A-PSPNet: A novel segmentation method of renal ultrasound image*

Pengceng Wen[1], Yu Guan[1], Jianqiang Li[1], Xi Xu[1], Haoran Peng[1],
Jijiang Yang[2], Yanhe Jia[3], Xianghui Xie[4], Minglei Li[4], Xiaoman Wang[4], Yue Xin[4], Yuzhu He[4]

***Abstract*— Hydronephrosis is a common renal disease in children which can lead to a series of complications, and ultrasonography is a basic examination usually performed on suspected hydronephrosis patients. If we can use deep learning approaches to judge and grade the disease in the ultrasonic examination stage, we can save a lot of manpower, medical resources, money, and help the suffered patients. For the semantic segmentation of kidney ultrasound image, we designed an attention-based Pyramid Scene Parsing Network, the core of which is the basic feature extraction network combining CBAM attention module and pyramid analysis module. Experiments were carried out on a hydronephrosis dataset containing 1850 annotated ultrasound images, including the arrangement of attention units, statistical computing power, and comparison of the effectiveness (MIoU and MPA) between the benchmark and our proposed method. Our constructed model achieved better segmentation performance than benchmarks with only little extra overhead, which validated the lightness and effectiveness of the model.**

I. INTRODUCTION

Hydronephrosis is a common kidney disease in children, it has an incidence rate of more than 1%. As a common cause of hydronephrosis in children, ureteropelvic junction obstruction (UPJO), which indicates obstruction at or along the pelvic-ureteral junction, can lead to a series of complications from abdominal mass, hematuria, to uremia, hypertension, or even renal rupture. Ultrasonography is a basic examination that is commonly performed on suspected hydronephrosis patients, which is convenient, time-saving, economical, and nonradiative.

The combination of medical images and artificial intelligence is prevailing in recent years. AI medical imaging technology has largely alleviated the shortage of medical resources and the imbalance between different regions[12], plus facilitate doctors' diagnostic abilities[10, 16, 8, 5]. Semantic segmentation is a mainstream AI-assisted diagnosis technology of medical images nowadays which is a pixel-level classification where all pixels belonging to the same category are grouped. If we can use deep learning approaches to judge and grade the disease in the ultrasonic examination stage, we can save a lot of money, manpower, medical resources, and help the suffered patients.

With the popularity of deep learning methods, Full Convolutional Network (FCN) emerges as the times require[19], yet its sampling operation reduces the image resolution, therefore, resulting in the weakening of the position information. However, since the false positive rate is very low in medical image processing, rich location information is needed to increase the reliability of the results or better explain the global context of the image[22]. Further on the basis of FCN, the Pyramid Scene Parsing (PSP) method which takes residual network (ResNet)[13] as the backbone is proposed in order to obtain the advanced feature layer with more global context information by its pyramid pooling module[28]. Attention mechanism mimics the way people look at objects, focusing on the prominent part of an image rather than giving equal weight to all areas of the image, started to play a more important role in medical image processing since the crucial organ regions or lesion regions features can directly draw its attention. Woo *et al*. [24] proposed a method to embed the attention module into the CNN network, but it destroyed the structure of the original network, made the pre-training weight file unusable, and increased the network complexity, thus increasing the computational burden of the model. We, therefore, combined PSPNet with the Convolutional Block Attention Module (CBAM) in a skillful way, which means that in order to reduce the complexity of the model, only one layer of attention is added before and after the basic feature extraction network respectively. **Contribution** Our main contribution is three-fold.

1) We propose an improved network structure and apply it to the semantic segmentation of ultrasonic images of hydronephrosis, and achieved a good segmentation performance.
2) We successfully combine CBAM with ResNet in a skillful way, which can be widely used to improve the characterization ability of the basic feature extraction layer, and the method does not destroy the original network structure of ResNet, so the pre-training model can still be used.
3) We show that better performance is possible with CBAM at minimal cost, and we verify that the added burden of the model is negligible.

