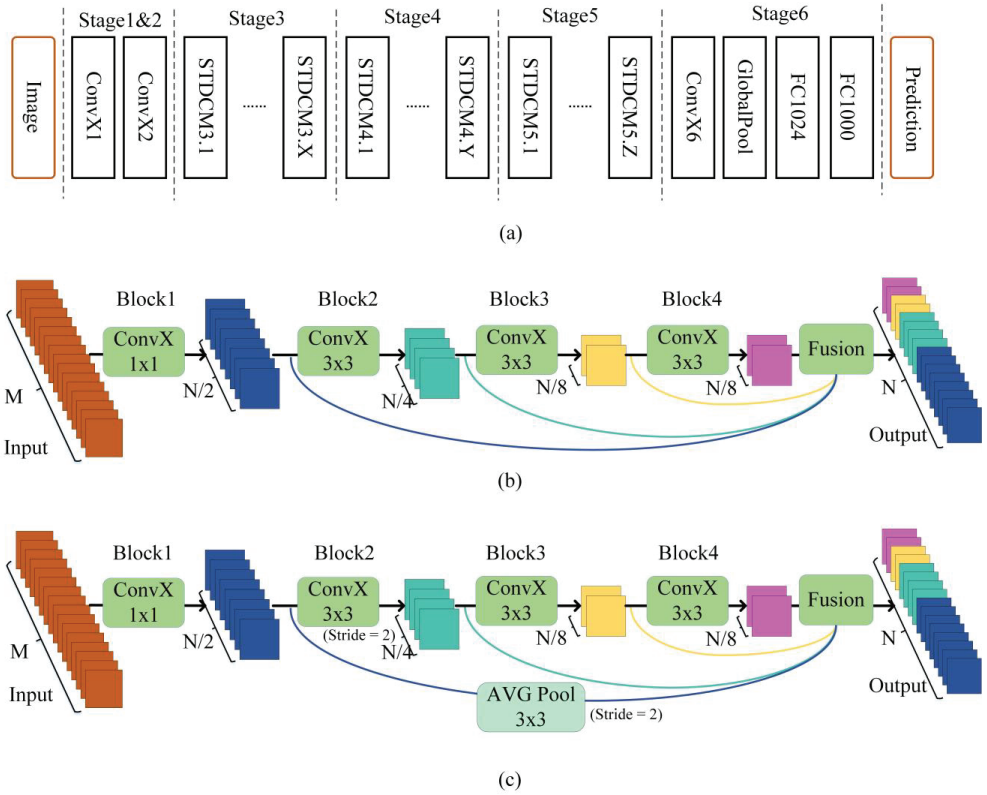

Fig. 2. (a) General STDC network architecture. (b) Short-Term Dense Concatenate module (STDC module). (c) STDC module with stride=2.

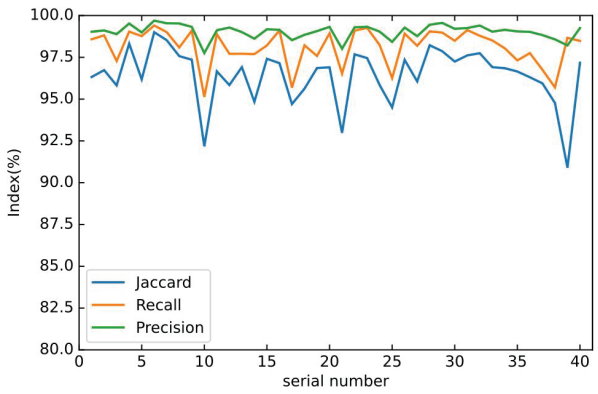

Fig. 3. The segmentation index results of STDC network

We applied ENet, BiSeNet, STDC813 and STDC1446 to the same kidney image, and overlaid the segmentation results on the original image, as shown in the example of segmentation results in the Fig. 4.

Due to the unclear edge between kidney noise and background noise, the results of ENet and BiSeNet showed under segmentation and over segmentation. Additionally, the edge in the segment is not smooth enough. Compared with the above two networks, the segmentation results of our proposed network show fewer rough edges and significantly improved accuracy.

As shown in Table I, we also reported the quantitative evaluation results of different segmentation methods in terms of network parameter quantity, FLOPs, FPS, Jaccard, Precision, and Recall. FPS represents the number of segmented images per second for different methods under experimental conditions. From the table, we can see that our method achieved the best results on three segmentation evaluation indicators. Compared with the segmentation results of other methods, our method has fewer parameters, higher segmentation accuracy, and achieves satisfactory segmentation speed, effectively balancing segmentation speed and accuracy.

