Learning Atomic Defects Without Real Data: YOLO Trained on Fully Synthetic STEM Imagery

Victoria Augoustides, Jed Doman, Mahmoud Hawary, Tatiana Proksch, Jingyun Yang

Abstract:

Here we describe a method to synthesize STEM images of lattice structures with vacancy, insertion and gradient boundary defects. We then train an object detection model on this synthetic data and use it to identify these defects on an experimental image. This model achieves an accuracy of 99.5% on its synthetic training dataset, and in 0.36s/image, generates promising detections on an experimentally acquired CdTe STEM image.

Introduction:

Scanning transmission electron microscopy (STEM) is an advanced technique that enables atomic-resolution imaging [1]. STEM is used to image the structure and chemical composition in a sample, which influence material properties. The identification of defects in lattice structures, such as vacancies, interstitials, and stacking faults, is often performed manually and qualitatively, which is time-consuming and subject to personal bias. The technical progress of STEM now makes the microscopist rather than the microscope the rate-limiting factor in materials characterization. Limitations of manual analysis have been addressed via machine learning and artificial intelligence platforms [2]. For example, Ayyubi et al. [3], [4] use a conditional variational autoencoder (CVAE) to cluster spatial distributions of defect populations in smaller subregions of a larger STEM. In tandem, recent work from Kalinin et al. [5] showed the potential of LLMs to screen images for features of interest circumventing the need to train a dedicated model from scratch. Each approach highlights the potential of AI to be used to address specific limitations of automated STEM analysis. The CVAE model requires some amount of labeled image sub-regions to learn a feature space that delineates defects from the bulk lattice pattern, but its classifications are based on clustering of manually selected image features. In contrast, the LLM model can provide a text description of defects in images with very few examples of labeled data, but cannot localize them. In this work, we describe Defect Detector, a software suite that can generate and label large amounts of synthetic atom lattice data for training a lightweight object detection model with real time inference potential on real STEM data, along with a Vision-Language model to screen defect types in STEM images for further analysis by the object detection model **(Figure 1a)**.

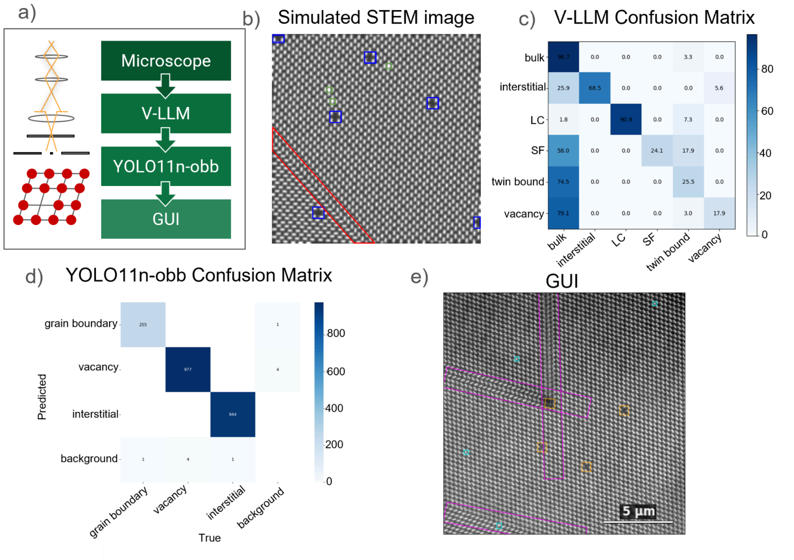

Figure 1: a) Workflow of the defect detection pipeline b) Simulated STEM image with bounding boxes. c) Confusion matrix for the V-LLM. d) Confusion matrix for the YOLO11n-obb model. e) Object detection results on a simulated automated STEM.

Methodology:

Data Simulation: To generate synthetic STEM images, we construct an idealized CdTe crystal lattice with dumbbell atomic columns [6], then introduce three defect types: grain boundaries, vacancies and interstitials **(Figure 1b)**. Grain boundaries were modeled by varying the dumbbell tilt across boundaries. Vacancies were created by randomly removing atoms, while interstitials were added at lattice midpoints. Defects were labeled with polygonal bounding boxes automatically generated during image creation. The image was then blurred and noised to mimic finite microscope resolution. The image and defect labels were augmented by rotating and reflecting each original image, and the entire dataset was exported to YOLO-formatted bounding boxes and metadata.