[1]P. W., Y. G., J. L., X. Xu, H. P. are with Faculty of Information Technology, Beijing University of Technology, Beijing, 100124, China wpc18581311918@outlook.com,guanyu0010@126.com, lijianqiang_bjut@126.com, gcaxuxi@163.com, HaoranPeng@emails.bjut.edu.cn

[2]J. Y. with the Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing, 100084, China jijiang_yang@126.com

[3]Y. J. is with School of Economics and Management, Beijng Information Science Technology University, Beijing, 100192, China yhejia@bistu.edu.cn

[4]X. Xie, M. L., X. W., Y. X., Y. H. are with Beijing Children's Hospital, Capital Medical University, National Center for Children's Health, Beijing, 100045, China xiexianghui@bch.comn.cn, dr.liminglei@foxmail.com, pcchty@yeah.net, pcchng@yeah.net, 13699179332@163.com

For a better understanding of this article, our organization is as below: In the second section, we briefly review the literature on feature extraction networks and attention mechanisms. In the third section, we present our overall network structure and the optimization method adopted. In the fourth section, we introduce the details of our data set, evaluation criteria, and network training. In the fifth section, we compare our model with other networks and perform ablation experiments, then explain the results. Finally, we discuss and summarize the whole paper.

II. RELATED WORK

Artificial intelligence has long been applied to various fields including medical image processing and has greatly facilitated doctors' diagnostic abilities. It constantly evolves, especially in these years.

1) Network engineering.: Extensive architectures have been proposed since the successful implementation of a large-scale CNN[17]. Bengio *et al.* [1] state that if more complex problems are to be solved, either depth or width needs to be increased. Delalleau *et al.* [6] show that for some specially constructed polynomial functions, the shallow network needs the number of neurons growing exponentially so that its fitting effect can match the polynomial growing deep network, while the generalization ability of the wide and shallow network is poor. Due to the difficulty of gradient propagation, the simple increase of depth reaches saturation and even reduces the performance of the network, while ResNet proposes a simple identity skipping connection to alleviate the optimization problem of the deep network.

2) Advanced feature extraction.: Low-level features such as contours, edges, colors, textures, and shapes contain more positional and detailed information but less semantic knowledge compared to higher-level features. While semantic segmentation is a pixel-level classification task that requires learning more advanced features to improve the performance of the network. Many multi-scale feature fusion methods are proposed accordingly[4, 2, 11, 25]. Chen *et al.* [4] introduces a scheme to deal with pixel classification problems by combining CNN and a probability graph model, but this mode has high complexity and high training cost. Hariharan et al. present a hyper-column pattern to achieve fine-grained localization of the target and simultaneous segmentation, yet it cannot capture the whole semantic information. Liu *et al.* [18] suggest an approach of the Single Shot Detector (SSD) that can separately predict multi-scale features and then synthesize the results, but it doesn't incorporate multi-scale features. Zhao *et al.* [28] design a structure of Pyramid Scene Parsing Network which uses a pyramid pooling layer to obtain advanced features and fuse them with low-level features into the full connection layer and achieved better performance.

3) Attention mechanism: The attention mechanism ignores irrelevant information and focuses on the crucial portion. It is widely used in various types of deep learning tasks such as natural language processing, image recognition, and speech recognition. According to whether it is differentiable or not, attention mechanism can be divided into two types, soft attention[26, 15, 23, 7] and hard attention[27, 21], which have distinct emphasis. When selecting information, soft attention calculates the weighted average of multiple input information and then inputs it into the neural network for calculation. On the contrary, hard attention selects the information at a single position in the input sequence, such as randomly selecting a piece of information or selecting the information with the highest probability, and this step is indifferentiable[20]. However, soft attention is differentiable, which means that the weight of attention can be learned by using a neural network to calculate the gradient and to propagate forward and feedback backward.

The CBAM, an extension of squeeze-and-excitation networks, is a kind of soft attention mechanism combining channel domain and spatial domain[14]. Its main advantage is the attention in spatial domains, that is, it not only teaches the network what to look at via the channel-wise attention but also where to look via the spatial-wise attention, moreover, it is extremely lightweight and easy to deploy end-to-end.

