Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information)
19/1 (2026), 77-88. DOI: http://dx.doi.org/10.21609/jiki.v19i1.1622

Optimization of 3D U-Net Using Attention Mechanism for Accurate Protein Object Identification in Cryo-Electron Tomography

Ghessa Theniana, Sendi Setiawan, Nurjannah Syakrani*, Yudi Widhiyasana

Informatics Engineering, Bandung State Polytechnic, West Bandung, 40559, Indonesia

*E-mail: *nurjannahsy@jtk.polban.ac.id*

Abstract

Object segmentation in 3D tomograms is a key problem in Cryo-Electron Tomography (Cryo-ET) analysis. In this work, performance for the 3D U-Net architecture, and its variants with three attention mechanisms (Attention Gate (AG), Squeeze-and-excitation (SE), and Convolutional Block Attention Module (CBAM)) was evaluated. Experiments were conducted on a publicly available Cryo-ET dataset comprising two tomogram samples using a combination of (16,32,64,128) and (32,64,128,256) channel configurations, and with patch sizes of (32,64,64), (32,96,96), and (32,128,128) respectively. Model performance was evaluated with the F-Beta Score metric. The results of the analysis show that larger patch sizes significantly improve performance, and deep channel configurations do not always lead to better performance. Compared to the baseline 3D U-Net, which achieved a best score of 0.670, 3D U-Net + SE led to the best model performance with the highest F-Beta Scores at 0.718, representing an improvement of 0.048. 3D U-Net + CBAM was second with F-Beta Scores at 0.707, improving by 0.037 over the baseline, while 3D U-Net + AG exhibited prediction inconsistency, with its accuracy falling below the baseline in multiple settings. Overall, these results show that incorporating either SE or CBAM, is a better approach to improve segmentation accuracy for 3D tomogram analysis.

Keywords: *object identification, cryo-et, 3D U-Net, attention mechanism, squeeze and excitation*

1. Introduction

Cryo-electron tomography (Cryo-ET) is a state-of-the-art imaging approach used to study the structure of macromolecules under conditions that more closely resemble their native environment, and it has tremendous potential and advantages in structural biology and molecular studies [1]. Cryo-ET allows researchers to visualize subcellular architecture and proteins interacting with their environments [2], [3]. In contrast to standard electron microscopy, Cryo-ET provides three-dimensional (3D) reconstructions of cellular structures at nanometer resolution. Although Cryo-ET has great potential in various applications, the process of object identification in Cryo-ET data still faces various challenges. One of the main challenges is the very high degree of sample heterogeneity, where the observed biological samples vary greatly in shape and size. In addition, Cryo-ET image quality is often affected by low contrast and low signal to noise ratio, which makes image processing and object segmentation a complex task [4]. These issues make it difficult for conventional image processing techniques to achieve accurate and reliable segmentation results.

Because of these difficulties, most Cryo-ET datasets have so far relied on labor-intensive annotation workflows. For example, initiatives such as the Chan Zuckerberg Initiative have implemented hybrid methods. The hybrid method uses manual processes with in-house and semi-automated tools, and uses template matching techniques. The identification of the tomogram datasets required months of craft and a multitude of technical steps, collaborative discussions, and combinations of tools to annotate as a complete tomogram dataset. The intended result of this process was an annotation of nearly five hundred experimental tomograms for six different particle classes with different shapes and sizes [5]. While effective, these approaches are extremely time consuming and difficult to scale to larger datasets.

To address these limitations, deep learning has emerged as a powerful solution for biomedical image segmentation, especially in the context of volumetric data such as Cryo-ET. Among various deep learning architectures, 3D U-Net is one of the most popular because it is able to model spatial hierarchies, preserving details through the use of skip connections [6]. The 3D U-Net architecture is very effective for medical and biological imaging, where contextual information across slices is necessary for capturing accurate segmentations. Moreover, Heebner et al. [7] show how segmentation efficiency is necessary in Cryo-Et context. Their research illustrates that employing a multi-slicing approach could particularly enable more efficient approaches to segmentation while increasing accuracy as well. Nevertheless, while the 3D U-Net is a successful architecture it suffers from noise and weak feature representation failures for different segmentations in Cryo-ET tomograms, leading to false-positive or false-noise false segmentations [8]. Essentially Cryo-ET data presents with a proxy lower signal to noise index, creating inherent noise to be addressed. For example, the network often cannot discern very low protein features from the high noise of their backgrounds under Cryo-ET. In addition, standard 3D convolution applies equal importance to all parts of the tomogram, in other words preprocessing ignores the spatial context, applying computational weight to empty space and cellular debris, and the same weight to the actual protein structure itself—in turn hurting the model's overall accuracy and efficiency [9]. These challenges highlight the need for methods that can help the model selectively focus on informative features while ignoring irrelevant background.

One promising way to enhance 3D U-Net performance is through the use of attention mechanisms. Attention mechanisms, or modules, enable the model to focus only on features that are relevant reducing noise and irrelevant details [10]. Attention Gate (AG), Squeeze-and-Excitation (SE), and Convolutional Block Attention Module (CBAM) are some trustworthy methods that aim to enhance feature selection, diminish irrelevant noise, and provide higher object localization [11], [12], [13]. The literature surveyed evidence that attention mechanism can enhance the quality of segmentation by refining feature representations and consequently increasing interpretability. The application of attention-enhancing methods in the context of the missing wedge artifact and low SNR problem of Cryo-ET, can have a significant role in reducing the impact on segmentation of cellular components, related to missing wedge artifacts and low SNR [14]. The Attention Gate, proposed by Oktay et al. [11] in their work on Attention U-Net, was proven to improve the segmentation precision of small biological structures by filtering out background interference. In the context of Cryo-ET, this principle can be extended. The implementation of spatial attention, such as in the CBAM by Woo et al. [13], can direct computational focus to sub-volumes with high molecular activity. Meanwhile, channel attention, as seen in the Squeeze-and-Excitation (SE) block by Hu et al. [12], strengthens the representation of feature channels that reflect unique protein characteristics. Therefore, integrating these various attention mechanisms is expected to significantly improve the model accuracy for protein identification in Cryo-ET data.