TABLE I. PERFORMANCE INDICATORS OF FOUR NETWORKS

Method	***Params***	***FLOPs***	***FPS***	***Jaccard(%)***	***Precision(%)***	***Recall(%)***
ENet	**0.35M**	1.77G	56.6	96.4	98.03	98.27
BiSeNet	23.07M	33.1G	**138.2**	95.63	97.06	**98.46**
STDC813	8.4M	**813M**	93.8	96.62	98.29	98.24
STDC1446	12.5M	1446M	68.4	**96.66**	**98.37**	98.21

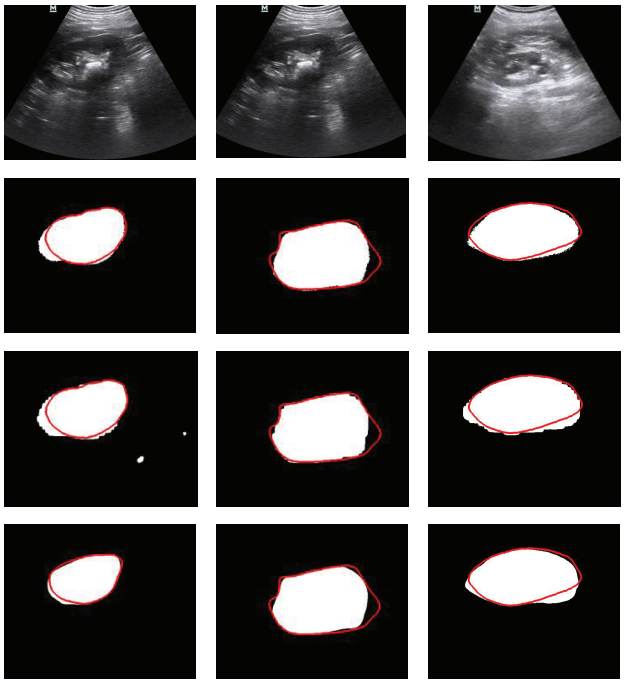

Fig. 4. Above: Original ultrasound image; Second row: ENet; Third row: BiSeNet; Last row: our method; Red line: Grand Truth; White area: segmentation result.

IV. CONCLUSION

In this paper, we address the demand for real-time segmentation of KDU images and present a deep learning segmentation method based on the STDC network. The proposed network utilizes the STDC module as the base block, learns spatial information in low-level layers through a single-stream approach, incorporates the Detail Aggregation module, and ultimately fuses low-level and deep features to predict the final segmentation results. Our experimental demonstrate that our method excels in multiple aspects, outperforming current networks in various metrics. As a future work, we plan to extend this method to the segmentation of dynamic ultrasound images of other organs, such as the liver, heart and breast, to enhance the generalizability and applicability of the method. Furthermore, we contemplate integrating KDU images with MRI and CT scans for joint segmentation. This integration aims to enhance segmentation precision and provide a more comprehensive diagnostic insight into diseases.

ACKNOWLEDGMENT

We would like to acknowledge Weite Feng, Yuming Liu, Jingjing Yin and other labmates for helpful discussions.

REFERENCES

[1] Xu Kewen, Xu Bo, Wu Ying, Xu Haoran.Overview of the application of machine learning in ultrasound images [J] .Computer Engineering and Application, 2021,57 (04): 11-17.

[2] Zhang Ning. Kidney ultrasound image segmentation based on deep learning[D].2022.DOI:10.27262/d.cnki.gqdau.2022.001758.

[3] CANNY J. A computational approach to edge detection [J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1986, PAMI-8(6): 679-698. DOI: 10.1109/TP AMI.1986.4767851.

[4] OTSU N. A threshold selection method from gray-level histograms[J/OL]. IEEE Transactions on Systems, Man, and Cybernetics, 1979, 9(1): 62-66. DOI: 10.1109/TSMC.1979. 4310076.

[5] Jonathan Long, Evan Shelhamer and Trevor Darrell, "Fully Convolutional Networks for Semantic Segmentation", arXiv:1411.4038 [cs.CV]

[6] V. Badrinarayanan, A. Kendall and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 12, pp. 2481-2495, 1 Dec. 2017, doi: 10.1109/TPAMI.2016.2644615.

[7] O. Ronneberger, P . Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," arXiv: 1505.04597 [cs.CV].

[8] Adam Paszke, Abhishek Chaurasia, Sangpil Kim and Eugenio Culurciello, "ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation," arXiv:1606.02147 [cs.CV].

[9] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu and Nong Sang, "BiSeNet: Bilateral Segmentation Network for Real-time Semantic Segmentation," arXiv:1808.00897 [cs.CV].

[10] M. Fan et al., "Rethinking BiSeNet For Real-time Semantic Segmentation," arXiv:2104.13188 [cs.CV]

[11] Wang Linlu.The Research and Development Trend of Medical Ultrasonic Image Division Technology [J].Imaging Research and Medical Application, 2018,2 (24): 252-254.

[12] Y. Hu et al., "Fully Automatic Pediatric Echocardiography Segmentation Using Deep Convolutional Networks Based on BiSeNet," 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 2019, pp. 6561-6564, doi: 10.1109/EMBC.2019.8856457.

[13] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh and Jianming Liang, "UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation," arXiv:1912.05074 [eess.IV]