III. METHODS

For the renal ultrasound image segmentation, we develop an attention-based Pyramid Scene Parsing Network (A-PSPNet), which can focus on the critical portion, aggregate contextual information, and well accomplish the semantic segmentation task. The core of the A-PSPNet is a basic feature extraction network combined with the CBAM and the pyramid parsing module. The general architecture of this model is shown in Fig. 1 and detailed interpretations are as follows.

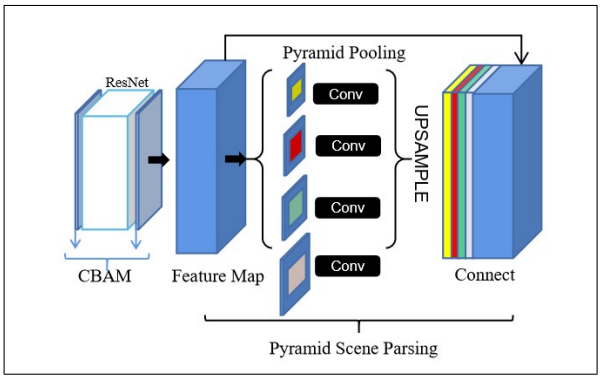

Fig. 1: The architecture of attention-based pyramid scene parsing network (A-PSPNet).

A. Improved ResNet

ResNet can retain the depth of the deep network meanwhile avoid the degradation problem. It also has low complexity, which means faster processing. These features grant that it is quite suitable for medical image semantic segmentation, especially for blurry ultrasonic images. About the structure of ResNet, its basic unit is a bottleneck (also called a building block), which consists of an identity block and a residual block, and is illustrated in Fig. 2. To better

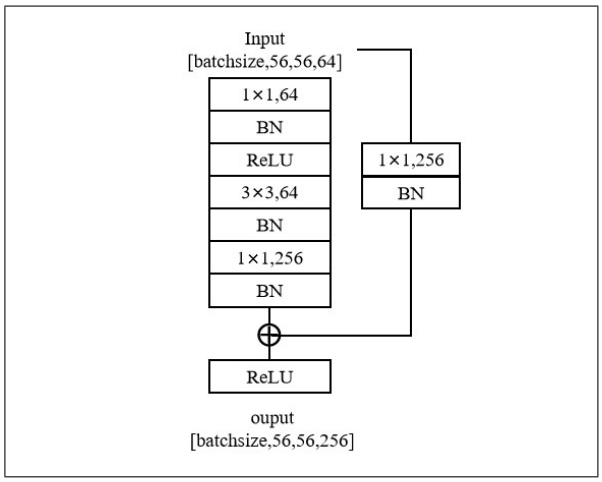

Fig. 2: The structure of a building block (a bottleneck). The identity block refers to the branch part and the residual block refers to the main part.

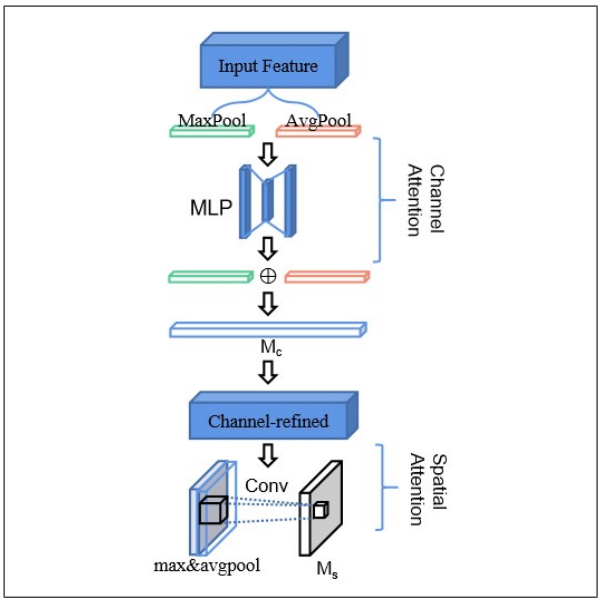

Fig. 3: The overview of a typical CBAM module. The module has both channel and spatial attention units, which are combined in a serial way

