Guided Proofreading of Automatic Segmentations for Connectomics

Supplemental Material

Anonymous CVPR submission

Paper ID 0947

	Traditional Network		Residual Network	
Conv. Layers	2	4	5	13
Dropout Reg.	y	y	y	n
Cost [m]	27.5	383	5080	1094
Test. Acc.	0.925	0.94	0.93	0.90
Prec./Recall	0.93/0.93	0.94/0.94	0.7/0.53	0.74/0.66
F1 Score	0.93	0.94 *	0.39	0.64

Table 1: Traditional CNN Architecture versus Residual Network Architecture [1]. All configurations are compared using the same parameters. Our final choice (indicated by *) trains relatively fast and performs better.

Parameter	Search Space
Filter size:	**3x3**, 5x5, 9x9, 13x13
No. Filters 1:	32, 48, **64**
No. Filters 2-4:	32, **48**, 64
Dense units:	256, **512**
Learning rate:	0.00001, 0.0001, 0.001, 0.01, **0.03-0.00001**
Momentum:	0.9, 0.95, **0.9-0.999**
Mini-Batchsize:	10, 100, **128**

Table 2: Brute force parameter search for the split error classifier. The final parameters are highlighted.

1. Classifier

1.1. Architecture

We explored different architectures for the convolutional neural network (CNN) for split error detection. We compare traditional CNN architectures versus residual networks [1] (Tab. 1). The traditional architecture with dropout regularization generalized better than residual networks on unseen testing data.

1.2. Training Parameters

We performed a limited brute force parameter search to tune the split error classifier (Tab. 2). This resulted in 3240 different CNN configurations which were evaluated on 10% of our training data. Learning rate and momentum ranges are defined linearly across 2000 epochs.

1.3. Automatic Method Threshold p_t

For automatic selection, we observed a threshold $p_t = 0.95$ as stable when evaluating on previously unseen testing data (Mouse S1 AC3 Open Connectome Project dataset). This means that automatic selection stops once all borders with $p_t \geq 0.95$ were proofread. Figure 4 shows split error classification on three randomly selected subvolumes ($700 \times 700 \times 2$ voxels) of AC3. In all cases, the threshold $p_t = 0.95$ reduces VI.

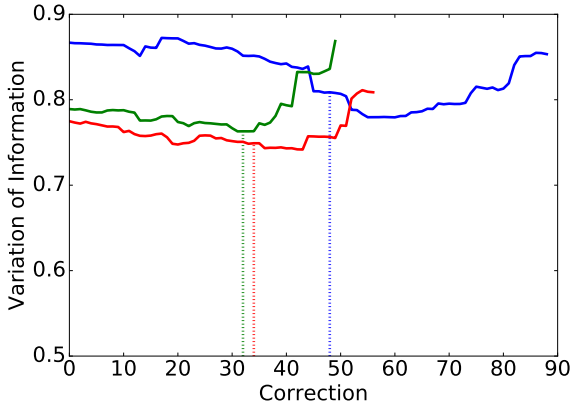


Figure 1: Observations of probability thresholds p_t during automatic selection on three different subvolumes of previously unseen testing data. The dashed lines show when $p_t = 0.95$ is reached.

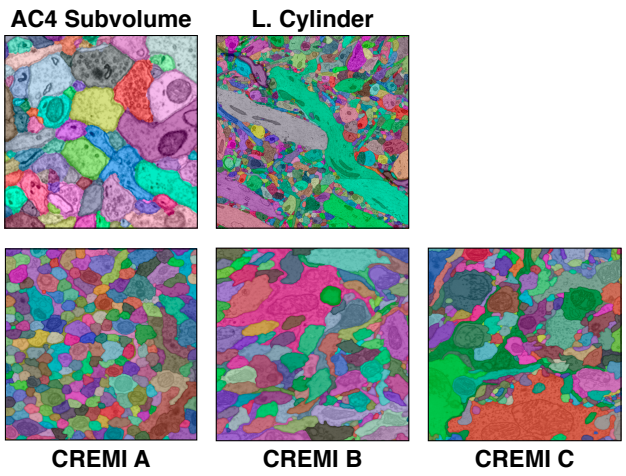


Figure 2: The five different datasets we use for evaluation. The top row shows the first slice of the AC4 and L. Cylinder mouse brain datasets as reported in the paper. The bottom row shows the first slice of the CREMI A/B/C fruit fly datasets which we used for additional experiments.

