Guided Proofreading of Automatic Segmentations in Connectomics

Anonymous CVPR submission

Paper ID ****

Abstract

Automatic cell image segmentation methods in connectomics can lead to split *and* merge *errors, which require correction through proofreading. To aid error correction, we develop two classifiers that are able to recommend candidate errors and their corrections to the user. These classifiers are informed by training a convolutional neural network with known errors in automatic segmentations by comparison to expert-labeled ground truth. Our network architecture is able to detect potentially erroneous regions by considering a large context region around a segmentation boundary. With recommendations, proofreading of mouse cortex electron microscopy image segmentations can reduce VI scores from 0.476 to 0.426, which we find is an improvement on pure manual and pure automatic cases that both cause VI to increase.*

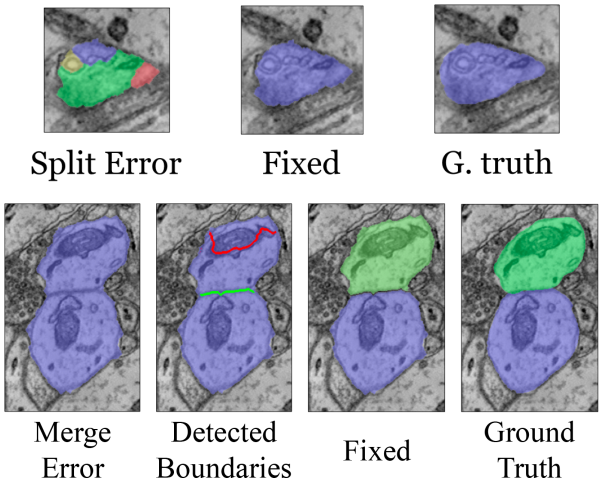


Figure 1: Split and merge error examples, their corrections, and their ground truths.

1. Introduction

In connectomics, neuroanatomists build 3D reconstructions of neurons and their connectivity to gain insight into the functional structure of the brain. Rapid progress in automatic sample preparation and electron microscopy (EM) acquisition techniques has made it possible to image large volumes of brain tissue at $\approx 6\,nm$ per pixel to identify cells, synapses, and vesicles. For $25\,nm$ thick sections, a $1\,mm^3$ volume of brain contains 10^{15} voxels, or 1 petabyte of data: manual annotation is infeasible, and automatic methods are needed [7, 13, 15, 11].

Automatic segmentation and classification of brain tissue is challenging [1], so learning-based methods are common. The state of the art uses supervised learning with convolutional neural networks [5], or potentially even using unsupervised learning [4]. Typically, cell membranes are detected in 2D images and the resulting region segmentation is grouped into geometrically-consistent cells across registered sections, or cells are segmented across registered sections in 3D directly. Using dynamic programming techniques [14], and a GPU cluster, these classifiers can segment ≈ 1 terabyte of data per hour [10], which is sufficient to keep up with the 2D data capture process on state-of-the-art electron microscopes (though 3D registration is still an expensive offline operation).

All automatic methods make errors, and we are left with large data which needs *proofreading* by humans. This crucial task serves two purposes: 1) to correct errors in the segmentation, and 2) to provide large corpora of labeled data to train better automatic segmentation methods. Recent proofreading tools provide intuitive user interfaces to browse segmentation data in 2D and 3D and to identify and manually correct errors [17, 8, 12, 6, 9, 18]. Many kinds of errors exist, such as inaccurate boundaries, but the most common are *split errors*, where a single cell is labeled as two, and *merge errors*, where two cells are labeled as one (Fig. 1). With user interaction, split errors can be joined, and the missing boundary in a merge error can be defined with manually-seeded watersheds [6]. However, even with semi-automatic correction tools, the visual inspection of the data to find the errors in the first place takes the majority of the time.

Our goal is to add automatic detection of split and merge

errors to proofreading tools. Instead of the user visually inspecting the whole data volume carefully to spot errors, we design automatic classifiers that detect split and merge errors in 2D segmentations. Then, a proofreading tool can recommend regions with a high probability of an error to the user, and suggest corrections to accept or reject.

