One-shot Learning with Attention-guided Segmentation in Cryo-Electron Tomography

Bo Zhou[1], Haisu Yu[2], Xiangrui Zeng[2], Xiaoyan Yang[2], Jing Zhang[3], and Min Xu[2,*]

[1] Department of Biomedical Engineering,
Yale University, New Haven, CT, USA

[2] Computational Biology Department,
Carnegie Mellon University, Pittsburgh, PA, USA

[3] Computer Science Department,
University of California - Irvine, Irvine, CA, USA

[*] Corresponding author emails: mxu1@cs.cmu.edu

Abstract

Cryo-electron Tomography (cryo-ET) generates 3D visualization of cellular organization that allows biologists to analyze cellular structures in a near-native state with nano resolution. Recently, deep learning methods have demonstrated promising performance in classification and segmentation of macromolecule structures captured by cryo-ET, but training individual deep learning models requires large amounts of manually labeled and segmented data from previously observed classes. To perform classification and segmentation in the wild (i.e. with limited training data and with unseen classes), novel deep learning model needs to be developed to classify and segment unseen macromolecules captured by cryo-ET. In this paper, we develop a one-shot learning framework, called cryo-ET one-shot network (COS-Net), for simultaneous classification of macromolecular structure and generation of the voxel-level 3D segmentation, using only one training sample per class. Our experimental results on 22 macromolecule classes demonstrated that our COS-Net could efficiently classify macromolecular structures with small amounts of samples and produce accurate 3D segmentation at the same time.

Keywords: One-Shot Learning, Cryo-ET, Macromolecule Classification, Macromolecule Segmentation, Attention, Squeeze-and-Excitation

1 INTRODUCTION

Cryo-Electron Tomography (cryo-ET) has made possible the observation of cellular organelles and macromolecular structures at nano-meter resolution with native conformations [14]. Without disrupting the cell, cryo-ET can visualize both known and unknown cellular structures *in situ*[1] and reveals their spatial and organizational relationships [15]. Using cryo-ET, it is possible to capture 3D structural information of diverse macromolecular structures inside a given scanned sample.

To analyze the macromolecular structures in cryo-ET, two major subsequent steps need to occur. First, we need to extract the subtomograms[2] and average those that belong to the same macromolecular class, in order to generate a high Signal-to-Noise Ratio (SNR) subtomogram for clear visualization [20]. Second,

[1] At their original locations.

[2] Small cubic subvolumes containing one macromolecular structure

it is desirable to obtain the macromolecule segmentation in subtomograms to analyze the macromolecular structure parameters such as size distribution and shape. However, the macromolecular structures are highly heterogeneous and contain large quantities of subtomograms. In the past, biologists would spend large amounts of time on a set of tomograms to manually classify and segment subtomograms, but manual annotation is time-consuming and susceptible to the biases of individual biologists. Therefore, it is desirable to automatically classify the extracted subtomograms into subset of macromolecule with similar structure, and automatically generate the macromolecular segmentation.

To automate the process as well as to achieve objective analysis, deep learning methods for classification [18, 2, 6, 21, 11, 10] and segmentation [4, 12, 22] have been developed for cryo-ET. xu2017deep proposed to use Inception3D network and DSRF3D network for cryo-ET subtomogram classification. Then, che2017improved further improved the DSRF3D network with residual connection design. guo2018model developed a cryo-ET classification model compression technique to reduce the model size while maintaining the classification performance. zhao2018respond developed a classification model visualization technique for explaining the model's attention on the classified subtomograms. For cryo-ET segmentation, chen2017convolutional utilized independent 2D CNNs for cryo-ET tomogram components segmentation. liu2018deep built a SSN3D net for subtomogram segmentation via supervised training with large amounts of segmentation data. While previous deep learning models on cryo-ET improved the accuracy and efficiency on classification and segmentation, there are still two major bottlenecks: 1) as supervised classification methods, previous algorithms still require large amount of manually annotated training data for deep model's training, and 2) previous algorithms need to be trained again to apply to a new dataset of different classes. The open question is: Is it possible to design a generalizable cryo-ET subtomogram classification model that requires only a small reference dataset (such as one manually picked sample in each class) and match the given subtomogram to a reference class, while performing generalizable subtomogram segmentation?

Inspired by one-shot learning models which aim to learn information about object categories from one, or only a few training images [5, 8], In this work, we develop a Cryo-ET One-Shot Network (COS-Net) that is able to 1) classify macromolecular structure using only a very small amount of samples, 2) simultaneously segment structural regions in a subtomogram based on the classification network, and 3) be readily and directly applied to classify and segment novel structures without needing to be re-trained. Using our COS-Net, biologists can classify and segment thousands of subtomograms by only manually picking a few representative subtomograms as support classes. When there is a need to classify new subtomogram datasets with novel structures, the support classes can be readily changed to accommodate without the need to train the model again. Moreover, unlike previous one-shot learning and few-shot learning algorithms that only address the classification task, our COS-Net can generate both classification and 3D segmentation with application in 3D imaging data of cryo-ET.

Our COS-Net is a Siamese network with pairs of volume encoders, volume decoders, and feature encoders. Given a support set of subtomograms and a target subtomogram, volume encoders first extract the volume's feature presentations. Then, the feature encoders transform the feature presentations for the next stage: one-shot learning. In the meantime, the volume decoders decode the feature presentations to generate the coarse attention/segmentation of the subtomograms. Our COS-Net with additional attention guidance from segmentation information allows better feature embedding for one-shot learning, and thus could provide better one-shot classification performance. During the test stage, we also developed a customized subtomogram processing pipeline to refine the coarse attention/segmentation from COS-Net based on 3D Conditional Random Field (3D-CRF) [9]. Our experimental results demonstrated that our method can effectively classify observed or novel macromolecular structures and produce accurate segmentation mask.

