Title

A machine learning pipeline for membrane segmentation of cryo-electron tomograms

Permalink

https://escholarship.org/uc/item/9sk89395

Authors

Zhou, Li
Yang, Chao
Gao, Weiguo
et al.

Publication Date

2023

DOI

10.1016/j.jocs.2022.101904

Copyright Information

Contents lists available at ScienceDirect

Journal of Computational Science

journal homepage: www.elsevier.com/locate/jocs

A machine learning pipeline for membrane segmentation of cryo-electron tomograms

Li Zhou [a], Chao Yang [b,*], Weiguo Gao [a,d], Talita Perciano [c], Karen M. Davies [f], Nicholas K. Sauter [e]

[a] School of Mathematical Sciences, Fudan University, Shanghai, 200433, China
[b] Applied Mathematics and Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[c] Scientific Data Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[d] School of Data Science, Fudan University, Shanghai, 200433, China
[e] Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[f] Electron Bio-Imaging Center, Diamond Lightsource, Harwell Science and Innovation Campus, Didcot, OX11 0DE, UK

ARTICLE INFO

Keywords:
Image segmentation
Machine learning
Membrane structure
Reinforcement learning
Gaussian process
Uncertainty quantification

ABSTRACT

We describe how to use several machine learning techniques organized in a learning pipeline to segment and identify cell membrane structures from cryo electron tomograms. These tomograms are difficult to analyze with traditional segmentation tools. The learning pipeline in our approach starts from supervised learning via a special convolutional neural network trained with simulated data. It continues with semi-supervised reinforcement learning and/or a region merging technique that tries to piece together disconnected components belonging to the same membrane structure. A parametric or non-parametric fitting procedure is then used to enhance the segmentation results and quantify uncertainties in the fitting. Domain knowledge is used in generating the training data for the neural network and in guiding the fitting procedure through the use of appropriately chosen priors and constraints. We demonstrate that the approach proposed here works well for extracting membrane surfaces in two real tomogram datasets.

1. Introduction

Despite the tremendous progress made in biological imaging that has yielded tomograms with ever-higher resolutions, the interpretation of data (e.g., the segmentation of cell tomograms into organelles and proteins) remains a challenging task. The difficulty is most extreme, in our experience, in the case of cryo-electron tomography (cryo-ET), where the samples exhibit inherently low contrast due to the limited electron dose that can be applied during imaging, before radiation damage occurs. The resulting tomograms thus have a low signal-to-noise ratio (SNR), as well as missing-wedge artifacts caused by the limited sample tilt range that is accessible during imaging [1]. Fig. 1 shows two cryo-EM tomogram slices from two different datasets. These tomogram slices show a number of circularly shaped membrane structures with proteins (visible as small dots) inside and outside the membrane surfaces.

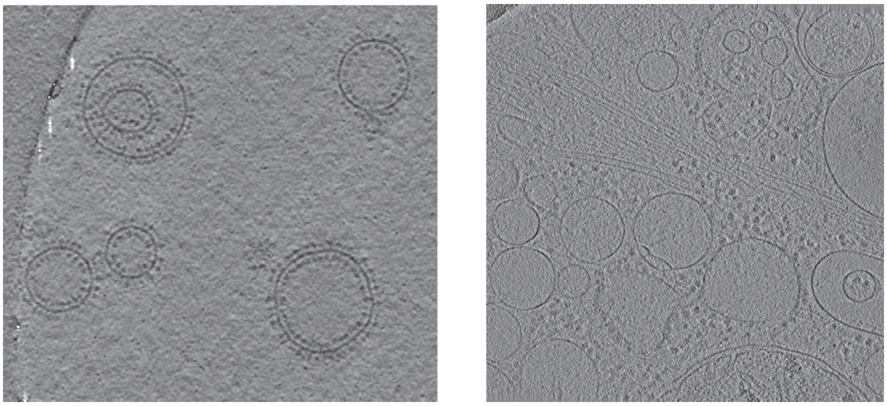
Our objective is to identify and isolate from such tomograms multiple cellular substructures such as membranes and protein complexes that can be analyzed further. This objective is often achieved through an image segmentation procedure. Currently, such a procedure is performed, in most cases, by a human expert manually tracing or highlighting specific features in a tomogram, which are then extracted and analyzed for length, curvature, volume, distance, etc. This is an extremely time-consuming and labor-intensive process.

Although a number of automated segmentation algorithms and tools have been developed in the last few decades for high contrast medical 3D imaging [2–8], most of them perform poorly on cryo-ET datasets due to low SNR and missing wedge artifacts. SNR can be partially improved by applying contrast enhancement and edge detection algorithms, such as nonlinear anisotropic diffusion, wavelet transforms, or Sobel filters. Nevertheless, these algorithms can also generate false connectivity and additional artifacts that degrade the results produced by automatic segmentation methods.

A human scientist can do a much better job than a computer program at segmenting and extracting membrane structures because they have prior knowledge (size, shape, etc.) about the biological object to be segmented. However, we could train a machine to learn such knowledge. In that case, it may be possible to develop a more

* Corresponding author.
E-mail addresses: lzhou11@fudan.edu.cn (L. Zhou), CYang@lbl.gov (C. Yang), wggao@fudan.edu.cn (W. Gao), TPerciano@lbl.gov (T. Perciano), karen.davies@diamond.ac.uk (K.M. Davies), NKSauter@lbl.gov (N.K. Sauter).

https://doi.org/10.1016/j.jocs.2022.101904
Received 25 March 2022; Received in revised form 28 July 2022; Accepted 2 November 2022
Available online 12 November 2022



(a) A cryo-EM tomogram slice of membrane-bound ATP synthase proteins reconstituted into liposomes [21].

(b) A cryo-EM tomogram slice of in-tact P19 cells in Mus musculus [22].

Fig. 1. Slices from two different cryo-EM tomogram datasets.

reliable automated segmentation tool that can be used to improve the throughput of the visualization and analysis and tie the structure to function, etc.

In recent years, there has been tremendous progress in the development of machine learning tools for image analysis and segmentation. In particular, convolutional neural network (CNN) based models such as U-Net [9] have been developed for cryo-ET segmentation, where any arbitrary feature may be selected from the tomogram to be used as CNN training data [10]. Although the output is promising, this automatic machine learning algorithm still suffers from problems similar to pixel-based density thresholding algorithms used to assist manual segmentation. In addition, the success of this approach is hampered by the limited number of existing segmented structures to be used for training. Even though the recent development of cryo-electron tomography has produced many tomograms, high-quality substructure segmentations that can be used to train a neural network are still scarce and will likely continue to be so. CNN models are also applied to localize several macromolecular species in cellular cryo-electron tomograms [11,12] where these methods will output only the location of macromolecules, not a segmentation map. For these methods, high-quality labeled data is also needed for the training process.

Given the complexity of the segmentation task and the inherent challenge in obtaining high-quality tomograms, it is unlikely a single image processing or machine learning technique can produce satisfactory results. However, multiple machine learning techniques can be combined to enhance the segmentation results produced by a CNN-based procedure. Among these are (1) reinforcement learning algorithms that can be used to connect multiple segmented pieces that belong to the same membrane structure (2) classification algorithms that can separate different membrane structures and place fragments of the same structure into the same group (3) parametric and non-parametric fitting algorithms that produce a smooth and continuous surface representation of membranes.

In this paper, we will illustrate how these methods can be combined in an image analysis and segmentation pipeline that can significantly enhance the fidelity of segmentation of cryo-tomograms. Although some of these methods can be directly applied to 3D tomograms [12], the large data volume of cellular cryo-tomograms makes direct 3D segmentation computationally costly in practice. Therefore, we choose to perform 2D segmentations of tomogram slices first and refine the segmentation results in 3D by taking into account the correlation among images in adjacent slices of the tomogram.

This paper is organized as follows. In the next section, we will provide an overview of the main workflow of the overall segmentation procedure and how they fit together to meet the ultimate structure analysis goal. This is followed by detailed discussions of each individual component of the methodology, including the initial segmentation by U-Net (Section 3), the refinement of the segmentation in 2D using reinforcement learning, classification, and parametric/non-parametric fitting (Section 4), as well as 3D refinement (Section 5).

2. Main workflow

Fig. 2 depicts the overall workflow of the machine learning-based tomogram segmentation strategy we propose to analyze cryo-ET images. We first preprocess the tomogram slices to enhance the image contrast if needed. We then generate training data for a U-Net by taking into account prior knowledge of the type of membrane structure we plan to segment and analyze. The training data generation combines simple 2D geometric motifs with measured signal and noise features in the tomogram.

