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

Thank you all for your time and considered feedback—We hope that the open source nature of the GP framework and the introduced proofreading benchmark are of value to the connectomics community and enable future research.

R2: Superhuman Performance Lee *et al*.'s arXiv paper indeed reports fantastic segmentation performance, but as a future direction they still state the need for "guiding focused human proofreading" with supervised learning (Sec. 8.2)—our work is evidence that this idea is viable.
"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.

R2: Rand Error We chose variation of information (VI) to overcome previously reported limitations of adapted Rand Error (aRE) [1, 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

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. The volume was quantitatively chosen to be representative for the full AC4 dataset with respect to the distribution of object sizes.

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 that our method will interest researchers working beyond connectomics, as segmentation proofreading for labeled dataset collection and correction is widely applicable in computer vision. However, edges in natural images are usually softer and not as prominent as in our data; we are optimistic but also have to be careful to not overclaim our contribution.

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 dilated mask of the border covers the gap of extra-cellular space.
R4: Adding the dilated mask of the border decreases VI.

Table 2: Automatic selection on the AC4 subvolume ($p_t = 0.95$) using our GP classifier; median VI reduction in ascending order. The combination of all four input channels performs best.

Input channels	VI reduction
*Prob. + Label + Border (**R3** request)*	TBA
Image + Prob.	-0.094
*Prob. + Border (**R4** request)*	-0.080
Image + Prob. + Border	-0.045
Label + Border	-0.008
Image + Prob. + Label	0.038
Image + Prob. + Label + Border	0.065

R4: 2D Slices only We report this limitation and a proposed solution in the supplemental material lines 129–133, but we will add a direct reference back in to the manuscript. 2D processing enables segmentation and proofreading in parallel to any expensive 3D alignment,

R4: No benefit from merge error detection? Only in the automatic case. In the guided proofreading case, the expert is able to judge given the candidate edge we generate. That said, merge correction is simply a harder visual task than split correction. 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. For this reason, the initial over-segmentation is tuned to try to find all possible cell boundary edges, such that mostly split errors remain.

References

[1] 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), 2013. 1