The application of such attention-guided segmentation has shown promise. For instance, a study by Zhou et al. introduced a one-shot learning framework with attention mechanisms to classify and segment macromolecular structures in Cryo-ET data, achieving notable improvements in segmentation accuracy [14]. However, despite these successes, there remains a need for a systematic evaluation and optimization of different foundational attention architecture for this specific task.

To address this gap, this study provides the first comprehensive comparison of three widely used attention modules—Attention Gate (AG), Squeeze-and-Excitation (SE), and Convolutional Block Attention Module (CBAM)—when integrated into a 3D U-Net framework for Cryo-ET protein segmentation. While prior studies have applied attention in isolated or task-specific ways, they have not systematically contrasted the strengths and weaknesses of different attention strategies under the unique challenges of Cryo-ET data. In this work, performance is rigorously evaluated on publicly available Cryo-ET datasets, with a focus on the F-Beta Score to emphasize sensitivity in identifying true protein instances. By explicitly analyzing how each attention mechanism interacts with low signal-to-noise ratios and structural heterogeneity, this research contributes novel insights into which architectural enhancement is most effective, thereby extending and refining the scope of existing approaches.

2. Methodology

This study proposes an optimized deep learning architecture based on 3D U-Net with attention mechanisms to improve the accuracy of protein object identification in Cryo-Electron Tomography (Cryo-ET) images. The overall research methodology, illustrated in Figure 1, is divided into four main phases: data preparation, architectural modifications, training procedures, and evaluation metrics.

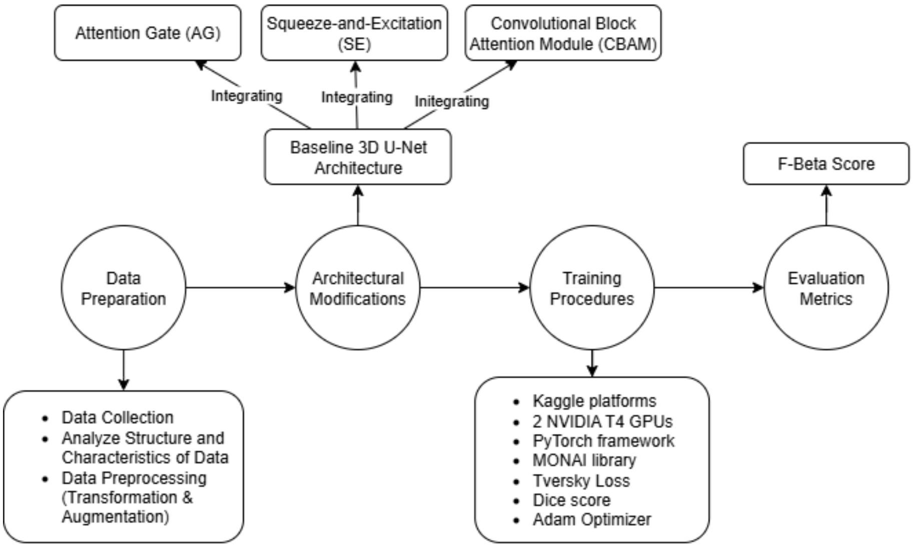

Figure 1. Research methodology.

2.1 Data Preparation

The study utilized the publicly available Cryo-ET dataset from the CZ Imaging Institute's Cryo-ET Object Identification competition on Kaggle. The dataset, sourced from the Cryo-ET Data Portal under deposition ID 10310 [5]. This dataset is consists of volumetric Cryo-ET tomograms along with point annotations indicating the locations of six types of biological particles. Table 1 lists these particles along with their difficulty level and scoring weight.

Table 1. Particle classes with their difficulty level and scoring weight.

No	Protein Name	Level	Weight
1	*Apo-ferritin*	Easy	1
2	*Beta-amylase*	Impossible (not scored)	0
3	*Beta-galactosidase*	Hard	2
4	*Ribosome*	Easy	1
5	*Thyroglobulin*	Hard	2
6	*Virus-like-particle*	Easy	1

The weight column indicates the scoring weight assigned to each particle class in the evaluation process. These weights are directly used in the calculation of the F-Beta score (with β = 4), which serves as the primary performance metric in this study. The use of weights reflects the relative difficulty of detecting different protein particles: "easy" particles (apo-ferritin, ribosome, and virus-like particles) are assigned a weight of 1, while "hard" particles (thyroglobulin and Beta-galactosidase) are assigned a higher weight of 2. This ensures that correct identification of more challenging particles contributes more strongly to the final score. An exception is Beta-amylase, which is included in the dataset but not evaluated. It is assigned a weight of 0, meaning that whether or not the model predicts this particle has no impact on the scoring outcome. This design choice is based on the fact that Beta-amylase is considered too difficult for reliable evaluation under the current framework.

The dataset consists of 79 tomograms, with each tomogram containing a resolution of 184 × 630 × 630 voxels. The average number of particles per tomogram is presented in Table 2. The class-wise distribution is reported as the average number of particles per tomogram for each particle class, providing an overview of dataset balance (see Figure 2).

Table 2. Particle classes with their difficulty level and scoring weight.

Protein Name	Average number per tomogram
Apo-ferritin	47
Beta-amylase	5
Beta-galactosidase	6
Ribosome	80
Thyroglobulin	13
Virus-like-particle	6

One of the tomogram slices is displayed in Figure 2. The dataset includes high-quality tomograms and ground truth annotations, allowing for supervised learning for protein segmentation tasks.