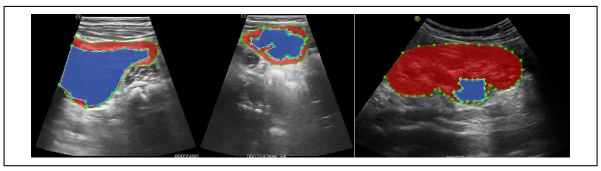

Fig. 4: Large differences of contours exist even in the same category in this dataset. In the left and middle renal ultrasound images, the lesion (blue part) occupies a large portion of the organ area (red part), while in the right image, the lesion area is small, and all the edge information is also fuzzy.

introduce the principle of the combination of CBAM and ResNet, which will be formally explained later, here we first define a building block at any depth as:

$$\mathbf{y_n} = \sum_{n-1}^{i=0} \mathbf{x_0} + \mathbf{F}\left(\mathbf{x_i}, \mathbf{W_i}\right), \mathbf{F} = \mathbf{W_2 ReLU}\left(\mathbf{W_1 x}\right) \quad (1)$$

where $\mathbf{ReLU}$ denotes the ReLU function, $\mathbf{F}$ denotes the residual function to learn.

We find a skillful way to add the CBAM without destroying the original structure of ResNet, that is, adding the same CBAM consist of the channel and spatial attention units in the front and back of ResNet, respectively. Since CBAM is a lightweight generic module, we seamlessly integrate it into the architecture without any extra overhead. The structure of a typical CBAM module is illustrated in Fig. 3. About our constructed UPJO ultrasonic dataset, there exists a non-negligible problem, that the same category in almost all images has quite different contours, for example in Fig. 4. Thus adding the CBAM module in the stage of extracting basic features also helps better identifying our interested part in the ultrasonic image. The CBAM module compute the attention map in turn along two independent dimensions (channel and spatial). Since the arrangement of how the attention units locate in the CBAM module can be changed while the form of the front and end CBAM is kept consistent, there arise four patterns to design a CBAM attention module: a spatial unit comes before a channel unit, a channel unit comes before a spatial unit, or adopt only one of them. We conducted experiments to compare these four patterns, and the best performed one, which is channel first and spatial second pattern as illustrated in Fig. 5, is finally preferred. Given an intermediate feature map $\mathbf{F} \in R^{C\times H\times W}$ as input, CBAM sequentially infers a 1D channel attention map $\mathbf{M_c} \in R^{C\times 1\times 1}$ and a 2D spatial attention map $\mathbf{M_s} \in R^{1\times H\times W}$. The overall attention process can be summarized as:

$$\mathbf{M(F)} = \{\mathbf{M_c}; \mathbf{M_s}\} \otimes \mathbf{F} \quad (2)$$

where $\otimes$ denotes element-wise multiplication. Steps for the output of channel attention, $\mathbf{M_c}$, are as below. We first aggregate spatial information of a feature map by using both average-pooling and max-pooling operations, then we obtain two C-dimensional pooling feature maps: $\mathbf{F^c_{avg}}$ and $\mathbf{F^c_{max}}$. Then, they are sent into the Multilayer Perceptron (MLP) containing a hidden layer to obtain two $1 \times 1 \times C$ channel attention maps. To reduce the number of parameters, the number of neurons in the hidden layer is C/R, which is also known as the compression ratio. Finally, the corresponding elements of the channel attention maps obtained through MLP are added and activated to get the final channel attention map $\mathbf{M_c}$. In short, the channel attention is computed as:

$$\mathbf{M_c(F)} = \sigma(\mathbf{W_1}(\mathbf{W_0}(\mathbf{F^c_{avg}})) + \mathbf{W_1}(\mathbf{W_0}(\mathbf{F^c_{max}}))), \quad (3)$$

where σ denotes the sigmoid function.$\mathbf{W_0} \in R^{C/r\times C}$, and $\mathbf{W_1} \in R^{C\times C/r}$ denote that the MLP weights. The output of spatial attention unit, $\mathbf{M_s}$, is computed as:

$$\mathbf{M_s(F)} = \sigma(f^{7\times 7}([\mathbf{F^s_{avg}}; \mathbf{F^s_{max}}])), \quad (4)$$

