Improving U-Net Confidence on TEM Image Data with L2-Regularization, Transfer Learning, and Deep Fine-Tuning

Supplementary Material

Aiden Ochoa Xinyuan Xu Xing Wang
Ken and Mary Alice Lindquist Department of Nuclear Engineering, Penn State University
{aio5165, xkx5062, xvw5285}@psu.edu

A. Supervised Holistic Metrics

Segmentation is an intermediate processing step where desired features, such as defect structures, are identified from input transmission electron microscopy (TEM) images. Machine learning metrics like loss, precision, recall, and F1-score, focus on the similarity between human annotations and model predictions. Therefore, these metrics may not fully capture the realistic usefulness of a model. In Secs. 4 and 5 of the main text, we evaluated the practical usefulness of a model using the total number of defects detected from model predictions. The detected defect count is a holistic metric because it intrinsically depends on all ML metrics. That is, a model must not only have good classification performance, but also must avoid noisy and uncertain predictions, as measured by the proposed pixel value histogram metrics, certainty and abundance.

Detected defect count can act as an effective and unsupervised indicator of practical model performance, but it does not measure the ML model accuracy. For example, the post-processing algorithm CHAC [4] could detect the same number of grains in a ground truth image and a model prediction image, but their shapes and locations could be very different. Therefore, we also included two additional supervised holistic metrics to capture model accuracy and to provide a more complete evaluation of practical model performance. As discussed in Sec. 4.3, the use of the human annotation as ground truth does introduce some bias due to the presence of human errors, but the insights gained are nonetheless valuable.

A.1. Intersection-over-Union

A common metric used in object detection, intersection-over-union (IoU) can be presently used to measure the similarity between defects detected in ground truth images and model prediction images [2]. IoU provides information about the accuracy of both defect shape and location. Although it is

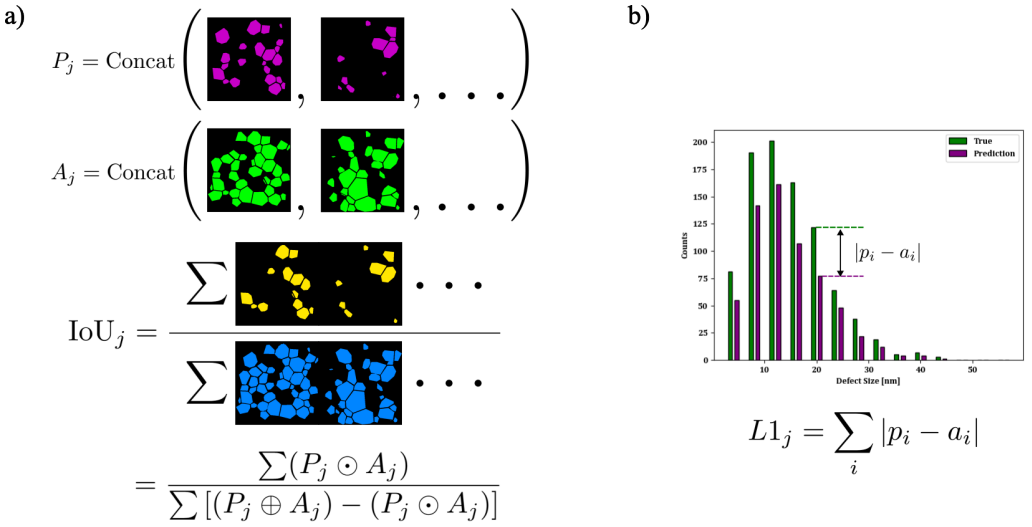

Figure A.1. Visualization of holistic metrics. (a) shows intersection-over-union (IoU) using the grain boundary dataset. j corresponds to a specific cross-validation model. P, A corresponds to concatenated validation predictions and corresponding ground truth annotations. Colors are shown only for visualization where purple are predictions, green are ground truth annotations, yellow are their intersection, and blue are their union. (b) shows L1-distance calculated, in this case, using grain diameter data. Again, j corresponds to a specific cross-validation model. i corresponds to a specific size bin