1.4. Limitations

Guided proofreading works on 2D image sections. This enables error correction without a computationally expensive alignment process. However, the output requires an additional (block-)merging step prior to 3D analysis. Several software packages exist for this purpose.

As described in section 3, the guided proofreading classifier has to be retrained if used on a different species than mouse. Parameters do not need to be changed.

2. L. Cylinder Results

We report experiments and results on the L. Cylinder dataset in the paper. Figure 3 and 4 visualize the reported results measured as variation of information (VI). We compare automatic selection with threshold and selection oracle using focused proofreading and guided proofreading.

Best possible VI. The selection oracle using guided proofreading does not reach the best possible VI score. We calculate this score by intersecting the initial segmentation and the ground truth. In theory, the classifier should be able to reach this lower bound. However, the membrane probability maps we use include a 30 pixel frame region because of the used classification patch size. Guided proofreading ignores all segments within this frame region and can not reach the best possible VI in some datasets.

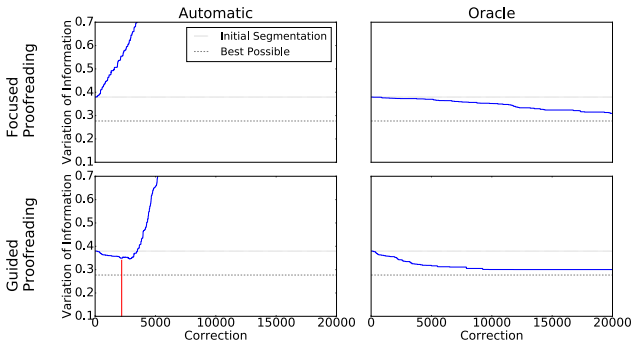


Figure 3: Performance comparison of Plaza's focused proofreading and our guided proofreading on the L. Cylinder dataset as reported in the paper. All measurements are shown as median VI, the lower the better. We compare automatic selection with threshold ($p_t = 0.95$, red line) and the selection oracle for accepting or rejecting corrections using each method. Guided proofreading yields better results faster with fewer corrections.

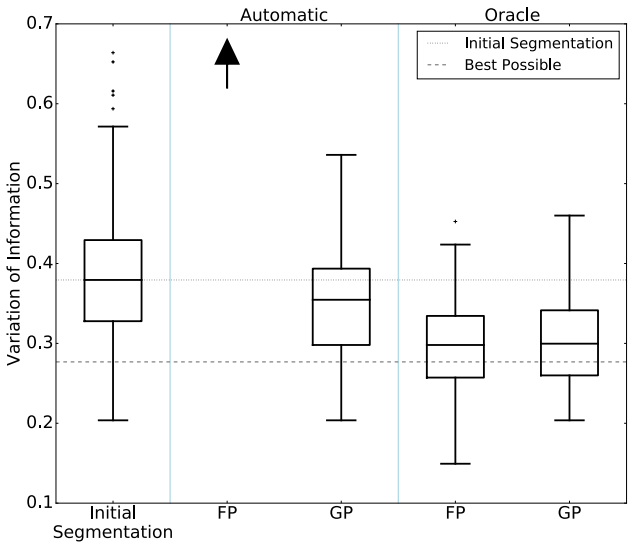


Figure 4: VI distributions of guided proofreading (GP) and focused proofreading (FP) output across slices of the L. Cylinder dataset, with different error correction approaches. The variation resulting from performance of FP with automatic selection is $7.8\times$ higher than GP (as indicated by the arrow), with median VI of 2.75 and $SD = 0.789$.

3. Additional Experiments

CREMI A/B/C. As part of the MICCAI 2016 challenge on circuit reconstruction from electron microscopy images (CREMI), six ssTEM datasets were made publicly available[1], each $1250 \times 1250 \times 125$ voxels. Since only three

[1] http://www.cremi.org

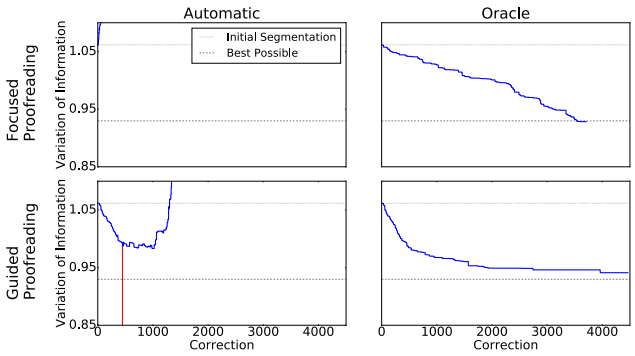


Figure 5: Performance comparison of Plaza's focused proofreading and our guided proofreading on 5 sections of the CREMI A dataset (split error correction only). All measurements are reported as median VI, the lower the better. The threshold for automatic selection is $p_t = 0.95$ (red line). The slope of the selection oracle shows that guided proofreading reduces VI faster.