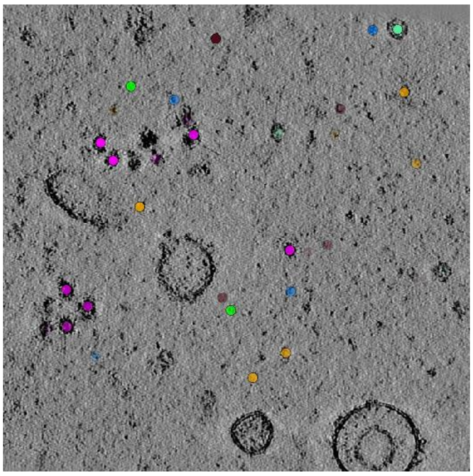

Figure 2. Slice of a tomogram.

The dataset is divided into two sets for independent training and validating runs, containing 35 tomograms each with a common test set of 9 tomograms. The first sample has the following number of annotations: 4282 for the train data and 1339 for the validation data. The second sample, with respect to the same test data used for both samples, has the following number of annotations: 2800 train annotations and 1072 for validation annotations for each sample providing a total of 1247 annotations.

To ensure data consistency and build model robustness, a combination of data transformation and augmentation process was performed using the MONAI library [15]. A series of transformations were performed to ensure format and features of each tomogram were consistent when presented to the model.

1. EnsureChannelFirstd: The tomograms' dimension format was modified from [Depth, Height, Width] to [Channel, Depth, Height, Width] which was appropriate for dimensionality expected by the 3D U-Net framework.
2. NormalizeIntensityd: Z-score normalization [16] is performed on the data, which standardizes the data to mean zero and standard deviation one [17]. This preserves the model against bias from differences in intensity scales between volumes. The transformation is applied to each voxel respectively according to equation (1).

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

In equation (1), x is the intensity of the voxel. μ denotes the mean intensity and σ the standard deviation which are calculated across all the voxels of the input volume. The process normalizes all intensity values to a measure of how many standard deviations each intensity value is from the exhibited mean.

3. Orientation: All volumes and their accompanying labels were oriented to the RAS (Right-Anterior-Superior) coordinate system to maintain geometric consistency in the dataset.

Data augmentation was then applied to artificially increase the diversity of the training data helping the model generalize to never seen data more effectively.

1. RandCropByLabelClasses: This was done to ensure extracted training patches were retaining the target label location at its center. This prevents the model from learning false representative examples about all protein classes by learning about the background data.
2. RandFlip: Volumes were randomly flipped along the depth, height or width.
3. RandRotate90: Random rotations of 90, 180 or 270 degrees were applied.
4. RandAffine: Small random rotations and small random scaling was applied to effectively mimic the natural variation in protein sizes and/or orientations.
5. RandShiftIntensity: Randomly changed contrast and brightness of images. This trains the model to detect structures based in structural patterns and no based on an intensities that represents absolute values.
6. RandStdShiftIntensity: Randomly changed contrast and brightness of images. This is also based on standard deviation.
7. RandGaussianNoise: Synthetic Gaussian noise was added to images, which further improves the models resilience to the generally high and variable noise in Cryo-ET.

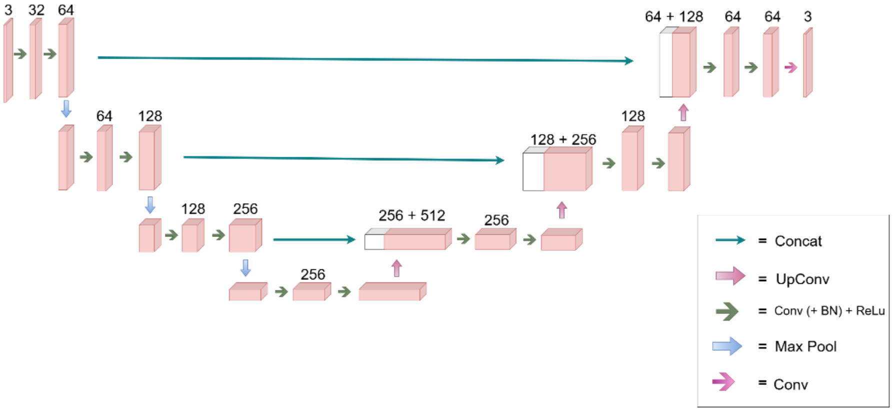

Figure 3. 3D U-Net architecture.

2.2 Architectural Modifications

The baseline architecture used is 3D U-Net. 3D U-Net is a well-established model for volumetric biomedical image segmentation tasks, and is particularly suitable for this application due to its encoder-decoder architecture with skip connections, shown in Figure 3. However, this model still has limitations recognizing objects with lower contrast and areas that are less facile to recognize. In order to improve localization accuracy and robustness in low-contrast Cryo-ET images, three attention mechanisms were added to the baseline 3D U-Net:

1. Attention Gate (AG)

The attention gate (AG), developed by Oktay et al. [11], is yet another mechanism designed to help convolutional networks focus on the relevant part of the input data, as depicted in Figure 4. Integrated between the encoder and decoder blocks, attention gates learn to suppress irrelevant features and allow important features to stand out, helping the model segment the data more accurately. AG trains by computing attention coefficients by using a gating signal and input features to learn to suppress irrelevant features and allow important features to stand out. The attention coefficients are used to control the modulated feature maps which are passed through the network.

