Segmentation and Annotation of CryoET Data with Machine Learning

Trivedi JOSH
Zahir AHMAD
Amgad KHALIL

Université Jean Monnet
Faculté des Sciences et Techniques
Master Machine Learning and Data Mining

CZII - CryoET Object Identification for Deep Learning II Project
December 16th, 2024

What are Protein Complexes

- Protein complexes are groups of proteins that work together to perform specific tasks in a cell.
- They are essential for processes such as energy production, DNA repair, and cell signaling.
- Understanding these complexes is crucial for improving our health and developing new treatments for diseases.

What is Cryo-Electron Tomography (CryoET)?

- Advanced 3D imaging technique that produces tomograms (3D images) of cellular structures.
- Captures biological structures in their natural state, preserving their true shape and function.
- Provides critical insights into how cells function and how diseases affect these processes.

Example:

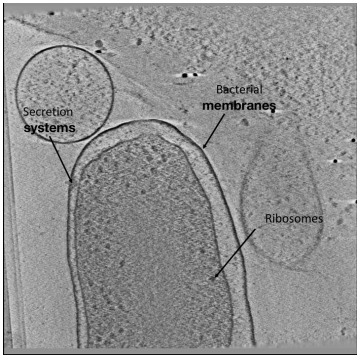


Figure 1: CryoET tomogram highlighting bacterial structures.

CryoET Architecture Diagram

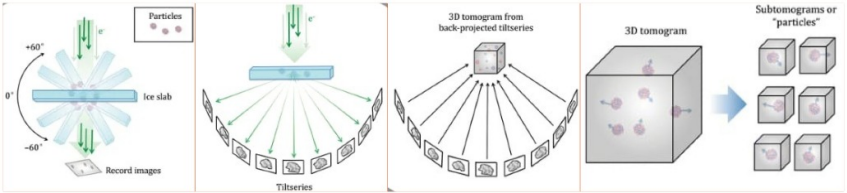


Figure 2: Architecture diagram of Cryo-Electron Tomography.

What is Segmentation?

Segmentation is the process of dividing an image into meaningful parts or regions. In traditional 2D images, segmentation identifies objects like cars, people, or animals.

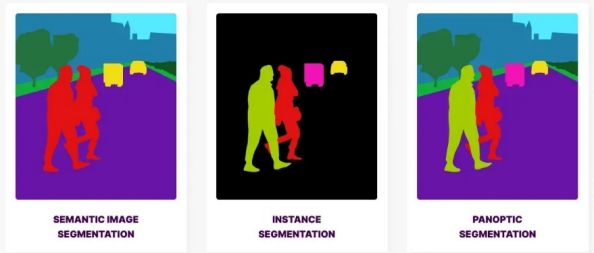


Figure 3: Segmentation Types

Note*

Segmentation in 3D Tomograms is different.

Problem Understanding

CryoET generates high-resolution 3D tomograms that reveal cellular structures like protein complexes. However, manually annotating these tomograms is **slow**, **labor-intensive**, and requires **domain expertise**. With only **5%** of over 15,000 publicly available tomograms annotated, there is an urgent need to automate this process using machine learning techniques.

Key Challenges

- **Noisy Data:** CryoET imaging produces datasets with a low signal-to-noise ratio, making analysis difficult.
- **Small Object Size:** Protein complexes are very small and require precise segmentation in large tomograms.
- **Sparse Labels:** Only a small percentage of available tomograms are annotated, limiting training data for supervised models.

Key Challenges Continued

- **Complex Segmentation:** Unlike traditional segmentation tasks, CryoET involves:
  - Tiny, overlapping structures in dense 3D environments.
  - Low contrast and noisy regions that complicate detection.
- **High Dimensionality:** Tomograms are 3D datasets requiring models capable of handling volumetric data efficiently.

Dataset Overview:

- **CryoET tomograms:** 3D images showing proteins in their natural environment.
- **Classes of interest:** 5 protein complexes (ribosome, virus-like particles, apo-ferritin, thyroglobulin, -galactosidase).
- **Training Data:** Includes RAW tomogram slice, along with 4 processed images in 7 experimental setups and 3 quality settings.

Challenge: Automate annotation to identify particles and evaluate performance using the F-beta metric $\beta = 4$

Synthetic vs Real-World Data:

- **Synthetic Data:** Simulated tomograms with annotated particles. (External dataset)
- **Real Data:** Captured tomograms with crowding and noise. (Provided on Kaggle)

Key Features:

- Spatial resolution of particles.
- Diverse and noisy environments.
- Structural and compositional variability.
- Multi-Class scenarios.

