

000  
001  
002  
003  
004  
005  
006  
007  
008  
009  
010  
011  
012  
013  
014054  
055  
056  
057  
058  
059  
060  
061  
062  
063  
064  
065  
066  
067  
068  
069  
070  
071  
072  
073  
074  
075  
076  
077  
078  
079  
080  
081  
082  
083  
084  
085  
086  
087  
088  
089  
090  
091  
092  
093  
094  
095  
096  
097  
098  
099  
100  
101  
102  
103  
104  
105  
106  
107

# 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:	<b>3x3, 5x5, 9x9, 13x13</b>
No. Filters 1:	32, 48, <b>64</b>
No. Filters 2-4:	32, <b>48</b> , 64
Dense units:	256, <b>512</b>
Learning rate:	0.00001, 0.0001, 0.001, 0.01, <b>0.03-0.00001</b>
Momentum:	0.9, 0.95, <b>0.9-0.999</b>
Mini-Batchsize:	10, 100, <b>128</b>

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

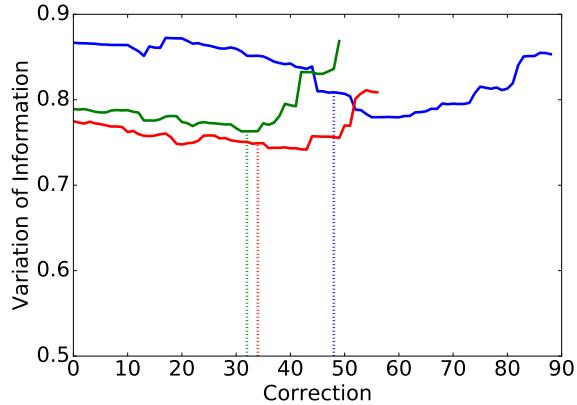


Figure 1: Observations of probability thresholds  $p_t$  during automatic selection on three different subvolumes of previously unseen testing data. The dashed lines show when  $p_t = 0.95$  is reached.

## 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.

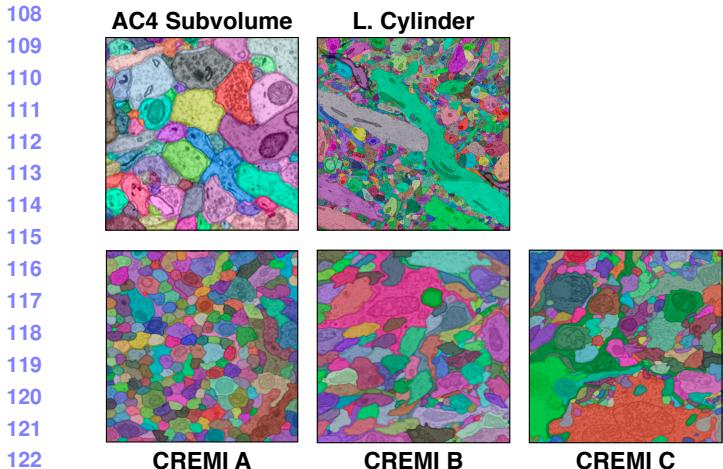


Figure 2: The five different datasets we use for evaluation. The top row shows the first slice of the AC4 and L. Cylinder mouse brain datasets as reported in the paper. The bottom row shows the first slice of the CREMI A/B/C fruit fly datasets which we used for additional experiments.

#### 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. In our experiments, parameters did 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, due to the classification patch size, the membrane probability maps we used included a 30 pixel frame region. Guided proofreading ignores all segments within this frame region, and so cannot reach the best possible VI in some datasets.

