**Authors’ response to reviews**

We appreciate the editor and reviewers for their constructive suggestions on our manuscript entitled “A-PSPNet: A novel segmentation method of renal ultrasound image” (ID: 250). Corresponding revisions are made and highlighted in the manuscript. Comments of reviewers (in *italic*) are responded (in blue) one by one.

**Reviewer #15293:**

*It is an interesting research. The subject is relevant. The novelty is clear. The writing is OK. Although, I have a few comments about the text and the methodology.*

1. *Abstract: Clarify the abbreviations. “A-PSPNet”, which is in the title, is not fully understandable in the abstract, as “attention-based” is not capitalized and neither the abbreviation is followed inside parentheses. So far, “CBAM” is also unknown. I believe MIoU and MPQ is not worth being written in the abstract.*

Thank you for reminding us. CBAM represents for Convolutional Block Attention Module. We have standardized our writing and made other corresponding revisions..

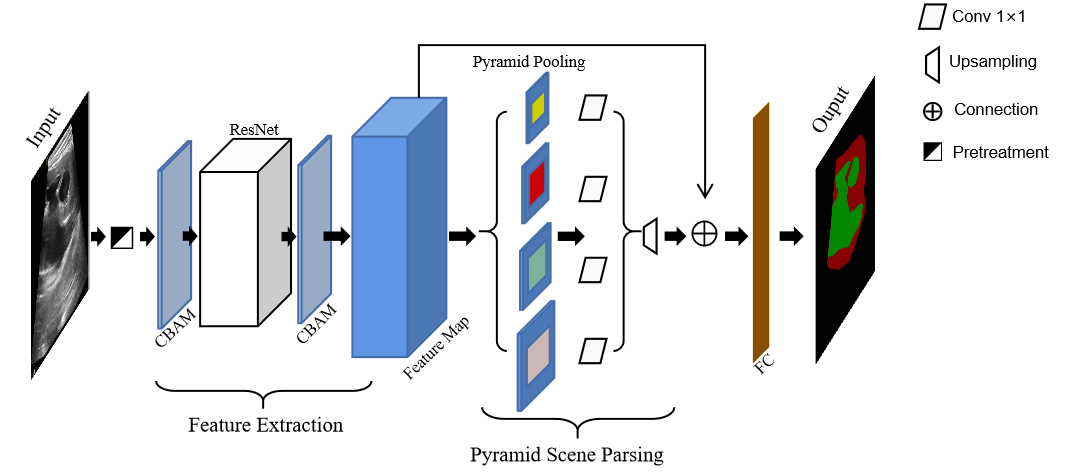
1. *Introduction: I do not understand what the authors mean with the phrase: “FCN emerges as the times require”. Also, I believe the bold word “Contribution” should be removed.*

Sorry for being unclear. We rephrased our expression as “Full Convolution Network (FCN) came into being for the semantic segmentation problem”. Also we removed the bold “Contribution”.

1. *Related work: There is a missing citation in “...training cost. Hariharan et al. present a hyper-column...”.*

Thank you for reminding us. We have added the corresponding citation.

1. *Fig.1: It is hard to see the whole picture in this figure. There is no input or output. The bottleneck is the unit of a ResNet which is inside a CBAM? How is each pyramid pool different from each other? After the pyramid parsing, what happens?*
2. Considering the suggestion, we have re-made the Fig.1 to solve the ambiguity of the combination of “CBAM” and “ResNet”.In addition, considering the various permutation and combination, inorder to simplify the experiments, the CBAM before and back are the same.