Next, we use this training data as input for a U-Net, a CNN-based segmentation tool that identifies membrane structures to match the geometric motifs used in the training data from tomogram slices. The output from the U-Net is typically imperfect and may contain fragmented components and artifacts. Consequently, a 2D refinement procedure is used to correct this output by identifying components that belong to the same membrane structure using algorithms based either on reinforcement learning or region merging.

A parametric or non-parametric nonlinear fitting procedure is then used to create smooth and continuous boundaries for the membrane structures. Finally, the corrected 2D sections are combined in 3D and refined through a non-parametric fitting procedure to produce the final 3D segmentation.

3. Segmentation by U-Net

U-Net [9] is a convolutional neural network (CNN) [13] based segmentation tool that has enjoyed tremendous success in biomedical image segmentation. The letter U in the name characterizes the layout of the CNN, which consists of a contracting path (the left half the U) and an expansion path (the right half of the U.) The contracting path maps the input image to a set of features through successive layers of convolution, rectified linear unit (ReLU) and max pooling operations. The expansion path upsamples the feature channels before convolving them with weighting matrices, concatenating them with feature maps produced in the contracting path, and feeding them into the ReLU layer.

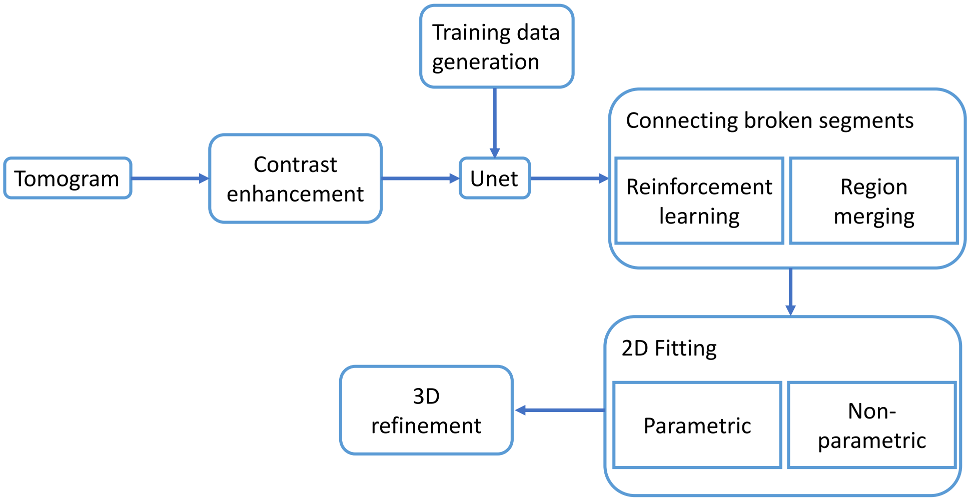


Fig. 2. The main workflow of a machine learning-based approach that combines a number of techniques for segmenting cryo-EM tomograms and improving the segmentation.

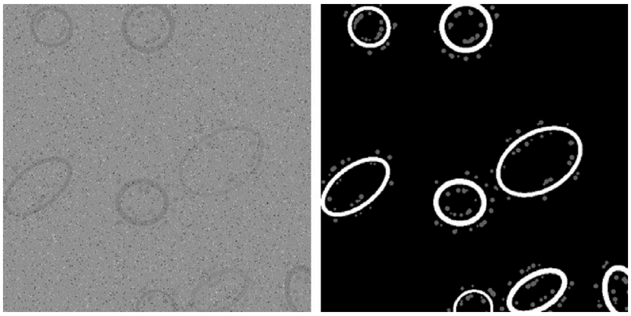


Fig. 3. The simulated membranes and protein particles (left) and their labels (right).

One of the algorithmic ingredients that make U-Net robust is the ability to use excessive data augmentation generated by applying elastic deformations to the available training images. This algorithmic feature allows us to use a few well-defined geometric motifs (such as circles and ellipses) to generate training data without relying on manual segmented data that are difficult to obtain.

We generate training data through simulation. Our simulated 2D images combined simple geometric motifs (such as ellipses and circles) with a simulated noisy background. The type of geometric motifs we generate will depend on the visible structural features in the observed tomogram. The intensity profiles of both the geometric motifs and the background are chosen to match those in the tomogram to be segmented. (See Section 1 in the Supplementary Material.)

In addition to the membranes, we also generate small solid circles near the membrane to mimic membrane proteins (e.g., ATP synthase in Fig. 1), with globular domains adjacent to the membrane. We label the membranes and proteins separately. The use of three distinct labels, i.e., 0 for background, 1 for the protein, and 2 for the membrane (see Fig. 3) improves the segmentation result.

4. Connecting broken segments

We should note that the use of a CNN-based deep learning method to segment and annotate 2D or 3D images is not new. This type of approach has been used successfully in a number of cellular applications [10,12]. However, its success depends largely on the quality of the training data as well as proper tuning of the hyperparameters (such as the number of neurons and layers) of the network, which is non-trivial. Therefore, we believe it is not realistic to rely completely on a CNN based approach to obtain a satisfactory segmentation. The CNN-based approach can be used as a preliminary segmentation tool followed by other refinement procedures to be described below.

Although the U-Net we use does a remarkable job at identifying membranes of subcellular structures, some of the identified membrane segments are disconnected. The gaps in the segmented membrane result from (1) low contrast and signal to noise ratio in the tomogram (2) incomplete tomogram reconstruction due to the missing wedge problem.

However, human vision can easily recognize how some of the disconnected components should be joined. With some prior knowledge of the possible shapes of the targeted membrane structure, we can deduce how the disconnected components should be connected and how open boundaries can be closed. In this section, we discuss how to use other learning algorithms to join all disconnected membrane segments that should lie on the same subcellular membrane surface.

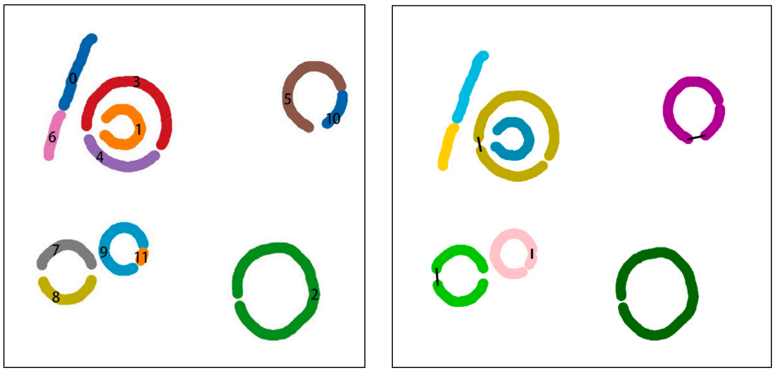
Our goal is to perform this type of postprocessing in an automated fashion with as little human intervention as possible. The challenge is that a tomogram may contain multiple membrane structures, each enclosed by a membrane. If there were only one such structure, we could possibly use a curve or model fitting procedure to connect the membrane segments identified by the U-Net. Other contour completion algorithms may also be used [8].

In the presence of multiple membrane structures, we need to determine, in an automated fashion, which labeled pixels belong to the same membrane segment, and which segments belong to the same membrane surface. While the first question is relatively easy to address by grouping labeled pixels within an ϵ-distance from each other, the second question is much harder to address because the labeled membrane segments can have different shapes, lengths, and curvatures.

We present two strategies for achieving this task. The first strategy is based on reinforcement learning. The second strategy is based on region-based pixel merging.

4.1. Reinforcement learning

Our first strategy is to train an agent to walk along segmented components and make connections with other segmented components with the goal of returning to the point it started from without crossing any segmented components that have already been traversed. Once the agent successfully returns to the starting point, the traversed segments are selected for further processing, and the agent can start again from a segmented component that has not been traversed. Otherwise, the agent is allowed to backtrack or start a new exploration trip (episode) if the existing journey is unlikely to be successful. The learning process is terminated when the number of attempts to traverse and return to the



(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.

Fig. 4. Applying RL algorithm to a U-Net segmented tomogram slice.

starting point exceeds a preset number. This type of learning algorithm is often referred to as *reinforcement learning* (RL).

The RL algorithm we use consists of two phases. In the first phase, an agent is trained to traverse through U-Net segmented pixels that are sufficiently close with the goal of creating ordered lists of pixels that are connected. The walker starts at a labeled pixel on a segmented component that has not been included in any of the connected membrane surfaces. It moves around by taking one of the eight actions (move up, down, left, right, up then left, down left, up right, down right by one pixel) until no action can be taken. Fig. 4(a) shows the outcome of the first phase of the RL algorithm when applied to a 2D slice of the liposome tomogram that has already been segmented by a U-net. The RL algorithm identified 12 connected segments, each labeled by a distinct color and number in the figure.