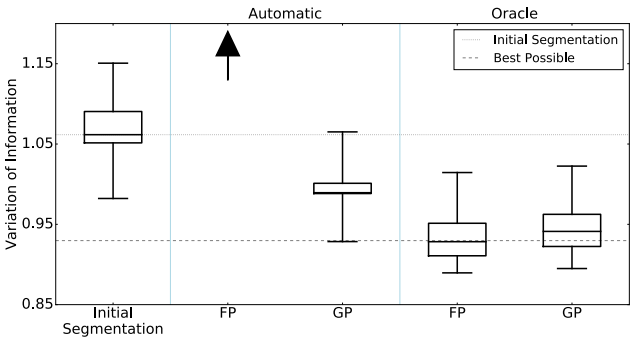


Figure 6: VI distributions of guided proofreading (GP) and focused proofreading (FP) output across slices of the CREMI A dataset, with different error correction approaches. The variation resulting from performance of FP with automatic selection is $5.4\times$ higher than GP (as indicated by the arrow), with median VI of 5.32 and $SD = 0.009$. GP does not reach the best possible VI as discussed in the text.

datasets include manually-labeled 'ground truth', we use these three volumes for our experiments. The volumes are part of an adult fruit fly (Drosophila melanogaster) brain. The resolution of all three datasets is $4 \times 4 \times 40$ nm^3/voxel.

Retraining. Since the CREMI data is a different species, we simply retrain our split error classifier as well as focused proofreading by Plaza [2]. For this, we use the first 100 sections of each of the three CREMI datasets combined as training data. All parameters are unchanged and left as reported in the paper.

Error correction. In all three datasets merge error detection found over 300 merge errors. Unfortunately, merge error correction crashed because of a software error on our part. Therefore, we only evaluate split error detection and correction on subvolumes of CREMI A/B/C with the dimensions $1250 \times 1250 \times 5$ voxels. The subvolumes were cut from the last 25 sections of each of the three datasets and unseen during training. We compare focused proofreading and guided proofreading with automatic selection ($p_t = 0.95$) and selection oracle.

3.1. CREMI A

Figure 5 and 6 compare Plaza's focused proofreading and guided proofreading on the 5 sections of the CREMI A dataset.

Selection oracle. With focused proofreading, the selection oracle reduces median VI to 0.928, $SD = 0.043$ from an initial median VI of 1.06 ($SD = 0.055$). 532 corrections out of 3707 were accepted. Guided proofreading does not reach the best possible VI, however, reduces VI faster with less corrections to 0.941 ($SD = 0.04$). Out of 4463 corrections, 1275 were accepted.

Automatic selection with threshold. Not surprisingly, focused proofreading performs poorly when ran automatically (VI of 5.32, $SD = 0.009$). Guided proofreading is able to reduce VI to 0.989 ($SD = 0.043$) with $p_t = 0.95$.

3.2. CREMI B

Figure 7 and 8 show the results on the CREMI B dataset.

Selection oracle. Focused proofreading is able to reduce median VI to 1.29, $SD = 0.031$ from an initial median VI of 1.63 ($SD = 0.025$). Out of 1959 corrections, the selection oracle accepted 517. With guided proofreading, the median VI is reduced to 1.30, $SD = 0.03$ while accepting 1111 corrections out of 3073.

Automatic selection with threshold. Focused proofreading results in a VI of 4.25 ($SD = 0.07$). Guided proofreading reduces median VI to 1.43 ($SD = 0.038$).

3.3. CREMI C

The results of split error correction using focused proofreading and guided proofreading on the CREMI C subvolume are shown in Figure 9 and 10.

Selection oracle. With focused proofreading, the initial median VI of 1.75 ($SD = 0.086$) is reduced to 1.45 ($SD = 0.056$) with 670 accepted corrections out of 2694. Guided proofreading is able to reduce the VI to 1.47 ($SD = 0.06$). Here, the oracle accepted 1531 out of 4332 corrections.