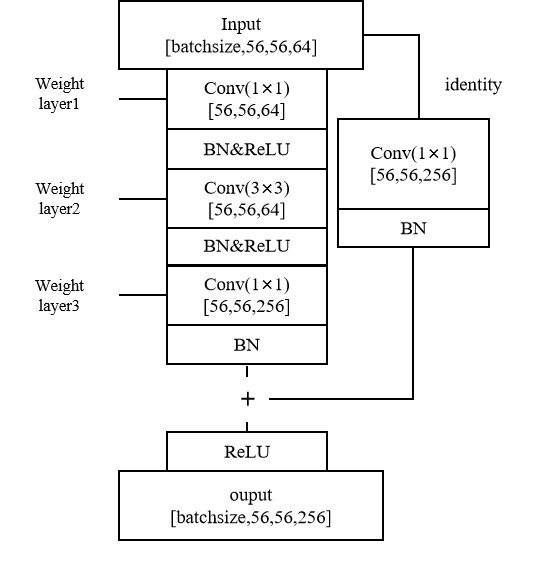


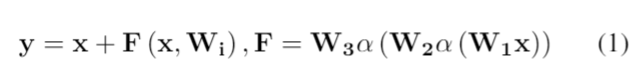
2. We have added “There are two core ... result output” in the 1st paragraph of Section 3 to better introduce the entire network structure (include your fourth question).

3. We have added “in the pyramid ... 6×6” in the 1st paragraph in of Section 3-B

1. *Fig.2 & Eq.1:*
2. *Specify the information of each dimension of the input and output.*
3. *I do not fully understand it. Some variables are not defined.*

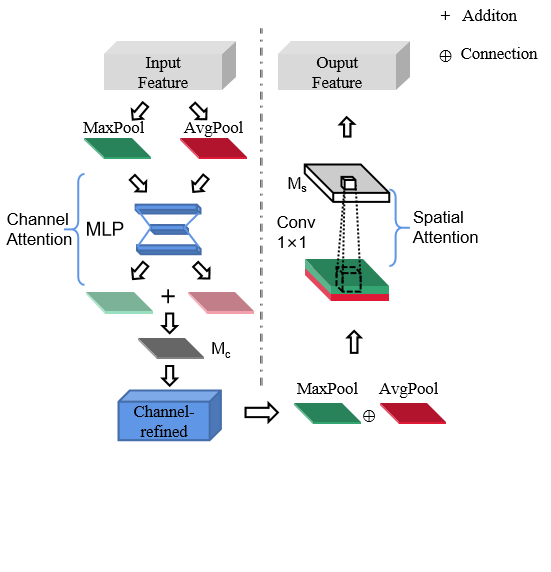
We have re-made Fig.2 according to the comments and added text labels on it to allow readers to

better understand Eq.1. Considering the suggestion, we have defined all variables and re-written the Eq.1. The W1~W3 correspond to the weight layers in Fig.1



1. *Fig.3 & Eq.3 :*
2. *Isn’t the MLP in the wrong direction?*
3. *Does the “plus symbol” mean sum? Fig. 3 indicates an operation of concatenation.*

We are sorry that we mistyped the plus sign on Fig.3(The plus symbol in Eq.3 means sum.),



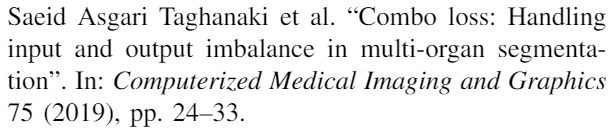
and we have re-written related description according to the suggestion and re-made Fig.3.

1. *III-B. A-PSPNet: how the upsampling is computed?*

We are very sorry for neglecting the describtion of upsampling’s computition, and we have added “by the method of bilinear interpolation” in the 1st paragraph of Section 3-B.

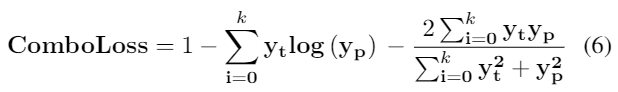
1. *III-C. Combo Loss: Combo Loss: Interesting function, a reference should be included for readers that would like more details about it.*

We have added citation according to the comments.



1. *Eq.6: Is the equation complete? Isn’t it missing some parameters of the sum symbols?*

We are very sorry that our carelessness caused your confusion and we have re-written the equation. Instructions to equation are as follows: yt and yp respectively denotes the true and predicted values, and k denotes the number of classes in the semantic segmentation area(In the experiment are organs area, lesion area and background respectively).