2 METHODS

The general structure of our COS-Net is shown in Figure 1. The COS-Net is a Siamese network with two encoding-decoding streams. First, each stream consists of one volume encoder, one volume decoder, and one feature encoder. The volume encoders, volume decoders, and feature encoders shared weights between the dual streams. The design of our volume encoders, volume decoders, and feature encoders are

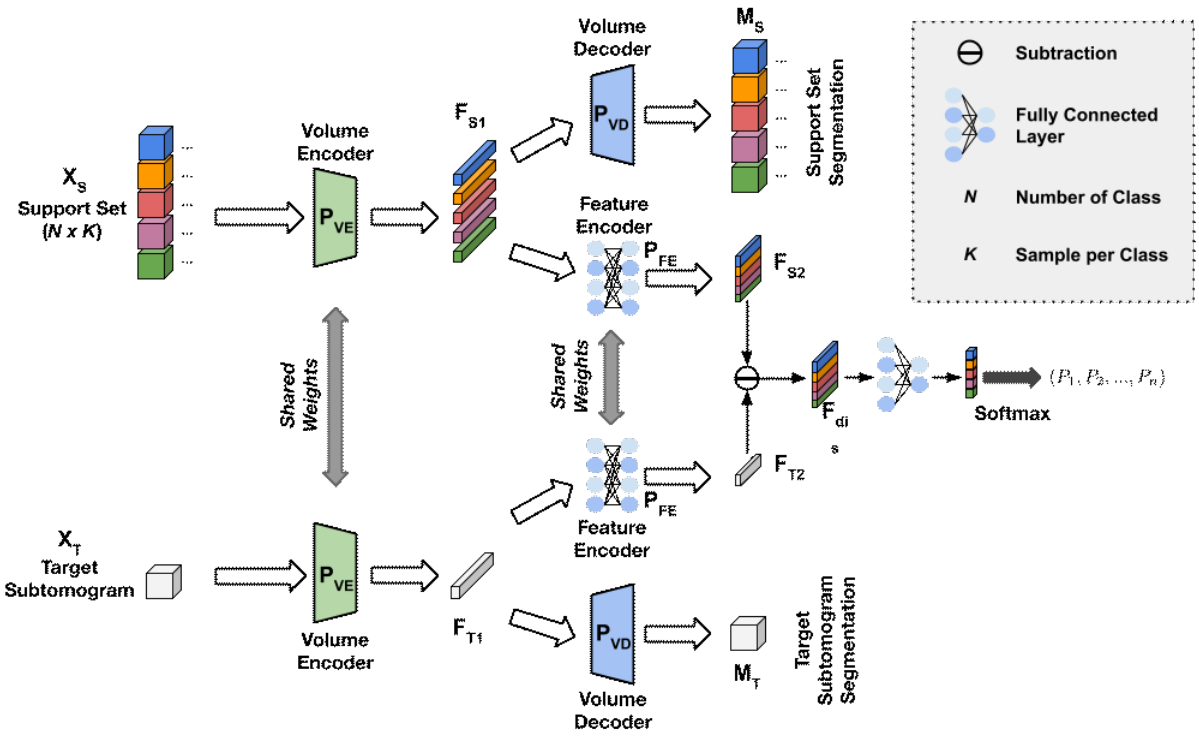

Figure 1: Illustration of our **C**ryo-ET **O**ne-**S**hot **Net**work (**COS-Net**) structure. The data input consists of subtomogram support set and target subtomogram. The network consists of pairs of volume encoders $\mathcal{P}_{VE}$, volume decoder $\mathcal{P}_{VD}$, and feature encoder $\mathcal{P}_{FE}$ with details illustrated in Figure 2.

illustrated in Figure 2 and are discussed in detail in our next section. Denoting the input for the upper stream as X_S that is our support set with dimensions of $N \times K$, where N is the number of classes and K is sample per class, support set X_S consists of N classes of macromolecules with K samples per class. In our one-shot learning scheme, $K = 1$. The upper volume encoder takes the support set X_S as input and generates the latent representation of the support set with:

$$F_{S_1} = \mathcal{P}_{VE}(X_S) \tag{1}$$

where F_{S_1} is the latent representation of the support set X_S and $\mathcal{P}_{VE}$ is the volume encoder function. Then, the support set's latent representations F_{S_1} are simultaneously fed into the volume decoder $\mathcal{P}_{VD}$ and feature encoder $\mathcal{P}_{FE}$:

$$M_S = \mathcal{P}_{VD}(F_{S_1}) \tag{2}$$

$$F_{S_2} = \mathcal{P}_{FE}(F_{S_1}) \tag{3}$$

where M_S is the predicted segmentation of the support set, and F_{S_2} is the feature for next stage one-shot learning. Similarly, denoting the input for the lower stream as X_T that is our target set with dimensions of $1 \times K$, target set X_T consists of 1 classes of macromolecules with K samples per class. In our one-shot learning scheme, $K = 1$. Similarly, the same volume encoder $\mathcal{P}_{VE}$ takes the target set X_T as input and generates the latent representation of the target set with:

$$F_{T_1} = \mathcal{P}_{VE}(X_T) \tag{4}$$

where F_{T_1} is the latent representation of the target set X_T. Then, the target set's latent representations F_{T_1} are simultaneously fed into the shared weights volume decoder $\mathcal{P}_{VD}$ and feature encoder $\mathcal{P}_{FE}$:

