particleSeg: A semantic segmentation tool for cryo-EM particles

Abstract

Cryo-EM Single Particle Analysis, as one of the key techniques of solving protein structures, efficiently provides atomic level insight into macromolecules. The process of reconstructing a high-resolution electron density map is based on alignments and averages of high-quality protein particles in both 2D and 3D spaces. With the average and alignment information provided by diversified datasets on EMPIAR, we built a deep-learning based semantic segmentation tool, particleSeg, for raw cryo-EM particle images. Source code is available at https://github.com/nzhou26/particleSeg

Introduction

Traditional Single Particle Analysis consists the steps of collecting raw data, particle picking and extraction, cleaning particles based on 2D and 3D averages, and reconstruction of final 3D density maps. The whole process is based on selecting large number of particles and averaging these particles regardless the quality of each particle. Normally, when a sample condition is not optimized, more particles are needed to be initially picked to reach a high resolution in final reconstruction. Without 2D or 3D averaging, it is not possible to determine the quality of each particle. However, averaging is a biased process. Low-quality particles are very likely to be misclassified and affect those good classes.

Commonly used particle picking software have their own evaluation score to estimate the quality of picking particles. For template matching particle picker, the evaluation score is called figure-of-merit, indicating the cross-correlation coefficient between low pass filtered particles and 2D templates(relion_autopick). For deep learning-based particle picker, the evaluation score is softmax value in the activation layer of the model (WARP). Both scores are calculated by low-pass filtered particles and not accurate enough.

We decided to use the semantic segmentation of each particle to estimate its quality. Recently, there are software trying to segment entire cryo-EM micrographs into useful and deleterious area. MicrographCleaner (2020) took manually segmented micrographs and trained them with U-net-like model. The model was capable of detecting commonly undesired regions like ice contamination and carbon area. Warp (2018), came up with BoxNet, which is based on classical ResNet model, could also detect high contrast artifacts including ethane. Mean IOU, the metrics to evaluate accuracy of segmentation, of MicrographCleaner and WARP are 0.78833 and 0.57297, respectively.

Since recently developed packages show the feasibility and accuracy of semantic segmentation for cryo-EM micrographs, we developed particleSeg, an automatic deep learning segmentation tool to efficiently segment a raw cryo-EM particles into signal, edge and background. The software doesn't rely on manually label micrographs or manually picked particles to train the model, but rather gathers data from high resolution reconstruction on EMPIAR database.

Therefore, the future training process will be easy and the potential accuracy could be improved to a higher degree.

Methods

Data Generation

particleSeg used raw particles as input data and projection of high-resolution reconstruction density map as labeled mask (Figure 1). To acquire these data, first we filtered out datasets on EMPIAR that meet these requirements: map resolution better than 3.5 A and particle coordinates provided by author. The scrapping was done with the help of REST API of EMPIAR. Then we sorted these datasets by their sizes and downloaded from smallest to largest. Smaller dataset size means it reaches high resolution with fewer particles. Ideally, in this kind of dataset, most particles are in high quality and have accurate Euler angles. Although particles or particles coordinates on micrographs have been provided, we still need to do a 3D reconstruction in relion 3.0 to acquire values of three Euler angles (AngleRot, rlnAngleTilt, rlnAnglePsi) and shift in X and Y axis (OriginX, OriginY) in metadata with respect to each particle. With these values known, a 2D projection of its mask can be generated by relion_project. To label each pixel with three classes, signal, edge and background, a threshold of 10 was chosen so that pixels larger than the threshold were labeled as signal, pixels smaller than the threshold were labeled as background. Contour was found by scipy.convolve, which leaves the edge between signal and background 5 pixels width. Finally, to read and write data faster, a copy raw particle image and its mask projection has been saved into npy format. Currently, there are eight datasets have been pre-processed and used to train the model (Figure 2). Seven of them are dataset on EMPIAR and one was collected in-house.