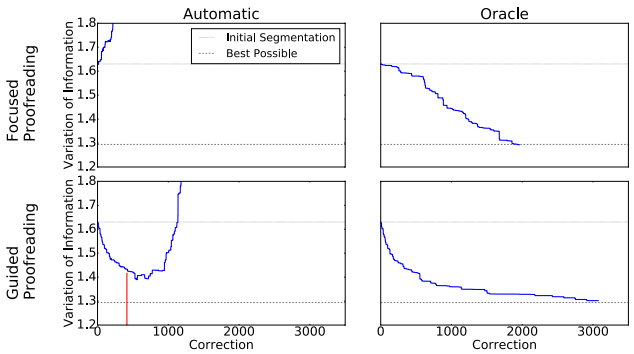


Figure 7: Split error correction by Plaza's focused proofreading and our guided proofreading compared on the CREMI B dataset. All measurements are reported as median VI, the lower the better. Automatic selection with threshold (red line) yields reasonable performance using guided proofreading.

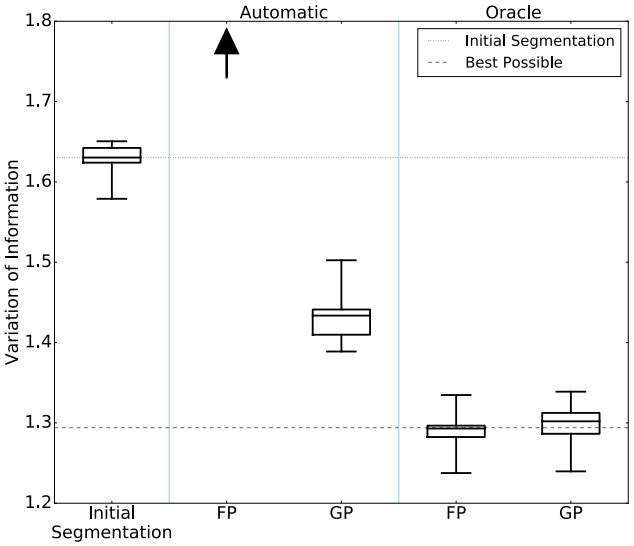


Figure 8: VI distributions of guided proofreading (GP) and focused proofreading (FP) output across 5 sections of the CREMI B dataset. We compare automatic selection and oracle selection. The variation resulting from performance of FP with automatic selection is $3\times$ higher than GP (as indicated by the arrow), with median VI of 4.25 and $SD = 0.07$.

Automatic selection with threshold. Focused proofreading results in a VI of 4.81 ($SD = 0.03$). Guided proofreading with $p_t = 0.95$ reduces median VI to 1.57 ($SD = 0.081$).

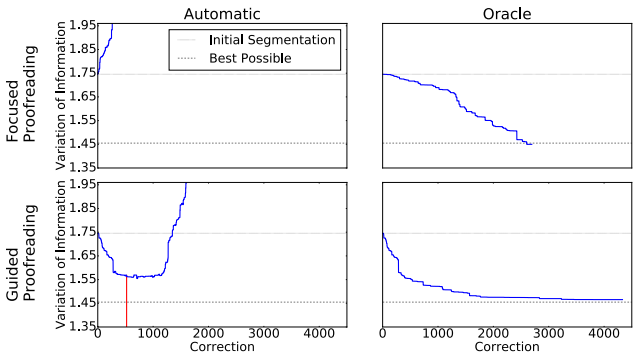


Figure 9: Performance comparison of Plaza's focused proofreading and our guided proofreading on the CREMI C dataset (only split error correction). Lower VI scores are better. Guided proofreading corrects the initial segmentation faster with less corrections than focused proofreading.