1. *IV-A. Dataset: the database is very interesting and I believe the authors should make it publicly available, if possible. As I understood, the database holds several images from the same 17 patients, most of them extracted from frames of videos, so it might have images very similar and, if the sets were randomly divided, they might have very similar images from the same video of the same kidney into both training and validation sets, making the segmentation easier. Also, the experiments were conducted with just one holdout. Considering a private database, possibility of similar images of the same kidney in training and validation set and, finally, just one holdout, I believe the results are slightly questionables. The experiments should be executed with cross-validation or several random holdouts. Also, the same kidneys should be either the training or the validation set, never in both of them. More explanations about the definition of the experimental protocol should be included in the paper, to answer questions like:*

*-> Why is the database not publicly available on some repository such as Figshare?*

*-> Why is it only one holdout? Why not several holdouts or cross-validation? Is it a computational time issue? Is it a common practice in (medical) image segmentation?*

*-> Why use the rate 9:1 and how was it divided? Randomly?*

Thank you very much for your meticulous comments. The answers are as follows

1. We are very sorry that the data cannot be made public. In order to reassure your worries, the reasons are as follows: due to the sensitivity of the data, researchers cannot download or back up the data privately.

2. We are very sorry that we did not use a more rigorous and excellent cross-validation method to prove our results. Due to the non-publicity of the data, we cannot directly compare the accuracy of other deep learning models to our data, so we have to use all models train the data once, and ensure that the parameters of all models are consistent and the data are consistent during the training process, and then select the best. As you said, if cross-validation is used, it will be very computationally expensive.

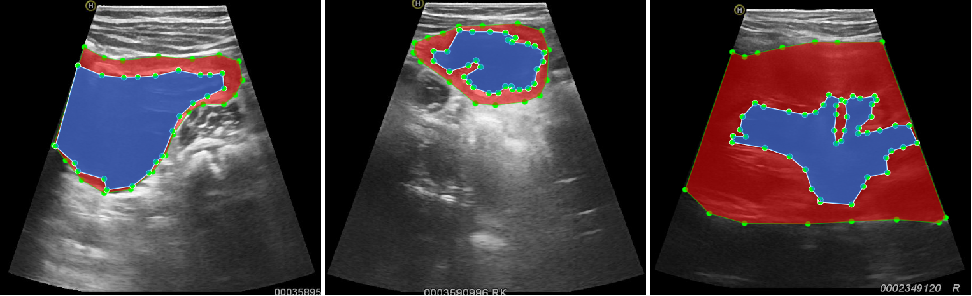
3. The ratio of 9:1 is a common way when there is only one holdout, if the amount of data is small (less than 1000), there is also a choice of 8:2. And as you might guess, the data is divided randomly. In order to appease your confusion about data similarity, here are some explanations, although we can't directly manipulate the source data, we have explained the requirements and labeling requirements with professional doctors before, including as far as possible to ensure that each frame of image is different from each other. Therefore, the method chosen by the doctor is to extract images at intervals of 5 frames and discard blurry and unmarkable images.

All above explanation I have added in the Section 4 selectively

1. *III-B. Implementation details: Fig. 4 indicates the images have different sizes. How is the input resolution fixed? The initial weights were generated by a general purpose database?*

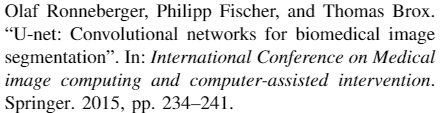


To ease your doubts, firstly, Fig 5 is a part of the pictures I get from the dataset. The original picture is shown above, so the size of all pictures is the same. Our intention is to show the marks of the data, thus cutting unnecessary boundary, but in the experiment, we do not cut out pictures as shown in Fig 4. Then, considering the comments, we have added ”Therefore, in the stage of loading data, we will reset the size of the image. Since the length and width of the image are not equal, in order to avoid distortion, we fill the image into a square by adding gray borders and then reset the size” in Section 4-B.



1. *IV-C. Performance Indicators: it would be good to have references of other segmentation researches that use the same indicators.*

In the field of medical image processing, reference rate of the U-Net is the highest, the indicator they use is exactly the MIOU, and MIOU has also been widely used in the field. so we quote this method. And we have made added ”With reference ... image segmentation” and “In actual applications...we take the calculation of Giga” in the 3rd paragraph of Section 4-C according to the comments.