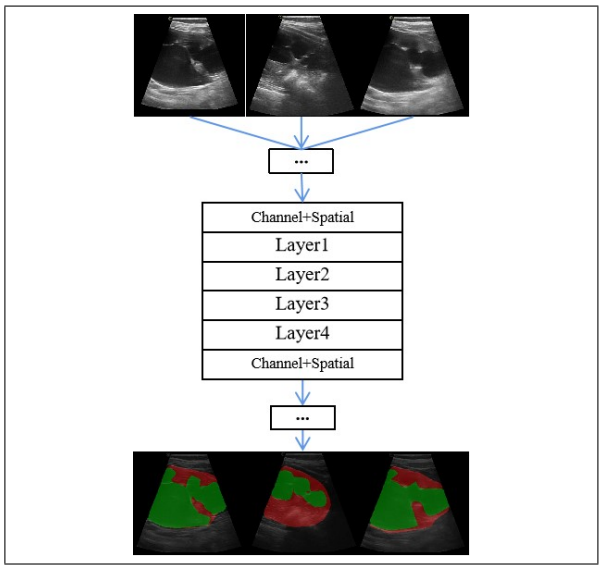

Fig. 5: The structure of ResNet with CBAM. Each layer consists of several bottleneck modules, and CBAM that is composed of the channel attention unit and spatial attention unit are embedded at the front and back ends to form the improved basic feature extraction network structure.

where σ denotes the sigmoid function and $f^{7\times7}$ represents a convolution operation with the filter size of 7×7. Finally, the two attention maps are multiplied with the input feature map for adaptive feature optimization.

Such an insertion can improve the capabilities of the model with almost no additional complexity. The final output is computed as:

$$\mathbf{f_{out}} = \mathbf{M_{end}}\left(\mathbf{y_n}\left(\mathbf{M_{start}}\right)\right) \tag{5}$$

where $\mathbf{M_{end}}$ and $\mathbf{M_{start}}$, respectively, denotes the CBAM embedded in ResNet. Each module consists of a channel attention unit followed by a spatial attention unit in series.

B. A-PSPNet

Using the improved ResNet, we preliminarily analyze the underlying features, including contour, edge, chroma, texture, and shape features. However, in ultrasonic image processing, we need to judge the organ region and lesion region in the image more accurately. Besides, since great differences of contours exist even in the same category, take Fig. 4 for example, the characteristics of organs and lesions in ultrasound images vary greatly from patient to patient, which means high-performance image segmentation cannot be carried out while there is only basic feature information.So we need more semantic information, and we send the feature vectors to the advanced feature extraction layer, which is the pyramid scene parsing module. As shown in Fig. 1, we divide the feature map into four different sub-regions and get pooled feature representations at different locations. Then using 1×1 convolution layer after each pyramid level to reduce the dimension of context features into low-dimensional feature representations. After that, the low-dimensional feature map is directly up-sampled to make its scale the same as the original feature map. Finally, the feature maps of different levels are stitched together into the final pyramid pooled global feature. When the segmentation layer has more global information, the probability of false segmentation will be lower.

C. Combo Loss

In addition to the previous problem, another problem is that in the ultrasonic image data, the number of lesion regions is not balanced with the number of organ regions, which leads to the imbalance of the data set. Therefore, we seek solutions in the part of the loss function and form a combo-loss function (Combo Loss), which consists of the Cross Entropy (CE) loss function and Dice loss function. We leverage the Dice similarity coefficient to deter model parameters from being held at bad local minima and meanwhile gradually learn better model parameters by penalizing for false positives/negatives using a cross entropy term. In short, we formally defined it as:

$$\mathbf{ComboLoss} = 1 - \sum \mathbf{y_t log}\left(\mathbf{y_p}\right) - \frac{2\sum \mathbf{y_t y_p}}{\sum \mathbf{y_t^2} + \mathbf{y_p^2}} \tag{6}$$

where $\mathbf{y_t}$ and $\mathbf{y_p}$ respectively denotes the true and predicted values. This loss function can effectively alleviate the problem of multi-party dominant learning caused by unbalanced datasets.