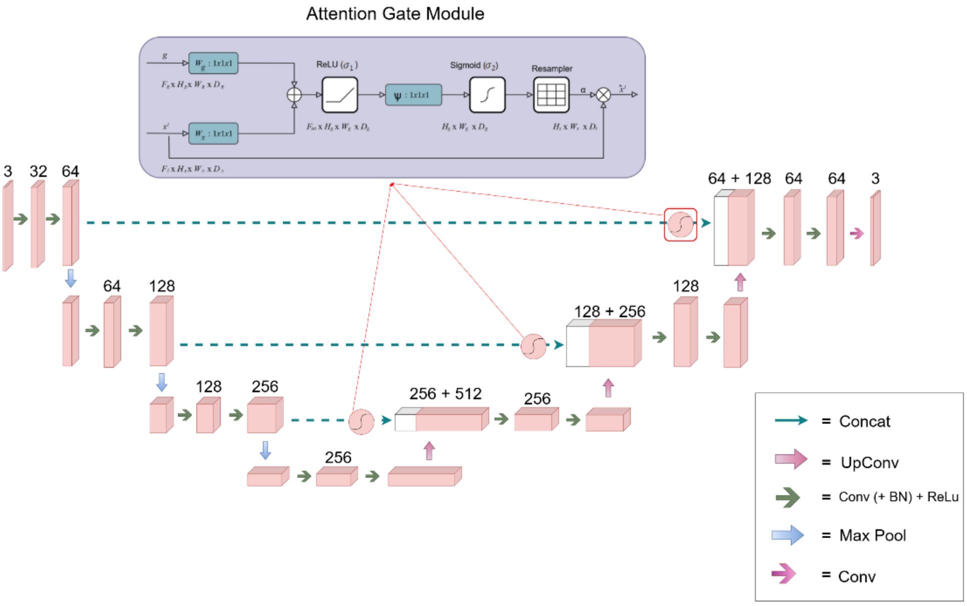

Figure 4. Attention gate architecture .

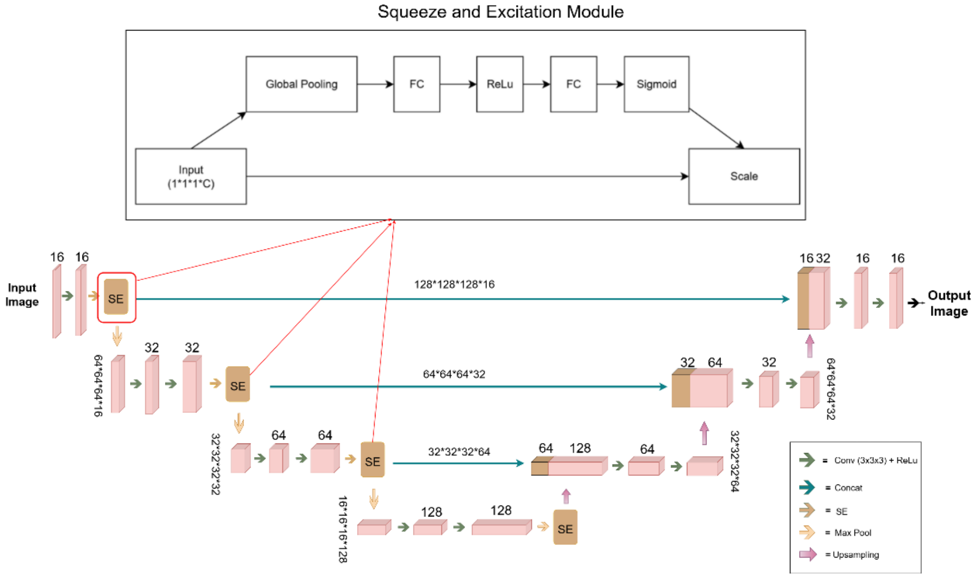

Figure 5. Squeeze-and-Excitation architecture.

2. Squeeze-and-Excitation (SE)
   Squeeze-and-Excitation, propounded by Hu, Shen and Sun [12] applies channel-wise attention to adjust feature maps to emphasize informative feature channels and suppress less informative feature channels to exploit inter-dependencies across different feature channels. While initially proposed for 2D convolutional networks, the SE component has been used in 3D U-Net architectures to capture volumetric data. For example, Yu et al. [18] leveraged a 3D U-Net with Squeeze and Excitation, the overall architecture is illustrated in Figure 5.

3. Convolutional Block Attention Module (CBAM)
   Convolutional Block Attention Module designed by Woo et al. [13] improves feature representation by using both channel attention and spatial attention one after another (see Figure 6). The two attention mechanisms provided in CBAM first finds out what to pay attention to (the form of channel attention) and then finds out where it is located (the form of spatial attention) to narrow the feature maps. Next, we present an overview of the two sub-modules, channel and spatial, in Figure 7.

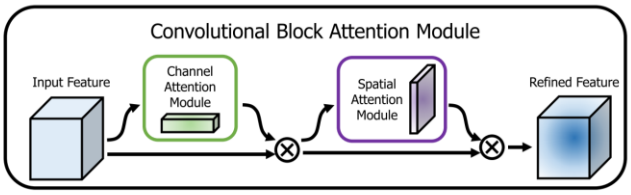

Figure 6. Convolutional block attention module [13]

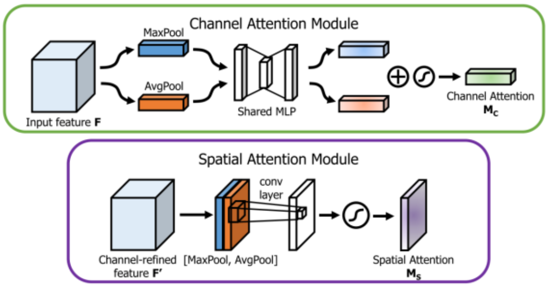

Figure 7. Channel and spatial attention module diagram [13].

In summary, the three attention mechanisms differ in their focus of feature refinement. AG emphasizes spatially filtering features passed through skip connections, targeting relevant regions during decoding. SE, on the other hand, operates at the channel level, globally recalibrating feature maps to capture channel dependencies. It highlights what channels are more informative, helping the network prioritize useful features across channels. CBAM combines both strategies by sequentially applying channel attention to determine what features are important, followed by spatial attention to determine where they are located.

