Reviews Response:

Review 1:

1. **Running author is still *F. Author et al.* in every even page from the template**

   I have changed the running author.

2. **It is better to change the citation *[6] developed* to *Luo et al. [6] developed* for readability**

   The citation has been changed.

3. **The paper depicts that the previous approaches dealt only with a small number of classes but provides no evidence of them performing worse in a large number of classes. Also, the paper doesn't introduce a novel way to address this issue.**

   In the text I do not mention that they perform worse with a larger number of classes, but rather that they are not tested in datasets with many classes (as is the case of this challenge) and that the complexity and size of the proposed frameworks will rise importantly. Hence the characteristic of the current challenge is different than the characteristics of the problem solved by the proposed frameworks.

4. **The citations like *2D U-Net ronneberger2015u models* don't provide any information on what the paper was. Please use proper citation style.**

   The citation to the paper has been corrected.

5. **No explanation on why 2D models were used for a 3D dataset**

   Due to memory and computational limitations, I implemented a 2D network. Moreover, as the challenge also focused on efficiency, the 2D CNN helped in this metric. This has been added to the manuscript.

6. **The equation for cross-entropy seems to be of binary case. Is the segmentation task formulated as a multi-label classification? A pixel can only belong to one class, so the CE loss equation should be used rather than its binary counterpart.**

   The cross-entropy equation has been updated to a multi-label classification task.

7. **No explanation for *a deep supervised layer with an auxiliary segmentation loss 12Lee2015 is located in the second-last up-sampling block***

   The reference and explanation of use has been added.

8. **No explanation on why only 3 models were selected among 5-fold cross-validated models. The obvious way would be to use models trained on all five folds.**

   Using only 3 models provided a better validation dice score than using the 5 models. Hence, we only used 3 models. This reason has been added to the manuscript.

9. **No explanation on what *ỹic* is. Although it can be inferred to be a pseudo-label, explicitly mentioning it would be clearer for the readers.**

   The definition has been added in the manuscript.

10. **No explanation why an additional MSE loss was used in the second phase for pseudo labels.**

    A previous work showed that adding a L1 loss incentivizes the segmentations to be consistent. This has been added to the manuscript.

11. **I don't think making the spacing the same is needed because either way there are sampled to the same resolution. If not, please explain it.**

    The resolution of the images in the x,y and z axis is not the sample. Unless you make the spacing similar on all of them with a pre-processing operation, there is no way the images will be sampled to the same resolution.

12. **No explanation why no post-processing was applied**

    I applied a largest component post processing operation, but it reduced the validation dice. Therefore, it wasn´t implemented.

13. **How was the subset of 750 selected from the unlabeled images?**

    Due to limitations in the available memory, the images were selected based on the weight of the .zip file.

14. ***Table ??* should be changed to *Table 5* ?; the table is hyperlink missing**

    The table number has been modified.

Review 2:

1. **In Section 1, Paragraph 3, Line 2, the description of "teacher" network should match its description in abstract. By the way, please avoid using Arabic numerals at the beginning of acronyms. (e.g. 5-network in abstract)**

   The abstract has been changed so that it matches the methodology description. The Arabic numerals have been eliminated.

2. **For Fig.1, there is still room to zoom in the image. Horizontal composition is recommended. Besides, the additional explanation of the block in the proposed network is highly expected.**

   I have zoomed Fig 1. I did not make it a horizontal image because it will decrease the size and the letters are not visible. The explanation of the network block is presented in the subsection Phase One and in Fig. 2.

3. **Some references do not appear to be cited correctly, please check the code in your latex file. (e.g. Page 3, Line 2, "ronneberger2015u"; Page 3, Line 19 "12Lee2015" etc.)**

   All references have been revised.

4. **In Section 2.1, Paragraph 3, what's the link between five-networks and three-networks? And, what rules are the selection process based on? Besides, an equation to represent the final loss function of Phase one is well recommended.**

   I trained five networks using a five-fold division. However, the final ensemble is formed by only three networks as this combination provided a better validation loss than the five-network ensemble. This explanation has been included in the manuscript. The loss for phase 1 is presented in equation 1.

5. **For equation 2, the definition of all the variables needs to be added.**

   All the definitions of the variables in Equation 2 have been included.

6. **Lack of qualitative analysis of segmentation results.**

   The section qualitative results on the validation set has been included. This comment refers to the first submission of the paper.

7. **In Page 7, Line 5, the Table reference failed, please check the code in your latex file.**

   The references for all tables have been checked.

Review 3:

1. **The ensemble size is inconsistent in the text, is it 3 or 5?**

   The ensemble is formed with 3 networks. The abstract has been modified to show this information, and the text now has a consistent size.

2. **It would be nice to have a bit more elaborate figure captions.**

   I have added more added more description to the captions.

3. **In preprocessing, you set to a fixed size of 256x256x123, does that mean you are cropping images which are larger than this size?**

   This is correct.

4. **Table references are not correctly written everywhere**

   All table references have been checked.

5. **The main issue that I am missing in the discussion is the big gap in performance between this model (0.53 DSC) and top leaderboard entries (0.88 DSC). A baseline using only the labeled data should get at least 0.75 DSC I think.**

The reason for this under performance might be the use of a 2D network instead of a 3D network, which also exploits intra slice information. Moreover, due to memory limitations not all the unlabeled were used for training. This discussion has been added to the results section.

Review 4:

1. **Five networks are trained in phase one, but why do you select only three networks to form an ensemble? How do you ensemble the chosen three networks?**

   The final ensemble is formed by only three networks as this combination provided a better validation loss than the five-network ensemble. This explanation has been included in the manuscript.

2. **In the proposed method, 2D U-Net is chosen to construct teacher model and student. How do you set the input of 2D U-Net? Do you regard one slice of CT image as a sample?**

   Yes, the 2D U-Net is trained only with slices. This information has been added in the manuscript.

3. **It is unclear if you consider the neighbor information between CT slices.**

   Interslice information is not considered as we only train with slices.

4. **It is better to use abbreviation of 'Convolutional Block' in Fig.2.**

   I prefer to keep the phrase complete.

Review 5:

1. **This work proposes a simple teacher-student approach for semi-supervised organ segmentation using a 2D U-Net. The final Dice on the test set is 0.5272 using the entire training dataset. There is no novelty in the proposed method, and the results are low compared to other submissions in this validation set. Due to these factors, this work is of limited significance.**

   As the challenge did not require the use of a novel approach, this wasn´t covered in the present work. The reasons for the underperformance have been included in the manuscript.

Review 6:

1. **References are not shown correctly, requires fixing in latex.**

   All references have been fixed in the manuscript.

2. **Figures 1&2 are a bit small and hard to read.**

I I have increased the size of Figures 1 and 2.

3. **Missing info on inference requirements in terms of GPU memory, speed, etc. as required by organizer's checklist.**

This information has been included in Table 6.

4. **Missing qualitative comparison as required by organizer's checklist.**

   The qualitative comparison has been included in the manuscript.

5. **Dice performance below nnU-Net baseline.**

   This might be caused by the fact that nnU-Net network uses 3D network, while I use a 2D network. Furthermore, the nnU-Net has a memory consumption higher than the one required by the challenge.