IV. EXPERIMENTS

A. Dataset

All of our experiments are carried out on the same UPJO dataset which contains 1850 ultrasonic images from 17 outpatients. The source data are collected by Beijing Children's Hospital from Sept. 10, 2019 to Mar. 10, 2020, which consists of 174 screenshots taken by a professional sonographer in the coronal and transverse planes of the kidney (the standard positions), as well as 120 screen videos recording sonographers' operations during the ultrasonic scan. All the source videos are sliced into images at a frame rate of size 5 and images that are too vague to be distinguished are discarded. Then experienced doctors annotate each image with hydrops and kidney tags by using an image annotation software Labelme. Finally, 1850 annotated UPJO ultrasonic images are obtained after such preprocessing, and we further divide this dataset into a training set and a validation set in a 9:1 ratio in our experiments.

B. Implementation Details

We choose PSPNet as the basic model through comparative experiments, and combine ResNet with CBAM, then we use attention-based ResNet50 as the backbone of our model. We set the learning rate at 0.001 and the training epoch at 100, the input resolution is 473×473, and use the weight of the pre-trained model to optimize the learning process. In addition, other models of feature extraction networks based on ResNet are trained for comparison, and the same learning rate, optimizer, and training algebra are maintained.

All of our experiments are conducted using Intel CPU and NVIDIA GPU. Since our GPU memory size is 8GB, thus we use batch training and the batch size is set to be 8. Then we train our models until convergence by using the adaptive moment estimation (ADAM) optimizer (a common optimization algorithm). Some samples we predict with our model are shown in Fig. 6.

C. Performance Indicators

We evaluate the effectiveness of the hydronephrosis ultrasonic image segmentation model by calculating Mean Intersection over Union (MIoU) of true and predicted values and Mean Pixel Accuracy (MPA), which can be defined as:

$$\mathbf{MIoU} = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ij}-p_{ii}} \quad (7)$$

$$\mathbf{MPA} = \frac{\sum_{K}^{i=1}\sum_{n}(R_n - R_{n-1})P_n}{K} \quad (8)$$

where p_{ij} denotes the number of j that is predicted to be the true value of i, and $k+1$ is the number of classes, the R_n denotes the value of recall and P_n denotes the value of precision

To thoroughly describe the performance of a specific deep learning model, in addition to the accuracy, the complexity of the model like the number of parameters and the amount of calculation should also be considered. To increase its performance in accuracy and speed, we integrate only independent CBAM in the ResNet model at the top and bottom positions, respectively. To find out how the attention units locate in the CBAM module influence the overall results, we conduct the previously mentioned comparison experiments on two single units and two tandem combinations, respectively. Results are shown in Table I. The MIoU and the MPA of the final model that was trained with 50 epochs (the remaining parameters are consistent in all experiments) are used as the evaluation indices. Besides, to confirm that the network structure of our proposed attention-based ResNet is lighter than the original ResNet, we compare the number of parameters, and the calculation of Giga multiply and add per second (GMACs) to measure the computing power of each method in Table II.

V. RESULTS

A. comparison of the arrangement of attention units

We conduct experiments to compare the four arrangements of attention units in CBAM and receive results.

TABLE I: The Comparison of The Arrangement of Attention Units

Description	MIoU	MPA
ResNet50 + Spatial attention unit	87.39	93.03
ResNet50 + Channel attention unit	87.16	93.05
ResNet50 + Spatial&Channel attention units	87.17	92.78
ResNet50 + Channel&Spatial attention units	**87.43**	**93.53**

As is in Table I, we evaluate the performance of the final model with mixed domain attention units in different combinations. Through comparison, it is found that the performance gap of each method is not obvious, and the largest difference value in MIoU is only 0.27. Relatively speaking, the attention model that goes through the channel domain first and then the spatial domain has the best effect.