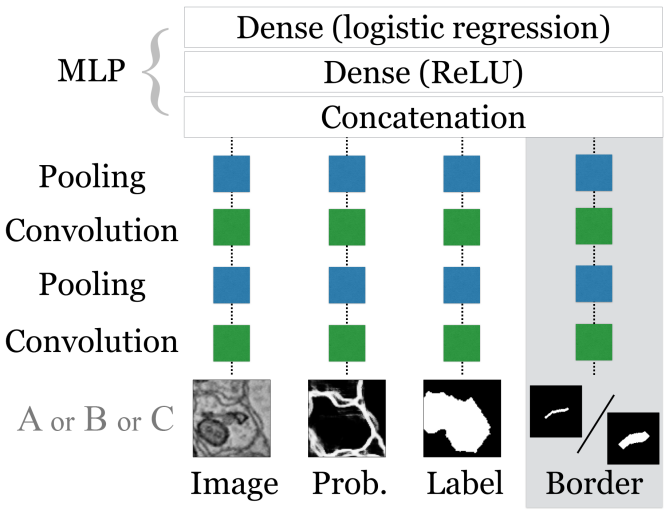
The initial automatic segmentation is constrained by the data rate of the microscope. Given an membrane segmentation from a fast automatic method, our classifiers operate on whole cell regions, which relaxes the constraint on speed: compared to techniques that must analyze every input pixel, this boundary assessment focus reduces the data analysis to the boundaries only, and so allows us to employ wider convolutional neural networks that take regional context and multiple input channels into account. One reason to classify errors on 2D images lies with the cost of 3D registration. This is often slow as it requires non-linear image alignment [2, 16]. However, typically segmentation results are local decisions at the cell level. In this case, 3D reconstruction is unnecessary and, instead of waiting for the 3D output, proofreading can start immediately to maximize error correction before cell grouping occurs across sections.

We quantitatively validate our approach with variation of information (VI) versus ground truth expert segmentations. We compare our approach in an experiment against an existing proofreading tool that provides only semi-automatic merge error correction [6]. Here, with a simulated user, our automatic error suggestions and corrections decrease VI from 0.476 to 0.426, which is in contrast to pure manual or pure automatic methods that can both increase the VI. As a consequence, we are able to provide tools to proofread segmentations more efficiently, and so better tackle large volumes of connectomics imagery.

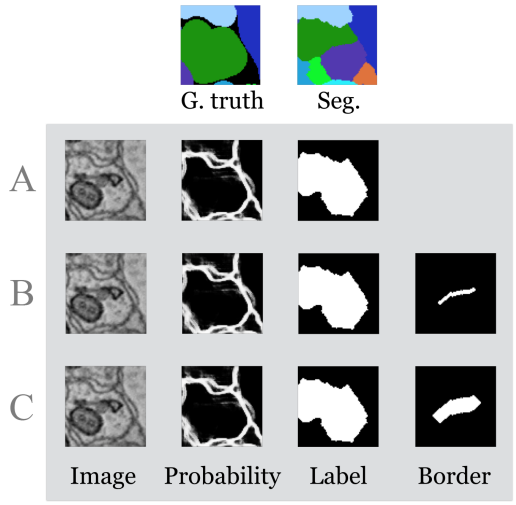
2. Method

We build a split error classifier using a convolutional neural network (CNN) to check the boundaries of an existing automatic segmentation. For each boundary, the CNN provides a probability that points sampled along the boundary caused a split error. For each boundary, we sample up to 10 decision points, where the decision points are spread evenly over the boundary length, so long as their context windows do not overlap. These probabilities are then weighted by the length of the boundary within the context over the total boundary length, and averaged. A greedy algorithm then merges neighboring regions sequentially, starting with the highest probability score. Following each merge, neighboring boundaries are re-evaluated for split errors. Correcting a split error is as simple as merging the two bordering labels.