In the second phase of the RL algorithm, our goal is to connect segmented components (ordered lists of pixels) identified in the first phase if they belong to the membrane of the same membrane structure. The walker is initialized to exit from an endpoint of a segmented component. We train the walker to find another segmented component to connect to in order to have the best chance to return to the same component it starts from along a smooth path without revisiting any segments that have already been traversed. Once the decision is made, it picks one of the endpoints of the next component to enter and exit from the other endpoint. Fig. 4(b) shows how disconnected segments belonging to the same membrane in Fig. 4(a) are connected in second phase of the RL algorithm.

4.2. *Classification via region merging*

The RL algorithm presented above classifies the U-Net segmented components into separate classes that represent membranes of distinct membrane structures. Segmented components belonging to the same class can be connected via a number of geometric fitting procedures to be discussed in Section 4.3.

In this section, we consider another strategy to perform this classification. Our classification scheme is a variant of the statistical region merging method originally developed in [14]. In this approach, each pixel identified by U-Net to be part of the membrane initially forms its own region. Regions that are sufficiently close to each other are then merged successively. The distance between two regions R_1 and R_2 is defined by

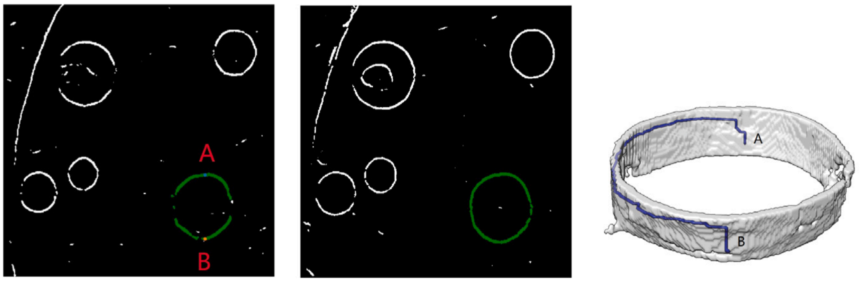
$$d(R_1, R_2) = \min_{x_i^{(1)} \in R_1, x_j^{(2)} \in R_2} \|x_i^{(1)} - x_j^{(2)}\|, \qquad (1)$$

where $x_i^{(1)}$ and $x_j^{(2)}$ are the coordinates of two points in regions R_1 and R_2 respectively. In addition to using d as a metric to decide whether to merge two adjacent regions, other visual cues such as curvature of the existing region can be used to define a predicate for reaching a merging decision. Such a predicate may also take into account uncertainty in the data (due to the presence of noise, artifacts and missing information) to allow merging decisions to be made on a statistical basis [15]. At the end of the merging process, each distinct region represents a distinct (membrane) class which is assigned a unique label.

Suppose we perform region merging within each 2D tomogram slice. In that case, disconnected components with a relatively large gap will remain in different regions and thus could be disconnected after the merging process. However, suppose we allow merging to be performed in 3D, i.e., allowing pixels in different slices to be merged into the same region. In that case, two disconnected segments on the same membrane can be combined into the same region when each of these segments contains separated pixels that can be connected to other pixels in an adjacent slice that have already been merged into a common region. That is possible because segmented components that belong to the same membrance surface may be disconnected at different locations in different slices. By exploiting the continuity of a membrane structure among different tomogram slices, we can successfully place two disconnected segments in a single tomogram slice into the same class.

Fig. 5 shows that even though segments A and B are disconnected in slice 90 of the liposome tomogram, they contain pixels that can be connected (via the path shown in the right subfigure) to other pixels in an adjacent slice that has been merged into a common region on slice 105.

Fig. 6 shows the final six regions created by the region merging procedure when it is applied to the initial 2D segmentations of dataset 1 produced from U-Net. We assign a different color to each region, which corresponds to one membrane structure (except the sheet in the upper left corner of the 3D rendering). Although these regions can be viewed as a 3D segmentation of the tomogram, the segmented structures contain visible artifacts such as extra voxels protruding from a membrane surface and gaps in the membrane surface. We will show in Section 5 that this problem can be fixed by a 3D fitting and refinement procedure.



(a) The U-Net segmentation performed on slice 90 of the liposome tomogram yields two disconnected segments *A* and *B*.

(b) Slice 105 of the liposome tomogram shows that the A and B segments on slice 90 are likely to be connected.

(c) A 3D path that connects pixels in *A* and *B* on slice 90 to pixels in other slices that have been merged in to a common region identified in slice 105.

Fig. 5. Placing two disconnected segments *A* and *B* in slice 90 into the same class through 3D region merging.

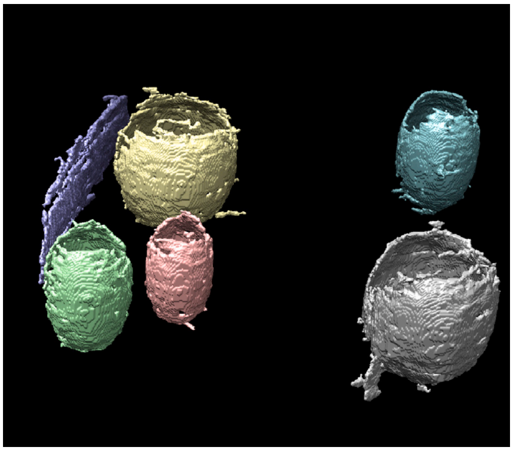


Fig. 6. 3D rendering of six regions obtained at the end of the region merging procedure applied to the entire tomogram. Each region is labeled by a unique color.

4.3. Connecting segmented components via 2D parametric and non-parametric fitting

Once the segmented components have been classified, the pixels belonging to the same class can be connected in 2D via a parametric and non-parametric fitting scheme by taking into account prior knowledge of the membrane structure to be examined.

4.3.1. Parametric fitting

If the object to be segmented has a simple geometry, we can use a parametric fitting scheme to deduce the missing pieces between disconnected components that have already been segmented out by U-Net. If, for example, the horizontal slice of vesicle membranes in Fig. 7(b) all have a elliptical shape, we can fit the visible points to a parameterized elliptic equation. Fig. 7(c) shows such a fitting result in which the red curve represents the elliptic fitting, whereas the blue curve represent the segmentation produced by U-Net.

4.3.2. Non-parametric fitting via Gaussian process

When the membrane of the subcellar structures cannot be easily described by simple geometric objects that admit an analytic parameterization, we use a non-parametric fitting procedure based on the Gaussian process (GP) formalism [16] and the implicit surface [17] formulation.

The basic idea is to view the 2D curve that encloses a membrane structure as the zero level set of a smooth 2D scalar function $f(x, y)$. Our goal is to construct this non-parametric function $f(x, y)$ such that $f(x, y) = 0$ for $(x, y) \in U$, where U contains pixels in segmented components that have been identified and connected by the algorithms presented in Sections 3, 4.1 and 4.2. In addition to the segmented pixels, the set U also includes the pixel coordinates of a number of anchor points both inside and outside the expected membrane surface so that a smooth convex or concave function $f(x, y)$ can be constructed. The choice of these anchor points represents our prior belief that certain parts of the image should belong to the exterior of the membrane. In contrast, the other parts should belong to the interior even though we have yet to determine the precise location of the interior/exterior separation in the region of interest.

We initially set the values of f to negative and positive constants at these anchor points as shown in the example given in Fig. 8. If $f(x, y)$ is continuous and sufficiently smooth, the pixels in the zero level set of $f(x, y)$ that have not been included in the set U defined above will fill in the gaps of the partially segmented membrane components returned from the algorithms described in Sections 4.1 and 4.2, thus forming a continuous and smooth boundary (surface) as shown by the example given in Fig. 10.

A GP is a prior of the distribution of functions f that is generally defined by a multivariate Gaussian $\mathcal{N}(\boldsymbol{\mu}, \mathbf{K})$, with a mean function $\boldsymbol{\mu}$ and covariance function $\mathbf{K}$. Since we are only interested in function values at the n pixels of a 2D image, $\boldsymbol{\mu}$ is a vector of length n, and $\mathbf{K}$ is a $n \times n$ matrix. We denote the function values of f on n pixels by $\mathbf{f}$.