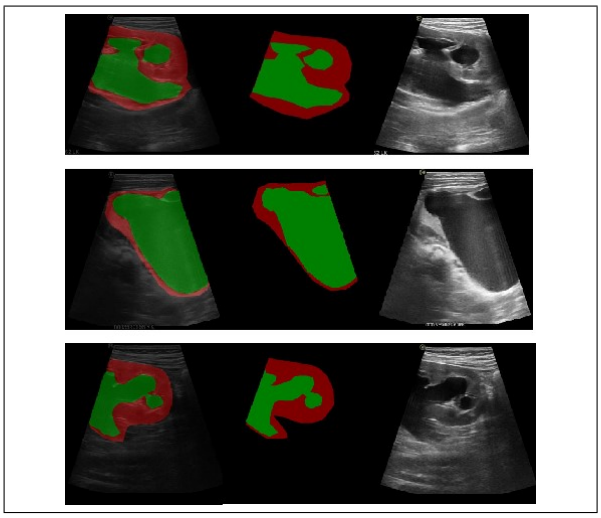

Fig. 6: The labeled images are in the middle (the red part refers to the kidney area, the green part refers to the hydronephrosis area), the predicted results are on the left, and the original images are on the right.

B. comparison of complexity

We evaluate the computing performance of some classic models and compared them with the improved Resnet50. Besides, we separately evaluated the computational parameters of the spatial attention unit and the channel attention unit to demonstrate their lightness.

TABLE II: The Comparison of Complexity

Description	Params(M)	MACs(G)
ResNet50	25.55703	18.84624
Spatial attention unit	0.00020	0.00004
Channel attention unit	0.52480	0.00378
A-PSPNet	**26.08203**	**18.85144**
ResNet101	45.07416	35.96785
ResNet152	60.7178	53.10086
AlexNet	61.10084	3.09258
VGG16	138.35754	68.2407
DenseNet	7.97886	12.74194

We evaluate and compare the complexity between baselines and our proposed model. Besides, we also calculate the computational statistics of the spatial attention unit and the channel attention unit, and both experimental results are given in Table II. In general, compared with the original ResNet structure, the spatial attention unit is 5 orders of magnitude smaller in terms of the number of parameters, the channel attention unit differs by 2 orders of magnitude, and the calculation amount differs by 6 and 4 orders of magnitude, respectively, which demonstrate their lightness. The overall attention-based ResNet has only about 2% more

parameters and 0.05% more computations than the original structure. Compared with other benchmark network structures, the lightness is also a major advantage of this model.

C. comparison of effectiveness

We selected some other baseline methods that also take ResNet for feature extraction [9, 3], and conduct all the experiments. The model that outperforms others was finally obtained to combine with our lightweight CBAM, and then used as the final image segmentation tool. Experimental results are shown in Table III. Overall, we chose PSPNet to be the final image segmentation tool and improve it by combining it with two lightweight attention units, and we performed contrast experiments to thoroughly evaluate the effectiveness of the final model. We verify that ResNet with CBAM outperforms the other same baseline without bells, results shown a 1.32% improvement in performance compared to the original PSPNet and even greater increment than other popular segmentation networks.

TABLE III: The comparison of Effectiveness

Methods	MIoU	MPA
PSPNet	86.61	92.01
Deeplabv3	85.47	91.58
DANet	80.54	83.14
A-PSPNet	**87.93**	**93.52**

VI. CONCLUSIONS

In the application of renal ultrasound image segmentation, we hope to improve the segmentation accuracy of the model at a very small cost, therefore, we propose an attention-based residual network structure and applied it to the pyramid scene parsing network model. Given an input and obtain the initial feature map, our model inferred the attention map in turn along the channel and spatial dimensions, then multiplied the attention-map by the input feature map and input it to the residual network structure, and repeated the process of attention at the output and then served as the input of the next module. Because of the lightness of attention module, its overhead can be ignored, and it can effectively improve the accuracy of the model. We validated our model through experiments on data sets obtained and labeled from Beijing Children's Hospital. Our experiment shows that compared with the model before adding the attention module, the calculation amount and the number of parameters of the model increase less and the performance improves, which proves the lightness and effectiveness of the model.