1. *Eqs.7,8: are they corrected? There is a parameter missing in a sigma. “k+1 is the number of classes”, what classes?*

1、Sorry for our careless mistake and we have corrected the equation. Actually, k is the number of area classes in the labeled data (+1 denotes the class of background additionally), and we have amended our statement.

2、We are very sorry for your incomprehension due to our negligence. We have made correction according to the comments, and the classes is the number of the divided areas (“Organs, diseased areas”+”background” )

1. *Results: check the capitalization of the name of the subsections. The phrase “Relatively speaking, the attention model that goes through the channel domain first and then the spatial domain has the best effect.” demonstrate a quite simple idea, but it is really hard to understand.*

We are very sorry that we can't give you a causality on this issue. However the creator of CBAM just give an empirical relationship, they only used experiment to surface a result. From a spatial viewpoint, the channel attention is globally applied, while the spatial attention works locally, the might guess is that use channel attention firstly can obtain more accurate parameters.

1. *Table2: it does not seem fair to compare one attention unit with the whole ResNet50.*

We very much agree with your point of view. It is indeed unfair to compare the attention module with the complete network structure ResNet, we also thought about this issue before submitting the paper, which forced us to take five decimal places. However, our original intention is to exchange the smallest possible price for a larger improvement, which is also the meaning of this part of the content. Therefore, we wish a reservation of this comparison.In addition, we have added experiments on DANet and Deeplabv3 on this indicator, and added “...as shown in Table 2, A-PSPNet increases the amount of network parameters by about 1.13% and the amount of calculation by 0.01% compared with PSPNet...” in last paragraph of the Section V-C.

1. *V-C. Comparison of effectiveness: Is the PSPNet from table 3 somehow related to the ResNet50 from table 2? If they are, the gain in the indicators are almost the same as the overhead. If they are related, it should be discussed. If they are not related, it should be clarified. Why are the networks from table 3 not discussed in table 2? If the lightness is so important, at least the models that present the top 2 indicators in table 3 should be compared in table 2. Also, I believe it should be 1.32 percentage point (pp), not “a 1.32% improvement”.*

We are very sorry that we wrote the ambiguous "A-PSPNet" in Table 2. Its original meaning refers to the basic feature extraction modules in A-PSPNet (exclude Pyramid analysis module) to compare other models, which is the improved ResNet50(Feature Extraction Module). Then we have made correction according to the comments.

1. *VI-Conclusion:*

*-> the overhead can not be ignored, but it can be compared to the gain in the results.*

*-> “...experiments on data sets obtained...” I believe they were obtained with one database.*

*-> “...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.” I believe it should be “Our experiments show”. The parameters of the model Increase less than what?*

*-> I believe the results “indicate” or ”demonstrate” the effectiveness, due to the issues of the experimental methodology and the database, also due to the fact that the experiments were conducted in only one non-openly database and without statistical tests comparing results from other well-known models.*

Thank you very much for your meticulous comments. We have re-written them and made correction according to the comments.

Conclusion:

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. We tested several models, and finally applied our method on the basis of PSPNet, and successfully obtained a performance improvement of 1.32 percentage points at a computational cost of 0.05%. We validated our model through experiments with one databse obtained and labeled from Beijing Children's Hospital. Our experiments show that compared with the model before adding the attention module, the calculation amount and the number of parameters of the model increase a little. and the performance improves, which demonstrate the lightweight and effectiveness of the model.

1. *References: they are not completely in the ieeetrans bibliography style.*

We are very sorry about the mistake.

**Reviewer 15307:**

*The experimental part is strong but the novelty of the proposed method could be further reinforced.*

*The authors are suggested to better define "lightness".*

Thank you very much for your affirmation of our research, we have studied your comments carefully and have made revision which marked in red in the paper. We have tried our best to revise our manuscript according to the comments, and we have replaced “lightness” to “lightweight” and added descriptions in Section 4-C. Attached please find the revised version, which we would like to submit for your kind consideration.

**We would again like to thank the reviewers for their constructive comments that greatly help us revise and improve our manuscript.**