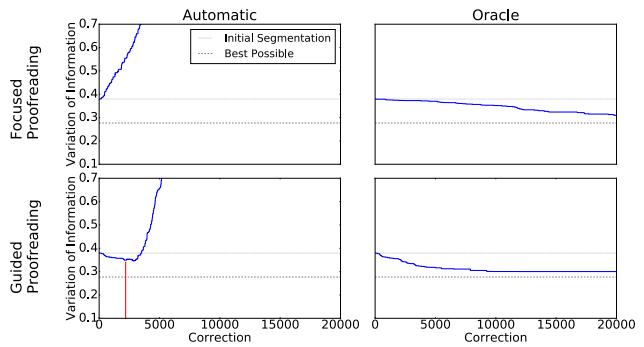


Figure 3: Performance comparison of Plaza’s focused proofreading and our guided proofreading on the L. Cylinder dataset as reported in the paper. All measurements are shown as median VI, the lower the better. We compare automatic selection with threshold ( $p_t = 0.95$ , red line) and the selection oracle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

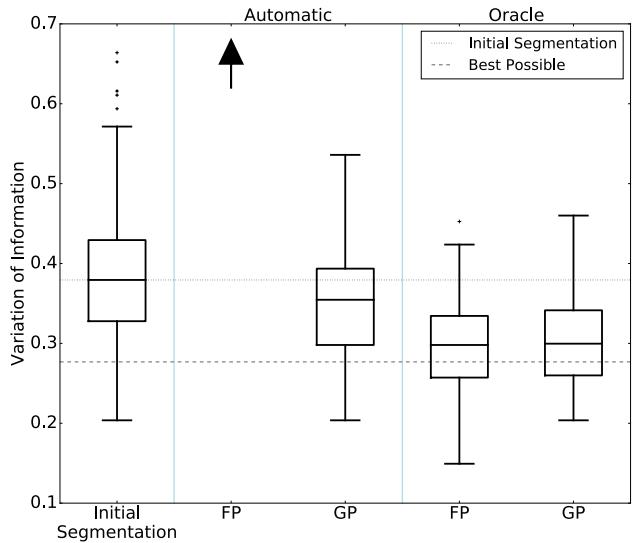


Figure 4: VI distributions of guided proofreading (GP) and focused proofreading (FP) output across slices of the L. Cylinder dataset, with different error correction approaches. The variation resulting from performance of FP with automatic selection is  $7.8 \times$  higher than GP (as indicated by the 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<sup>1</sup>, each  $1250 \times 1250 \times 125$  voxels. Since only three

<sup>1</sup><http://www.creml.org>

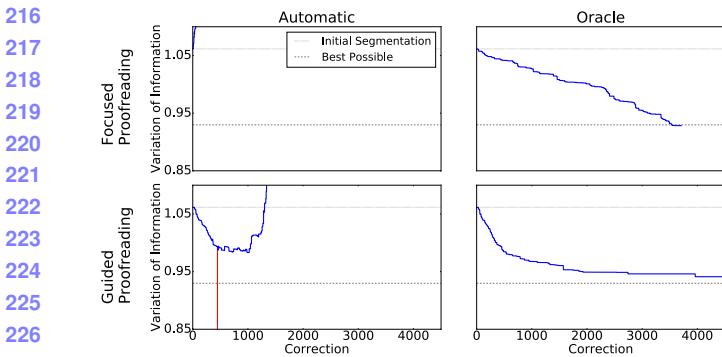


Figure 5: Performance comparison of Plaza’s focused proofreading and our guided proofreading on 5 sections of the CREMI A dataset. All measurements are reported as median VI, the lower the better. The threshold for automatic selection is  $p_t = 0.95$  (red line). The slope of the selection oracle shows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

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<sup>3</sup>/voxel. We evaluate 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.

**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.

### 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

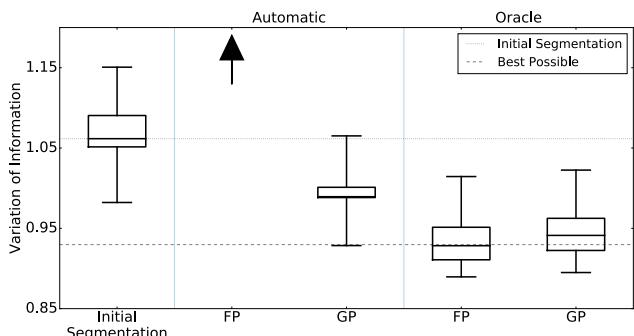


Figure 6: VI distributions of guided proofreading (GP) and focused proofreading (FP) output across slices of the CREMI A dataset, with different error correction approaches. The variation resulting from performance of FP with automatic selection is  $5.4 \times$  higher than GP (as indicated by the arrow),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

(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.