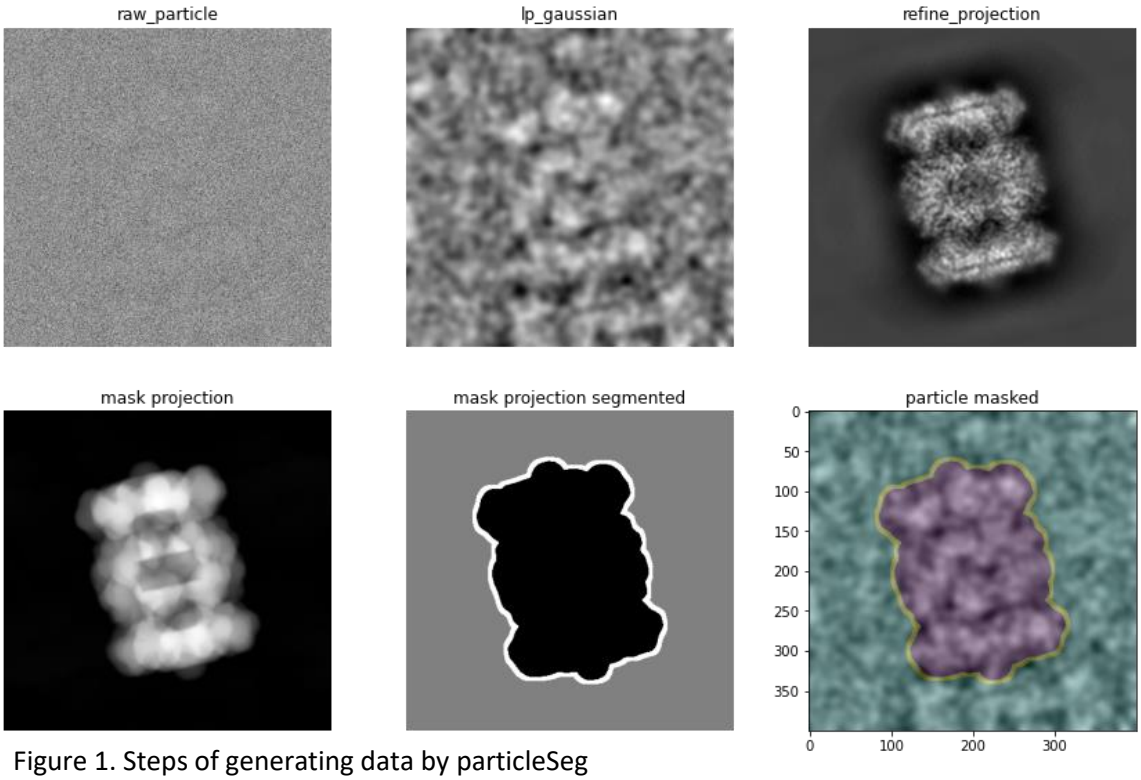


Figure 1. Steps of generating data by particleSeg

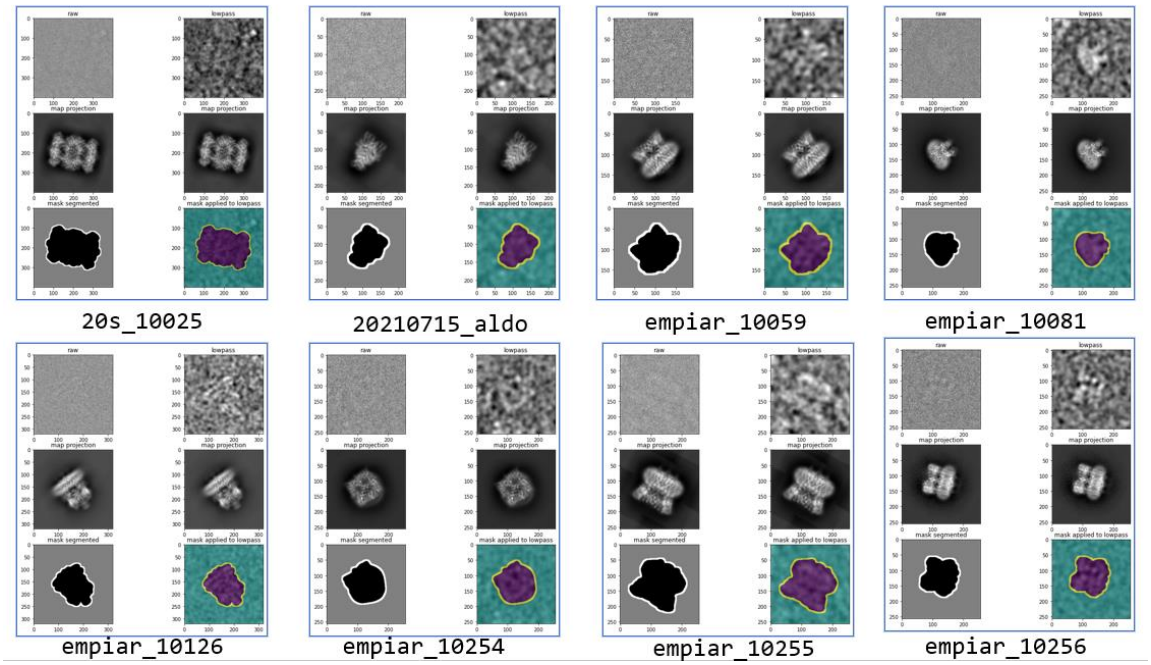


Figure 2. Overview of current datasets (290,000 particles in total) used for training.

U-Net Like Model

Raw particles images and labeled masks were feed into a U-Net like model. Both of them had been resize to a designated size of 256x256. The model consists of a down-sampling block and an up-sampling block (Supplementary Figure 1). The down-sampling block can be constructed in two ways, a custom architecture without pretrained weights (S. Fig. 1.1), or layers from popular architectures with pretrained weights from imagenet(S. Fig. 1.2 to S. Fig. 1.5). The up-sampling block was constructed by pix2pix upsample function, which consisted 8, 16, 32, 64, 128 kernels. Optimizer and loss function used in model compilation were rmsprop and sparse categorical crossentropy. The training was set to be stopped when loss of validation set was not to decrease in three epochs.

Evaluation metrics