Identification and correction of merge errors is more challenging, because we must look inside segmentation regions for missing or incomplete boundaries and then propose the correct boundary. However, we can reuse the same trained CNN for this task. For each segmentation label, we generate 30 potential boundaries through the region by placing watershed seed points at opposite sides of the label boundary and generating the corresponding split. Then, we check to see whether any potential edge is classified as a split error. If the CNN detects a boundary with a very low split error score, then the boundary should have been in the segmentation and the region is a candidate for a merge error.



(a) CNN Layers



(b) Network configurations

Figure 2: (a) Our network architecture with up to four input channels, each with two convolutional and two pooling layers. (b) We trained three different network configurations with three and four inputs: A) image, boundary map probability, and merged binary mask; B) A configuration extended with a small border mask, to focus on the specific boundary in question; C) A configuration extended with a large border mask.

2.1. Convolutional Neural Network Design

To train a CNN for split error detection we take multiple channels of context information of the boundary into consideration for the decision making process. We pass multiple inputs into the CNN windowed around a particular decision point or pixel: the input grayscale image patch, the corresponding boundary probability map patch, and two corresponding binary mask patches for the segmented regions at either side of the boundary. Following Bogovic et al. [4], these two masks can be combined into a single mask with comparable performance (configuration A, Fig. 6b). The network then leverages these multiple input patches to identify and correct errors made by the previous membrane detection network and automatic segmentation pipeline.

One way to combine these inputs is to treat them as a 3-channel input, so that alignment between the input image and the segmentation masks are not lost throughout the convolutions. However, training a boundary-classifying network can be difficult due to rigid ground-truth segmentations, which often differ substantially from automatic segmentation regions in ambiguous extra-cellular space. To cope with this variation, our network is based on multiple separate input channels (Fig. 6a). Each of the input patches is connected individually to a 2-layer network, with each layer consisting of convolutional and pooling layers. The output of these networks is then combined by a fully connected multi-layer perceptron (MLP) with one hidden layer and a two class logistic regression output layer. The intuition for this multiple input channel approach is that we want to allow variation in the input and masks independently, to accommodate potential error, and then for the hidden layers to discover appropriate combinations of the relevant features learned separately for the different input channels.

To better direct the network to train on the true boundary edge, which in many cases is missing from the boundary probability map and hence is the cause of merge errors, we additionally pass as input a second binary mask. This mask contains the true boundary edge (configuration B, Fig. 6b). To consider slight edge ambiguities, we also test a version of this network where the true boundary mask has been dilated by 5 pixels (configuration C).

2.2. Training

To train the network, we use a mouse cortex data set ($1024 \times 1024 \times 75$ voxels). The tissue is dense mammalian neuropil from layers 4 and 5 of the S1 primary somatosensory cortex of a healthy mouse. The resolution of our data set is $6\,nm$ per pixel, and the section thickness is $30\,nm$. A manually-labeled expert segmentation is available as a ground truth for the entire data set. We use the first 65 sections of the data for training, the next 5 for validation, and the last 5 for testing. To generate training data, we identify correct regions and split errors in the automatic segmentation by intersecting with the ground truth regions. From these regions, we sample 79,828 correct regions and 79,828 split error patches.

We train our network using the following parameters: learning rate $lr = 0.00001$, momentum $m = 0.9$, filter size $fs = 13 \times 13$ and number of filters $fn = 16$. We assume that the training has converged if the validation loss does not decrease for 30 epochs. The network is specified using the deep learning libraries Lasagne and Theano [3], and trained on a Tesla K40m graphics card.

Table 1 presents training and test loss function scores (cross validation), and test accuracy percent. Based on these performances, we select configuration C to evaluate against human performance in a VI improvement experiment.

