

# Synthetic data generation and Mask-RCNN for Transmission Electron Microscope Image Segmentation

Natalia Da Silva De Sa<sup>1</sup>, Dr. Andrew Stewart<sup>2</sup>

<sup>1</sup>University of Limerick, Limerick, Ireland, <sup>2</sup>University College London, London, United Kingdom

## Background incl. aims

The rapid advancement of neural networks (NN) has led to a diverse array of innovations, including autonomous driving, image and text generation. Notably, within image processing, object detection and classification has been significantly improving with a variety of NNs leading the way, particularly U-Net, Fast-RCNN, Mask-RCNN. A significant challenge while training a NN is the substantial volume of training data required to achieve satisfactory results. ImageNet or COCO are typical datasets, with 14M and 200K labelled images (image and ground truth pairs), respectively, used to train NNs to detect everyday objects, people and animals within images. The Transmission Electron Microscope (TEM) field stands to benefit from machine learning image analysis techniques, which can reduce the time and effort of researchers analysing images, especially for particle size and morphology. Furthermore, automation of image analysis will lead to an increase in reproducibility when compared to manual analysis, however, there is a lack of publicly available data which is segmented and labelled appropriately for such applications. Here we present an alternative approach using synthetic data as a method for overcoming the lack of segmented and labelled data, which differs from simulated data and is a pictorial representation rather than a specifically generated simulation of a TEM image and use Mask-RCNN for instance segmentation, a form of image segmentation that detects individual objects in an image. The application shown here is for nanoparticle analysis. The principles can be generalised to all image data produced in a transmission or scanning electron microscopes.

## Methods

Our research introduces an innovative approach to generating synthetic images for training machine learning algorithms in nanoparticle detection and classification. Unlike traditional simulation methods (multislice, Bloch waves), our method utilises Python packages to generate images, bypassing the need for costly computing resources. We illustrate this approach through two distinct examples: the creation of polylatex spheres and silica particles on holey carbon substrates with ultra-thin continuous carbon layers and expand to gold nanoparticles. By randomising various parameters such as magnification, particle size, illumination, and contrast, we generate synthetic data for training an instance segmentation machine learning algorithm. Specifically, we employ a Mask-RCNN model pretrained on the COCO dataset and refine it using our synthetic transmission electron microscopy (TEM)

images, a technique known as transfer learning. This approach significantly improves the model's performance by leveraging knowledge from a broader dataset for a specialised task. Subsequently, we apply the trained model to segment experimental TEM images sourced from Certified Reference Materials (CRMs), which are industry standards used for analytical validation and microscope calibration within the PAT4Nano standardization project. To enhance the algorithm's segmentation accuracy and reduce false positives, we explore augmenting the training data with synthetic particles overlaid on experimental TEM grid substrate images. The efficacy of this approach is demonstrated through comparative analysis and visualisation. Expanding our data generation efforts, we simulate various scenarios including continuous carbon structures and diverse shapes of gold nanoparticles. Additionally, we adjust intensity values to fall within the 1-99 percentile range, further enriching the dataset. This augmented dataset is utilised to train a secondary Mask-RCNN model, which is then deployed for predicting segmentation of experimental images.

## Results

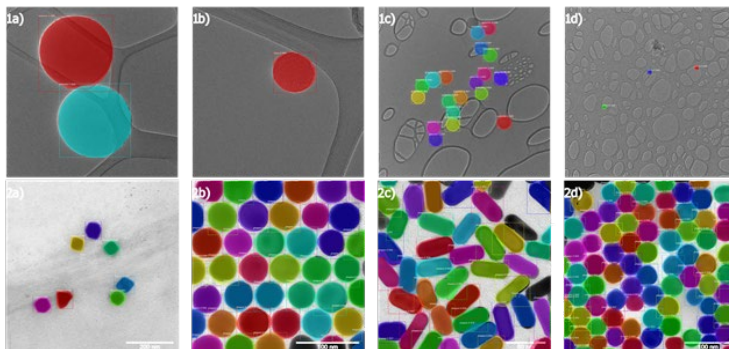
Our evaluation of the models was conducted separately, utilising the Mean Average Precision (mAP) metrics. The Average Precision (AP) is quantified as the area under the precision-recall curve (PR curve), where precision denotes the ratio of true positive predictions to all predictions, and recall represents the proportion of correct positive predictions to all true positive cases. The mAP aggregates all AP values across various classes or categories. The initial model, trained on the polylatex/silica dataset with approximately 500 synthetic images, achieved a mAP of 0.40 when tasked with predicting 90 experimental images. Subsequent augmentation of the training dataset with an additional 500 synthetic images of particles overlaid on experimental backgrounds resulted in a notable improvement, raising the mAP to 0.80 for the same set of 90 experimental images. In parallel, the second model, trained on the gold nanoparticles dataset, attained a mAP of 0.84 when predicting 19 experimental images. Both model predictions were subjected to a minimum detection confidence threshold of 0.9. The graphics section visually depict the predictions generated by models 1 and 2, respectively, providing tangible insights into their performance.

## Conclusion

Our research demonstrates the successful training of an instance segmentation algorithm using synthetic data generation, with notable enhancements are achieved by incorporating experimental backgrounds into the training process. Additionally, our findings illustrate the adaptability of the algorithm across diverse datasets characterised by varied backgrounds and particle shapes. Moving forward, our focus will centre on refining particle edge detections to achieve pixelwise accuracy and advancing nanoparticle measurement techniques. Furthermore, we aim to develop an open-source,

user-friendly interface for the generation of synthetic data adaptable to a wide variety of transmission electron microscopy (TEM) data. This interface will include a built-in trainable model which can be tuned and refined with user generated synthetic data, facilitating broader accessibility to automated image segmentation for researchers with electron microscope data.

**Graphic:**



**Keywords:**

ML, Instance Segmentation, Synthetic Data

**Reference:**

He, Kaiming, et al. "Mask r-cnn." IEEE international conference on computer vision. 2017.

Janique Hupperetz, et al. (2023). Database on Certified Reference Materials measured with PAT tools for validation and verification purposes. Zenodo.

Salley, D., et al. A nanomaterials discovery robot for the Darwinian evolution of shape programmable gold nanoparticles. (2020).