$$M_T = \mathcal{P}_{VD}(F_{T_1}) \tag{5}$$

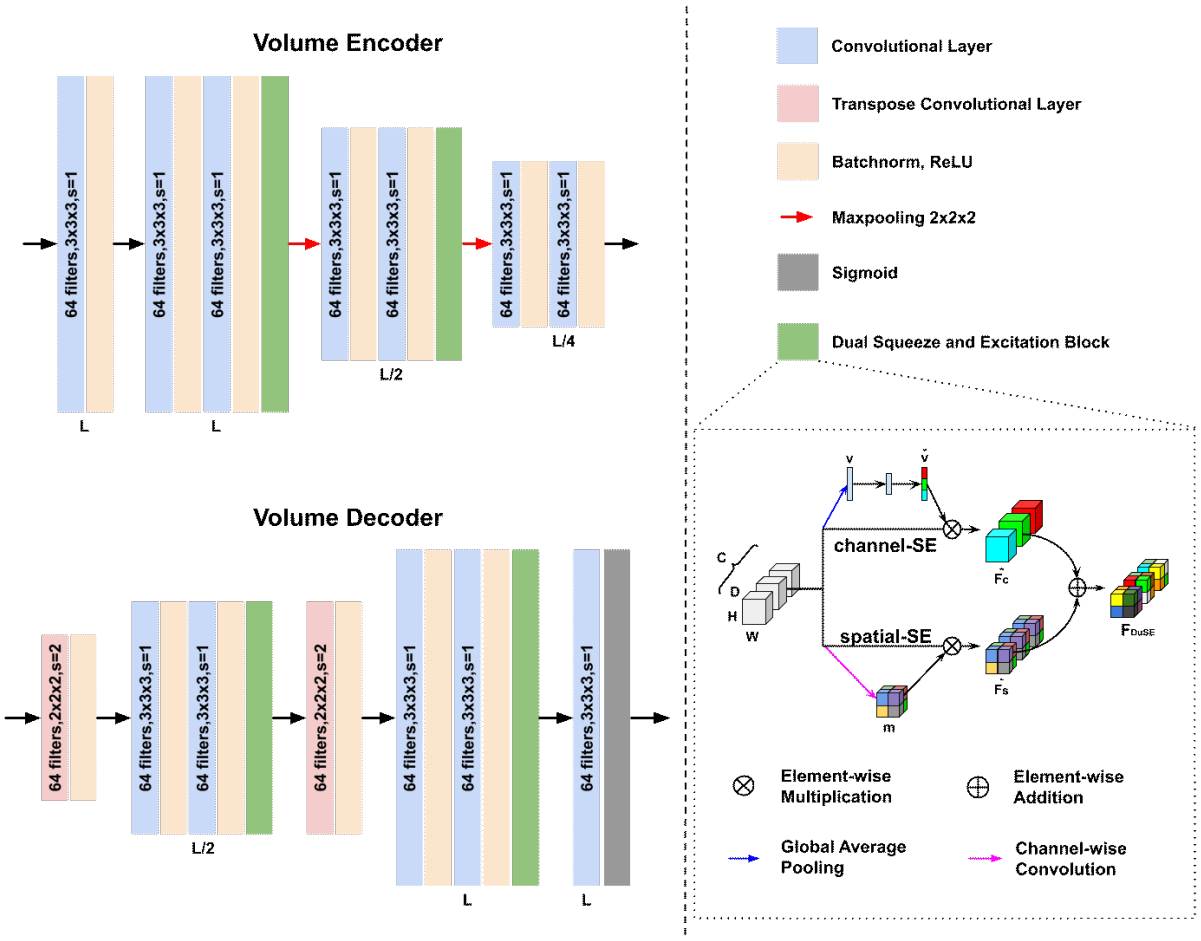

Figure 2: Architectures of our volume encoder and volume decoder in Figure 1. The **Du**al **S**queeze-and-**E**xcitation (**DuSE**) block is illustrated on the bottom right.

$$F_{T_2} = \mathcal{P}_{FE}(F_{T_1}) \tag{6}$$

where M_T is the predicted segmentation of the target set, and F_{T_2} is the feature for next stage one-shot learning. Given the features F_{S_2} from support set and the features F_{T_2} from target set, we compute the L1 distance between the features to calculate the similarity between the support set features F_{S_2} and the target set features F_{T_2} with:

$$F_{dis} = |F_{S_2} - F_{T_2}| \tag{7}$$

where F_{dis} is the feature distance. F_{dis} is then input into a fully connected layer followed by a softmax function:

$$F_{out} = softmax(\mathcal{P}_{final}(F_{dis})) \tag{8}$$

where F_{out} is the final output with one-shot prediction indicating that the target data matches with which specific class in the support set.

Sub-networks Design: We use a 512×512 fully connected layer as our feature encoder. The volume encoder and decoder design are shown in Figure 2. Our volume encoder and volume decoder consist of three level of 3D convolution layers. Unlike conventional convolutional encoder and decoder, we concatenate a Dual Squeeze-and-Excitation (DuSE) block at each level's output in order to re-calibrate the features channel-wise and spatial-wise. More specifically, as illustrated in Figure 2 bottom right, our DuSE block contains two 3D Squeeze-and-Excitation branches for spatial-Squeeze-channel-Excitation (scSE) and channel-Squeeze-spatial-Excitation (csSE), respectively [7, 17].

For scSE, we spatial-wise squeeze the input feature map using global average pooling, where the feature map is formulated as $F = [f_1, f_2, \ldots, f_C]$ here with $f_n \in \mathbb{R}^{H \times W \times D}$ denoting the individual feature channel. We flatten the global average pooling output, generating $v \in \mathbb{R}^C$ with its z-th element:

$$v_z = \frac{1}{H \times W \times D} \sum_{i}^{H} \sum_{j}^{W} \sum_{k}^{D} f_z(i, j, k) \tag{9}$$

where vector v embeds the spatial-wise global information. Then, v is feed into two fully connected layers with weights of $w_1 \in \mathbb{R}^{\frac{C}{2} \times C}$ and $w_2 \in \mathbb{R}^{C \times \frac{C}{2}}$, producing the channel-wise calibration vector:

$$\hat{v} = \sigma(w_2 \eta(w_1 v)) \tag{10}$$

where η and σ are the ReLU and Sigmoid activation function, respectively. The calibration vector is applied to the input feature map using channel-wise multiplication, namely channel-Excitation:

$$\hat{F}_{sc} = [f_1 \hat{v}_1, f_2 \hat{v}_2, \ldots, f_C \hat{v}_C] \tag{11}$$

where $\hat{v}_i$ indicates the importance of the i-th feature channel and lies in $[0, 1]$. With scSE embedded into our network, the calibration vector adaptively learns to emphasize the important feature channels while playing down the others.

In csSE, we formulate our feature map as $F = [f^{1,1,1}, \ldots, f^{i,j,k}, \ldots, f^{H,W,D}]$, where $f^{i,j,k} \in \mathbb{R}^C$ indicates the feature at spatial location (i, j, k) with $i \in \{1, \ldots, H\}$, $j \in \{1, \ldots, W\}$, and $k \in \{1, \ldots, D\}$. We channel-wise squeeze the input feature map using a convolutional kernel with weights of $w_3 \in \mathbb{R}^{1 \times 1 \times 1 \times C \times 1}$, generating a volume tensor $m = w_3 \circledast F$ with $m \in \mathbb{R}^{H \times W \times D}$. Each $f^{i,j,k}$ is a linear combination of all feature channel at spatial location (i, j, k). Then, the spatial-wise calibration volume that lies in $[0, 1]$ and can be written as:

$$\hat{m} = \sigma(m) = \sigma(w_3 \circledast F) \tag{12}$$

where σ is the Sigmoid activation function. Applying the calibration volume to the input feature map, we have:

$$\hat{F}_{cs} = [f^{1,1,1}\hat{m}^{1,1,1}, \ldots, f^{i,j,k}\hat{m}^{i,j,k}, \ldots, f^{H,W,D}\hat{m}^{H,W,D}] \tag{13}$$

where calibration parameter of $\hat{m}^{i,j,k}$ provides the relative importance of a spatial information of a given feature map. Similarly, with csSE embedded into our network, the calibration volume learns to stress the most important spatial locations while ignores the irrelevant ones.