3. Evaluation

We evaluate our split and merge error detection and correction recommendation in the context of interactive proofreading tools: to direct users to regions with a high probability of error and to suggest corrections (Fig. 3). For comparison, we take publicly available mouse cortex data of the same kind as our training data. This data is part of the ISBI 2013 challenge training dataset ($1024 \times 1024 \times 100$ voxels) which was acquired using a serial section scanning electron microscope (ssSEM) with a resolution of $6 \times 6 \times 30\,nm$ per voxel. We use the available manually-labeled ground truth to score our approach using the variation of information (VI) metric, which is closely related to mutual information. VI is a measure of the distance between two clusterings, where lower VI numbers are better. Since our classifiers are trained on 2D image slices, we perform all evaluations on slices rather than 3D volumes.

Interactive proofreading. Recently, Haehn et al. discussed requirements for interactive proofreading and evaluated three different tools on connectomics data in a study with naive users [6]. This study asked users to spend 30 minutes proofreading with the different tools, to correct split and merge errors to improve the automatic segmentation. The best performing tool in their evaluation was Dojo. We use their findings and their user-generated proofreading result data, which they kindly provided, as a baseline for the evaluation of our method. Haehn et al. perform their user study on the most representative sub-volume ($400 \times 400 \times 10$ voxels) in terms of distribution of object size. For optimal comparison, we use exactly the same data. We asked a novice and two experts to perform the proofreading task using our system (Fig. **??**).

In addition, we simulate a user for proofreading correction. We assume that all classification has been computed ahead of time, and that the user is presented with a stream of error corrections to assess. The assessment is simulated

Table 1: Network design training evaluation. Adding an extra channel containing a binary mask of just the border slightly increases performance. We take configuration C to evaluate VI against human performance.

Network	Training loss	Test loss	Test acc. (%)
A. Image + boundary prob. + seg. label	0.3853	0.4163	81.15
B. A config. + small border overlap($d = 1$)	0.3798	0.3843	82.34
C. A config. + large border overlap ($d = 5$)	0.3703	0.3919	83.02

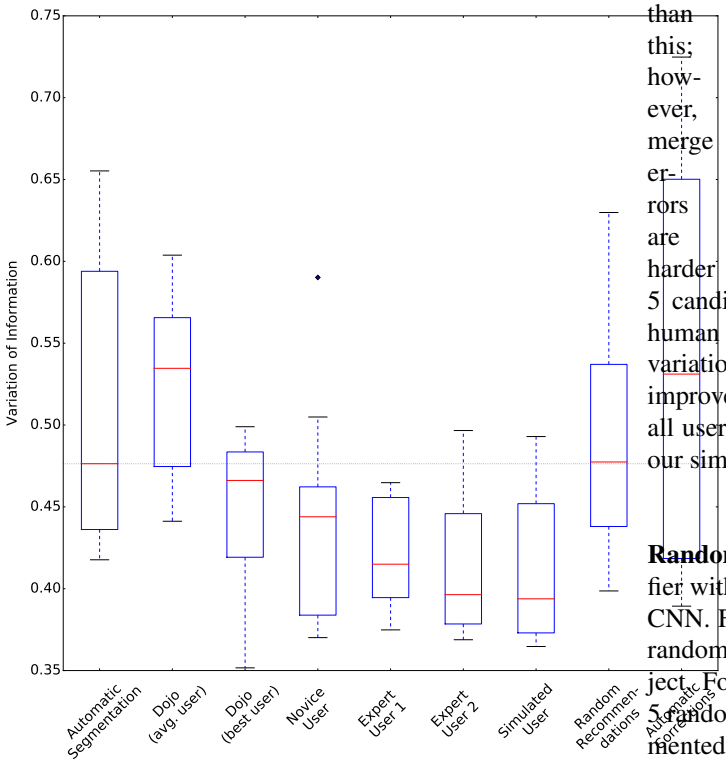


Figure 3: (a) We compare distributions of VI measures across 10 sections for the initial automatic segmentation, the average and best users in the Haehn et al. experiment with Dojo, a novice and two experts using our system, our simulated user, our simulated user when presented with random recommendations, and finally a fully-automatic correction of recommended errors based on a threshold of acceptance. Lower scores are better.