F-beta Metric: Prioritizing Recall Over Precision

- A performance metric that balances **precision** and **recall**.
- The parameter β determines the weight of recall relative to precision:

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}}$$

- In this competition, $\beta = 4$, making **recall** significantly more important than precision.

Why $\beta = 4$?

- Missing a true particle (**low recall**) is heavily penalized.
- False positives (**low precision**) are less critical, as their impact is reduced.

Real Data Visualization:

- CryoET tomograms with identified regions of interest.
- Challenges: noise and overlapping particles.

Example:

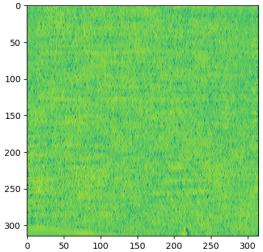


Figure 4: Visualization of real RAW CryoET tomograms.

Experiment Data Visualization

- Different directories with Experiments on the CryoET database (train).
- Below are the 4 types of tomogram slices available for Experiment TS-86-3 (VoxelSpacing10).

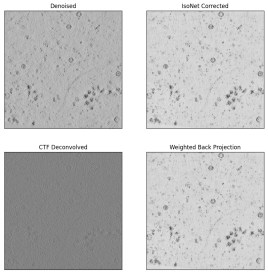


Figure 5: Experiment Results on the CryoET Real Data.

A Machine Learning Pipeline for Membrane Segmentation

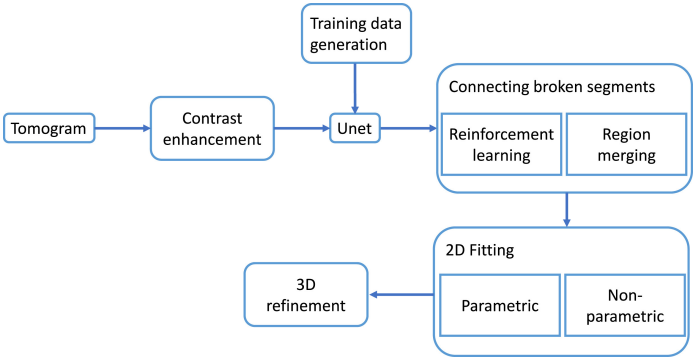


Figure 6: Framework proposed by Li Zhou et al. (2023) for membrane segmentation of cryo-ET tomograms.

Our Proposed Framework

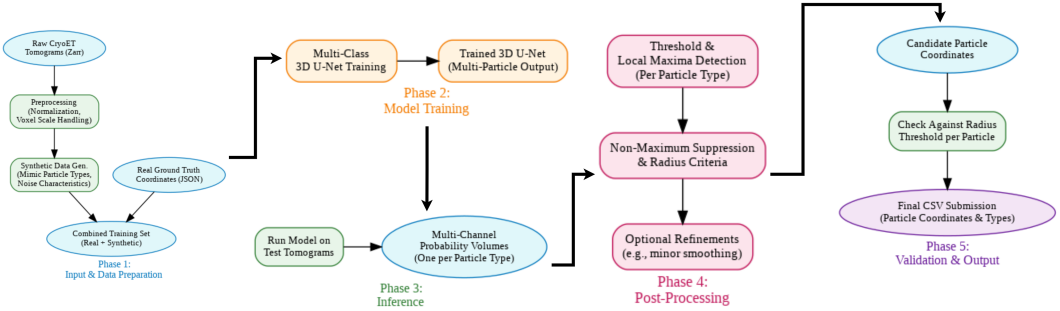


Figure 7: High-level Overview of the Multi-Phase Segmentation and Detection Framework

How We Propose to Handle the Challenges

- **Noise-Resilient Preprocessing:** denoised tomograms and normalization to handle SNR
- **Augment with Synthetic Data:** Combine real annotated data with synthetic training samples that mimic particle geometry and noise profiles.
- **Multi-Channel U-Net:** Train a 3D U-Net architecture that outputs probability volumes for each particle type simultaneously.

How We Propose to Handle the Challenges continued

- **Refined Post-Processing:** Use thresholding, local maxima detection, and optional smoothing rather than complex geometric fits better suited for continuous structures.
- **Radius-Based Validation:** Internally validate predictions against known particle radii to ensure reliable coordinate predictions before final submission.