Finally, channel-wise calibration and spatial-wise calibration are combined via element-wise addition: $F_{DuSE} = \hat{F}_{sc} + \hat{F}_{cs}$. With the two SE branch fusion, feature at (i, j, k, c) possess high activation only when it receives high activation from both scSE and csSE. Our DuSE encourages the networks to re-calibrate the feature map such that more accurate and relevant feature map can be learned.

Training Strategy and Losses: We design a customized training strategy to train our COS-Net, such that the training procedure matches the inference at test time. Specifically, two support set are randomly generated during the training procedure. Within N classes, the same n classes are randomly sampled for each support set. 1 subtomogram is randomly sampled from these classes to form a *n-way-1-shot* scheme. The ground-truth one-shot classification label is generated by matching the class labels from the two support set, i.e. 1 for matched class label and 0 for unmatched class label.

Our training loss consists of two parts, including a Binary Cross Entropy (BCE) loss for one-shot classification learning and a Dice Similarity Coefficient (DSC) loss for one-shot segmentation. Denoting the ground-truth one-shot classification label as F_{gt}, the BCE loss can be written as:

$$\mathcal{L}_{bce} = -F_{gt} log(F_{out}) - (1 - F_{gt}) log(1 - F_{out}) \tag{14}$$

Denoting the ground-truth subtomogram segmentation for the two support set as M_{gt1} and M_{gt2}, the segmentation loss can be written as:

$$\mathcal{L}_{dsc} = 2 - \frac{2 \times |M_{gt_1} \cap M_{S_1}|}{|M_{gt_1}| + |M_{S_1}|} - \frac{2 \times |M_{gt_2} \cap M_{S_2}|}{|M_{gt_2}| + |M_{S_2}|} \tag{15}$$

where M_{S_1} and M_{S_2} are the predicted segmentation from COS-Net. The total loss thus can be formulated as:

$$\mathcal{L}_{tot} = \mathcal{L}_{dsc} + \mathcal{L}_{bce} \tag{16}$$

In testing, one of the support sets during training can be replaced with the target subtomogram for direct inference.

Attention-guided Segmentation: The segmentation predicted from COS-Net is a probability distribution, which is used for guiding our final segmentation. Specifically, the volume decoder's output is a probability distribution ranging between 0 and 1. We use a 3D Conditional Random Field (CRF) to refine and generate the final 3D subtomogram segmentation. The CRF aims to optimize the following objective function:

$$E(x) = \sum_i \psi_u(x_i) + \sum_{i,j} \psi_p(x_i, x_j) \tag{17}$$

where ψ_u is the unary potential that encourages the CRF output to be loyal to the probability distribution from the COS-Net. ψ_p is the pairwise potential between label on voxel i and j and can be expanded as:

$$\psi_p = \mu(x_i, x_j)\left[w_1 exp\left(-\frac{|p_i - p_j|^2}{2\sigma_\alpha^2} - \frac{|I_i - I_j|^2}{2\sigma_\beta^2}\right) + w_2 exp\left(-\frac{|p_i - p_j|^2}{2\sigma_\gamma^2}\right)\right] \tag{18}$$

where $\mu(x_i, x_j)$ is the compatibility transformation and depends on the labels x_i and x_j such that $\mu(x_i, x_j) = 1$ if $x_i \neq x_j$, and 0 otherwise. I_i and I_j are the intensity value at voxel location i and j. p_i and p_j are the spatial coordinates of voxel i and j. w_1, w_2, σ_α, σ_β, and σ_γ are learnable parameters for CRF. This term penalizes pixels with similar position p and intensity x but with different label.

3 EXPERIMENTS AND RESULTS

3.1 Data Preparation

We prepared a realistically simulated dataset with known macromolecular structures by reconstructing the tomographic image using the projection images [16]. The limiting factors of cryo-ET, such as noise, missing wedge, and electron optical factors (Modulation Transfer Function, Contrast Transfer Function) were all properly included. The simulation process mimicked the experimental cellular sample imaging condition and tomographic reconstruction process. We took into account the randomness of macromolecule structural poses. The packed volume containing macromolecular structures were projected to a series of 2D projection images with specified tilt angle steps. The resulting projection images were convolved to include optical factors and then back-projected to obtain the reconstructed 3D simulated tomogram. 22 distinct macromolecular structures are chosen from the Proterin Databank (PDB) with their PDB ID information [1] of atomic coordinates and connectivity, and secondary structure assignments. We choose very representative macromolecules such as ribosome (4V4Q), proteasome (3DY4), and RNA polymerase (2GHO), which are well studied due to their abundance and importance in cellular functions. Each simulated tomogram of $600 \times 600 \times 300$ voxels contains 10000 randomly distributed macromolecules. Given the true position of these macromolecules inside tomograms, we collected 5,835 subtomograms of size $32 \times 32 \times 32$, belonging to 22 structural classes. The dataset with 22 distinct classes was split into a training set with 14 classes and a test set with 8 classes. Three datasets with different levels of signal-to-noise ratio (SNR) were used, including SNR=∞, SNR=1000, and SNR=0.5.

3.2 Classification Results

Table 1 summarizes the one-shot classification performance with different sub-network setup. We evaluated the one-shot classification accuracy under different noise level and various one-shot training schemes. First, comparing the COS-Net with and without volume decoder for guiding the one-shot classification, with volume decoder can significantly improve the classification accuracy for sub-networks with or without DuSE block. For example, using the SNR=1000 dataset, the 2way-1shot COS-Net with DuSE improve the accuracy from 0.928 to 0.939 by adding the volume decoder. Second, comparing the COS-Net with and without DuSE block, adding DuSE block to volume encoder/decoder can also improve the classification accuracy. However, the classification accuracy decreases as the SNR decreases, due to

Table 1: The one-shot classification accuracy on three dataset with three different SNR levels. 2way-1shot, 4way-1shot, 6way-1shot, and 8way-1shot learning scenarios are included. The highest accuracy for each learning scenario is marked in blue.