**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 ( $SD = 0.081$ ).

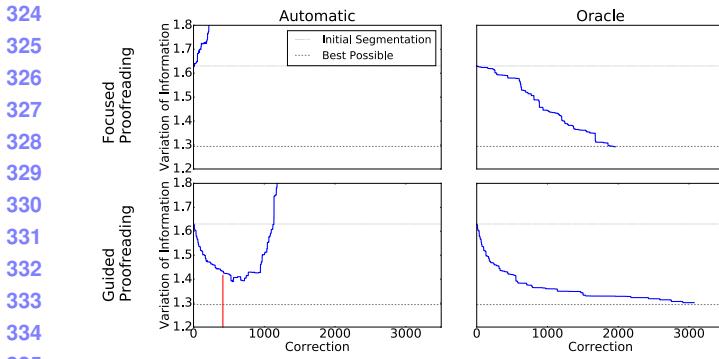


Figure 7: Split error correction by Plaza’s focused proofreading and our guided proofreading compared on the CREMI B dataset. All measurements are reported as median VI, the lower the better. Automatic selection with threshold (red line) yields reasonable performance using guided proofreading.

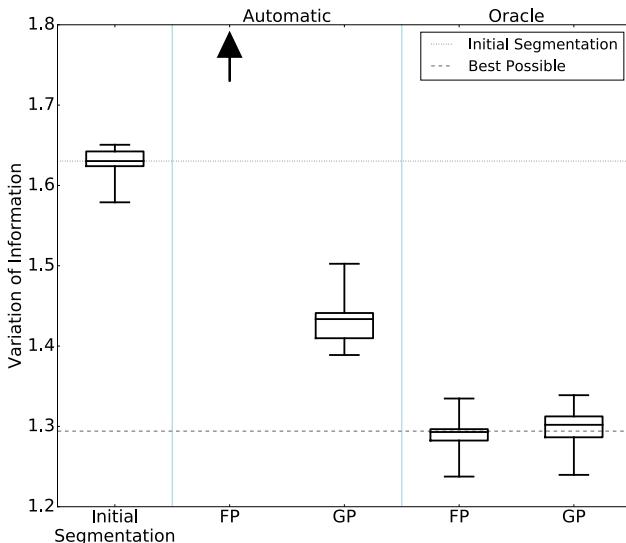


Figure 8: VI distributions of guided proofreading (GP) and focused proofreading (FP) output across 5 sections of the CREMI B dataset. We compare automatic selection and oracle selection. The variation resulting from performance of FP with automatic selection is 3× higher than GP (as indicated by the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

## 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.

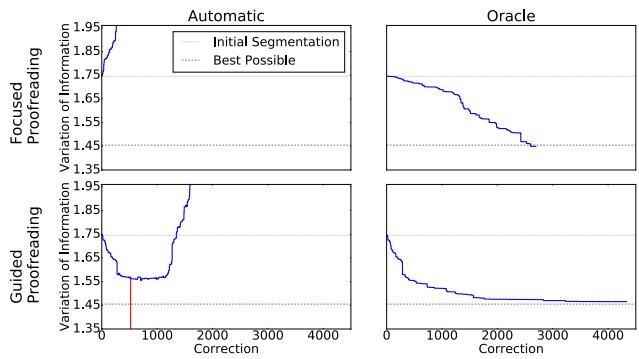


Figure 9: Performance comparison of Plaza’s focused proofreading and our guided proofreading on the CREMI C dataset. Lower VI scores are better. Guided proofreading corrects the initial segmentation faster with less corrections than focused proofreading.