by comparing the VI before and after each performed correction. Corrections are accepted only when VI reduces, and we test this across different user error rates (Fig. 3). In Haehn et al., the proofreading time was limited to 30 minutes, and human participants performed 59 corrections on average ($\approx$ 30 seconds per correction). In our scenario, users do not need to visually find errors and manually correct them, and so instead we assume each correction assessment takes 15 seconds (120 assessments in 30 minutes). Split errors are likely to take less time than this; however, merge errors are harder to assess, as the user must select between the top 5 candidate boundaries. Since the performance between human participants of Haehn et al.'s user study shows large variation, we present both the best performing user (VI improvement: 0.0102) and the average performance among all users (VI improvement: -0.0582) as our baseline. For our simulated user, the VI improvement is 0.0502 (Fig. 3).

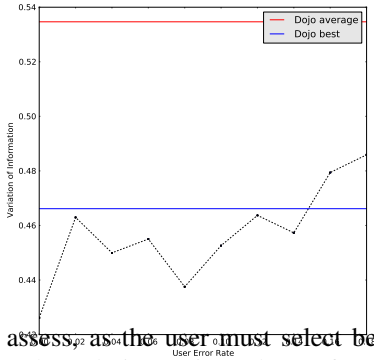
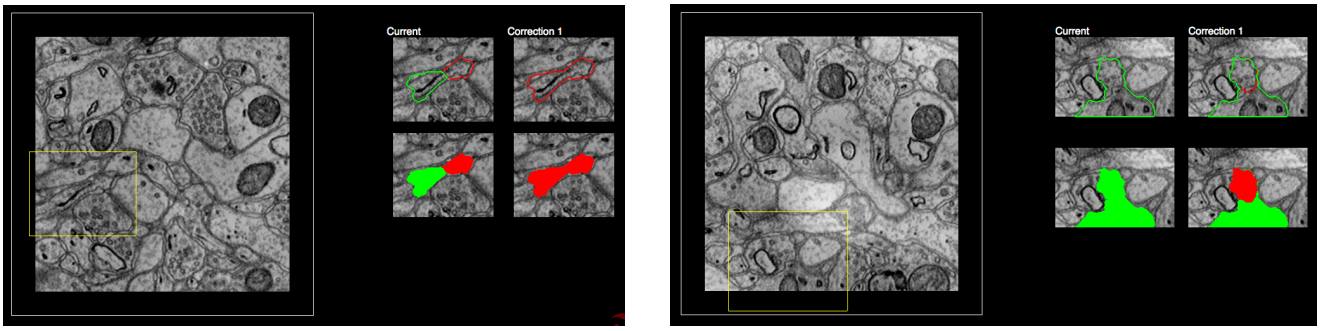


Figure 4: VI vs. simt. user error rate.

Random recommendations. We decided to test a classifier with random performance in comparison to our learned CNN. For split errors, the simulated user is presented with randomly picked boundaries, which they can accept or reject. For merge errors, the simulated user is presented with 5 randomly selected boundaries from the interior of the segmented region. The significantly worse performance of this approach demonstrates that our network is informative to the user.

Automatic correction. As a comparison, we also perform automatic correction. During training, we define a probability threshold $p_t = 0.95$ for automatic split correction based on CNN probability from the test set. Then, for automatic correction, we apply both classifiers to produce lists of split and merge errors sorted by confidence. First, we correct merge errors with $\max(1 - p)$, followed by split error correction using p_t. The total time for correcting all errors was 17 minutes on a 3.2 GHz Quad-core Intel Xeon with an NVIDIA GeForce Titan (merge error correction 15min, split error correction 2min). The median VI improvement in comparison to the ground truth was negative, at -0.0552 (Fig. 3). This is not surprising, as the problem is very challenging, and this motivates the need for human-in-the-loop proofreading tools.



(a) Split error **(b)** Merge error

Figure 5: Our web-based user interface includes a slice overview with the relevant area highlighted in yellow. The interface shows (a) a split error with a suggested correction as well as (b) a merge error with correction. The user selects whether to accept a correction or to skip it.