ACKNOWLEDGMENT

This study is supported by the National Key RD Program of China with project no. 2019AAA0104904.

REFERENCES

[1] Yoshua Bengio, Yann LeCun, et al. "Scaling learning algorithms towards AI". In: *Large-scale kernel machines* 34.5 (2007), pp. 1–41.

[2] Liang-Chieh Chen et al. "Attention to scale: Scale-aware semantic image segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 3640–3649.

[3] Liang-Chieh Chen et al. "Rethinking atrous convolution for semantic image segmentation". In: *arXiv preprint arXiv:1706.05587* (2017).

[4] Liang-Chieh Chen et al. "Semantic image segmentation with deep convolutional nets and fully connected crfs". In: *arXiv preprint arXiv:1412.7062* (2014).

[5] Jie-Zhi Cheng et al. "Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans". In: *Scientific reports* 6.1 (2016), pp. 1–13.

[6] Olivier Delalleau and Yoshua Bengio. "Shallow vs. deep sum-product networks". In: *Advances in neural information processing systems* 24 (2011), pp. 666–674.

[7] Yang Du et al. "Interaction-aware spatio-temporal pyramid attention networks for action classification". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 373–389.

[8] Andre Esteva et al. "Dermatologist-level classification of skin cancer with deep neural networks". In: *nature* 542.7639 (2017), pp. 115–118.

[9] Jun Fu et al. "Dual attention network for scene segmentation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 3146–3154.

[10] Varun Gulshan et al. "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs". In: *Jama* 316.22 (2016), pp. 2402–2410.

[11] Bharath Hariharan et al. "Hypercolumns for object segmentation and fine-grained localization". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 447–456.

[12] Jianxing He et al. "The practical implementation of artificial intelligence technologies in medicine". In: *Nature medicine* 25.1 (2019), pp. 30–36.

[13] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[14] Jie Hu, Li Shen, and Gang Sun. "Squeeze-and-excitation networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7132–7141.

[15] Max Jaderberg et al. "Spatial transformer networks". In: *arXiv preprint arXiv:1506.02025* (2015).

[16] Daniel S Kermany et al. "Identifying medical diagnoses and treatable diseases by image-based deep learning". In: *Cell* 172.5 (2018), pp. 1122–1131.

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25 (2012), pp. 1097–1105.

[18] Wei Liu et al. “Ssd: Single shot multibox detector”. In: *European conference on computer vision*. Springer. 2016, pp. 21–37.

[19] Jonathan Long, Evan Shelhamer, and Trevor Darrell. “Fully convolutional networks for semantic segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.

[20] Volodymyr Mnih et al. “Recurrent models of visual attention”. In: *arXiv preprint arXiv:1406.6247* (2014).

[21] Marijn Stollenga et al. “Deep networks with internal selective attention through feedback connections”. In: *arXiv preprint arXiv:1407.3068* (2014).

[22] Saeid Asgari Taghanaki et al. “Deep semantic segmentation of natural and medical images: a review”. In: *Artificial Intelligence Review* (2020), pp. 1–42.

[23] Fei Wang et al. “Residual attention network for image classification”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 3156–3164.

[24] Sanghyun Woo et al. “Cbam: Convolutional block attention module”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 3–19.

[25] Fangting Xia et al. “Zoom better to see clearer: Human and object parsing with hierarchical auto-zoom net”. In: *European Conference on Computer Vision*. Springer. 2016, pp. 648–663.

[26] Tianjun Xiao et al. “The application of two-level attention models in deep convolutional neural network for fine-grained image classification”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 842–850.

[27] Bo Zhao et al. “Diversified visual attention networks for fine-grained object classification”. In: *IEEE Transactions on Multimedia* 19.6 (2017), pp. 1245–1256.

[28] Hengshuang Zhao et al. “Pyramid scene parsing network”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2881–2890.