YOLO Model: A YOLOv11 model was trained on the synthetically generated STEM images exclusively. Specifically, we trained a YOLOv11 Nano Oriented Bounding Box model (yolo11n-obb) which offers (1) oriented bounding box labels that match the orientation of the gradient boundaries, and (2) real-time inference supporting integration into an experimental workflow with an STEM. The dataset consisted of 80 unique synthetic images each one having 16 different rotational and reflective augmentations. This created a dataset of 1280 labeled images. The model trained for 100 Epochs, roughly 30 minutes on a mid-range GPU (NVIDIA RTX 3090).

Vision-language model: We also explored whether a modern vision-language model (VLM) could perform the same classification task using prompt-based guidance, rather than training, by leveraging its pretrained multimodal knowledge to reason about lattice patterns without explicit fine-tuning. In this VLM setup, smaller patches of the STEM image were provided, due to limited experimental training data, along with definitions, visual feature descriptions, and example image-text pairs corresponding to the six defect categories in Ayyubi et al. [3]. The model then infers the most likely defect type directly from these prompts, and outputs a list of possible defects along with a confidence score. These can be visualized with a confusion matrix **(Figure 1c)** to tune the VLM for screening novel experimental images before they are inputted to YOLOv11.

Results and Discussion:

Despite the limited real data available for evaluation, the synthetically trained YOLO model performed well. It correctly identifies all 10 defects (3 interstitial, 3 grain boundary and 4 vacancy) **(Figure 1d)**, but incorrectly labels one region of bulk as an interstitial defect. This approach is expected to work well for other atomic lattices with some adjustment to the synthetic image generation parameters to better fit the images to the targeted real data. The model is able to run inference on the target image size of 768 x 768 pixel in under 500 milliseconds on a mid range GPU enabling this detection to be integrated into live analysis workflows **(Figure 1e)**.

For the vision–language model, text-only prompting showed limited effectiveness: zero-shot performance was near chance, adding defect definitions increased accuracy to only 20%, and including crystallographic descriptions sometimes reduced performance, indicating that textual information alone cannot be reliably mapped to atomic-scale patterns. Providing six representative image-text examples significantly improved accuracy to 50%, particularly for bulk, interstitial, and LC classes, demonstrating the importance of visual grounding, though smaller defect categories remained challenging. Adding targeted examples improved vacancy detection (0% to 40%) but also increased false positives, highlighting that while in-context augmentation can help, unbalanced or non-representative examples may lead to overgeneralization. Inference time was approximately 3 seconds per image using the ChatGPT 4o-mini model, suggesting that vision–language models may be suitable for coarse defect pre-screening, but require further optimization for integration into real-time detection pipelines.

References:

[1] A. V. Crewe, J. Wall, and J. Langmore, “Visibility of Single Atoms,” *Science*, vol. 168, no. 3937, pp. 1338–1340, 1970.

[2] S. V. Kalinin *et al.*, “Machine learning in scanning transmission electron microscopy,” *Nat. Rev. Methods Primer*, vol. 2, no. 1, p. 11, Dec. 2022, doi: 10.1038/s43586-022-00095-w.

[3] R. A. W. Ayyubi, J. P. Buban, and R. F. Klie, “Automated Defect Detection in Atomic Resolution STEM Images: A Machine Learning Approach with Variational Convolutional Autoencoders,” *Microsc. Microanal.*, vol. 30, no. Supplement_1, p. ozae044.180, July 2024, doi: 10.1093/mam/ozae044.180.

[4] kingbedjed, *kingbedjed/Defect-Detector*. (Dec. 19, 2025). Jupyter Notebook. Accessed: Dec. 18, 2025. [Online]. Available: https://github.com/kingbedjed/Defect-Detector

[5] S. V. Kalinin *et al.*, “Machine learning in scanning transmission electron microscopy,” *Nat. Rev. Methods Primer*, vol. 2, no. 1, p. 11, Mar. 2022, doi: 10.1038/s43586-022-00095-w.