The vectors $\mathbf{f}$ and $\boldsymbol{\mu}$, and the covariance matrix $\mathbf{K}$ can be partitioned as

$$\mathbf{f} = \begin{pmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \mathbf{K} = \begin{pmatrix} \mathbf{K}_{11} & \mathbf{K}_{12} \\ \mathbf{K}_{21} & \mathbf{K}_{22} \end{pmatrix}, \tag{2}$$

where $\mathbf{f}_1$ corresponds to random variables associated with values of f defined on pixels contained in the set U described above, which includes both the coordinates of the segmented components and the coordinates of the anchor points, and $\mathbf{f}_2$ corresponds to random variables associated with values of f defined on the other pixels in the image. The vectors $\boldsymbol{\mu}_i$ are the means of $\mathbf{f}_i$, $i = 1, 2$ respectively. The partition of $\mathbf{K}$ is conformal to that of $\mathbf{f}$ and $\boldsymbol{\mu}$.

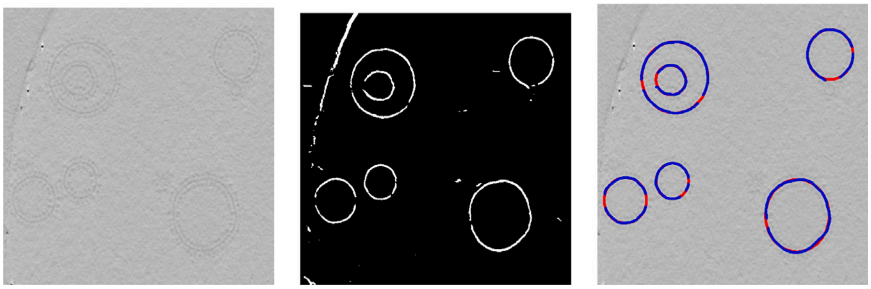

(a) Input images (b) Output of U-Net (c) Output of segments

Fig. 7. Fitting U-Net segmented components with ellipses for the 120th slice of the liposome tomogram. In (c), the blue curves are produced by the U-Net segmentation and the red parts are produced by parametric fitting.

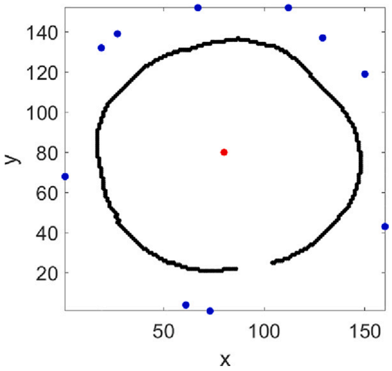


Fig. 8. The anchor points added to a partially segmented membrane slice. The value of $f(x,y)$ is set to −1 for blue anchor points (outside the membrane), and to 1 for the red anchor points (inside the membrane).

The conditional probability density function (PDF) of $\mathbf{f}_2$ given $\mathbf{f}_1$ yields the posterior PDF of $\mathbf{f}_2$ given $\mathbf{f}_1$. It is well known [18] that this PDF is a multivariate Gaussian also with the mean

$$\hat{\boldsymbol{\mu}}_2 = \boldsymbol{\mu}_2 + \mathbf{K}_{21}\mathbf{K}_{11}^{-1}(\mathbf{f}_1 - \boldsymbol{\mu}_1), \tag{3}$$

and covariance matrix

$$\hat{\mathbf{K}}_{22} = \mathbf{K}_{22} - \mathbf{K}_{21}^T\mathbf{K}_{11}^{-1}\mathbf{K}_{21}. \tag{4}$$

The mean $\hat{\boldsymbol{\mu}}_2$ yields a good estimate of the values of f on pixels outside of U. It allows us to reconstruct the missing components on the membrane surface by finding pixels $(x,y) \notin U$ that satisfy $|\hat{\boldsymbol{\mu}}_2| < \epsilon$ for some small constant ϵ. In practice, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are often set to 0. Therefore, (3) and (4) can be computed explicitly through the solution of a linear system, matrix–vector and matrix–matrix multiplications. Regularization may be needed when $\mathbf{K}_{11}$ is ill-conditioned.

In addition to providing a mean estimate of where the missing components of the segmented surface (curve) should lie, we can also quantify the uncertainty associated with the reconstructed surface by evaluating the marginal likelihood of a pixel being on the zero level set of f, i.e.

$$p(\mathbf{f}_2(i) = 0|\mathbf{f}_1) = \frac{1}{\sqrt{2\sigma_i^2}} \exp\left[-\frac{(0 - \hat{\boldsymbol{\mu}}_2(i))^2}{2\sigma_i^2}\right], \tag{5}$$

where $\mathbf{f}_2(i)$ and $\hat{\boldsymbol{\mu}}_2(i)$ denote the ith component of $\mathbf{f}_2$ and $\hat{\boldsymbol{\mu}}_2$ respectively, and σ_i is the ith diagonal element of $\hat{\mathbf{K}}_{22}$. This marginal PDF quantifies the uncertainty of a particular pixel being on the surface of the membrane. Fig. 9 shows the marginal PDF associated with the $\hat{\boldsymbol{\mu}}_2$ as a grayscale image. The darker the pixel, the higher the likelihood of the value of $\hat{\boldsymbol{\mu}}_2$ being zero (hence on the membrane) at that pixel. We exclude the previously segmented pixels by setting the color of these pixels to red.

Note that the GP prior on the distribution of $\mathbf{f}$ is largely determined by the covariance $\mathbf{K}$. The mean $\boldsymbol{\mu}$ does not play an essential role and is usually set to 0. The covariance describes how function values $f(x,y)$ are correlated for different (x,y)'s. It is often expressed in terms of the distance between different (x,y)'s. A commonly used covariance kernel is the Gaussian kernel.

However, this particular kernel does not work well in sufficiently constraining the zero level set of $\boldsymbol{\mu}_2$ by that of $\boldsymbol{\mu}_1$ through the smoothness of f. A more effective covariance kernel proposed in [19] has the form

$$\mathbf{K}(i,j) = 2r_{ij}^2 \log r_{ij} - (1 + 2\log R)r_{ij}^2 + R^2, \tag{6}$$

where $r_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$ and R is the maximum distance between any two pixels in the 2D image. Note that $K(i,i) = R^2$ for all i. This kernel function is the Green's function of a 4th order differential operator. It is related to smoothing splines interpolation [20] and the thin-plate spline regularizer [19].

Fig. 10 (left) shows the mean function defined on pixels outside of the segmented surface (curve) shown in Fig. 8. The segmented surface is shown in black. The zero level set that fills in the opening on the segmented surface is shown in blue. The figure on the right shows more clearly the reconstructed surface (curve) as a 2D contour.

The GP framework is flexible in allowing us to construct the mean of f (and the implicit surface associated with its zero level set) to match prior knowledge about certain biological structures. For example, by placing one anchor point within the inner ring of the double membrane structure present in, for example, Fig. 7(b), some anchor points between the inner and outer rings and some outside of the outer ring, and setting the values of anchor points to −1, 1 and −1, we can reconstruct the missing segments in both the inner and outer membrane as shown in Fig. 11.

5. Refinement in 3D

Although 2D parametric and non-parametric fittings allow us to fill in the missing pixels in the segmented component in each slice of the tomogram and produce smooth 2D curves in each slice, stacking these 2D curves together may produce a nonsmooth 3D surface with gaps or bumps along the vertical direction as shown in Fig. 12(a). This is to some extent inevitable when the segmentation is performed in 2D. Even a manual segmentation produces a nonsmooth 3D surface shown in Fig. 12(b).

The nonsmoothness of the surfaces can also be quantified by the maximum of distances between each membrane pixel and its nearest neighboring pixel on an adjacent slice. We plot in Fig. 13 a histogram of such maxima for all slices shown in Fig. 12(a). We can see from the histogram that most of the maximum distances are within 2 pixels, but there are quite a few between 2 and 5 pixels. There is even one that is 8 pixels long.

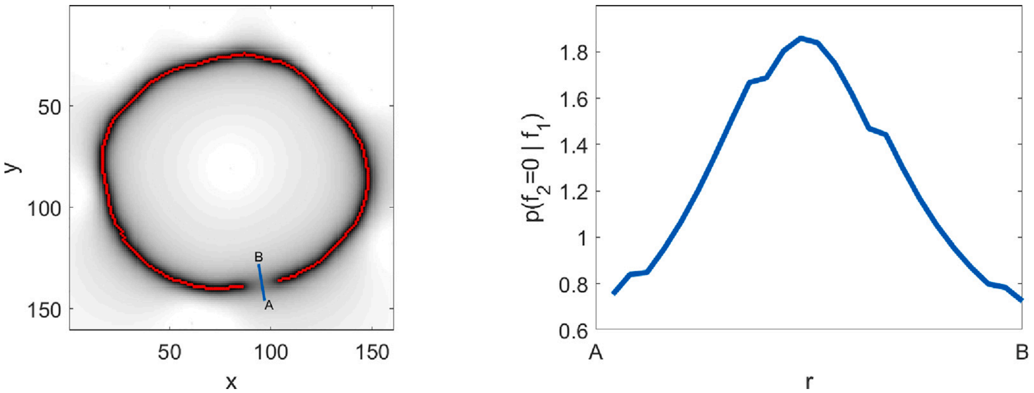


Fig. 9. The marginal likelihood of each pixel being on the zero level set of f, i.e. on the surface/boundary of the membrane (left). The marginal likelihood of pixels along the line segment AB (shown in the left figure) being on the zero level set of f (right).