4. Discussion and Conclusion

The task of automatic cell boundary segmentation is difficult, and trying to improve such segmentations automatically as a post-process through split and error correction is, in principle, no different than trying to improve the underlying cell boundary segmentation. This is shown by the approximately equivalent VI distributions of the initial segmentation and our automatic segmentation correction (Fig. 3). Due to the task difficulty, manual proofreading of connectomics segmentations is necessary, but it is a time consuming and error-prone task, as can be seen from the Dojo human trials: on average, participants actually made the segmentations worse. However, there is value in being able to recommend to users possible regions for correction, as the time cost of proofreading is dominated by the visual search for errors.

We have addressed this problem through training a CNN to detect ambiguous regions from labeled data—in effect, (re-)learning a confidence measure on boundaries. This allows us to recommend split and merge errors, and also to recommend their corrections, which is an improvement over existing systems which just provide semi-automatic merge error correction. Through simulating users with different error rates, we have shown that, for an equivalent 30 minutes of work, correction recommendation has the potential to reduce VI over existing proofreading tools. This helps reduce the proofreading bottleneck to the analysis of large connectomics datasets. To encourage testing of our proposed architecture on more data, we provide the trained networks and classifier code as free and open source software at (link omitted for review).

References

[1] IEEE ISBI challenge: SNEMI3D - 3D segmentation of neurites in EM images. http://brainiac2.mit.edu/SNEMI3D, 2013. Accessed on 31/03/2016. 1

[2] A. Akselrod-Ballin, D. Bock, R. Reid, and S. Warfield. Improved registration for large electron microscopy images. In *IEEE Int. Symp. Biomedical Imaging (ISBI)*, 2009. 1

[3] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. J. Goodfellow, A. Bergeron, N. Bouchard, and Y. Bengio. Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop, 2012. 3

[4] J. A. Bogovic, G. B. Huang, and V. Jain. Learned versus hand-designed feature representations for 3d agglomeration. *CoRR*, abs/1312.6159, 2013. 1, 3

[5] D. C. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. In *NIPS*, 2012. 1

[6] D. Haehn, S. Knowles-Barley, M. Roberts, J. Beyer, N. Kasthuri, J. Lichtman, and H. Pfister. Design and evaluation of interactive proofreading tools for connectomics. *IEEE Transactions on Visualization and Computer Graphics (Proceedings IEEE SciVis 2014)*, 20(12):2466–2475, 2014. 1, 2, 3

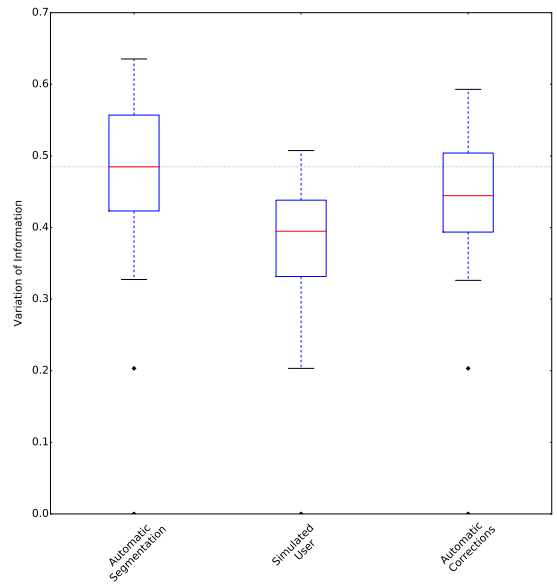
[7] V. Jain, B. Bollmann, M. Richardson, D. Berger, M. Helmstädter, K. Briggman, W. Denk, J. Bowden, J. Mendenhall, W. Abraham, K. Harris, N. Kasthuri, K. Hayworth, R. Schalek, J. Tapia, J. Lichtman, and S. Seung. Boundary learning by optimization with topological constraints. In *Proceedings of IEEE CVPR 2010*, pages 2488–2495, 2010. 1

