2996
doi:10.1017/S1431927621010424
Microsc. Microanal. 27 (Suppl 1), 2021

Construction Zone: a software package for building complex nanoscale atomic scenes for applications in machine learning data generation pipelines

Luis Rangel DaCosta[1] and Mary Scott[2]

[1]UC Berkeley, United States, [2]UC Berkeley, Berkeley, California, United States

Applied machine learning in the realm of atomic-resolution electron microscopy is becoming an increasingly prevalent technique. Training popular models like neural networks to state-of-the-art accuracy, though, often requires vast amounts of well-labeled data, on the order of tens to hundreds of thousands of images. Similarly, it is important to validate machine learning approaches against some form of ground-truth reference in order to develop a fuller understanding of a particular model or technique's performance. Utilizing ground-truth references can also help characterize a particular technique's robustness against label noise in training datasets, which can be detrimental to the training of a neural network [1].

These needs could potentially be well-addressed through simulation and perhaps easily done for simple problems of interest such as single crystal classification, where it is relatively easy to generate enough atomic models to well cover a relevant experiment sample space. However, many materials systems of interest have complex nanoscale structures where there is much more significant challenge in generating enough unique samples to represent experimental distributions and thus properly train a machine learning model. For example, in nanoparticle systems with multiply-twinned structures, there have been a variety of proposed synthetic routes for the formation of twinned regions [2,3]—studying such systems in an high-throughput manner with machine learning would require generating image data of many nanoparticles with varying grain boundary and twin structures.

Current popular software packages for generating and manipulating atomic models for materials science like pymatgen [4] and Atomic Simulation Environment (ASE) [5] are well-suited for ab-initio type problems but are ultimately not designed to handle larger nanoscale environments or features. Often, for image simulation of complex structures, researchers manually write bespoke scripts that are hard to scale and reuse. In this work, we discuss the development of a python-based software package designed to facilitate the generation of nanoscale atomic scenes for the purpose of large-scale image simulation efforts. We will also discuss how such a tool could be implemented into a practical workflow in a machine learning application, namely, segmentation of nanoparticles in HRTEM imaging.

Construction Zone (CZ) is a python package for building and generating nanoscale atomic scenes, primarily for use in conjunction with image simulation. It builds on popular open-source software in the materials simulation community, namely, pymatgen and ASE. The design philosophy in CZ is to separate scene construction into two main classes of objects—generators, which populate a region with atoms, and volumes, which describe boundaries on said regions—and provides easy methods and routines for establishing relationships between these objects.

At the most basic level, each generator is attached to a volume. Generators and volumes maintain their own spatial properties (such as origins and orientations) and can be either independently or jointly

manipulated. From these, CZ provides an interface for constructing complex scenes quickly by providing an easy interface to both create volumes and generators and perform key manipulations, such as object orientation and lattice alignments. CZ development is currently focused on substrated nanoparticle systems with twin defects; however, we look to be able to handle any arbitrarily complex system. With these methods, generating a large reference library of nanoscale structures would be feasible and scriptable.

Once we have generated a suitable reference structure library, we can proceed to the simulation of images. Here, we will discuss an example pipeline for image segmentation of nanoparticles on amorphous carbon in HRTEM. For stable, generalizable performance, it is important not only to curate a diverse set of structures but also a suitably diverse set of imaging conditions and noise levels. Given the nature of aberration in HRTEM, it is sufficient to simulate an aberration-free image and store a database of complex images—this can be accomplished easily and quickly with GPU resources through simulation software like Prismatic [6] and py_multislice [7]. With this database, one could then process simulated results to develop final training images either before training or on-the-fly for more active data augmentation. For each structure under each desired imaging condition, the aberration function must be applied to the exit wavefunction of the simulated result, which then can be averaged over frozen phonons and image tilts. Finally, noise features like Poisson shot noise and distortions can be applied. In segmentation, for example, we can apply this processing to both a simulation of the full structure and one without the substrate, from which we can suitably generate both the training data and ground truth output, as shown in Figure 1, which compares a simulated image of a gold nanoparticle on a carbon substrate with its corresponding image segmentation mask.