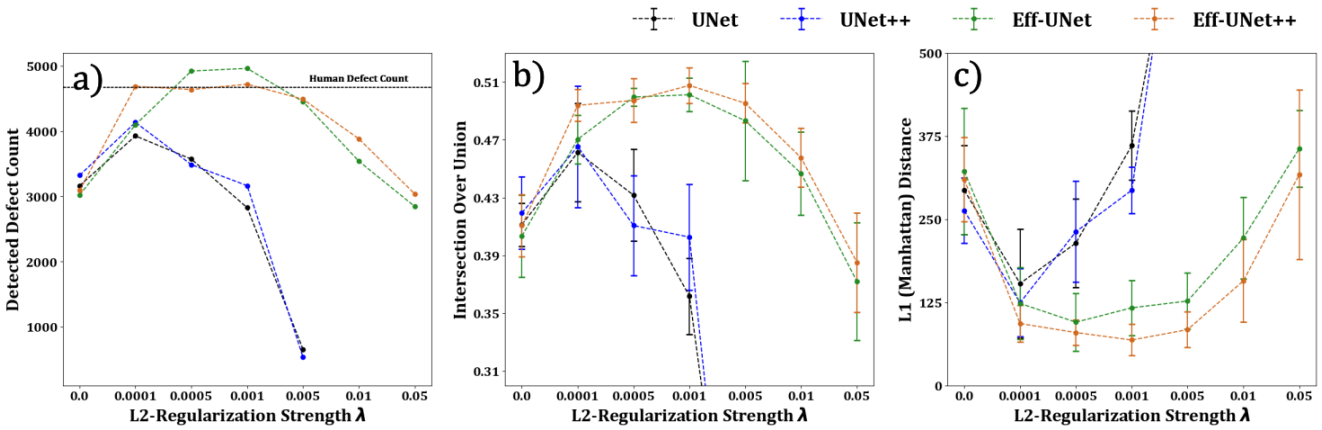

Figure A.2. Holistic metrics plotted by L2-regularization strength λ and U-Net architecture for the grain boundary dataset using the CHAC post-processing algorithm. (a) is the number of detected defects (grains) across the entire dataset, (b) is the intersection-over-union, and (c) is the L1-distance based on grain diameter

similar to F1-score, IoU is a holistic metric because it operates on the final, post-processed dataset, and thus describes the ability of a model to produce usable segmentations yielding defects consistent with ground truth data. It is calculated using the equation

$$\mathrm{IOU}(P, A) = \frac{\sum(P \odot A)}{\sum[(P \oplus A) - (P \odot A)]}$$

where P, A are the concatenation of all post-processed binary validation predictions and ground truth annotations, respectively, $\oplus, \odot$ are element-wise addition and multiplication, respectively, and $\sum \cdot$ is the sum over all pixel values for the resulting binary arrays. A pictorial visualization of IoU using the grain boundary dataset is shown in Fig. A.1a.

A.2. L1 (Manhattan) Distance

Defects are often studied statistically by computing the histogram of their sizes. Models which yield size histograms similar to the ground truth histogram will be of more practical value. Measuring the dissimilarity between histograms is a common mathematical procedure with many available techniques, the simplest of which being L1, or Manhattan, distance [1]. It is calculated using the equation

$$\mathrm{L1}(p, a) = \sum_i |p_i - a_i|$$

where p, a are prediction and ground truth annotation defect size histograms and i corresponds to a specific bin. If p and q are left unnormalized, L1-distance will depend on both sample shape accuracy and sample size. It therefore represents the overall ability of a model to yield quality defect statistics. It is visualized using the grain boundary dataset in Fig. A.1b.