[8] Janelia Farm. Raveler. https://openwiki.janelia.org/wiki/display/flyem/Raveler, 2014. Accessed on 31/03/2016. 1

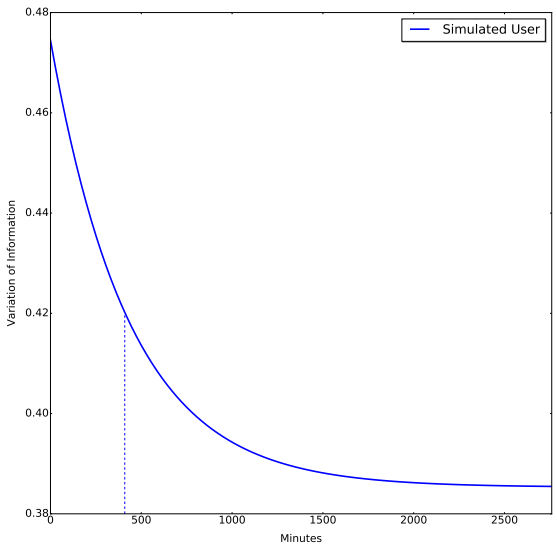
[9] A. Karimov, G. Mistelbauer, T. Auzinger, and S. Bruckner. Guided volume editing based on histogram dissimilarity. *Computer Graphics Forum*, 34(3):91–100, May 2015. 1

[10] N. Kasthuri, K. J. Hayworth, D. R. Berger, R. L. Schalek, J. A. Conchello, S. Knowles-Barley, D. Lee, A. Vázquez-Reina, V. Kaynig, T. R. Jones, et al. Saturated reconstruction of a volume of neocortex. *Cell*, 162(3):648–661, 2015. 1

[11] V. Kaynig, A. Vazquez-Reina, S. Knowles-Barley, M. Roberts, T. R. Jones, N. Kasthuri, E. Miller, J. Lichtman, and H. Pfister. Large-scale automatic reconstruction of



(a) Split error



(b) Merge error

Figure 6: Our web-based user interface includes a slice overview with the relevant area highlighted in yellow. The interface shows (a) a split error with a suggested correction as well as (b) a merge error with correction. The user selects whether to accept a correction or to skip it.

neuronal processes from electron microscopy images. *Medical image analysis*, 22(1):77–88, 2015. 1

[12] S. Knowles-Barley, M. Roberts, N. Kasthuri, D. Lee, H. Pfister, and J. W. Lichtman. Mojo 2.0: Connectome annotation tool. *Frontiers in Neuroinformatics*, (60), 2013. 1

[13] T. Liu, C. Jones, M. Seyedhosseini, and T. Tasdizen. A modular hierarchical approach to 3D electron microscopy image segmentation. *Journal of Neuroscience Methods*, 226(0):88 – 102, 2014. 1

[14] J. Masci, A. Giusti, D. C. Ciresan, G. Fricout, and J. Schmidhuber. A fast learning algorithm for image segmentation with max-pooling convolutional networks. In *ICIP*, 2013. 1

[15] J. Nunez-Iglesias, R. Kennedy, S. M. Plaza, A. Chakraborty, and W. T. Katz. Graph-based active learning of agglomeration (GALA): A python library to segment 2D and 3D neuroimages. *Frontiers in Neuroinformatics*, 8(34), 2014. 1

[16] S. Saalfeld, A. Cardona, V. Hartenstein, and P. Tomančák. As-rigid-as-possible mosaicking and serial section registration of large ssTEM datasets. *Bioinformatics*, 26(12):i57–i63, 2010. 1

[17] R. Sicat, M. Hadwiger, and N. J. Mitra. Graph abstraction for simplified proofreading of slice-based volume segmentation. In *EUROGRAPHICS Short Paper*, 2013. 1

[18] M. G. Uzunbas, C. Chen, and D. Metaxas. An efficient conditional random field approach for automatic and interactive neuron segmentation. *Medical Image Analysis*, 27:31 – 44, 2016. Discrete Graphical Models in Biomedical Image Analysis. 1