Review Response for 'Guided Proofreading of Automatic Segmentations for Connectomics'

Thank you all for your time and considered feedback.

R2: Venue Biological and Cell Microscopy Image Analysis is included in the call for papers of CVPR 2018, and the field of connectomics has previously given rise to interesting papers at CVPR (Kumar *et al*. 2010, Kaynig *et al*. 2010, Jain *et al*. 2010, Funke *et al*. 2012, Pape *et al*. 2017).

R2: Algorithmic Contribution/Innovation While we use a traditional CNN architecture, we believe that the GP framework can work with many classifiers and is a promising direction to proofread segmentations more efficiently.

R2: Superhuman Performance Lee *et al*.'s arXiv preprint indeed reports fantastic segmentation performance, but as a future direction they still state the need for "guiding focused human proofreading" with supervised learning [1, Sec. 8.2]—our work is evidence to the viability of this idea. *"Is manual proof-reading competitive with a superhuman automatic method? Is your method able to find mistakes still present in Lee et al.'s segmentation?"* Interesting questions, to which there are no concrete answers yet (we'd be happy to test this if Lee et al. release their software/segmentation). We can also look at this problem as one of raising the bar for manual proofreading. In their paper, Lee et al. state that "human accuracy depends on the procedures and software tools used to perform the reconstruction" (Sec. 1). Through this lens, we measure by how much our software tool improves performance given consistent human time/effort across skill levels; for this, our tool advantages both novices and experts.

R2: Generalization beyond the AC4 Subvolume We agree that this dataset is small; however, it was introduced by Haehn *et al*. 2014 for feasible interactive proofreading studies and is representative for the full AC4 dataset with respect to the distribution of object sizes. [JT: Evidence?]

R2: Rand Error We chose variation of information (VI) to overcome previously reported limitations of aRE [2, p. 5]; however, we include aRE numbers below (Table 1).

Table 1: Forced Choice User Experiment in adapted Rand Error (aRE) metric (lower is better). Novices and experts using GP perform better than using FP.

Slice	1	2	3	4	5	6	7	8	9	10
Init. Segm.	0.074	0.081	0.085	0.079	0.103	0.098	0.176	0.188	0.206	0.174
FP Novices	0.073	0.082	0.086	0.091	0.102	0.103	0.182	0.184	0.209	0.167
GP Novices	0.054	0.074	0.083	0.081	0.100	0.086	0.127	0.095	0.100	0.096
FP Experts	0.066	0.080	0.078	0.087	0.083	0.096	0.163	0.174	0.202	0.155
GP Experts	0.051	0.074	0.075	0.071	0.078	0.075	0.099	0.088	0.094	0.074

R3: U-Net training data The supplemental material includes this information (lines 140–161, Table 3). We will add a direct reference to the main paper (lines 492–493).

R3: Generalization to other segmentation problems We believe our method will interest researchers working beyond connectomics, as segmentation proofreading for labeled dataset collection and correction is widely applicable in computer vision. We state mandatory re-training of GP for other datasets as a limitation in the supplemental material (lines 134–138) but will further elaborate on general segmentation problems in this section. [JT: Discuss.]

R3+R4: Input channel contributions All four input channels help to reduce VI (Table 2). As identified by Bogovich *et al*., image data adds intracellular structures (e.g., vesicles) to the decision process, and membrane probabilities include global knowledge of the staining protocol to highlight cell membranes. Then, the label channel provides knowledge about neuron shapes while the border mask covers the gap of extra-cellular space. [JT: From the table, it might make sense to test just Label+Border, because the other two decrease VI...]

Table 2: Automatic selection on the AC4 subvolume ($p_t = 0.95$) using the GP classifier with different input channels and report median VI reduction. The combination of all four channels performs best.

Input channels	VI reduction
Image + Prob.	-0.094
Image + Prob. + Border	-0.045
Image + Prob. + Label	0.038
Image + Prob. + Label + Border	0.065

R4: Dilated mask of the border between two segments R4 raises the question of whether the dilated mask of the border is a crucial input to our classifier. The dilated border improves classifier performance measured as VI reduction from 0.038 to 0.065 (Table 2). We qualitatively describe the intention behind the border mask in lines 315-323 but will add the experiment above to the supplemental material and to the open source repository.

R4: 2D Slices only R4 observes that GP works in 2D. We report this limitation and a proposed solution in the supplemental material lines 129-133 and will add a direct reference to the manuscript. However, 2D processing enables

segmentation and proofreading in parallel to 2D image acquisition prior to any expensive 3D alignment.

R4: Merge error detection performance R4 expresses concerns regarding the performance contribution of merge error correction. Typical connectomics segmentations begin with an over-segmentation (lines 498-499), in which we attempt to find all possible cell boundary edges. At this stage, some edges are missed due to low contrast - these missing edges cause merge errors. To correct this, we need to imagine where any number of edges might be among empty space, and this is simply a harder visual task than assessing whether an identified edge is correct. This is true even for a human: on the AC4 dataset, our experts only agree with the selection oracle in two thirds of merge error cases.

References

[1] K. Lee, J. Zung, P. Li, V. Jain, and H. S. Seung. Superhuman accuracy on the SNEMI3D connectomics challenge. *CoRR*, abs/1706.00120, 2017. 1

[2] J. Nunez-Iglesias, R. Kennedy, T. Parag, J. Shi, and D. B. Chklovskii. Machine learning of hierarchical clustering to segment 2D and 3D images. *PLoS ONE*, 8(8):e71715+, 2013. 1