2.3 Experiment Design

We are interested in exploring how varying architectural decisions and hyperparameters impact segmentation performance, and conducted the following studies to investigate our research questions. Specifically to compare the baseline model with the other models that utilized attention and analysing the affect of architectural decisions, model complexity, and the size of the input data. This study compare the performance of four different model architectures:

1. Baseline 3D U-Net: the standard 3D U-Net architecture, with no attentional contributions model.
2. 3D U-Net + Attention Gate (AG): the baseline model, with attention gates applied to skip connections.
3. 3D U-Net + Squeeze-and-Excitation (SE): the baseline model, with SE blocks added in the encoder.
4. 3D U-Net + CBAM: the baseline model, has been augmented with Convolutional Block Attention Module (CBAM).

This plain comparison between models was done to determine which attentional contributions offer the greatest/least significant impact on the Cryo-ET segmentation task with the least variability.

In addition to comparing architectures, the study also assessed the impact of two hyperparameters: channel configuration and patch size. In order to highlight the model complexity effect, two channel configurations were tested, a standard configuration of (32, 64, 128, 256), represent a standard 3D U-Net architecture; and a pruned channel configuration of (16, 32, 64, 128), will determine whether it is possible to assess performance using fewer computational resources. Simultaneously, the selection of patch sizes was informed by prior work from Tang et al. [19], who demonstrated that larger patched improve 3D segmentation performance. While their study used a depth of 48, the patch depth for this research was set to 32, so that the patch size configuration is (32, 64, 64), (32, 96, 96), and (32, 128, 128) to accommodate the increased GPU memory requirements of the attention mechanisms. Since protein objects in tomograms are relatively small and sparsely distributed, the relevance of patch size lies in balancing the likelihood of capturing proteins against computational efficiency. Thus, different patch size were selected, smaller patches offer efficiency but may include fewer proteins, while larger patches increase contextual information at the expense of higher memory usage. Each architectural and hyperparameter combination was then subjected to the training and evaluation procedure.

2.4 Training Procedures

The experiments were conducted on a Kaggle environment with two NVIDIA T4 GPUs (dual GPU). The deep learning model was built and trained in the PyTorch library (version 2.5.1) with the usage of the MONAI (Medical Open Network for AI) library to facilitate the experiments in deep learning in the medical imaging space. The main loss function used was Tversky Loss, which is a generalization of the Dice score. Unlike Dice or Cross-Entropy loss, Tversky Loss introduces adjustable parameters to control the balance between false positives and false negatives. This makes it particularly suitable for highly imbalanced segmentation tasks, such as cryo-ET data, where the volume of protein voxels is very small compared to the background. In this context, prioritizing the reduction of false negatives is important, since missing a protein structure is more detrimental than predicting additional background voxels. Therefore, Tversky Loss was selected to better handle class imbalance and to emphasize sensitivity toward protein regions during training. The model weights were updated using a standard Adam Optimizer with a learning rate of 1e-3. The model was trained for a total of 200 epochs, with a batch size of 1 given the hardware available.

2.5 Evaluation Metrics

Model performance was assessed using the F-Beta score, which combines precision and recall using a weighting factor β. This study fixed β at 4 to place more attention on recall (reduce costs from false negatives). Recall is emphasized when β is greater than 1. Therefore, the model was highly penalized for failing to capture false negatives as compared to false positives. This choice reflects the biological significance of the task: failing to identify a protein particle means losing important structural information, while occasional false positives can still be filtered in later stages. The F-Beta score with β > 1 shifts the balance towards recall, and β = 4 provides a strong emphasis on minimizing missed detections. The F-Beta score [20] is defined in equation (2).

$$F_\beta = \frac{\left((1+\beta^2) * Precision * Recall\right)}{\beta^2 * Precision + Recall} \quad (2)$$

Where β is a positive parameter that tunes the balance between precission and recall. A β value greater than 1 gives more weight to recall, while a value less than 1 gives more weight to precision.

3. Results and Analysis

The performance of each model architecture was evaluated across 48 different experimental scenarios, using various configurations of the 3D U-Net architecture enhanced with different attention mechanism, including Attention Gate (AG), Squeeze-and-Excitation (SE), and Convolutional Block Attention Module (CBAM), channel configuration, and patch size. The complete results for the data sample 1 are presented in Table 3. To complement these results, an example of the output is illustrated in Figure 8 which shows the prediction with color legend for every class. The same set of experiments was then conducted on the second data sample. The complete results for the data sample 2 are presented in Table 4.

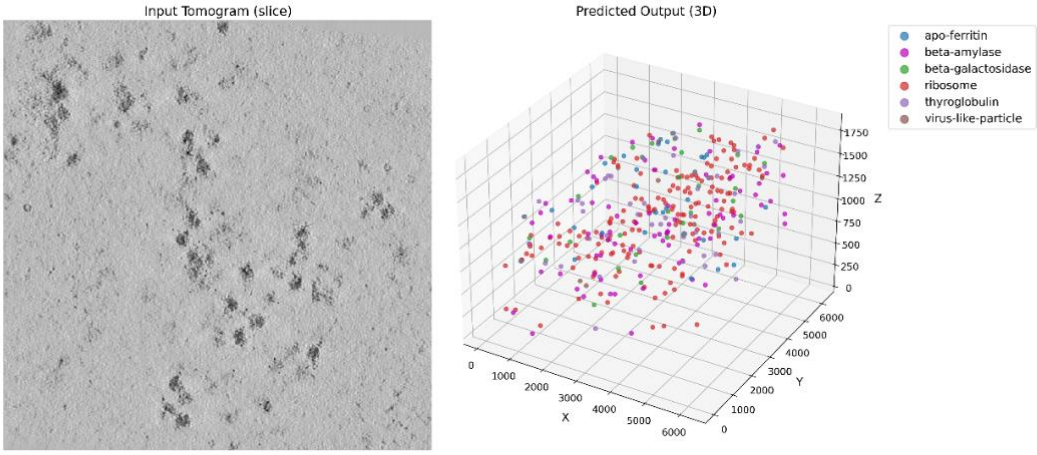

Figure 8. Examples of data input and predicted output.

