

Supplemental Information

For Automated Defect Analysis in Electron Microscopy Images

Wei Li, Kevin Field, Dane Morgan

Section I. Analysis on Systematic Issues

To obtain more detailed understanding in how errors in predictions correlate with test image characteristics, we evaluated the automated model recall and precision score for loop defects with varying brightness, contrast, and loop size. We measured the values for each loop bounding box region as follows: brightness was given by the mean value of pixel intensity, contrast was given by the standard deviation of pixel intensity, and size was given by an effective radius equal to the square root of area divided by π . We examined each bounding box with a loop in all six test images, divided all loop defects into 16 bins, and aggregated the recall and precision score for each bin. We didn't measure the focus of loops as it is hard to determine a single value to represent the level of focus by applying a computer-generated image filter such as a Gaussian blur. Note, that one could systematically apply a level of focus during the electron microscopy experiment, but such activity is prohibitively time consuming and rarely is focus the largest variant within a given image dataset. As such, variation in focus was not a primary focus, only brightness, contrast, and size were.

Figure S1 shows how different image characteristics affects the model performance. In general, the precision and recall curve drops when the number of ground truth in the bin decreases. The analysis revealed there was no direct correlation of model performance with brightness, contrast and loop size. The model performance was rather highly influenced by the frequency of loops in the dataset, i.e. the more frequent a certain type of loops appears (e.g. loops of typical brightness) in the dataset, the better automated model performs for that type.

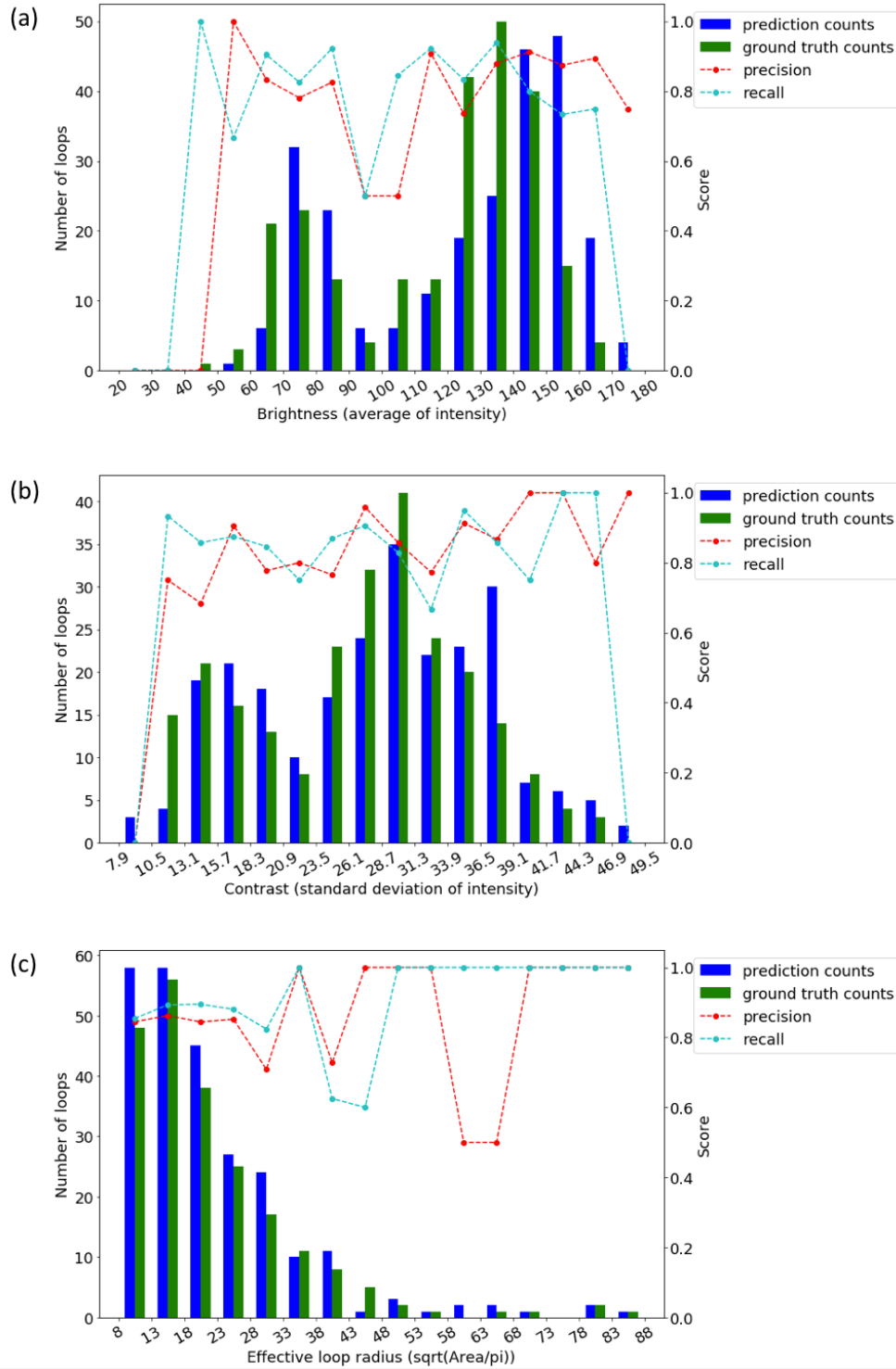


Figure S1 Precision and recall of automated model on loops with different brightness, contrast, and size. Blue bar height indicates the number of prediction counts in the bin range, and the green bar height indicates the number of ground truth counts in the bin range. The precision and recall is evaluated for the loops in each bin.

Section II. Loop Labeling Tutorial Video

The loop labeling tutorial is available as the supplementary video along with the publication of the article. (Also available online at <https://www.youtube.com/watch?v=OGHBY2XIjVs>.) The tutorial video was prepared by two experienced researchers (two of the authors, Li and Field) who prepared the training and test data set.