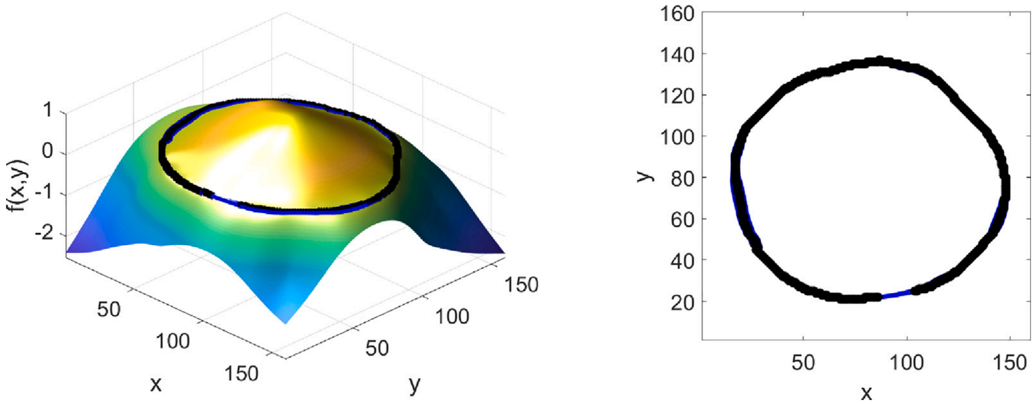


Fig. 10. The mean function produced by GP (left) and its zero-level set (right).

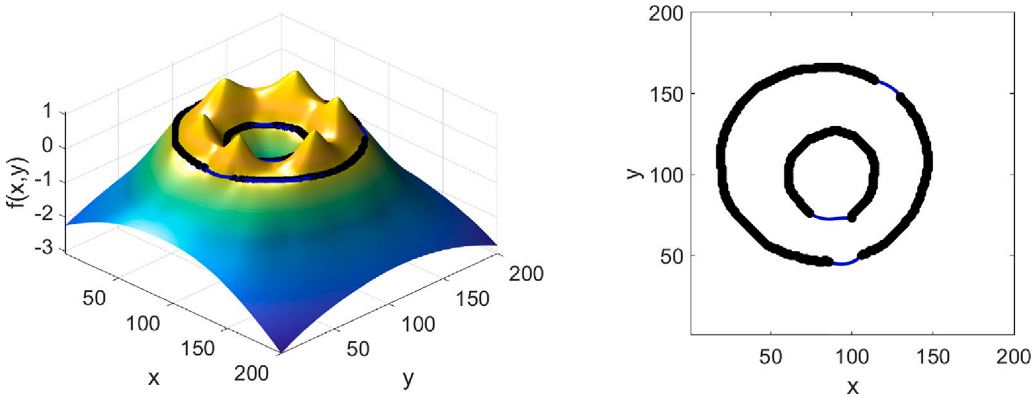


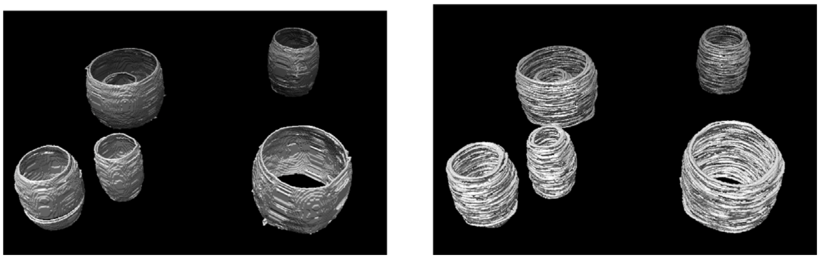
Fig. 11. The mean function produced by GP and the zero-level set for inner and outer surfaces.

The lack of smoothness along the vertical direction results partially from missing data in the tomogram, and partially from artifacts produced by the U-Net segmentation which picks up some spurious pixels that do not belong to the surface of the membrane. It may also be caused by either an ill-posed 2D fitting (due to the presence of only a few pixels grouped into the same class) or overfitting that tries to connect pixels on the membrane with spurious pixels. Neither procedure is properly constrained by the continuity and smoothness of the membrane surface across tomogram slices.

To address this problem, we develop a refinement procedure to first collect segmented pixels in different tomogram slices that belong to the same membrane surface. Spurious pixels are pruned. We then use the 3D Gaussian process formalism to construct 3D membrane surfaces that are zero level sets of a continuous function defined on a 3D volume and anchored by a few voxels both inside and outside the membrane surfaces to be reconstructed.

5.1. Voxel selection

In order to make effective use of the Gaussian process technique in 3D to construct the desired membrane surfaces, we need to identify as many voxels that lie on the same surface as possible. These voxels are collected from segmented pixels within each tomogram slice. Pixels that belong to the same class produced from the classification schemes discussed in Sections 4.1 and 4.2 are grouped together. However, in some tomogram slices, only a few pixels belonging to the surface are



(a) The isosurface of the 3D segmentation obtained by stacking the segmented slices (70 through 160) of the liposome tomogram.

(b) The isosurface of the 3D segmentation obtained by manual segmentations of slices 70 through 160 of the liposome tomogram.

Fig. 12. The isosurface of the 3D segmentation of the liposome tomogram.

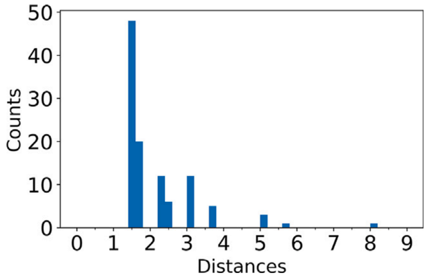


Fig. 13. Histogram of the maximum distance between a membrane pixel and its nearest neighboring membrane pixel on an adjacent slice for all slices shown in Fig. 12(a).

visible. Even fewer can be identified by the U-Net. Although parametric or non-parametric fitting can be used to reconstruct some of the pixels, the lack of visible pixels in these slices makes the fitting procedure ill-posed. For example, Fig. 14 shows that U-Net picked up a few pixels on the membrane of the inner vesicle in the upper left corner of the 100th slice of the liposome tomogram. Some of these pixels were filtered out during the classification procedure because they are isolated and not connected to other pixels. Only a small number of pixels at the top of the inner membrane are retained. When an elliptic parametric fitting procedure is applied, a small ellipse is produced, which does not correctly characterize the shape of the inner membrane.

However, the correct shape of the inner membrane is obtained when the parametric fitting procedure is applied to the 120th tomogram slice shown in Fig. 7. For that tomogram slice, many pixels can be seen to lie on the inner membrane. They are correctly identified by the U-Net segmentation. The parameters associated with the ellipse that closes the gap in the U-Net segmented inner membrane can be used to train the parameters to be optimized when a nonlinear least squares fitting is applied to the 100th tomogram slice. Specifically, we can use the parameters obtained from the least squares fit of the inner membrane for the 120th tomogram slice as the starting guesses to the parameters associated with the ellipse that fits the selected pixels in tomogram slice 100. The constrained optimization with a good starting guess yields an ellipse shown as the green curve in Fig. 14(c). Once this ellipse is constructed, all segmented pixels produced by U-Net that are sufficiently close to the ellipse are selected as voxels to be used in the 3D Gaussian process fit.

5.2. 3D fitting via Gaussian process

Once all valid voxels for each one of the membrane structures in the tomogram have been identified, we use the technique of Gaussian process discussed in Section 4.3.2 to construct an isosurface that connects all these voxels. This isosurface is defined as the zero level set of a 3D function $f(x,y,z)$ that is smooth. For each membrane, we choose an anchor point interior to membrane and set the function value of that point to 0. This point can typically be chosen as the centroid of all validated voxels that are considered to be on the membrane. A number of anchor points (x_i, y_i, z_i) in the exterior region of the membrane must also be chosen. The value of f is set to 1 at these anchor points. There are a few ways to choose these anchor points. These choices represent our prior knowledge of the shape of the membrane. For example, if the membrane is believed to have an ellipsoidal shape, we can enclose the validated voxels associated with this membrane by an ellipsoid with an appropriate size and orientation estimated from the selected voxels, and sample quasi-uniformly on the surface of the ellipsoid. Another possible way is to simply choose a few validated voxels that are well separated, and extend the ray connecting the centroid with the selected voxel proportionally to the distance between the selected voxel and the centroid (See Fig. 15). For example, point A is obtained by connecting the centroid at C with a validated voxel B and extending the ray so that $|AC| = \alpha|AB|$ for some $1 < \alpha < 2$. We chose $\alpha = 1.2$ in our selection of anchor points.