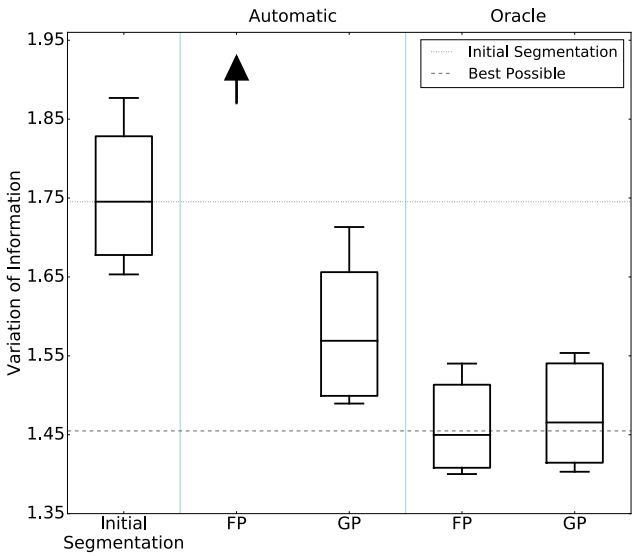


Figure 10: VI distributions of guided proofreading (GP) and focused proofreading (FP) output across the CREMI C subvolume, with different error correction approaches. The variation resulting from performance of FP with automatic selection is $3\times$ higher than GP (as indicated by the arrow), with median VI of 4.81 and $SD = 0.08$.

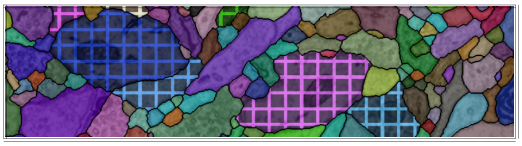
4. Forced Choice User Experiment

4.1. Recruitment and Participation

Novice participants were recruited via flyer (figure 11). An anonymized listing of all participants including demographic information is shown in table 3.

4.2. Example Classifications

During the user study, participants were asked to accept or reject potential errors and their corrections — some more



Get **$10** Cash!

And look at

Pretty Pictures

of the brain while

helping to **Advance Science**

We are looking for people who are 18+ and have no experience with nano-scale electron microscopy data of neurons (noobs).

The experiment will last less than 1 hour.

Starting NOW!

SIGN UP:

http://XXX/YYXXXXXZZZZ

Contact: Anon. <anon@anon>

Anon.

Figure 11: Participants were recruited with this flyer.

difficult than others. Figure 13 shows a selection of potential errors and their corrections.

4.3. Subjective Responses

After the experiment, we acquired subjective responses using the NASA-TLX task load index (figure 12). We performed ANOVA to test for statistical significance [3]. Mental, physical, and temporal demands were reported slightly higher for participants using focused proofreading but the analysis did not yield any significance.

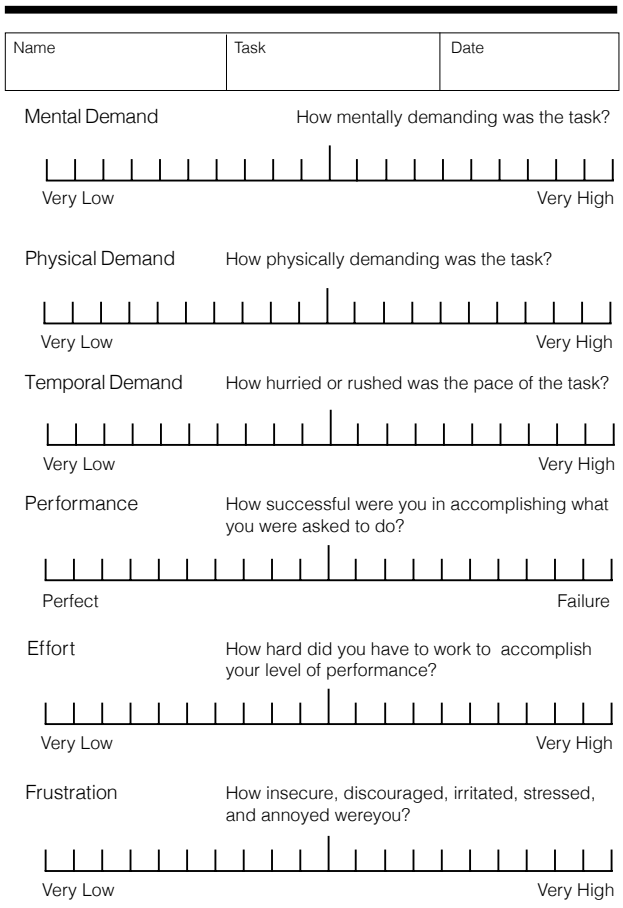
- **Mental Demand.** Participants using focused proofreading stated a higher mental demand $M = 11.5$ ($SD = 2.098$) than with guided proofreading $M = 8.1$ ($SD = 2.003$). This was not statistically significant ($F_{1,18} = 3.2574, p = 0.3695$).

- **Physical Demand.** While naturally physical demand was rated low, participants using focused proofreading stated it slightly higher $M = 5.4$ ($SD = 2.26$) than with guided proofreading $M = 2.9$ ($SD = 1.76$). This was not statistically significant ($F_{1,18} = 1.7507, p = 0.5454$).