Table 3. Experiment results on data sample 1.

Model	channel	patch_size	F-Beta Score	Training Time (minutes)
3D U-Net	(16, 32, 64, 128)	(32, 64, 64)	0.492	141
		(32, 96, 96)	0.524	162
		(32, 128, 128)	0.643	182
	(32, 64, 128, 256)	(32, 64, 64)	0.509	144
		(32, 96, 96)	0.560	165
		(32, 128, 128)	0.670	207
3D U-Net + AG	(16, 32, 64, 128)	(32, 64, 64)	0.524	146
		(32, 96, 96)	0.585	169
		(32, 128, 128)	0.617	207
	(32, 64, 128, 256)	(32, 64, 64)	0.484	159
		(32, 96, 96)	0.567	204
		(32, 128, 128)	0.609	330
3D U-Net + SE	(16, 32, 64, 128)	(32, 64, 64)	0.595	150
		(32, 96, 96)	0.637	176
		(32, 128, 128)	0.714	210
	(32, 64, 128, 256)	(32, 64, 64)	0.429	164
		(32, 96, 96)	0.657	205
		(32, 128, 128)	0.683	348
3D U-Net + CBAM	(16, 32, 64, 128)	(32, 64, 64)	0.545	151
		(32, 96, 96)	0.630	193
		(32, 128, 128)	0.683	212
	(32, 64, 128, 256)	(32, 64, 64)	0.552	166
		(32, 96, 96)	0.678	247
		(32, 128, 128)	0.685	402

Table 4. Experiment results on data sample 2.

Model	channel	patch_size	F-Beta Score	Training Time (minutes)
3D U-Net	(16, 32, 64, 128)	(32, 64, 64)	0.554	135
		(32, 96, 96)	0.626	152
		(32, 128, 128)	0.657	173
	(32, 64, 128, 256)	(32, 64, 64)	0.546	137
		(32, 96, 96)	0.643	154
		(32, 128, 128)	0.658	185
3D U-Net + AG	(16, 32, 64, 128)	(32, 64, 64)	0.555	151
		(32, 96, 96)	0.562	166
		(32, 128, 128)	0.560	200
	(32, 64, 128, 256)	(32, 64, 64)	0.463	165
		(32, 96, 96)	0.623	192
		(32, 128, 128)	0.638	310
3D U-Net + SE	(16, 32, 64, 128)	(32, 64, 64)	0.555	140
		(32, 96, 96)	0.665	167
		(32, 128, 128)	0.718	214
	(32, 64, 128, 256)	(32, 64, 64)	0.501	155
		(32, 96, 96)	0.625	222
		(32, 128, 128)	0.693	368
3D U-Net + CBAM	(16, 32, 64, 128)	(32, 64, 64)	0.618	148
		(32, 96, 96)	0.696	169
		(32, 128, 128)	0.694	207
	(32, 64, 128, 256)	(32, 64, 64)	0.584	161
		(32, 96, 96)	0.656	231
		(32, 128, 128)	0.707	422

Across all models and both samples tested, the best performing model was found to be 3D U-Net with SE attention, achieving F-Beta Scores of 0.714 from the first sample and 0.718 from the second one, under the configuration (16, 32, 64, 128) channel depth and patch size (32, 128, 128). To assess whether the performance improvement of the 3D U-Net + SE architecture over the baseline was statistically significant, the mean and standard deviation of F-Beta scores were evaluated for two runs and six configurations. The baseline 3D U-Net had an average F-Beta of 0.610 ± 0.064, while the 3D U-Net + SE average score was 0.638 ± 0.078. The 3D U-Net + CBAM version of the architecture had an average F-Beta of 0.651 ± 0.050. The AG mechanism resulted in an average F-Beta score of 0.567 ± 0.055 suggesting that this spatial gating mechanism did not perform as well as the baseline (3D U-Net). Collectively, channel-wise attention mechanisms (SE and CBAM) showed a consistent pattern of improved segmentation accuracy compared to 3D U-Net and the AG spatial gating mechanism did not improve segmentation performance in this sparse protein segmentation application.

3.1 The effect of patch size on accuracy and training time

Experimental results demonstrate that increasing the patch size from (32, 64, 64) to (32, 128, 128) significantly improves segmentation accuracy across all tested models. The largest patch size, which is (32, 128, 128), consistently yielded the highest F-Beta Scores 0.714 from the first sample data and 0.718 from the second one. This indicating its superior ability to capture more spatial context for identifying protein structure in low-contrast Cryo-ET images.

In sample 1, the baseline 3D U-Net model with (16, 32, 64, 128) channel configuration, achieved F-Beta Scores of 0.492, 0.524, and 0.643 for patch size (32, 64, 64), (32, 96, 96) and (32, 128, 128). This represent an improvement of +0.151 from the smallest to the largest patch size. A similar trend was observed in the other samples. Sample 2, showed that the baseline model with the same configuration also retrieved a consistent improvements. F-Beta Scores increased from 0.458 for patch size (32, 64, 64) to 0.579 for patch size (32, 128, 128). These results support that increased patch size can improve the model's ability to learn more complex structural patterns within volumetric data, thereby improving segmentation performance. However, this increase shows that there is a clear trade off between segmentation optimization and computational efficiency. If, for instance, we chose a larger patch size, then we have to process more voxels per batch, leading to increased memory and longer training.

3.2 The effect of channel configuration on accuracy values

Furthermore to patch size, channel configuration (i.e. channel count at each level of the encoder-decoder architecture) also can impact model performance. Test results demonstrated that channel depth is not a factor that guarantees increased accuracy and depends on the model architecture and patch size.

When comparing the two channel configurations, the deeper configuration, (32, 64, 128, 256), did not always outperform the shallower (16, 32, 64, 128) version. In some cases, the added model complexity led to an unsuccessful result, particularly when smaller patch sizes were used.