Data	**Networks**	2way-1shot	4way-1shot	6way-1shot	8way-1shot
SNR:∞	SCNN w/o Decoder	0.931	0.763	0.613	0.595
	SCNN w Decoder	0.945	0.798	0.663	0.636
	DuSE-SCNN w/o Decoder	0.934	0.772	0.618	0.603
	DuSE-SCNN w Decoder	0.957	0.831	0.672	0.646
SNR:1000	SCNN w/o Decoder	0.923	0.698	0.493	0.473
	SCNN w Decoder	0.935	0.706	0.493	0.473
	DuSE-SCNN w/o Decoder	0.928	0.701	0.504	0.479
	DuSE-SCNN w Decoder	0.939	0.718	0.534	0.513
SNR:0.5	SCNN w/o Decoder	0.812	0.599	0.501	0.387
	SCNN w Decoder	0.824	0.616	0.502	0.399
	DuSE-SCNN w/o Decoder	0.821	0.614	0.510	0.391
	DuSE-SCNN w Decoder	0.829	0.628	0.513	0.403

the structural details being degraded by noise. Meanwhile, the classification accuracy also decreases as the number of classes (way) increase.

Table 2: The segmentation results for all eight test classes on SNR$=1000$ dataset. The mean$\pm$standard deviation DSC are reported in the table. 2way-1shot, 4way-1shot, 6way-1shot, and 8way-1shot learning scenarios are reported at different rows. The macromolecular PDB ID is indicated for each classes.

SCNN	1A1S	1BXR	1EQR	1F1B	1FNT	1GYT	1KPB	1LB3
2way-1shot	.84 $\pm$ .07	.85 $\pm$ .02	.86 $\pm$ .02	.87 $\pm$ .01	.89 $\pm$ .01	.84 $\pm$ .01	.88 $\pm$ .01	.83 $\pm$ .01
4way-1shot	.84 $\pm$ .07	.85 $\pm$ .02	.86 $\pm$ .02	.87 $\pm$ .01	.90 $\pm$ .01	.85 $\pm$ .01	.88 $\pm$ .01	.84 $\pm$ .02
6way-1shot	.85 $\pm$ .08	.85 $\pm$ .02	.85 $\pm$ .02	.87 $\pm$ .01	.89 $\pm$ .01	.84 $\pm$ .01	.87 $\pm$ .01	.84 $\pm$ .01
8way-1shot	.85 $\pm$ .07	.84 $\pm$ .02	.86 $\pm$ .02	.87 $\pm$ .01	.90 $\pm$ .01	.85 $\pm$ .01	.88 $\pm$ .01	.83 $\pm$ .01
DuSE-SCNN	1A1S	1BXR	1EQR	1F1B	1FNT	1GYT	1KPB	1LB3
2way-1shot	.85 $\pm$ .08	.85 $\pm$ .02	.86 $\pm$ .02	.87 $\pm$ .01	.90 $\pm$ .01	.85 $\pm$ .01	.88 $\pm$ .01	.85 $\pm$ .01
4way-1shot	.85 $\pm$ .07	.85 $\pm$ .02	.85 $\pm$ .02	.87 $\pm$ .01	.90 $\pm$ .01	.85 $\pm$ .01	.88 $\pm$ .01	.85 $\pm$ .01
6way-1shot	.85 $\pm$ .08	.85 $\pm$ .02	.86 $\pm$ .02	.87 $\pm$ .01	.90 $\pm$ .01	.85 $\pm$ .01	.88 $\pm$ .01	.85 $\pm$ .02
8way-1shot	.84 $\pm$ .08	.85 $\pm$ .01	.86 $\pm$ .02	.87 $\pm$ .01	.90 $\pm$ .01	.85 $\pm$ .01	.88 $\pm$ .01	.85 $\pm$ .02

Table 3: The segmentation results for all eight test classes on SNR$=\infty$ dataset. The mean$\pm$standard deviation DSC are reported in the table. 2way-1shot, 4way-1shot, 6way-1shot, and 8way-1shot learning scenarios are reported at different rows. The macromolecular PDB ID is indicated for each classes.

SCNN	1A1S	1BXR	1EQR	1F1B	1FNT	1GYT	1KPB	1LB3
2way-1shot	.92 $\pm$ .08	.94 $\pm$ .03	.98 $\pm$ .02	.97 $\pm$ .02	.97 $\pm$ .03	.95 $\pm$ .03	.96 $\pm$ .01	.97 $\pm$ .02
4way-1shot	.92 $\pm$ .08	.95 $\pm$ .03	.98 $\pm$ .02	.97 $\pm$ .02	.97 $\pm$ .02	.95 $\pm$ .03	.96 $\pm$ .03	.97 $\pm$ .03
6way-1shot	.92 $\pm$ .08	.94 $\pm$ .04	.98 $\pm$ .01	.96 $\pm$ .02	.97 $\pm$ .02	.95 $\pm$ .03	.96 $\pm$ .01	.96 $\pm$ .02
8way-1shot	.92 $\pm$ .08	.94 $\pm$ .03	.98 $\pm$ .02	.96 $\pm$ .02	.97 $\pm$ .02	.95 $\pm$ .02	.96 $\pm$ .02	.96 $\pm$ .02
DuSE-SCNN	1A1S	1BXR	1EQR	1F1B	1FNT	1GYT	1KPB	1LB3
2way-1shot	.92 $\pm$ .08	.94 $\pm$ .03	.98 $\pm$ .02	.97 $\pm$ .02	.97 $\pm$ .02	.95 $\pm$ .03	.96 $\pm$ .02	.97 $\pm$ .02
4way-1shot	.93 $\pm$ .07	.96 $\pm$ .02	.98 $\pm$ .01	.97 $\pm$ .02	.97 $\pm$ .02	.95 $\pm$ .03	.96 $\pm$ .02	.97 $\pm$ .02
6way-1shot	.92 $\pm$ .08	.95 $\pm$ .03	.98 $\pm$ .02	.97 $\pm$ .02	.97 $\pm$ .02	.95 $\pm$ .02	.96 $\pm$ .02	.96 $\pm$ .02
8way-1shot	.92 $\pm$ .07	.94 $\pm$ .03	.98 $\pm$ .02	.96 $\pm$ .02	.97 $\pm$ .02	.95 $\pm$ .02	.96 $\pm$ .02	.96 $\pm$ .03

3.3 Segmentation Results

The segmentation performance of our attention-guided segmentation is evaluated using the same test set as in the classification section based on DSC:

$$DSC = \frac{2 \times |M_{gt} \cap M_{pred}|}{|M_{gt}| + |M_{pred}|} \tag{19}$$

where M_{pred} is our generated segmentation, and M_{gt} is the ground-truth segmentation. Segmentation results with different training schemes on SNR=1000 dataset are visualized in Figure 3. As we can see, our method can generate accurate 3D segmentation that does not rely on unseen classes' pixel-level or image-level training data. It is also worth notice that our method can achieve robust and consistent segmentation performance over different way one shot learning schemes. Besides, a comparison of segmentation results with and without DuSE block on eight different macromolecule classes is visualized in Figure 4. While segmentation with DuSE block does not significantly outperforms segmentation without DuSE block, they both produce reasonable segmentation of macromolecules.