For 3D fitting, each element of the covariance matrix associated with the joint Gaussian distribution is chosen as

$$K(i,j) = 2(r_{ij})^3 + 3Rr_{ij}^2 + R^3, \tag{7}$$

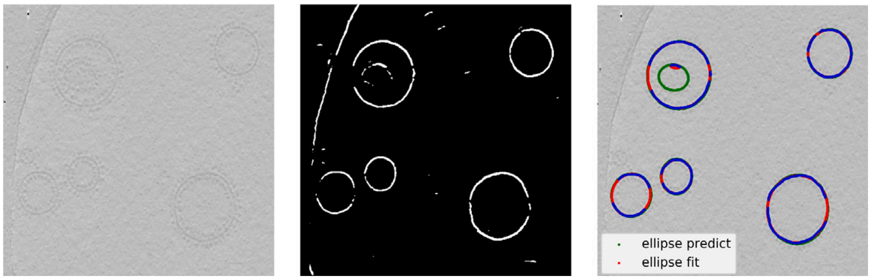
where $r_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}$, and R is the maximum distance between any two validated voxels that are considered to be on the same membrane surface.

6. Results

In this section, we demonstrate the effectiveness of the segmentation pipeline by applying it to two real datasets. One of them is the tomographic reconstruction of lipid vesicles reconstituted with monomeric mitochondrial F-type ATP synthase [21] and the other is the tomogram of an intact P19 embryonic carcinoma cells available at EMDataResource (EMD-10439) [22]. One slice of each tomogram is shown in Fig. 1.

6.1. U-Net output

For both datasets, we trained a U-Net using simulated images shown in Section 3. We generated a total 2000 simulated images. During the training process, 200 images are reserved for testing. We generated 10 additional simulated images for validation after the training process is completed.



(a) The 100th slice of the liposome tomogram.

(b) The U-Net segmentation output.

(c) A failed parametric fit and the correction.

Fig. 14. An elliptic fitting of the U-Net segmented the inner membrane in the 100th slice of the liposome tomogram and the correction made by constraining the fitting parameters using bounds obtained from the 120th tomogram slice (green).

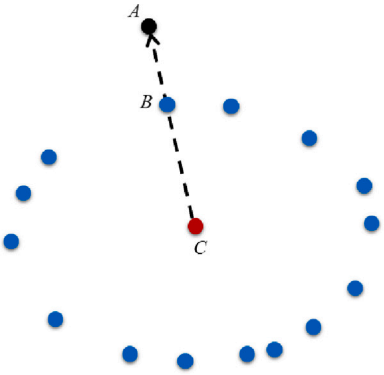


Fig. 15. The anchor point (C) is chosen by connection a validated voxel B with the centroid of all validated voxels C and extending the ray away from the centroid so that $|AC| = 1.2 \times |BC|$.

Fig. 16(a) shows that the average loss function for the training images decreases rapidly as the number of training epochs increases. Fig. 16(b) shows that the loss function associated with the testing images decreases in general also, but the change of E is not monotonic, and it is less smooth also. On average, 99% of the pixels in 10 validation images are correctly classified. We also provide the Dice–Sorensen coefficients [23,24] for each pixel class in the validation images. More details are provided in the supplementary material. Dice–Sorensen coefficients are used to measure the accuracy for each pixel class (membrane, background and protein).

Figs. 17 and 18 show the initial 2D segmentation output produced from the U-Net for the corresponding tomogram slices shown in Fig. 1. We show the membranes surfaces and proteins separately in (a) and (b) respectively. These two types of structures are combined in (c).

6.2. The final segmented 3D membranes and proteins

In this section, we show the final output produced at the end of the segmentation pipeline. Figs. 19(a) and 20(a) show the lower half of the reconstructed membranes within the tomogram as well as the validated voxels selected for fitting in the upper half of the tomogram, for both datasets. Because GP produces a different 3D density map corresponding to the mean of a Gaussian probability distribution for each of the substructures in the figure, using a single isosurface rendering threshold produces a variation in thickness. We apply thinning and dilation operations [25] to the images generated by GP to create the results with the same thickness.

The entire exterior membrane surfaces as well as some of the ATP synthase proteins obtained from liposome dataset are shown in Figs. 19(b). For the P19 (EMD10439) dataset, we only show segmented membranes with well defined shapes in Fig. 20(a). These are a subset of all the membranes structures identified by the RL algorithm shown in Figs. 20(b). Many of the membrane structures segmented by the RL algorithm cannot be easily fitted by GP.

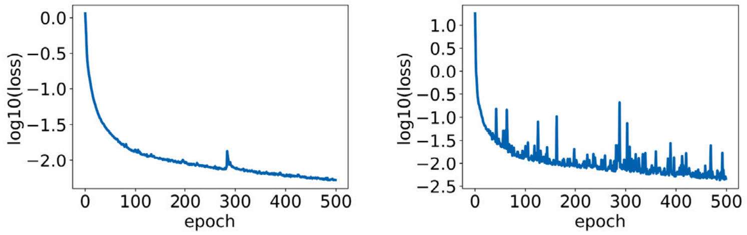
In Fig. 19(b), all membrane surfaces shown are closed surfaces even though no visible membrane structure can be detected in the top and bottom slices of the liposome tomograms shown in Figs. 19. Indeed, the top and bottom parts of the reconstructed membrane surfaces are deduced by the GP from the segmented membranes as those shown in Figs. 6 and 12(a) using the constraints implicitly defined by the correlation among neighboring voxels in a joint Gaussian probability distribution. The presence of ATP synthase proteins near the top and bottom part of the membrane surfaces serves as an indirect validation of the surface reconstruction in these regions because we know that ATP synthases proteins are often attached to the surfaces of the liposome membranes.

6.3. Validation

To demonstrate the performance of our segmentation pipeline, we compare the segmented structures presented in Fig. 19(a) with manual segmentation results presented in Fig. 12(b). We use Dice–Sorensen coefficients [23,24] to measure the similarity of the segmented structure. The Dice–Sorensen coefficient between two sets of voxels X and Y is defined as:

$$s = \frac{2|X \cap Y|}{|X| + |Y|}, \tag{8}$$

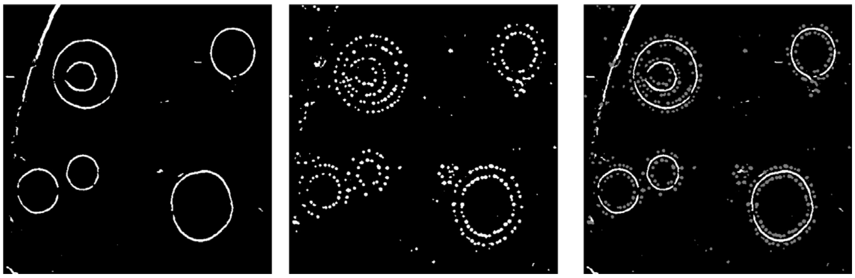
where $|X|$ is the number of voxels in set X and $X \cap Y$ is the intersection between X and Y. Here X is the set of voxels identified by the machine learning segmentation pipeline as voxels that lie on membranes and Y is the set of membrane voxels obtained by manual segmentation. Although we treat Y as the ground truth here, it should be noted that manual segmentation is not completely accurate due to low SNR in the tomogram, missing information and noise in the data. We compute Dice–Sorensen coefficients slice by slice. The Dice–Sorensen coefficients for different segmented slides of the liposome tomogram are plotted in Fig. 21. The mean Dice–Sorensen coefficients over all slices is 0.6065, which is above the 0.5 threshold. However, Fig. 21 shows that Dice–Sorensen coefficients are much higher for middle slices that contain more information than top and bottom slices that suffer from the missing wedge problem. We also plot the Dice–Sorensen coefficients associated with the segmentation of the liposome tomograph produced directly from the U-Net in blue in Fig. 21. We can clearly see that these coefficients are much lower than those associated with the segmentation produced by our machine learning based pipeline indicating that our machine learning pipeline can significantly improve the quality of the segmentation.



(a) The change of the loss function on the training images with respect to training epochs.

(b) The change of the loss function on the testing images with respect to the training epochs.

Fig. 16. The convergence of the training process.