Main evaluation metrics is average Intersection over Union (IOU) between predicted mask and true mask of particles in the testing set that was split from training data. IOU is calculated by area of overlap divided by area of union. (Figure 3.1) After a particle was segmented, the quality of this particle was evaluated by Edge to Signal Ratio. ETS was calculated by the area of edge divided by the area of signal. (Figure 3.2) By ranking the ETS value of a given dataset, particles with lowest ETS score could be dropped to clean that dataset to get a better 2D classification.

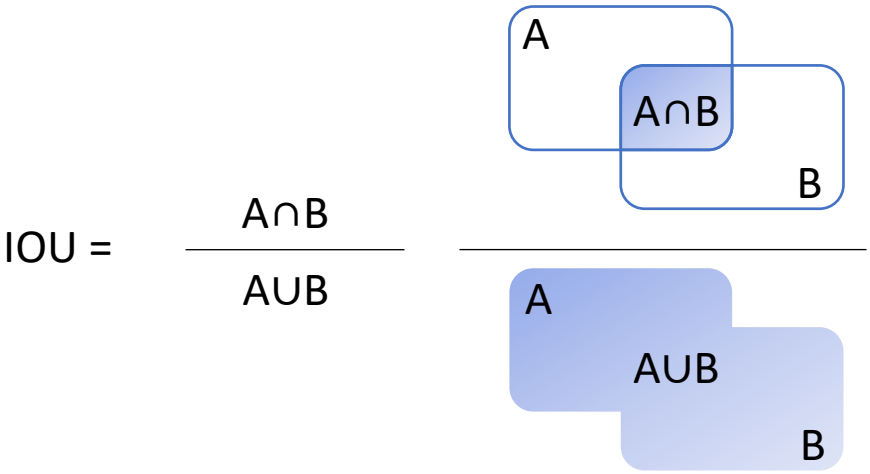


Figure 3.1. Formula of intersection over union (IOU)

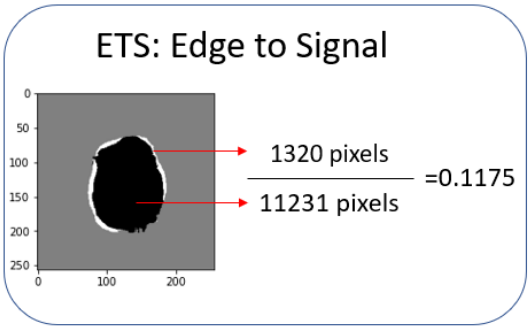


Figure 3.2. Formula of Edge to Signal (ETS)

Packages and Dependencies

particleSeg was implemented as a python3.x package. The installation could be done through a conda environment creation. Prerequisites and installation steps are listed in GitHub repo. Micrographs and metadata from relion were read by mrcfile and starfile. Image data and metadata were written by Numpy and Pandas respectively. Preprocessing and model generation took use of opencv and Tensorflow.