In summary, expanding machine learning approaches to more complex nanoscale challenges requires first being able to generate suitable reference structure libraries. With the python package Construction Zone, we demonstrate one such tool for structure generation, and when used in conjunction with accelerated image simulation packages, demonstrate how it can be used to create a machine learning data pipeline for segmentation tasks.

https://doi.org/10.1017/S1431927621010424 Published online by Cambridge University Press

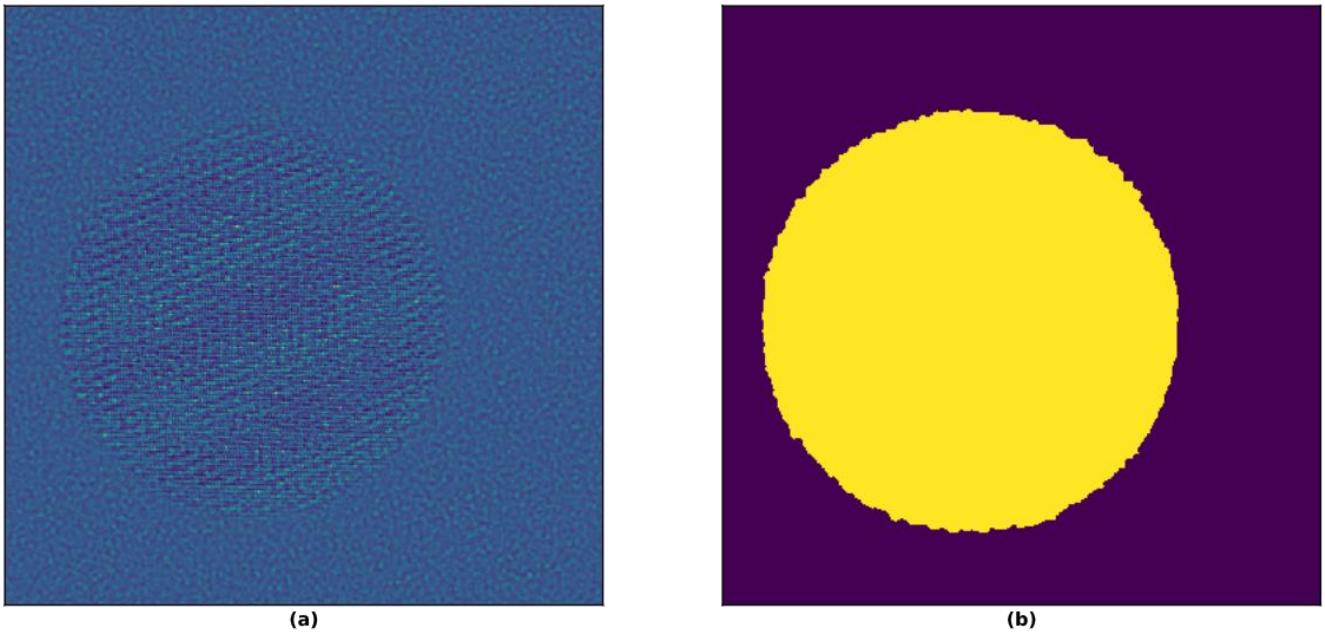

Figure 1. (a) Simulated HRTEM image of a tilted spherical gold nanoparticle on an amorphous carbon substrate (b) Corresponding segmentation mask for the nanoparticle

References

[1] Heller et. al, "Imperfect Segmentation Labels: How Much Do They Matter," *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, 2018.

[2] Ma et. al, "Unveiling Growth Pathways of Multiply Twinned Gold Nanoparticles by In Situ Liquid Cell Transmission Electron Microscopy," *ACS Nano*, 14, 9594−9604, 2020.

[3] Song et al., "Oriented attachment induces fivefold twins by forming and decomposing high-energy grain boundaries," *Science*, 10.1126/science.aax6511, 2019.

[4] Ong, et. al, "Python Materials Genomics (pymatgen) : A Robust, Open-Source Python Library for Materials Analysis," *Computational Materials Science*, 68, 314–319, 2013.

[5] Larsen, et. al., "The Atomic Simulation Environment—A Python library for working with atoms," *J. Phys.: Condens. Matter,* Vol. 29 273002, 2017.

[6] Pryor, et. al. "A streaming multi-GPU implementation of image simulation algorithms for scanning transmission electron microscopy," *Adv. Struct. Chem. Imag.,* 3, 15, 2017.

[7] Brown, et. al, "A Python Based Open-source Multislice Simulation Package for Transmission Electron Microscopy," *Microscopy and Microanalysis*, 26(S2), 2954-2956, 2020.

https://doi.org/10.1017/S1431927621010424 Published online by Cambridge University Press