- **Temporal Demand.** For temporal demand, participants using focused proofreading $M = 8.4$ ($SD = 1.95$) reported almost equal to guided proofreading $M = 8.3$ ($SD = 1.99$). This was not statistically significant ($F_{1,18} = 0.0033, p = 0.9987$).

- **Performance.** Here, participants were asked to rate their own performance. All participants rated their performance as pretty well (the lower, the better). For focused proofreading $M = 6.8$ ($SD = 1.97$) and for guided proofreading $M = 7.8$ ($SD = 2.04$). This was not statistically significant ($F_{1,18} = 0.3091, p = 0.8878$).

- **Effort.** Participants using focused proofreading stated higher effort $M = 13.0$ ($SD = 2.336$) than with guided proofreading $M = 10.6$ ($SD = 2.127$). This was not statistically significant ($F_{1,18} = 1.1459, p =$



NASA Task Load Index

Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.

Name	Task	Date

Mental Demand — How mentally demanding was the task?
Very Low — Very High

Physical Demand — How physically demanding was the task?
Very Low — Very High

Temporal Demand — How hurried or rushed was the pace of the task?
Very Low — Very High

Performance — How successful were you in accomplishing what you were asked to do?
Perfect — Failure

Effort — How hard did you have to work to accomplish your level of performance?
Very Low — Very High

Frustration — How insecure, discouraged, irritated, stressed, and annoyed wereyou?
Very Low — Very High

Figure 12: The NASA-TLX workload index to record subjective responses.

ID	Sex	Age	Classifier
S38	F	20	FP
S57	F	30	FP
S32	M	38	FP
S34	F	21	FP
S21	F	65	FP
S9	M	33	FP
S45	M	28	FP
S31	M	27	FP
S24	F	21	FP
S6	F	38	FP
S28	M	32	GP
S36	F	19	GP
S35	M	26	GP
S25	M	26	GP
S54	F	30	GP
S53	M	29	GP
S52	M	27	GP
S51	M	31	GP
S200	F	37	GP
S3	F	30	GP

Table 3: The novice participants ($N = 20$) of the forced choice user experiment. The table shows sex (20 female), age ($M = 30$) and the randomly assigned classifier (focused proofreading as FP, guided proofreading as GP).

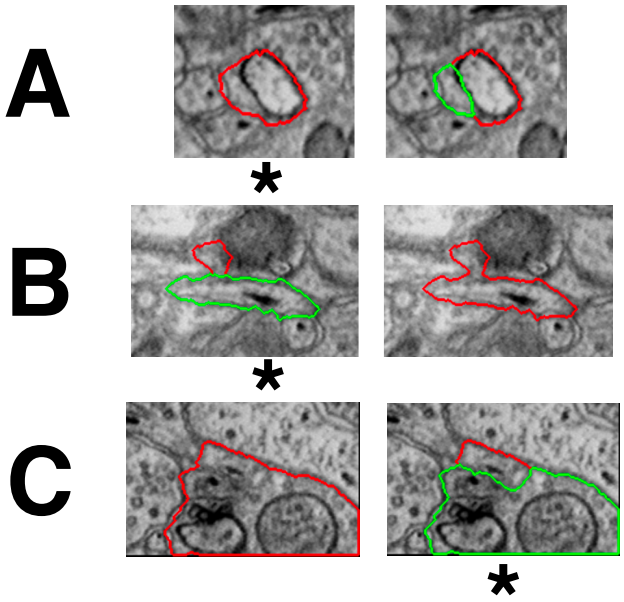


Figure 13: A selection of suggested errors and potential corrections during the forced choice user experiment. The star (*) indicates which choice reduces VI. While all participants were able to correctly choose for patch A, only few were able to correctly choose for patch B and C.

0.6599).

- **Frustration.** Participants overall reported low frustration. Reported were $M = 5.0$ ($SD = 1.90$) using focused proofreading and $M = 5.9$ ($SD = 185$) using guided proofreading. This was not statistically significant ($F_{1,18} = 0.3271, p = 0.8818$).

References

[1] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1

[2] S. M. Plaza. Focused Proofreading: Efficiently Extracting Connectomes from Segmented EM Images, Sept. 2014. 3

[3] J. P. Shaffer. Multiple hypothesis testing. *Annual Review of Psychology*, 46(1):561–584, 1995. 5