The quantitative results using SNR=1000 and SNR=∞ datasets are summarized in Table 2 and Table 3, respectively. As we can observe, for all 8 unseen classes, our COS-Net is able to generate reasonable 3D segmentation. For SNR=∞ data, the DSC of our COS-Net with DuSE are all > 0.92 for all classes, indicating accurate 3D macromolecule segmentation. For SNR=1000, the DSC of COS-Net with DuSE are > 0.84. The decrease in segmentation performance is due to the increased noise level that degrades the macromolecule structure details. However, as illustrated in Figure 3, our COS-Net can still generate reasonable 3D segmentation for unseen classes.

4 DISCUSSION and CONCLUSION

In this work, we developed a one-shot learning framework for cryo-ET where simultaneous classification and segmentation can be performed for seen or unseen macromolecule subtomograms. Specifically, we developed a COS-Net to learn the class matching between a support set consisting of multiple classes with only 1 sample per class and a target subtomogram. In COS-Net, the segmentation attention is utilized to better guide the one-shot classification. In the mean time, the volume decoder of COS-Net allows us to generate the coarse segmentation of the macromolecule in the subtomogram. Then, 3D CRF is utilized to refine the 3D macromolecule segmentation from COS-Net.

We demonstrated the successful application of our COS-Net on a cryo-ET dataset consisting of 22 macromolecule classes. First, our method demonstrated accurate one-shot classification performance over dataset with different noise levels. Even with SNR as low as 0.5, the classification accuracy is over 0.8 in a 2way-1shot classification scheme. As compared to previous supervised cryo-ET classification methods with classification accuracy of about 0.9, our method is able to achieve comparable performance without using large-scale high-quality labelled data [3, 12]. Second, our method can produce high-quality 3D segmentation for unseen macromolecules under different one-shot classification schemes. As we can observe in table 3, our COS-Net can produce 3D segmentation with DSC> 0.84 on all test macromolecules over all one-shot schemes. As compared to previous supervised segmentation methods, our segmentation performance is comparable to these supervised cryo-ET segmentation models with DSC of about 0.88, which require segmentation ground truth on seen macromolecule classes for training [3, 12]. Therefore, our method provides a solution of both accurate classification and segmentation for unseen macromolecule classes.

The presented work can potentially be further improved from the following perspectives. First of all, the classification accuracy decreases as the number of classes in the support set increases. As more classes are involved in the class matching procedure and only one sample is used for each classes, the classification difficulty will naturally increase. However, our COS-Net can be extended from one-shot to few-shot if more samples are available for each class, and this strategy could potentially improve the classification accuracy. Moreover, the macromolecule alignment is not considered in the current one-shot classification pipeline. The macromolecule in the support set and target set may not be aligned, i.e. they have different orientations before feeding into our network, which could potentially decrease the classification accuracy. Subtomogram pre-processing by alignment of macromolecule in subtomograms