Possible Implementation Challenges

- **Scalability and Computation:**
- **Complex Multi-Channel Output:**
- **Post-Processing Parameter Sensitivity:** Thresholds for probability maps and peak detection methods must be carefully chosen to balance false positives and false negatives, varying by particle type.
- **Generalization and Domain Shifts:**

Phases Deadlines

- **Data Preparation (1):** December 22
- **Model Training (2):** December 28
- **Inference (3)** January 6
- **Post Processing (4):** January 9
- **Validation & Output (5):** January 15
- **Tabulation and Documentation (6):** January 23

Thank you for your attention

Bibliography

- Li Zhou et al. (2023), ”A machine learning pipeline for membrane segmentation of cryo-electron tomograms”,
- J. Cell Biol. 202 (3) et al. (2013), ”Cryo-electron tomography: The challenge of doing structural biology in situ”,
- E. Moebel, A. Martinez-Sanchez, D. Lariviere, E. Fourmentin et al. (2020), Deep learning improves macromolecules localization and identification in 3D cellular cryo-electron tomograms, 2020”
- ”CZII - CryoET Object Identification.” Kyle Harrington*, Mohammadreza Paraan*, et. al. (2024) https://kaggle.com/competitions/czii-cryo-et-object-identification, 2024. Kaggle.
- ”Visualising CryoET dataset”, yoshio13 @ kaggle.com https://www.kaggle.com/code/yoshio13/very-easy-visualization (2024).

- O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234-241.
- M. Chen, W. Dai, S.Y. Sun, D. Jonasch, C.Y. He, M.F. Schmid, W. Chiu, S.J. Ludtke, Convolutional neural networks for automated annotation of cellular cryo-electron tomograms, Nature Methods 14 (10) (2017) 983.
- E. Moebel, C. Kervrann, 3D ConvNets improve macromolecule localization in 3D cellular cryo-electron tomograms, in: Quantitative BioImaging, QBI Conference, Vol. 2, 2019.

Synthetic Data Visualization:

- Annotated positions of particles in simulated tomograms.
- Visual representation of particle density and distribution.

Example:

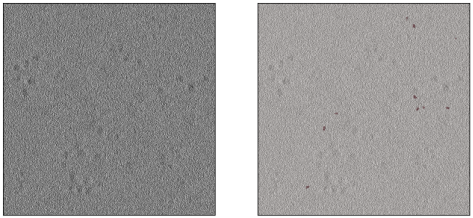


Figure 8: Annotated Beta-Amylase (Right) in synthetic tomograms (Left).

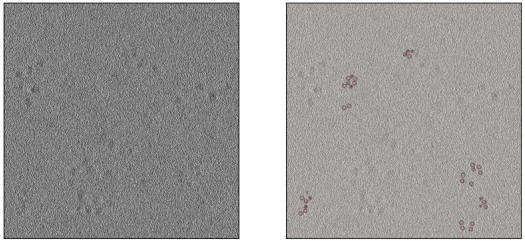


Figure 9: Annotated Apo-Ferratin (Right) in synthetic tomograms (Left)

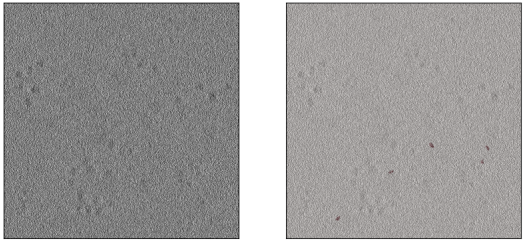
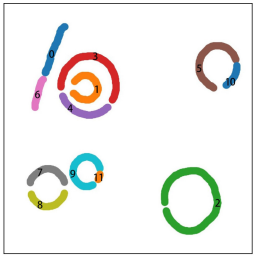
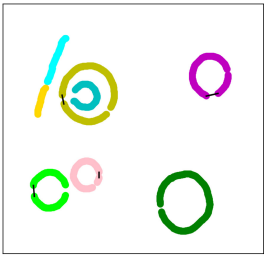


Figure 10: Annotated Beta-Galactosidase (Right) in synthetic tomograms (Left)



(a) Connected segments in a liposome tomogram slice identified by the RL algorithm. Each segment is labeled by a distinct number and color.



(b) All connections made by the RL algorithm. Each connection is marked by a black line in the figure.

Figure 11: Connected Region Segments