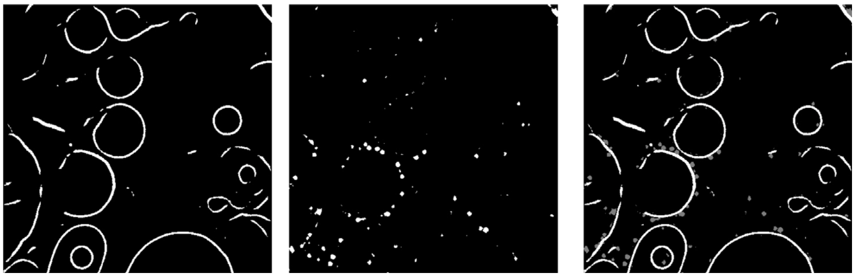


(a) Segmented membrane

(b) Segmented proteins

(c) Membranes and proteins combined.

Fig. 17. The U-Net segmentation output for the tomogram slice shown in Fig. 1(a).



(a) Segmented membrane

(b) Segmented proteins

(c) Membranes and proteins combined.

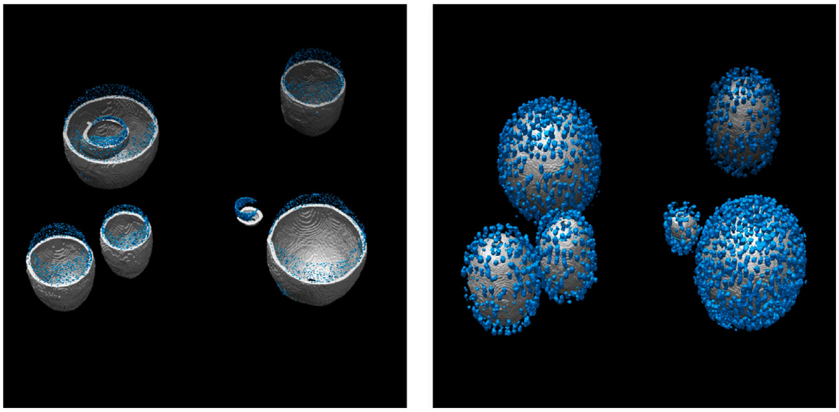
Fig. 18. The U-Net segmentation output for the tomogram slice shown in Fig. 1(b).

In Fig. 22, we plot the Dice–Sorensen coefficients associated with segmentations of the liposome tomogram obtained from different values of the α parameter used to place the outer anchor points for 3D Gaussian process fitting, discussed in Section 5.2. We can see that the choice of $\alpha = 1.2$ seems to yield the highest Dice–Sorensen coefficients for most tomogram slices. Although setting $\alpha = 1.1$ or $\alpha = 1.3$ produces slightly worse results, most of the slices still have Dice–Sorensen coefficients that are significantly above 0.5.

7. Discussion and conclusion

The extreme low contrast of cryo-electron tomograms and artifacts introduced by the limited sample tilt range that is accessible during imaging (the missing wedge problem) makes it difficult to use existing segmentation tools developed in the last few decades mainly for high contrast 3D medical imaging, to analyze the tomogram and identify important biological structures.

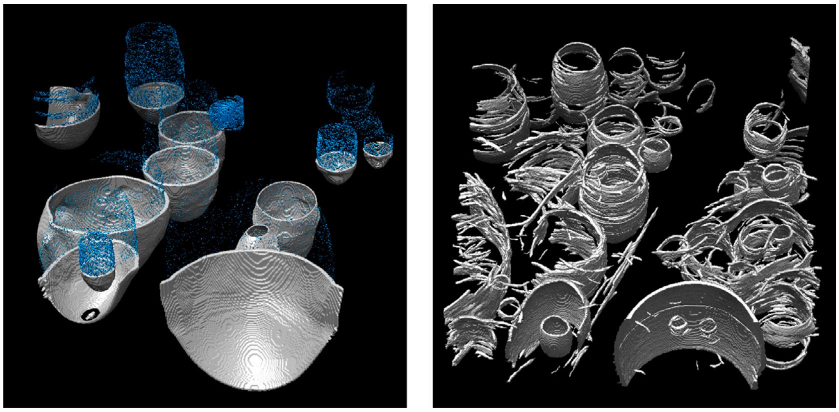
We presented a machine learning-based segmentation approach to overcome this difficulty. Our approach uses a variety of techniques organized in a learning pipeline to automate the segmentation process. The learning pipeline starts from supervised learning via a U-Net trained with simulated data. Although augmenting the simulated data with a few manual segmentation slices is likely to improve the quality of the initial segmentation, our final results show that this is not necessary. We should note that the U-Net step can be performed with other neural network based approaches, such as the mixed-scale dense convolutional neural network presented in [26], and other autoencoder based approaches [27]. The preliminary segmentation is followed by



(a) Lower half of the GP fitted membrane surfaces as well as validated voxels selected for the GP procedure.

(b) The exterior surfaces of the membranes tomogram as well as the membrane-bound ATP synthase proteins (blue).

Fig. 19. Final segmented 3D membrane surfaces and proteins for dataset 1.



(a) Lower half of the GP fitted membrane surfaces as well as validated voxels selected for the GP procedure.

(b) All segments identified by RL algorithm.

Fig. 20. Final segmented 3D membrane surfaces and proteins for dataset 2.

semi-supervised RL and/or the use of a region merging techniques to piece together disconnected components that belong to the same membrane structure. A parametric or non-parametric fitting procedure is then used to enhance the segmentation results and quantify uncertainties in the fitting. Domain knowledge is used in generating the training data for U-Net and in guiding the fitting procedure through the use of appropriately chosen priors and constraints (e.g., anchor points for GP). The generation of training data and the choice of priors and constraints can be problem dependent. We demonstrated that the approach proposed here worked well for extracting membrane surfaces of protein-reconstituted liposomes in a cellular environment that contains other artifacts, and in an additional dataset that contains a tomogram of intact P19 embryonic carcinoma cells. Although we have only demonstrated the effectiveness of our approach on two datasets, the approach itself is quite flexible and can be applied to different datasets with minimal modification. New domain knowledge for a different data can be incorporated by using a different set of simulated training data for U-Net, and new priors through the choice of different anchor points in GP fitting.

CRediT authorship contribution statement

Li Zhou: Methodology, Software, Validation, Visualization, Writing – original draft. **Chao Yang:** Conceptualization, Methodology, Validation, Investigation, Supervision, Writing – original draft, Writing – review & editing. **Weiguo Gao:** Methodology, Funding acquisition. **Talita Perciano:** Conceptualization, Methodology, Funding acquisition. **Karen M. Davies:** Conceptualization, Data curation, Validation, Writing. **Nicholas K. Sauter:** Conceptualization, Writing and editing, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

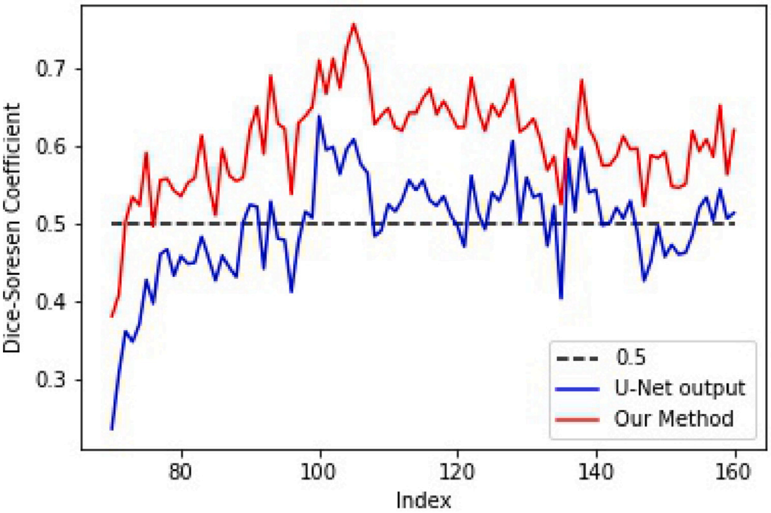


Fig. 21. The Dice–Sorensen coefficients between segmented results and manual segmented tomogram slices (70 to 160) of the liposome. We compare the ML segmented (red) and the U-Net output (blue).