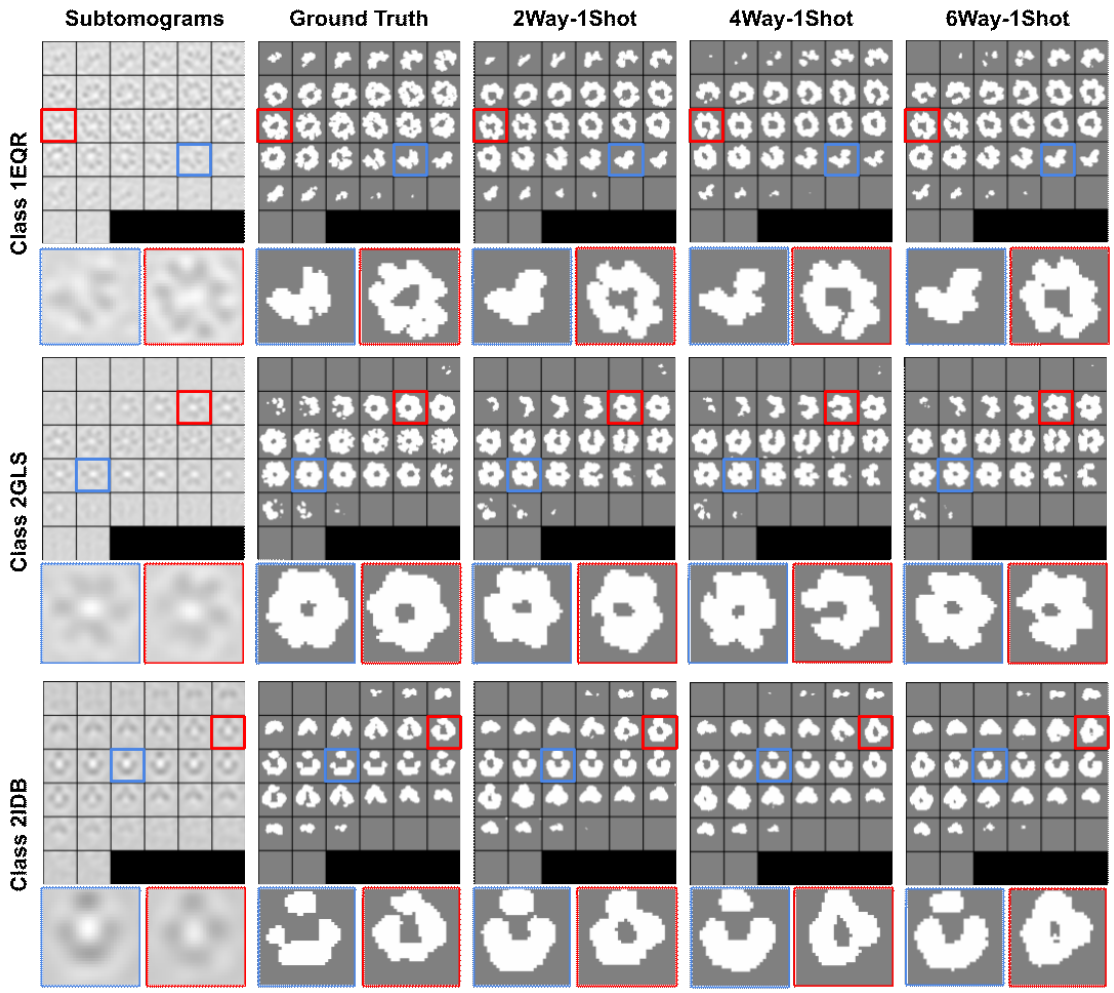

Figure 3: Illustration of segmentation results on all three test classes using COS-Net with DuSESCNN. The macromolecule PDB ID is indicated for each classes on the left. The ground truth segmentation (second column) is compared against COS-Net with 2way-1shot, 4way-1shot, 6way-1shot scenarios from second to fifth column. The enlarged images on selected 2D slices are visualized at the bottom.

could potentially further improve our classification accuracy and will be a focus in our future work [19, 13]. Second, the cryo-ET imaging data is reconstructed from limited angle conditions. The subtomogram image quality could be degraded by the limited angle reconstruction artifacts and potentially impact the downstream COS-Net's performance. Deep learning based limited angle reconstruction algorithms could be incorporated to mitigate these artifacts and potentially further improve our performance [23, 24]. Third, our study is performed based on realistically simulated cryo-ET dataset with sufficient amounts of macromolecule classes for one-shot learning studies. Currently, real cryo-ET data does not provide sufficient amounts of classes for one-shot learning studies, and we will include it in our future studies.

In summary, we developed a COS-Net for one-shot classification and segmentation in cryo-ET, which enables the classification and segmentation for unseen macromolecules in the wild. We believe our algorithm is an important step toward the large-scale and systematic in situ analysis of macromolecular structure in single cells captured by cryo-ET.

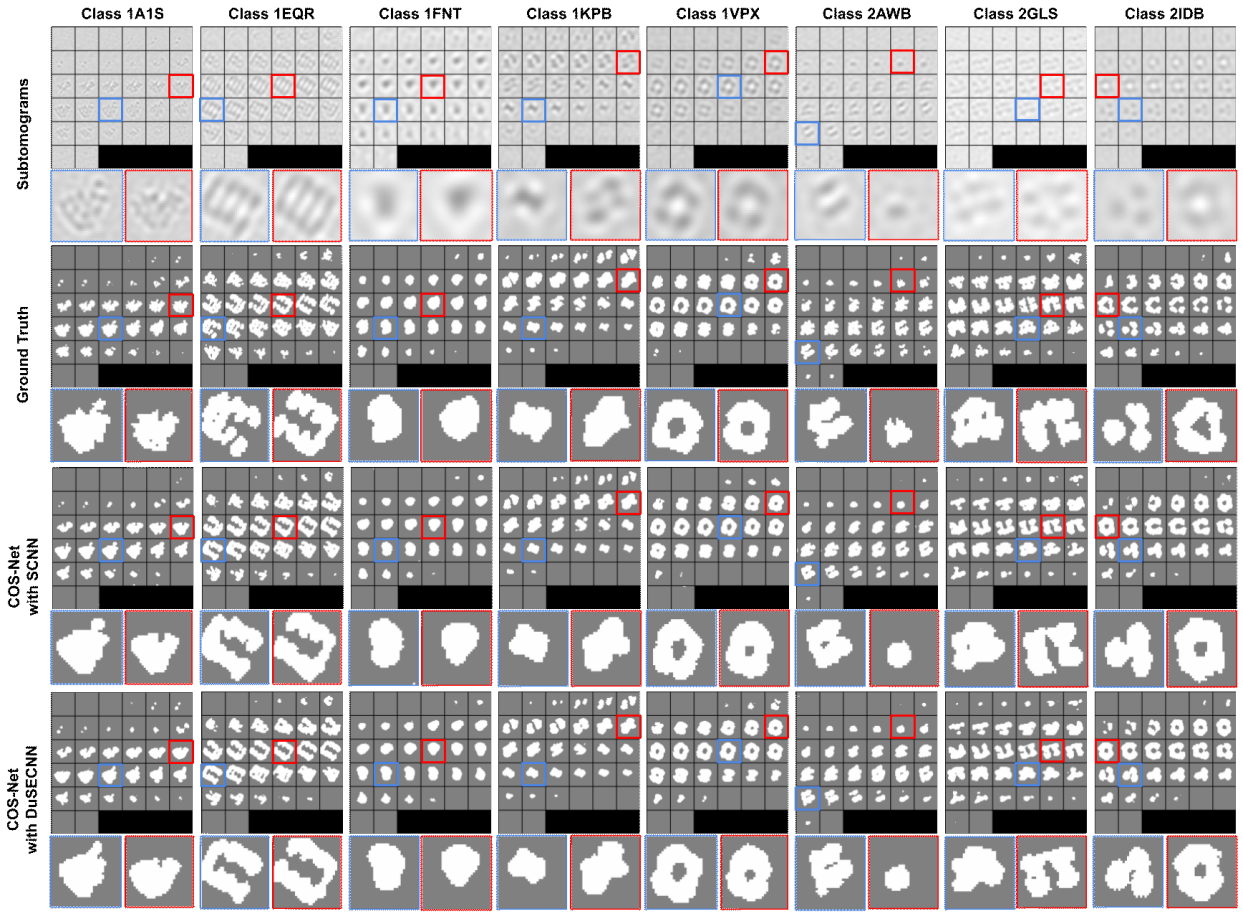

Figure 4: Illustration of segmentation results on all eight test classes using 2way-1shot. The macromolecular PDB ID is indicated for each classes on the top. The ground truth segmentation (second row) is compared against COS-Net with SCNN (third row) and COS-Net with DuSESCNN (fourth row). The enlarged images on selected 2D slices are visualized at the bottom.

Conflict of Interest Statement

The authors have no conflict of interest to declare.