A.3. Comparison of Holistic Metrics

Fig. A.2 shows results for the three holistic metrics discussed, applied to the grain boundary dataset. The prediction quality improvement from L2-regularization and the EfficientNetB7 encoder seen in Figs. 4 and 5 is repeated by IoU in Fig. A.2b, where an increase from 0.41 to 0.50 is observed. This implies that there is a greater agreement with the ground truth, but a substantial disagreement (0.50) still persists. Given that Fig. A.2a shows many more grains are being predicted, even surpassing the ground truth, most appear exclusively in either the prediction or ground truth, not both. This is demonstrated by a few representative examples taken from the Eff-UNet architecture (λ=1e-3) in Fig. A.3. Despite the mismatched position of many grains, L1-distance in Fig. A.2c shows that this architecture still yields more similar grain shapes on average than insufficiently regularized models. Furthermore, it indicates that Eff-UNet only marginally outperforms UNet when both are optimally regularized, despite the improvements to model self-confidence by the former. L1-distance is also one of the only cases where Eff-UNet++ clearly outperforms Eff-UNet, indicating some potential value in the U-Net++ architecture.

B. Generalization to Other TEM Datasets

In this work, it has been shown that EfficientNetB7 and L2-regularization contribute important prediction quality enhancements, at least for the grain boundary dataset used. Since every TEM dataset is unique, typically having a unique segmentation task, the methodology and conclusions must be able to generalize to other datasets. Here, the λ grid search and architecture comparison (Fig. 5) was repeated for two datasets with their own specific materials, environmental conditions, imaging conditions, and defect type of interest. The

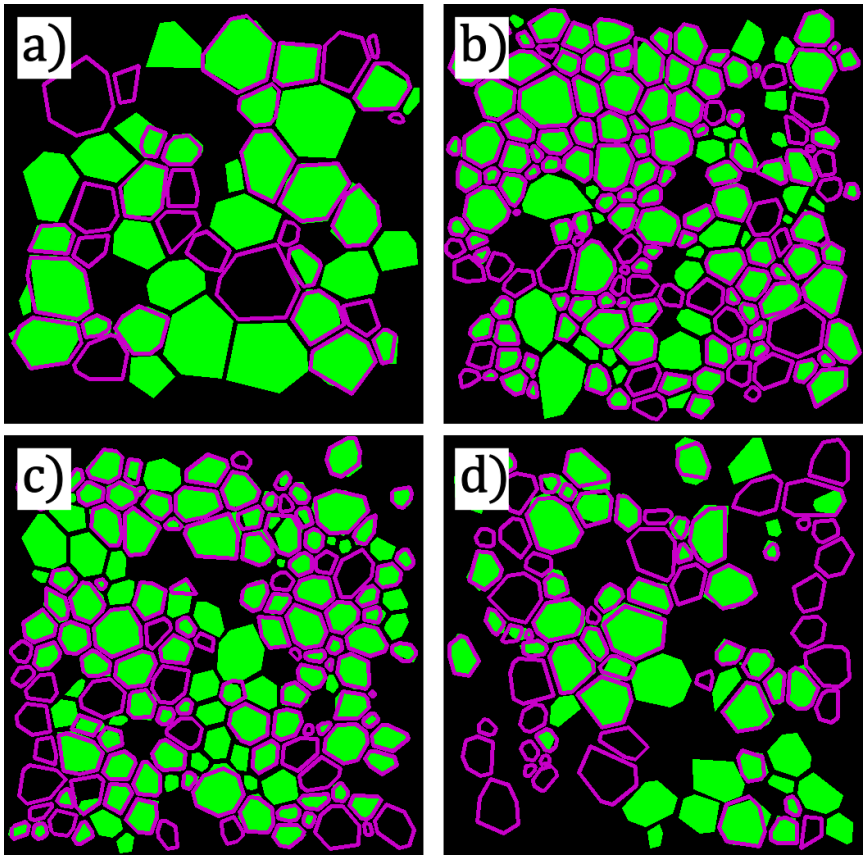

Figure A.3. Visualization of IoU results for validation predictions from the Eff-UNet, λ=1e-3 model. Purple polygons correspond to grains detected in prediction images, while green polygons correspond to grains detected in ground truth annotation images