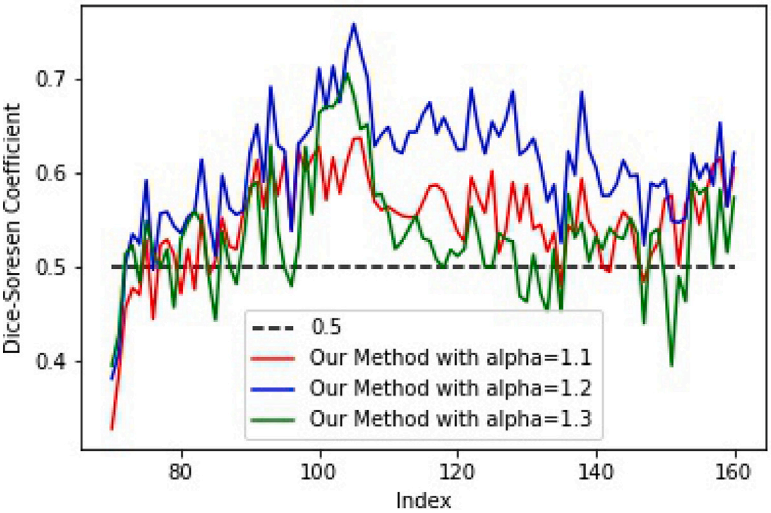


Fig. 22. The sensitivity of parameter α in 5.2. The Dice–Sorensen coefficients between the ML segmented and manual segmented tomogram slices (70 to 160) of the liposome.

Acknowledgments

This research was supported by the Laboratory Directed Research and Development (LDRD) Program award 20-122 of Lawrence Berkeley National Laboratory (LBNL), and the Center for Applied Mathematics for Energy Research Applications (CAMERA) funded by U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research (ASCR) under U.S. Department of Energy Contract No. DE-AC02-05CH11231. Z. Li would like to acknowledge support from both LBNL, USA and the Chinese Ministry of Education for carrying out this research during his visit to LBNL.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jocs.2022.101904.

References

[1] V. Lučić, A. Rigort, W. Baumeister, Cryo-electron tomography: The challenge of doing structural biology in situ, J. Cell Biol. 202 (3) (2013) 407–419.

[2] S. Beucher, F. Meyer, The morphological approach to segmentation: The watershed transformation, Math. Morphol. Image Process. 34 (1993) 433–481.

[3] K.J. Batenburg, J. Sijbers, Optimal threshold selection for tomogram segmentation by projection distance minimization, IEEE Trans. Med. Imaging 28 (5) (2008) 676–686.

[4] R. Kimmel, A.M. Bruckstein, Regularized Laplacian zero crossings as optimal edge integrators, Int. J. Comput. Vis. 53 (3) (2003) 225–243.

[5] V. Caselles, R. Kimmel, G. Sapiro, Geodesic active contours, Int. J. Comput. Vis. 22 (1) (1997) 61–79.

[6] S. Osher, J.A. Sethian, Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations, J. Comput. Phys. 79 (1) (1988) 12–49.

[7] X. Jiang, R. Zhang, S. Nie, Image segmentation based on level set method, Physics Procedia 33 (2012) 840–845.

[8] T.F. Chan, L.A. Vese, Active contours without edges, IEEE Trans. Image Process. 10 (2) (2001) 266–277.

[9] 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.

[10] 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.

[11] E. Moebel, C. Kervrann, 3D ConvNets improve macromolecule localization in 3D cellular cryo-electron tomograms, in: Quantitative BioImaging, QBI Conference, Vol. 2, 2019.

[12] E. Moebel, A. Martinez-Sanchez, D. Lariviere, E. Fourmentin, J. Ortiz, W. Baumeister, C. Kervrann, Deep learning improves macromolecules localization and identification in 3D cellular cryo-electron tomograms, 2020, http://dx.doi.org/10.1101/2020.04.15.042747, BioRxiv.

[13] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.

[14] R. Nock, F. Nielsen, Statistical region merging, IEEE Trans. Pattern Anal. Mach. Intell. 26 (11) (2004) 1452–1458.

[15] B. Peng, L. Zhang, D. Zhang, Automatic image segmentation by dynamic region merging, IEEE Trans. Image Process. 20 (12) (2011) 3592–3605.

[16] C.E. Rasmussen, Gaussian processes in machine learning, in: Summer School on Machine Learning, Springer, 2003, pp. 63–71.

[17] G. Turk, J.F. O'Brien, Variational Implicit Surfaces, Tech. Rep., Georgia Institute of Technology, 1999.

[18] K. Murphy, Machine Learning: A Probabilistic Perspective, MPI Press, 2012.

[19] O. Williams, A. Fitzgibbon, Gaussian process implicit surfaces, Gaussian Process. Pract. (2007).

[20] P. Green, B. Silverman, Non-Parametric Regression and Generalized Linear Models, Chapman and Hall, 1994.

[21] T.B. Blum, A. Hahn, T. Meier, K.M. Davies, W. Kühlbrandt, Dimers of mitochondrial ATP synthase induce membrane curvature and self-assemble into rows, Proc. Natl. Acad. Sci. 116 (10) (2019) 4250–4255.

[22] A. Martinez-Sanchez, Z. Kochovski, U. Laugks, J.M. zum Alten Borgloh, S. Chakraborty, S. Pfeffer, W. Baumeister, V. Lučić, Template-free detection and classification of membrane-bound complexes in cryo-electron tomograms, Nature Methods 17 (2) (2020) 209–216.

[23] L.R. Dice, Measures of the amount of ecologic association between species, Ecology 26 (3) (1945) 297–302.

[24] T. Sørensen, A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons, K. Dan. Vidensk. Selskab 5 (4) (1948) 1–34.

[25] S. van der Walt, J.L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J.D. Warner, N. Yager, E. Gouillart, T. Yu, the scikit-image contributors, Scikit-image: Image processing in Python, PeerJ 2 (2014) e453.

[26] D.M. Pelt, J.A. Sethian, A mixed-scale dense convolutional neural network for image analysis, Proc. Natl. Acad. Sci. 115 (2) (2018) 254–259.

[27] X. Zeng, M.R. Leung, T. Zeev-Ben-Mordehai, M. Xu, A convolutional autoencoder approach for mining features in cellular electron cryo-tomograms and weakly supervised coarse segmentation, J. Struct. Biol. 202 (2) (2018) 150–160.

Li Zhou is a researcher in the Theory Lab of Huawei. He received his Ph.D. from School of Mathematical Sciences, Fudan University in 2021. He was a visiting student at Lawrence Berkeley National Laboratory (LBNL) from 2018 to 2019. His research interests include machine learning for science, scientific computing and mathematical modeling for industry and engineering.

Chao Yang is a senior scientist in the Applied Mathematics and Computational Research Division at Lawrence Berkeley National Laboratory (LBNL). He received his Ph.D. from Rice University in 1998. He was a Householder fellow at the Oak Ridge National Laboratory from 1999 to 2000. He joined LBNL in 2000. His research interests include numerical linear algebra with applications in electronic structure calculations and quantum many-body problems, inverse problems, and high performance computing and machine learning. He is a member of SIAM.

Weiguo Gao is a professor with the School of Mathematical Sciences and the School of Data Science at Fudan University. He received his Ph.D. degree in computational mathematics from Fudan University in 1997. His research interests include numerical linear algebra and high performance computing with applications in physical sciences and data science.

Talita Perciano is a Research Scientist in the Machine Learning and Analytics group and the Computational Biosciences group, at Lawrence Berkeley National Laboratory (LBNL). She received her Ph.D. from the University of Sao Paulo, Brazil, in 2012. She joined LBNL in 2014 as a Postdoctoral Researcher before starting her position as a Research Scientist in 2016. She conducts research in the areas of image analysis, machine/deep learning, quantum image processing and machine learning, and high-performance computing motivated by the incredible challenges around scientific data generated by computational models, simulations, and experiments. She is an ACM, APS, and IEEE member.

Karen Davies is Principle Beamline Scientist at eBIC, the UK's national user Centre for Electron cryo-microscopy located at Diamond Light Source Ltd, UK. Karen received her Doctorate in Biophysics from St. Johns' College, Oxford University, UK and completed her Postdoctoral training in electron cryo-tomography at The Max Planck Institute of Biophysics, Frankfurt-am Main, Germany. In 2016, Karen started her own independent research group in cryoEM at Lawrence Berkeley National Laboratory (LBNL) before returning to the UK in 2021 to continue her independent research and oversee the running of the national CryoEM centre. Karen's research interest focuses on understanding how membranes influence bioenergy pathways in cells as well as how bacteriophages can be engineered to kill specific strains of bacterial.

Nicholas Sauter is a Senior Scientist in the Molecular Biophysics & Integrated Bioimaging Division at Lawrence Berkeley National Laboratory (LBNL). He received his Ph.D. from Harvard University in 1991 and held research positions at the University of California, San Francisco and the SLAC National Accelerator Laboratory, before joining LBNL in 2000. His research interests include X-ray crystallography and its application to structural biology, particularly with data-intensive experiments performed at X-ray free electron lasers.