Result

Mean IOU of Prediction

To estimate the accuracy of trained model, we first took 20% of particles (58,000 particles) in our total downloaded and pre-processed data to use them as testing data. Particles in testing set and training set are from the same eight datasets mentioned in Figure 2. Mean IOU when evaluating testing set using model based on DenseNet169 is 71.07. Figure 4 shows the predicted segmentation against true segmentation.

To test the versatility of our model, we also evaluated it with an unseen dataset (Figure 5). The dataset was downloaded from EMPIAR with id of 10482 and preprocessed to raw input image and pixel-labeled segmentation, totally 40,000 particles. Mean IOU when evaluating this testing set is 59.06.

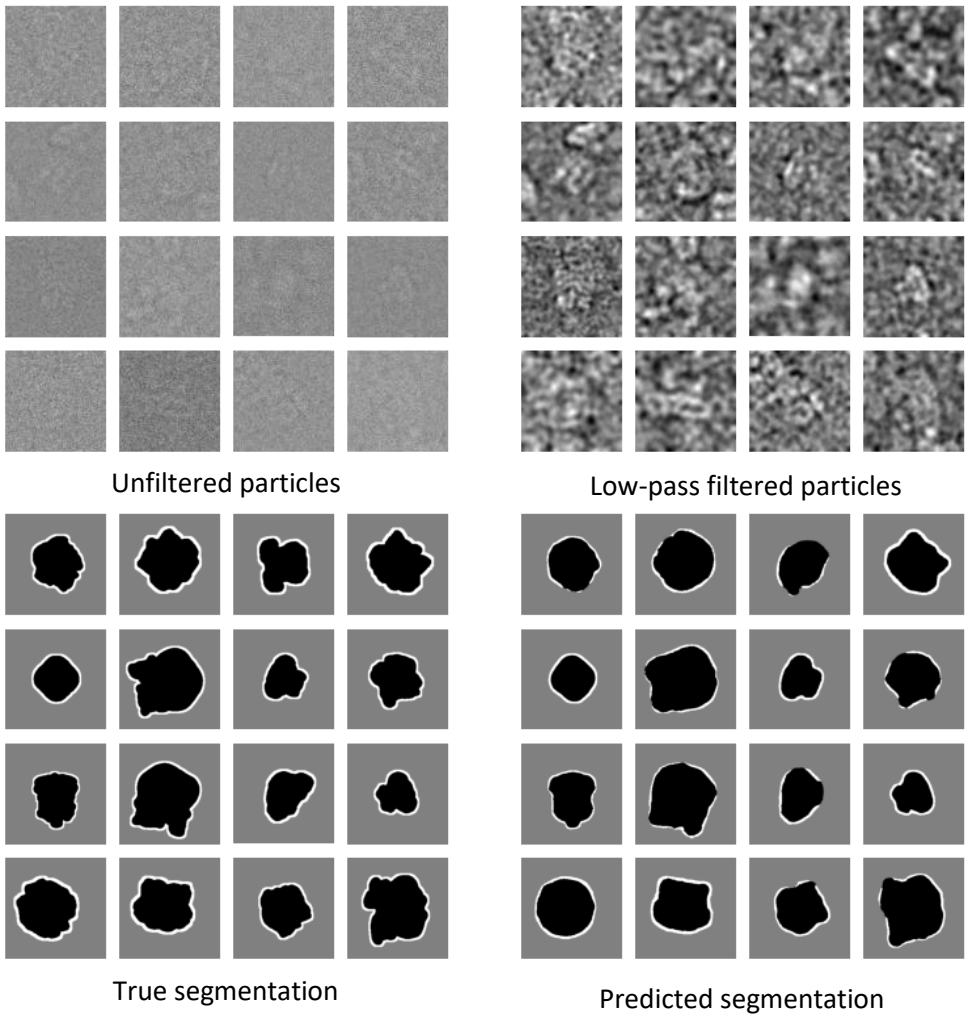


Figure 4. Prediction result of testing set pre-split from training set

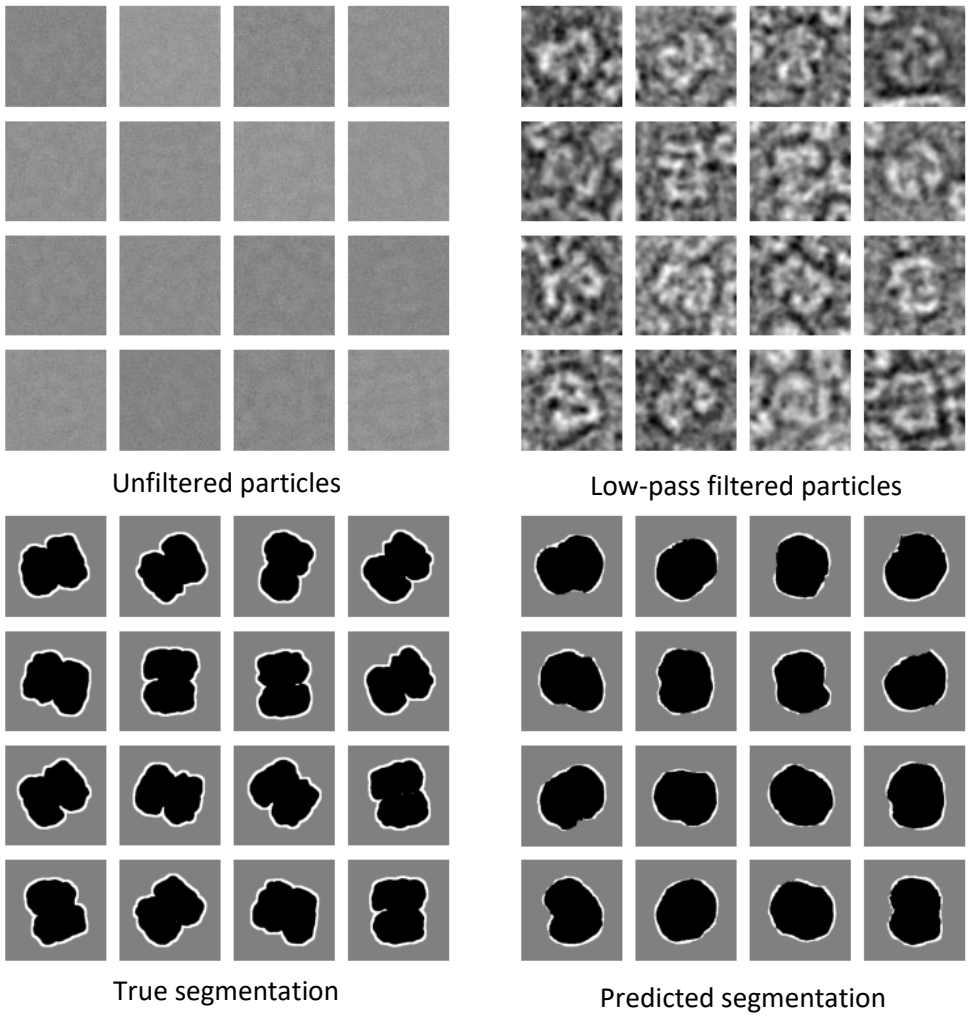
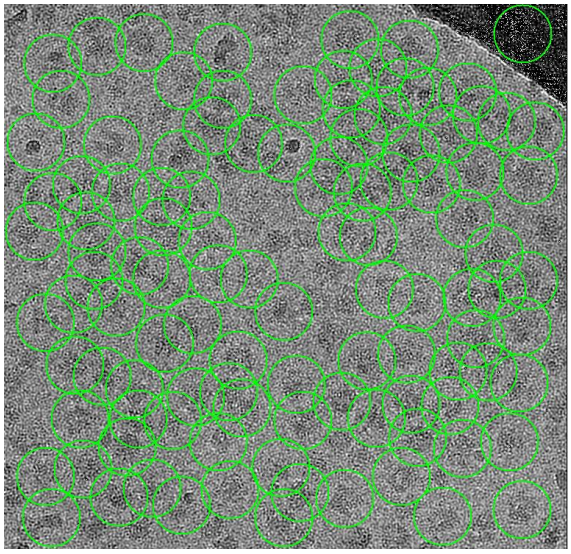


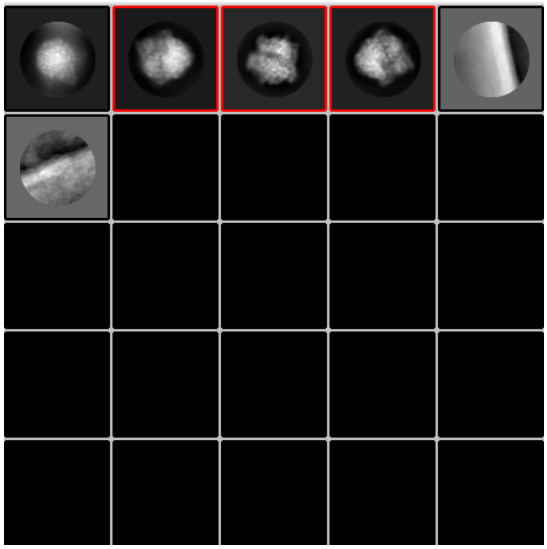
Figure 5. Prediction result of unseen dataset (EMPIAR 10482)

