We would like to thank the reviewer and the editor for their time spent reviewing our revised manuscript. We also thank you for your valuable suggestion about our answer to the overfitting concern. In the following, citations of reviewer passages are in black italics. Revised/added text is highlighted in yellow. Line numbers refer to the resubmitted manuscript.

Suggestion:

The newly added text in the manuscript does not fully substantiate their claim: "If overfitting to the training data was present, we would expect these scores to be different (much higher IoU for the training data than for the test data)." Including evidence of score differences would strengthen this argument. I encourage the authors to provide additional details, such as results from the CG-Net implementation by KS20 using a batch size of 10, to support this statement directly in the manuscript. Thank you.

To strengthen our claim, we deliberately overfitted the CG-Net implementation by KS20 by training the CNN for 100 epochs instead of 20. This leads to a large gap between IoU scores calculated on the training data (AR = 60.0%, TC = 53.3%, BG = 96.7%, AR-TC-BG mean = 70.0) and test data (AR = 37.8%, TC = 32.5%, BG = 94.4%, AR-TC-BG mean = 55.0%). We add the following lines at line 187 into our manuscript to discuss the deliberate overfitting:

Also, strong overfitting typically results in evaluation scores being distinctly better for the training data compared to the test data (e.g., Bishop, 2007). For example, for the CG-Net implementation by KS20, batch size 10, we obtain the following IoU scores for the training data: AR = 43.6%, TC = 37.7%, BG = 95.2%, AR-TC-BG mean = 58.8%. These scores are very close to those listed in Table 2 for the test data, indicating that no overfitting is present. In comparison, if we deliberately overfit CG-Net by training with 100 training-evaluation epochs (instead of 20), we obtain IoU scores of AR = 60.0%, TC = 53.3%, BG = 96.7%, AR-TC-BG mean = 70.0% for the training data and AR = 37.8%, TC = 32.5%, BG = 94.4%, AR-TC-BG mean = 55.0% for the test data.

Reference:

Bishop, C. M.: Pattern recognition and machine learning, 5. (corr. print.)., Springer, New York [u.a.], XX, 738 S. pp., 2007.

As a last point, we want to thank everyone involved in our review process again for your time and effort.

Best Regards,

Tim Radke, on behalf of all authors