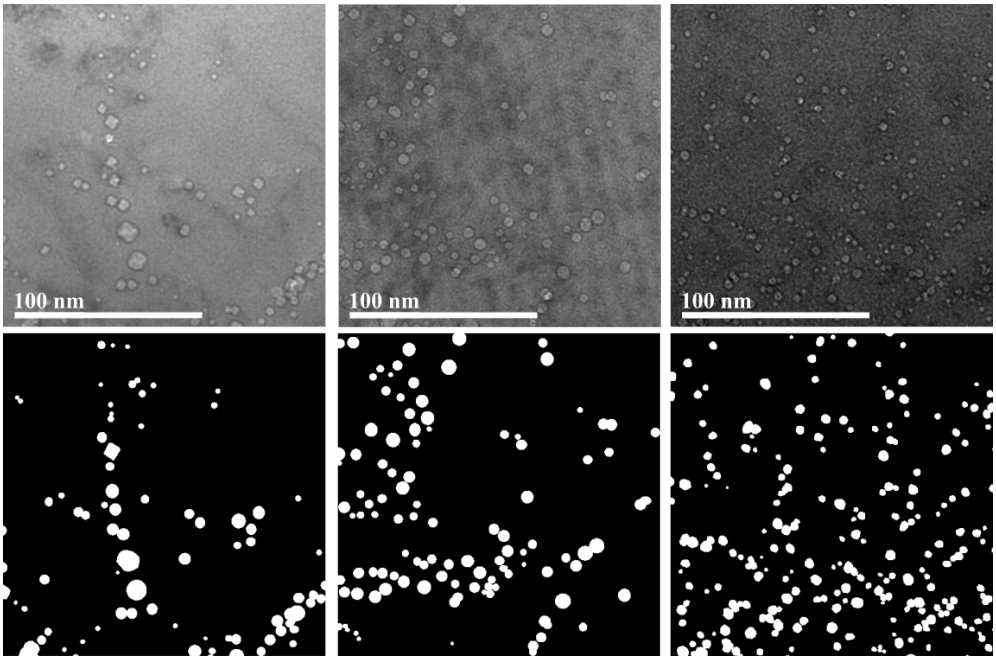

Figure A.4. Representative helium bubble dataset TEM images (top) and corresponding ground truth annotation images (bottom)

same dataset pre-processing procedure was applied to both datasets and the same hyperparameters were used for training. UNet++ and Eff-UNet++ architectures were neglected due to their minimal difference with UNet and Eff-UNet architectures.

B.1. Helium Bubble Dataset

The first dataset is composed of 10 bright field TEM images which have been downscaled to 1024×1024, then quartered into 4-512×512 images. The defect of interest is helium bubbles, which are a result of 200 keV He^+ irradiation to a fluence of 5e16 ions/cm2. The bulk materials imaged are Ni-based alloys with varying concentrations of minor alloying elements Fe, Cr, Co, etc. More details about material fabrication and dataset creation can be found in [3]. The post-processing algorithm we call BubbleFinder was used. It is designed to fit circles around segmented bubbles using LoG and Hough transforms, and was also developed in [3]. Some representative images and their ground truth annotations from the dataset are presented in Fig. A.4.

The results for the bubble dataset (all metrics) are shown in Fig. B.1, where familiar trends are seen: (1) precision and recall trade off as λ increases with F1-score showing little improvement, other than UNet having higher scores on average, (2) certainty and abundance are higher on average for Eff-UNet models, and (3) more defects tend to be detected by Eff-UNet models for all λ values. Interestingly though, Figs. B.1g-h show very similar performance for both architectures, and the ML metrics in B.2a-e are much higher numerically than for the grain boundary dataset. In addition to the less exaggerated trends, this behavior is result of much less intrinsic ambiguity in the dataset. Nearly every bubble is obvious and with clear boundary, leading to less human errors and higher model self-confidence. As a result, little performance gains are to be expected from L2-regularization and pre-trained EfficientNet encoders.