2D classification result of cleaned particles

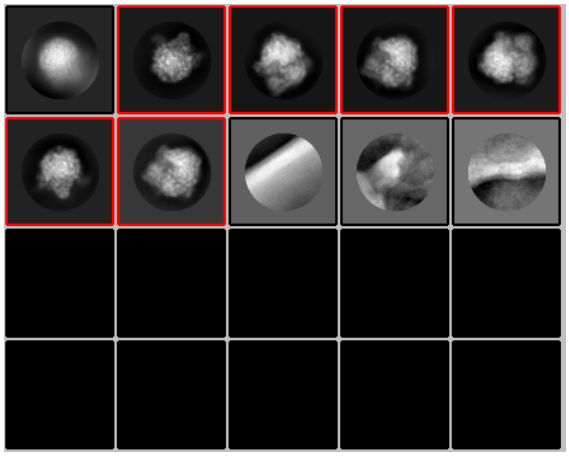
To effectively incorporate particleSeg into conventional cryo-EM data processing pipeline, we wrote a script that can read metadata from relion3.0 and automatically drop out particles with lower ETS score. Processing of EMPIAR 10406 was used to test the effectiveness (Figure 6). The micrographs were picked by relion_autopick using manually curated 2D templates. After first round of 2D classification in relion3.0, there are 83,319 particles can by selected from 3 good classed. The ratio of good particles to totally 339,043 particles is 24.57%. These particles were segmented and ranked by their ETS score using particleSeg. 30% of the particles with lower ETS score had been dropped. We run another 2D classification using remaining particles (237,331 in total) with exactly the same computation parameters. More particles can be selected out from good classes in this new 2D classification. The ratio of good particles (115,458) to total particles increased to 48.65%.