Author Contributions

Bo Zhou: Conceptualization, Methodology, Software, Visualization, Validation, Formal analysis, Writing original draft. **Haisu Yu**: Methodology, Software, Visualization, Validation, Formal analysis, Writing original draft. **Xiangrui Zeng**: Conceptualization, Methodology, Writing - review and editing. **Xiaoyan Yang**: Software, Visualization, Validation, Formal analysis. **Jing Zhang**: Writing - review and editing, Supervision. **Min Xu**: Conceptualization, Methodology, Writing - review and editing, Supervision.

Funding

This work was supported in part by U.S. National Institutes of Health (NIH) grants P41GM103712, R01GM134020, and K01MH123896, U.S. National Science Foundation (NSF) grants DBI-1949629 and

IIS-2007595, AMD COVID-19 HPC Fund, and Mark Foundation 19-044-ASP. BZ was supported by the Biomedical Engineering Ph.D. fellowship from Yale University. XZ was supported by a fellowship from Carnegie Mellon University's Center for Machine Learning and Health.

Acknowledgments

We would like to thank Erica Chiang at Carnegie Mellon University for improving the manuscript.

Data Availability Statement

The code for this study can be found via https://github.com/xulabs/aitom.

References

[1] Helen M Berman, Talapady N Bhat, Philip E Bourne, Zukang Feng, Gary Gilliland, Helge Weissig, and John Westbrook. The protein data bank and the challenge of structural genomics. *Nature Structural & Molecular Biology*, 7:957–959, 2000.

[2] Chengqian Che, Ruogu Lin, Xiangrui Zeng, Karim Elmaaroufi, John Galeotti, and Min Xu. Improved deep learning based macromolecules structure classification from electron cryo tomograms. *arXiv preprint arXiv:1707.04885*, 2017.

[3] Chengqian Che, Zhou Xian, Xiangrui Zeng, Xin Gao, and Min Xu. Domain randomization for macromolecule structure classification and segmentation in electron cyro-tomograms. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 6–11. IEEE, 2019.

[4] Muyuan Chen, Wei Dai, Stella Y Sun, Darius Jonasch, Cynthia Y He, Michael F Schmid, Wah Chiu, and Steven J Ludtke. Convolutional neural networks for automated annotation of cellular cryo-electron tomograms. *Nature methods*, 14(10):983, 2017.

[5] Li Fe-Fei et al. A bayesian approach to unsupervised one-shot learning of object categories. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1134–1141. IEEE, 2003.

[6] Jialiang Guo, Bo Zhou, Xiangrui Zeng, Zachary Freyberg, and Min Xu. Model compression for faster structural separation of macromolecules captured by cellular electron cryo-tomography. In *International Conference Image Analysis and Recognition*, pages 144–152. Springer, 2018.

[7] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[8] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*, volume 2, 2015.

[9] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, pages 109–117, 2011.

[10] Ran Li, Liangyong Yu, Bo Zhou, Xiangrui Zeng, Zhenyu Wang, Xiaoyan Yang, Jing Zhang, Xin Gao, Rui Jiang, and Xu Min. Few-shot learning for classification of novel macromolecular structures in cryo-electron tomograms. *PLOS Computational Biology*, 2020.

[11] Ran Li, Xiangrui Zeng, Stephanie E Sigmund, Ruogu Lin, Bo Zhou, Chang Liu, Kaiwen Wang, Rui Jiang, Zachary Freyberg, Hairong Lv, et al. Automatic localization and identification of mitochondria in cellular electron cryo-tomography using faster-rcnn. *BMC bioinformatics*, 20(3):75–85, 2019.

[12] Chang Liu, Xiangrui Zeng, Ruogu Lin, Xiaodan Liang, Zachary Freyberg, Eric Xing, and Min Xu. Deep learning based supervised semantic segmentation of electron cryo-subtomograms. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1578–1582. IEEE, 2018.

[13] Yongchun Lü, Xiangrui Zeng, Xiaofang Zhao, Shirui Li, Hua Li, Xin Gao, and Min Xu. Fine-grained alignment of cryo-electron subtomograms based on mpi parallel optimization. *BMC bioinformatics*, 20(1):1–13, 2019.

[14] Vladan Lučić, Alexander Rigort, and Wolfgang Baumeister. Cryo-electron tomography: the challenge of doing structural biology in situ. *J Cell Biol*, 202(3):407–419, 2013.

[15] Catherine M Oikonomou and Grant J Jensen. Cellular electron cryotomography: toward structural biology in situ. *Annual review of biochemistry*, 86, 2017.

[16] Long Pei, Min Xu, Zachary Frazier, and Frank Alber. Simulating cryo electron tomograms of crowded cell cytoplasm for assessment of automated particle picking. *BMC bioinformatics*, 17(1):405, 2016.

[17] Abhijit Guha Roy, Nassir Navab, and Christian Wachinger. Recalibrating fully convolutional networks with spatial and channel "squeeze and excitation" blocks. *IEEE transactions on medical imaging*, 38(2):540–549, 2018.

[18] Min Xu, Xiaoqi Chai, Hariank Muthakana, Xiaodan Liang, Ge Yang, Tzviya Zeev-Ben-Mordehai, and Eric P Xing. Deep learning-based subdivision approach for large scale macromolecules structure recovery from electron cryo tomograms. *Bioinformatics*, 33(14):i13–i22, 2017.

[19] Xiangrui Zeng and Min Xu. Gum-net: Unsupervised geometric matching for fast and accurate 3d subtomogram image alignment and averaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4073–4084, 2020.

[20] Peijun Zhang. Advances in cryo-electron tomography and subtomogram averaging and classification. *Current opinion in structural biology*, 58:249–258, 2019.

[21] Guannan Zhao, Bo Zhou, Kaiwen Wang, Rui Jiang, and Min Xu. Respond-cam: Analyzing deep models for 3d imaging data by visualizations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 485–492. Springer, 2018.

[22] Bo Zhou, Qiang Guo, Kaiwen Wang, Xiangrui Zeng, Xin Gao, and Min Xu. Feature decomposition based saliency detection in electron cryo-tomograms. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2467–2473. IEEE, 2018.

[23] Bo Zhou, Xunyu Lin, and Brendan Eck. Limited angle tomography reconstruction: Synthetic reconstruction via unsupervised sinogram adaptation. In *International Conference on Information Processing in Medical Imaging*, pages 141–152. Springer, 2019.

[24] Bo Zhou, S Kevin Zhou, James S Duncan, and Chi Liu. Limited view tomographic reconstruction using a deep recurrent framework with residual dense spatial-channel attention network and sinogram consistency. *arXiv preprint arXiv:2009.01782*, 2020.