B.2. Faulted Dislocation Loop Dataset

The second dataset is composed of 31 Rel-Rod TEM dark-field images of irradiated ion alloys. . The defect of interest is faulted dislocation loops of type $a/3\langle 111\rangle\{111\}$ which were imaged perpendicular to the habit plane such that they

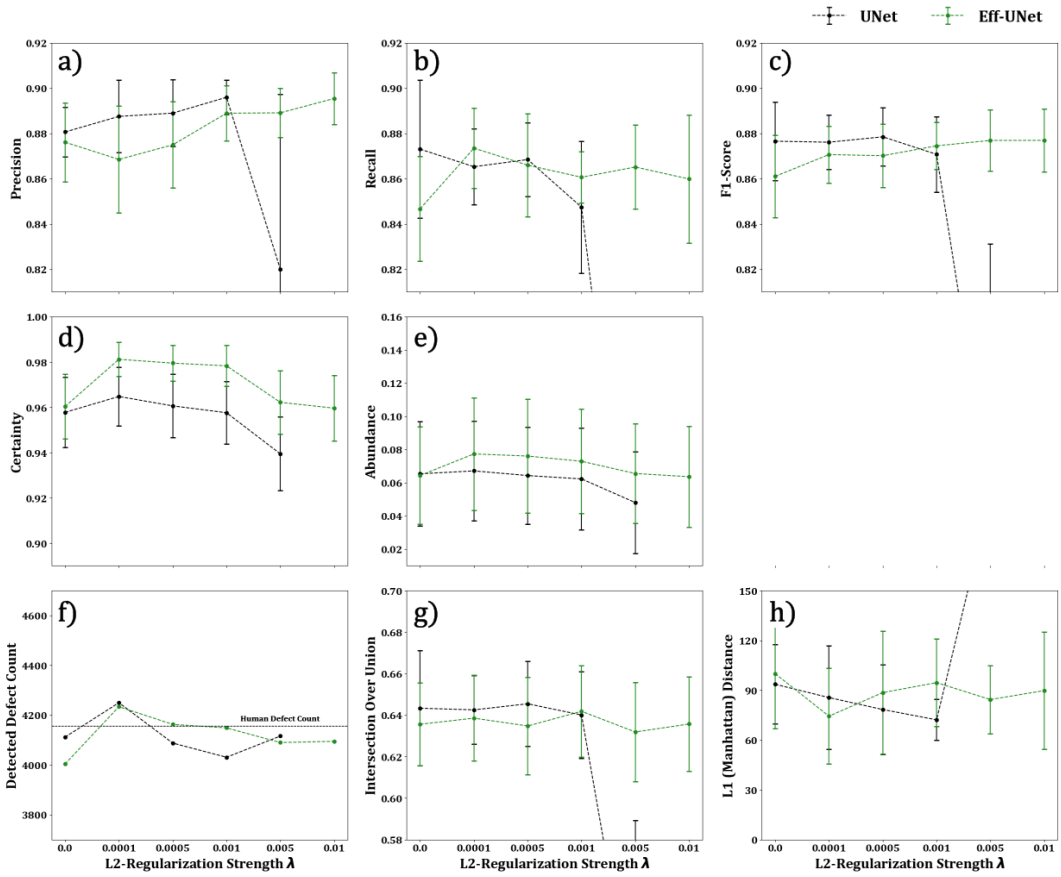

Figure B.1. All 8 metrics calculated for UNet and Eff-UNet architectures for the helium bubble dataset. (a-c) are classification metrics, (d-e) are histogram metrics, and (f-h) are holistic metrics