For example, in sample 1, 3D U-Net with SE model and patch size (32, 64, 64) has decreased F-Beta Score from 0.595, using the low channel configuration (16, 32, 64, 128) to 0.429 with the high-channel setup (32, 64, 128, 256). Similarly, in all patch sizes tested, 3D U-Net with AG model consistently performed worse with the higher-channel setup (32, 64, 128, 256). The lowest patch size (32, 64, 64),

This pattern was consistent in sample 2. The deeper channel configuration generally underperformed with small and medium patch sizes, for instance, both 3D U-Net + SE and 3D U-Net + CBAM has decreased F-Beta Score from 0.665 to 0.625 and 0.696 to 0.656. Performance only becoming comparable when using the largest parch size of (32, 128, 128). These results strongly suggest that the benefits of a larger model capacity, provided by more channels, are dependent on sufficient spatial context from larger patches. Without enough spatial information, the more complex model is likely to overfit, which reduces segmentation accuracy.

On the contrary, when a larger patch size were made available, having more channels was more effective and warranted the more complex channel configuration. For example, the 3D U-Net + CBAM model had a score of 0.683 on sample 1, and with the deeper channel setup, it had a slightly higher score of 0.685. This indicated that the channel count must be matched with the potential patch size, and therefore optimum performance depends on the balancing the channel count patch size. If there is an ill-fit, mostly in the case of deeper channels with small patch size (as previously mentioned), there is the potential for improvement versus worsening performance.

3.3 The influence and effectiveness of the attention mechanism on the model.

Experimental findings confirmed that attention gives an improvement in results with segmentation than the baseline 3D U-Net, but with deviations due to type of attention and architecture. Overall, the highest F-Beta Scores were obtained with the 3D U-Net with SE model consistently and using greater patch size settings of all varients tested. In the first sample, best performance was obtained using (16, 32, 64, 128) channel configuration and patch size (32, 128, 128), yielding an F-Beta Score of 0.714. Similarly, in Sample 2, the same configuration produced an even higher score of 0.718. This indicating that model with this configuration have a strong generalization capability and robustness in identifying protein structures in low-contrast Cryo-ET volumes.

The second-best performing model was 3D U-Net with CBAM model, which also showed consistent improvements over the baseline, although the accuracy remains slightly lower than 3D U-Net with SE. In Sample 1, the highest F-Beta Score collected from a configuration using 3D U-Net with CBAM model was 0.685, which used (32, 64, 128, 256) channels, due to using the (32, 128, 128) patch size. In Sample 2, the configuration could be seen to score 0.707 suggesting that this configuration is stable and occasionally provides better scores especially with a channel configurations near the greater depth and larger input patches. On the other hand, 3D U-Net with AG model performs even worse than the basic 3D U-Net model. The relatively poor performance of the 3D U-Net + AG architecture, in some cases falling below the baseline, can be attributed to the characteristics of Cryo-ET data and the inherent design of the Attention Gate mechanism. AG primarily operates through spatial attention, emphasizing regions deemed salient while suppressing background information. This strategy has demonstrated effectiveness in medical imaging domains where anatomical boundaries are relatively well defined. However, Cryo-ET tomograms present a substantially different challenge: they are dominated by low signal-to-noise ratios and high structural heterogeneity, which obscure the spatial boundaries that AG relies upon. Consequently, AG may mistakenly attenuate weak but relevant protein signals together with noise, thereby reducing the accuracy of object localization. In addition, unlike SE or CBAM, AG lacks channel-wise recalibration capabilities. Channel-level attention is particularly important in Cryo-ET, where different protein classes may share overlapping spatial patterns but can still be distinguished through channel-specific feature representations. Without this mechanism, AG provides limited robustness against noise and fails to fully exploit discriminative cues embedded in the feature channels.

These findings confirm that channel-wise attention (SE) offers the most effective enhancement to the 3D U-Net architecture for Cryo-ET segmentation tasks. The superior performance of SE over CBAM and AG can be explained by the characteristics of Cryo-ET data. Cryo-ET tomograms are dominated by high noise levels and low contrast, which makes it challenging for spatial attention mechanisms alone-as in AG-to effectively distinguish subtle protein features from background interference. While CBAM combines both channel and spatial attention, its added complexity may overemphasize local spatial cues that are unreliable in such noisy conditions, leading to less stable improvements.

In contrast, SE operates purely at the channel level by recalibrating feature maps according to their global importance. This mechanism is particularly well-suited for Cryo-ET because different protein classes often share overlapping spatial patterns but can still be distinguished through more discriminative channel representations. By amplifying the channels most relevant to protein structures and suppressing those dominated by noise or background, SE improves the signal-to-noise separation more effectively than CBAM or AG.

However, the integration of other attention mechanisms-AG and CBAM-also provides improvement although not in consistently and must be supported by the exact selection of configuration values for identifying protein objects in Cryo-ET images. These results highlight the importance of selecting an appropriate attention module based on dataset characteristics and balenceness between accuracy, computational cost, and model generalization.

4. Conclusion

This study aimed to enhance the performance of 3D U-Net through the integration of attention mechanisms, spesifically for accurate protein object identification in Cryo-ET images. Three types of attention modules that used-Attention Gate, Squeeze-and-Excitation, and Covolutional Block Attention Module-were systematically integrated into the baseline architecture to improve the feature localization and suppress irrelevant background noise. Experimental results demonstrated that the 3D U-Net with SE attention model achieved the highest performance, yielding F-Beta score 0.714 and 0.718 on two independent data samples in configuration channel (16, 32, 64, 128) and patch size (32, 128, 128), outperforming both the baseline and other attention variants in the same configuration.