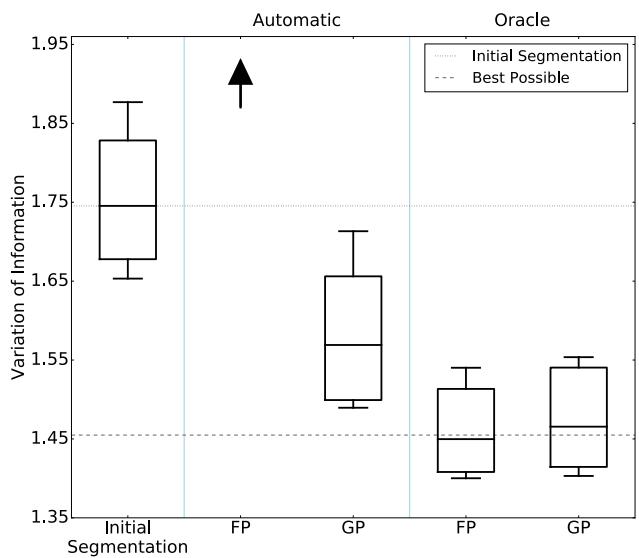


Figure 10: VI distributions of guided proofreading (GP) and focused proofreading (FP) output across the CREMI C subvolume, with different error correction approaches. The variation resulting from performance of FP with automatic selection is 3× higher than GP (as indicated by the arrow), with median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

### 4.2. Example Classifications

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

A horizontal strip of a colorful abstract painting. The composition includes a blue grid pattern on the left, followed by a large area of irregular, overlapping shapes in various colors like purple, green, yellow, and brown. A pink grid pattern is positioned in the center-right portion of the strip.

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.

The experiment will last less than 1 hour.  
**Starting NOW!**

**SIGN UP:**  
**http://XXX/YYXXXXXXZZZZ**

Contact: Anon. <[anon@anon](mailto:anon@anon)>  
Anon.

Figure 11: Participants were recruited with this flyer.

formed 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

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.*

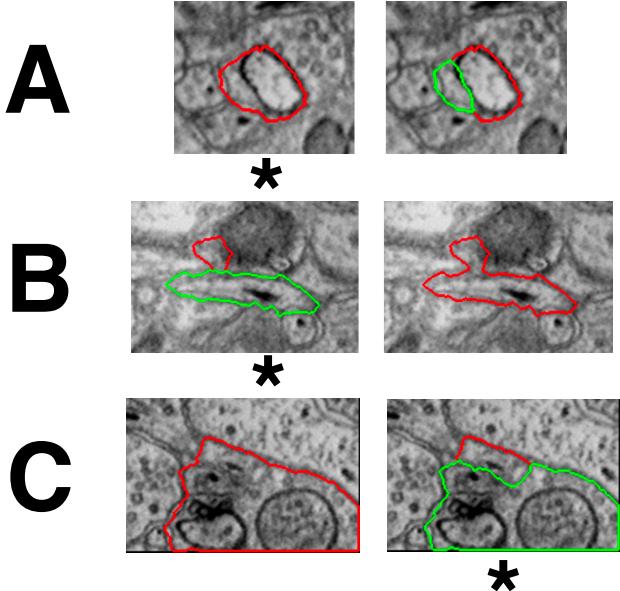
Figure 12: The NASA-TLX workload index to record subjective responses.

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 ( $F_{1,18} = 1.1459, p = 0.6599$ ).
  - **Frustration.** Participants overall reported low frustration. Reported were  $M = 5.0$  ( $SD = 1.90$ ) using focused proofreading and  $M = 5.9$  ( $SD = 1.85$ ) using guided proofreading. This was not statistically significant.

540	ID	Sex	Age	Classifier
541	S38	F	20	FP
542	S57	F	30	FP
543	S32	M	38	FP
544	S34	F	21	FP
545	S21	F	65	FP
546	S9	M	33	FP
547	S45	M	28	FP
548	S31	M	27	FP
549	S24	F	21	FP
550	S6	F	38	FP
551	S28	M	32	GP
552	S36	F	19	GP
553	S35	M	26	GP
554	S25	M	26	GP
555	S54	F	30	GP
556	S53	M	29	GP
557	S52	M	27	GP
558	S51	M	31	GP
559	S200	F	37	GP
560	S3	F	30	GP

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



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

594  
 595  
 596  
 597  
 598  
 599  
 600  
 601  
 602  
 603  
 604  
 605  
 606  
 607  
 608  
 609  
 610  
 611  
 612  
 613  
 614  
 615  
 616  
 617  
 618  
 619  
 620  
 621  
 622  
 623  
 624  
 625  
 626  
 627  
 628  
 629  
 630  
 631  
 632  
 633  
 634  
 635  
 636  
 637  
 638  
 639  
 640  
 641  
 642  
 643  
 644  
 645  
 646  
 647

## 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