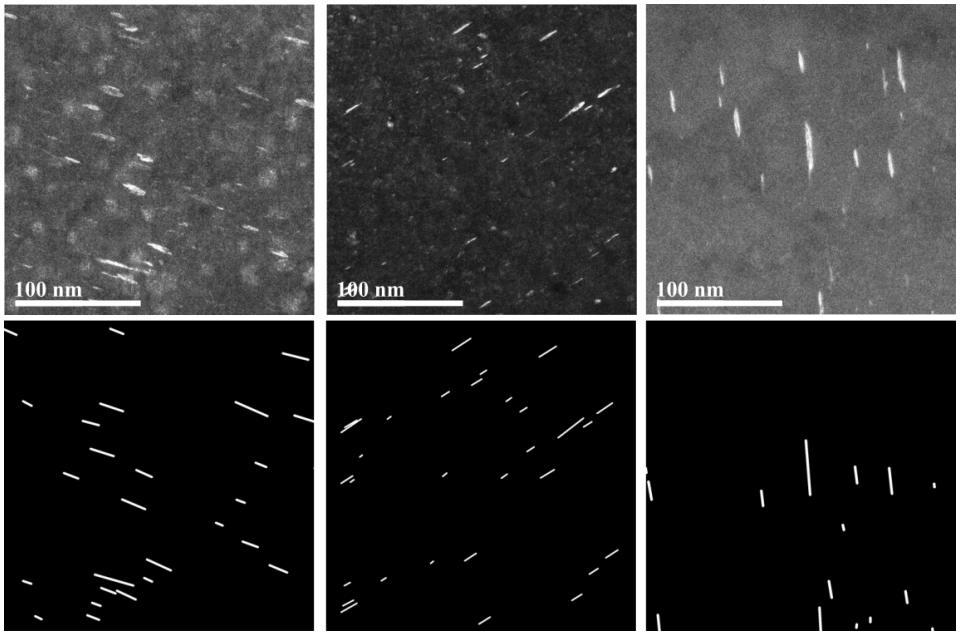

Figure B.2. Representative faulted dislocation dataset TEM images (top) and corresponding ground truth annotation images (bottom)