The inclusion of attention mechanisms exhibited to increase segmentation performance vastly, especially for the detection of low-contrast or more complex protein particles. The increments of patch size generally displayed a consistent impact on accuracy. Indeed, larger patch sizes (i.e. (32, 128, 128)), led to consistently greater F-Beta Scores in all configuration or models. This demonstrating that larger patches enable the model to capture more spatial and contextual information. However, this increase in accuracy came with longer training times, highlighting a trade-off between performance and efficiency. In contrast, the effect of channel configuration on accuracy was more variable. Although generally higher channel configurations (i.e. (32, 64, 128, 256)), promoted separation, not all of the higher configuration were associated with improved accuracy, particularly with smaller patches, 'higher' configurations were more susceptible to overfitting. A further observation was that the best practice (i.e. channel and patch size configurations) to maximize model performance is a higher channel configuration paired with larger patch size configurations. The implementation of attention mechanisms dramatically increased the computational overhead. For minimum, the integration of CBAM into 3D U-Net resulted in a near doubling the time and stopping tolerance; 207 minutes for baseline 3D U-Net to 402 minutes on 3D U-Net with CBAM model in the same configurations. The significant increase suggests that attention mechanisms would add considerable computational overhead. It is worth noting that any increases in segmentation accuracy should be balanced with training efficiency and configuration selection to best align with research aims.

Building on these findings, several directions for future research can be suggested. First, further exploration of alternative attention mechanisms, such as Transformer-based attention (e.g., Vision Transformers), non-local blocks, or hybrid attention that combines spatial and channel-wise focus, may provide additional gains in accuracy and robustness. Second, adaptive optimization methods, such as automated architecture search or hyperparameter optimization, could be employed to automatically determine the most effective patch size and channel configuration for a given tomogram dataset. Third, to evaluate generalization, it is important to test the proposed models on more diverse Cryo-ET datasets, including data from different sources or more complex biological samples. Such extensions would help validate the robustness of the approach and further advance the application of deep learning in structural biology.

References

[1] J. Frank, “Advances in the field of single-particle cryo-electron microscopy over the last decade,” Feb. 01, 2017, *Nature Publishing Group*. doi: 10.1038/nprot.2017.004.

[2] S. Sriram, “The cryo-EM revolution: Fueling the next phase,” 2019, *International Union of Crystallography*. doi: 10.1107/S2052252519000277.

[3] E. Moebel *et al.*, “Deep Learning Improves Macromolecule Identification in 3D Cellular Cryo-Electron Tomograms,” Apr. 16, 2020. doi: 10.1101/2020.04.15.042747.

[4] M. Turk and W. Baumeister, “The promise and the challenges of cryo-electron tomography,” Oct. 01, 2020, *Wiley Blackwell*. doi: 10.1002/1873-3468.13948.

[5] A. Peck *et al.*, “Annotating CryoET Volumes: A Machine Learning Challenge,” Nov. 06, 2024. doi: 10.1101/2024.11.04.621686.

[6] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Verlag, 2015, pp. 234–241. doi: 10.1007/978-3-319-24574-4_28.

[7] J. E. Heebner, C. Purnell, R. K. Hylton, M. Marsh, M. A. Grillo, and M. T. Swulius, “Deep Learning-Based Segmentation of Cryo-Electron Tomograms,” *Journal of Visualized Experiments*, vol. 2022, no. 189, Nov. 2022, doi: 10.3791/64435.

[8] Z. Li, “Advancements and Challenges in Medical Image Segmentation: A Comprehensive Survey,” 2025.

[9] T. Bepler *et al.*, “Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs,” *Nat Methods*, vol. 16, no. 11, pp. 1153–1160, Nov. 2019, doi: 10.1038/s41592-019-0575-8.

[10] J. Schlemper *et al.*, “Attention gated networks: Learning to leverage salient regions in medical images,” *Med Image Anal*, vol. 53, pp. 197–207, Apr. 2019, doi: 10.1016/j.media.2019.01.012.

[11] O. Oktay *et al.*, “Attention U-Net: Learning Where to Look for the Pancreas,” Apr. 2018, doi: 10.48550/arXiv.1804.03999.

[12] J. Hu, L. Shen, and G. Sun, “Squeeze-and-Excitation Networks,” 2017. doi: 10.1109/CVPR.2018.00745.

[13] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional Block Attention Module,” *ECCV*, Jul. 2018, doi: 10.1007/978-3-030-01234-2_1.

[14] B. Zhou, H. Yu, X. Zeng, X. Yang, J. Zhang, and M. Xu, “One-Shot Learning With Attention-Guided Segmentation in Cryo-Electron Tomography,” *Front Mol Biosci*, vol. 7, Jan. 2021, doi: 10.3389/fmolb.2020.613347.

[15] M. J. Cardoso *et al.*, “MONAI: An open-source framework for deep learning in healthcare,” Nov. 2022, doi: 10.48550/arXiv.2211.02701.

[16] R. E. . Walpole, R. H. . Myers, S. L. . Myers, and Keying. Ye, *Probability & statistics for engineers & scientists : MyStatLab update*. Pearson, 2017.

[17] S. Wahyuni Asrianda and S. Retno, “ITEJ Information Technology Engineering Journals Classification of Family Hope Program Assistance Recipients Using the C4.5 Algorithm with Z-Score Normalization (Case Study in Atu Lintang District),” vol. 10, pp. 160–173, 2025, doi: 10.24235/itej.v10i1.207.

[18] T. Yu, X. Wang, T. J. Chen, and C. W. Ding, “Fault Recognition Method Based on Attention Mechanism and the 3D-UNet,” *Comput Intell Neurosci*, vol. 2022, 2022, doi: 10.1155/2022/9856669.

[19] Y. Tang *et al.*, “High-resolution 3D abdominal segmentation with random patch network fusion,” *Med Image Anal*, vol. 69, Apr. 2021, doi: 10.1016/j.media.2020.101894.

[20] P. Christen, D. J. Hand, and N. Kirielle, “A Review of the F-Measure: Its History, Properties, Criticism, and Alternatives,” Mar. 31, 2023, *Association for Computing Machinery*. doi: 10.1145/3606367.