Particles picked by relion_autopick

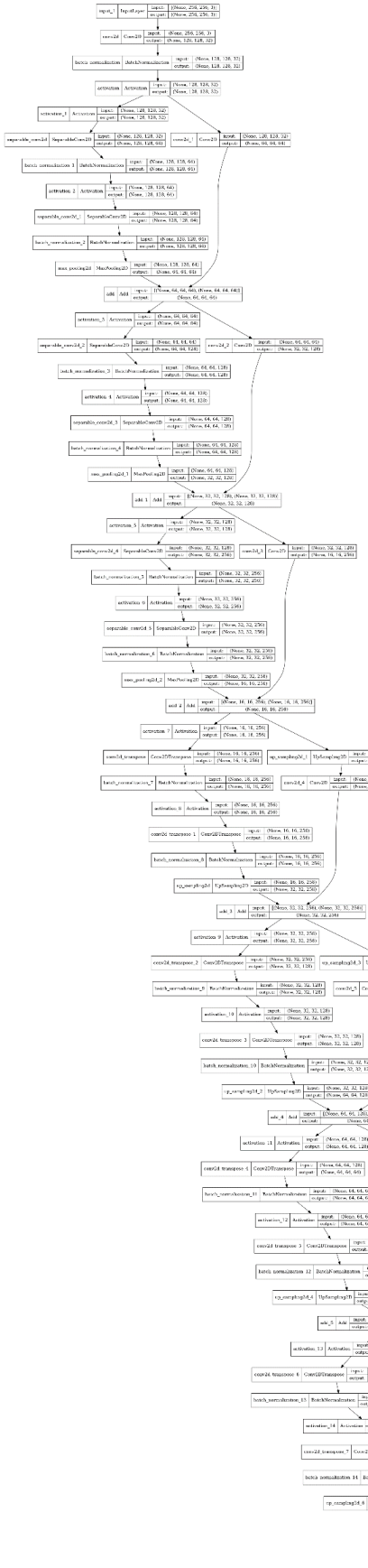


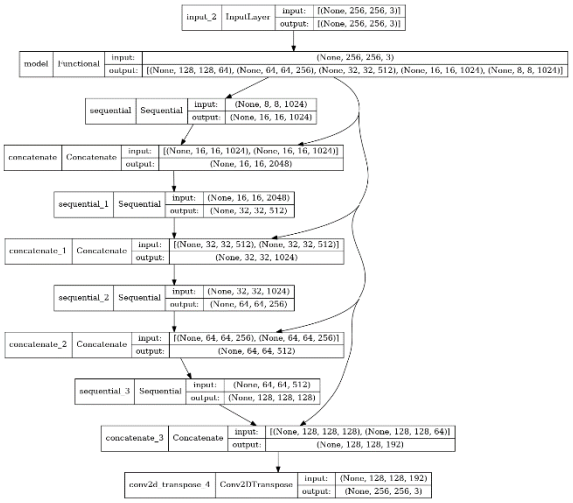
2D classification before cleaning



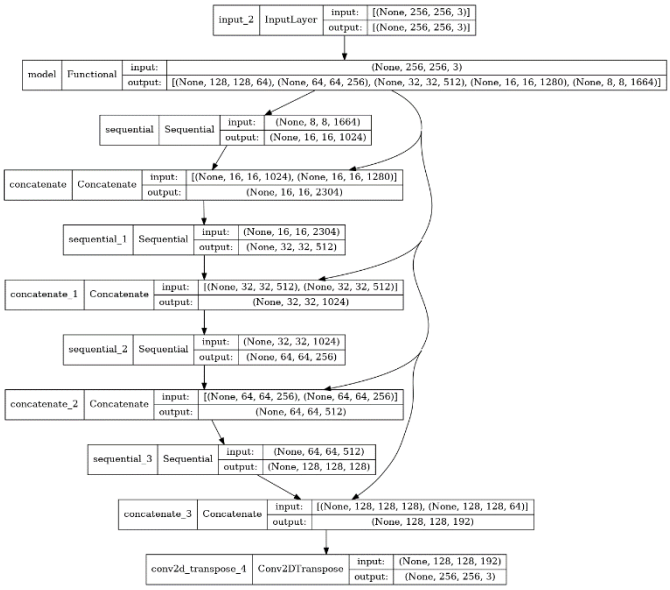
2D classification after cleaning

Figure 6. Cleaning bad particles using particleSeg with dataset of EMPIAR 10406

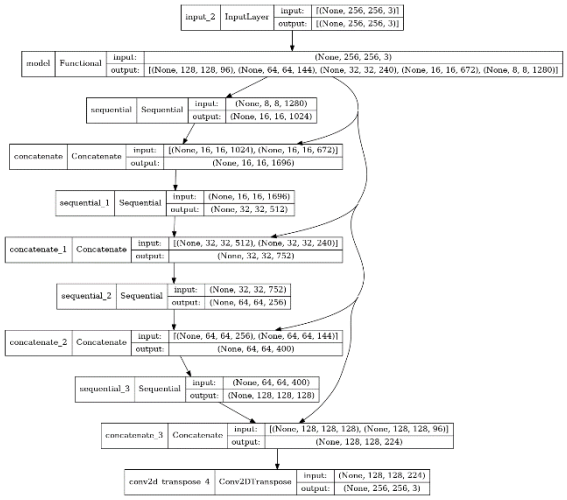




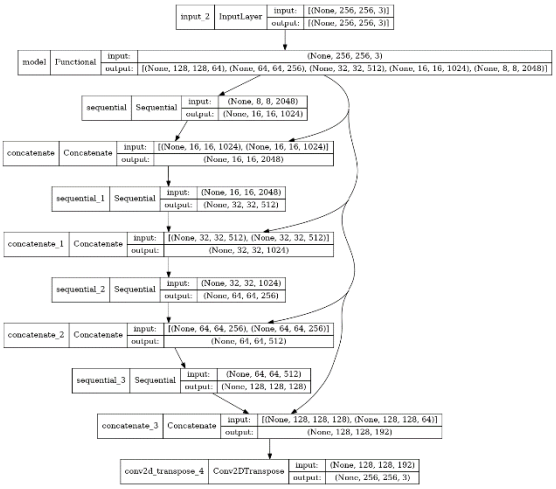
S. figure 1.2 U-Net based on DenseNet121



S. figure 1.3 U-Net based on DenseNet169



S. figure 1.4 U-Net based on EfficientNetB0



S. figure 1.5 U-Net based on ResNet101