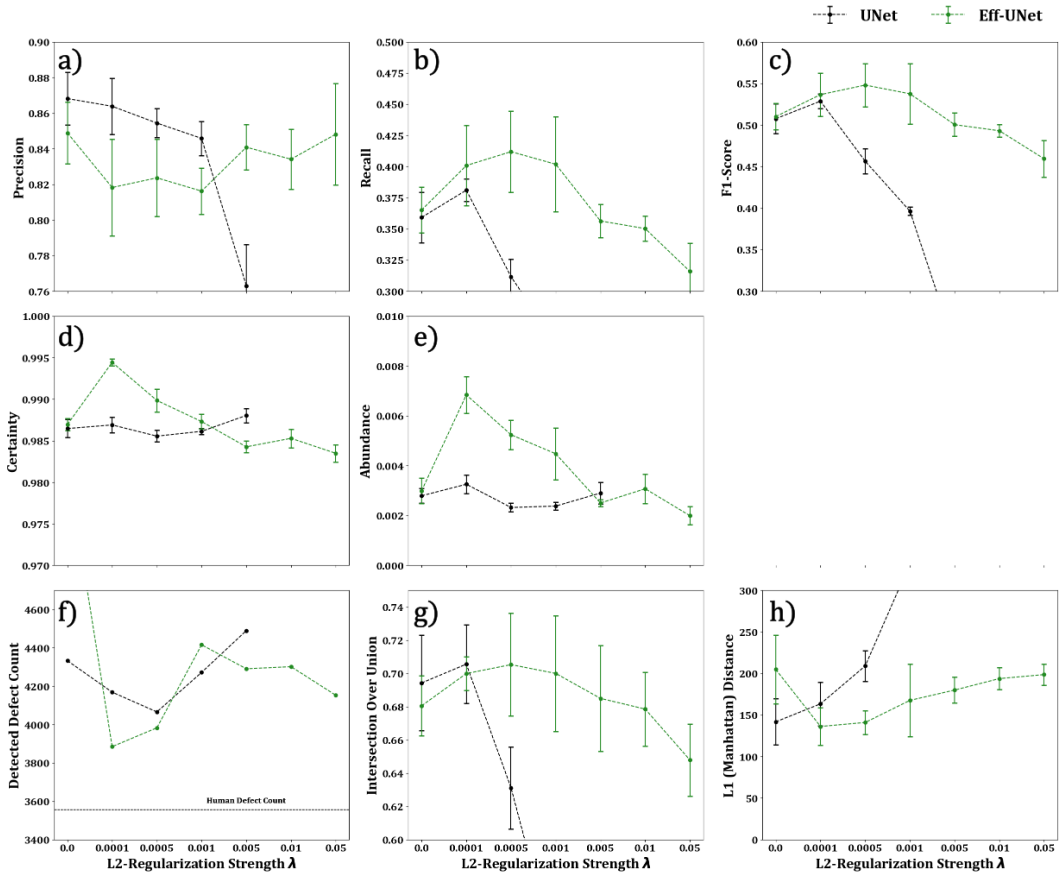

Figure B.3. All 8 metrics calculated for UNet and Eff-UNet architectures for the faulted dislocation dataset. (a-c) are classification metrics, (d-e) are histogram metrics, and (f-h) are holistic metrics

appear as lines not ellipses. The post-processing algorithm is a simple bounding box fitter which we call Bbox, allowing dislocation length to be extracted by measuring the bounding box diagonal length. Some representative TEM images and ground truth annotations from this dataset are shown in Fig. B.2.

The same trends with the grain boundary and bubble datasets are also present in the Fig. B.3 results for this dataset, only with some interesting differences: (1) precision and recall both show the expected tradeoff but F1-score is marginally, but consistently higher for Eff-UNet and recall is far lower at 0.38 for all models, (2) certainty and abundance are again much higher for Eff-UNet than U-Net, and it peaks at λ=1e-4 instead of λ=5-e4, 1e-3 and (3) Bbox detects less dislocations for more self-confident models, though it yields statistics more similar to ground truth data. This is because Bbox is a simple and naïve post-processor compared to CHAC and BubbleFinder, where it accepts any contiguous region of white pixels as a dislocation. Models with lower self-confidence, away from λ=1e-4, have noisier predictions, causing Bbox to detects many additional small ($<$5 nm), erroneous dislocations.

This dataset differs from both the grain boundary and bubble dataset because it has the lowest signal to noise ratio. In many cases it is difficult to differentiate a dislocation from precipitates or noise, leading to an ill-defined objective function. Despite this, notable improvements are still prevalent in models using optimal L2-regularization and EfficientNet. Together, the results for all three datasets (Figs. 4, B.1, and B.3) show that our methodology can generalize to other dataset, but the improvements are most significant when the objective function is not too ill-defined or well-defined.

Acknowledgments

We thank Xingyu Liu of Pennsylvania State University for providing the original TEM images and annotations of the dislocation loop data.

Code and Data Availability

The code developed for this work can be found on GitHub at https://github.com/psu-rdmap/unet-compare. All data, except for the dislocation TEM images, training results, and the accompanying analysis code can be found on ScholarSphere at https://scholarsphere.psu.edu/resources/b80356d7-6485-40aa-841d-8f598c4ee9e2.

References

[1] M. D. Malkauthekar. Analysis of euclidean distance and manhattan distance measure in face recognition. In *Third International Conference on Computational Intelligence and Information Technology (CIIT 2013)*, pages 503–507, 2013. 2

[2] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[3] Xing Wang, Ke Jin, Chun Yin Wong, Di Chen, Hongbin Bei, Yongqiang Wang, Maxim Ziatdinov, William J. Weber, Yanwen Zhang, Jonathan Poplawsky, and Karren L. More. Understanding effects of chemical complexity on helium bubble formation in ni-based concentrated solid solution alloys based on elemental segregation measurements. *Journal of Nuclear Materials*, 569:153902, 2022. 4

[4] Xinyuan Xu, Zefeng Yu, Wei-Ying Chen, Aiping Chen, Arthur Motta, and Xing Wang. Automated analysis of grain morphology in tem images using convolutional neural network with chac algorithm. *Journal of Nuclear Materials*, 588:154813, 2024. 1