Supplementary Material
BOE-ViT: Boosting Orientation Estimation with Equivariance in Self-Supervised 3D Subtomogram Alignment

Contents

List of Figures

List of Tables

A Background

What is Cryo-Electron Tomography (Cryo-ET)? Cryo-Electron Tomography (cryo-ET) is an advanced cellular imaging technique that produces three-dimensional views of cellular samples at nanometer resolution [1, 2]. In cryo-ET, biological samples are first vitrified at cryogenic temperatures below -150°C, preserving their native structures without chemical fixation or dehydration [3]. During imaging, the sample is tilted through a series of angles (typically ±60° with 1-3° increments) while being exposed to an electron beam, producing a tilt-series of 2D projection images [4, 5]. These projections are then computationally reconstructed into a 3D volume [6, 7] called a tomogram, which provides a comprehensive view of the cellular landscape. This technique is particularly powerful because it allows visualization of macromolecular structures in their native cellular context, making it fundamental for in situ structural biology [8, 9].

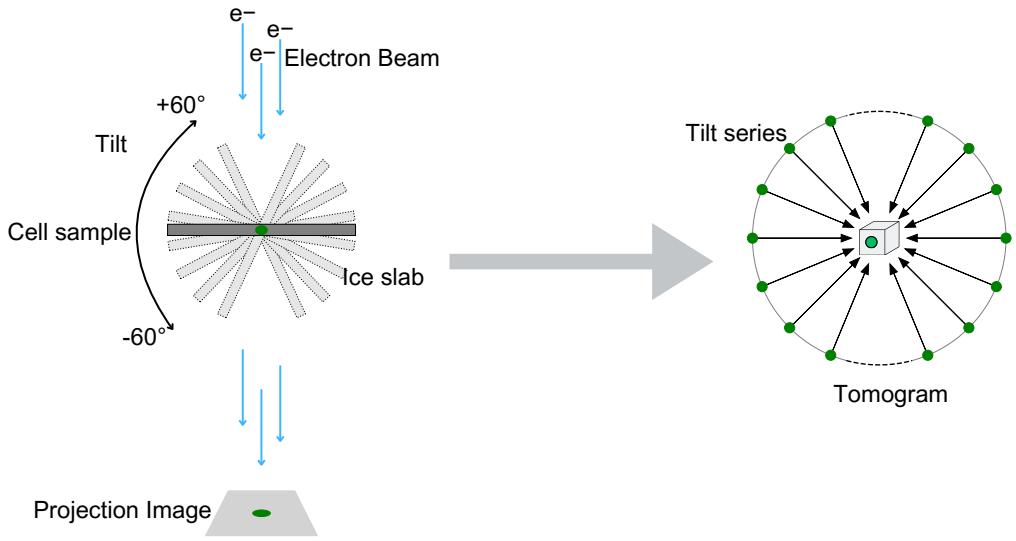

Figure 1: An illustration of cryo-ET imaging process.

What is Cryo-ET Subtomogram Alignment? Subtomogram alignment is a crucial computational process in cryo-ET analysis [10, 3, 4] that enables high-resolution structure determination from tomographic data. The process involves extracting multiple copies of similar particles (subtomograms) from tomograms and aligning them in 3D space to generate an averaged structure with improved signal-to-noise ratio [11]. This alignment presents significant computational challenges [12] due to several factors: (1) the complexity of three-dimensional rotational and translational alignment [13], (2) the inherently low signal-to-noise ratio of cryo-ET data compared to single-particle cryo-EM [14], and (3) the computational intensity of processing large volumetric datasets. Alignment algorithms typically employ sophisticated cross-correlation methods and can be performed using either angular searches or fast rotational matching approaches [15, 16]. The process is iterative, with each round of alignment improving the average structure until convergence is reached [15, 17]. Successful subtomogram alignment is essential for achieving higher resolution structural information from cryo-ET data and understanding macromolecular organizations in their native cellular context [8].

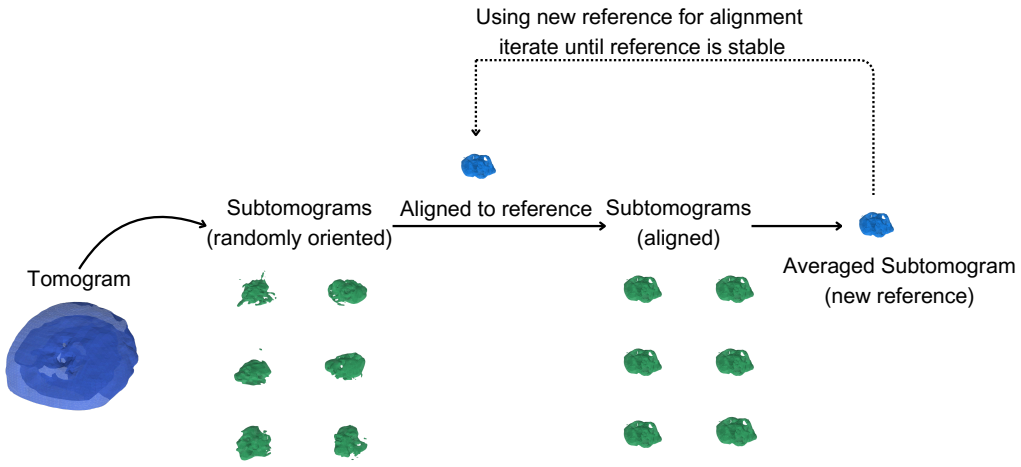

Figure 2: Workflow of Subtomogram Alignment.

B Subtomogram Visualization

B.1 Example of Input Subtomograms

To demonstrate the effectiveness of our method, we selected five representative macromolecular structures. As shown in Fig. 3, we visualized these structures through 2D slice representations of their corresponding subtomograms, simulated at varying signal-to-noise ratios (SNRs). The visualization spans from high-quality conditions (SNR = 100), where structural features are clearly visible, to extremely noisy scenarios (SNR = 0.01) that closely mimic challenging experimental conditions. This comprehensive visualization not only illustrates the challenging nature of subtomogram analysis but also provides a robust testbed for evaluating our method's performance across different noise conditions typically encountered in real experimental scenarios.

B.2 2D Slice Representation

As visualized in Fig. 4, we present the 2D slice representations of subtomograms at different processing stages. For each macromolecular structure, the visualization sequence shows the original input subtomogram, followed by three augmented versions, the target subtomogram, and the final aligned result. This visual representation effectively demonstrates the transformations that occur during our processing pipeline.

C Preliminaries

In this section, we establish our notation and revisit the definitions of group actions, equivariance and discuss the equivariance properties relevant to ViTs.

Notation. Let $\mathbf{H}^{(l)}$ denote the feature map at layer l (with $\mathbf{H}^{(0)}$ representing the input image). The c-th channel of this feature map is denoted by $\mathbf{H}_c^{(l)}$.

We represent the filters at layer l by $\Phi_m^{(l)}$, where $\Phi_m^{(l)}$ is the m-th filter. The n-th channel of this filter is denoted by $\Phi_{m,n}^{(l)}$.

Group Actions. Consider a group G acting on a set X via a function $\alpha : G \times X \to X$ satisfying the following properties:

1. For all $g, h \in G$ and $x \in X$,

$$\alpha(g, \alpha(h, x)) = \alpha(gh, x). \tag{1}$$

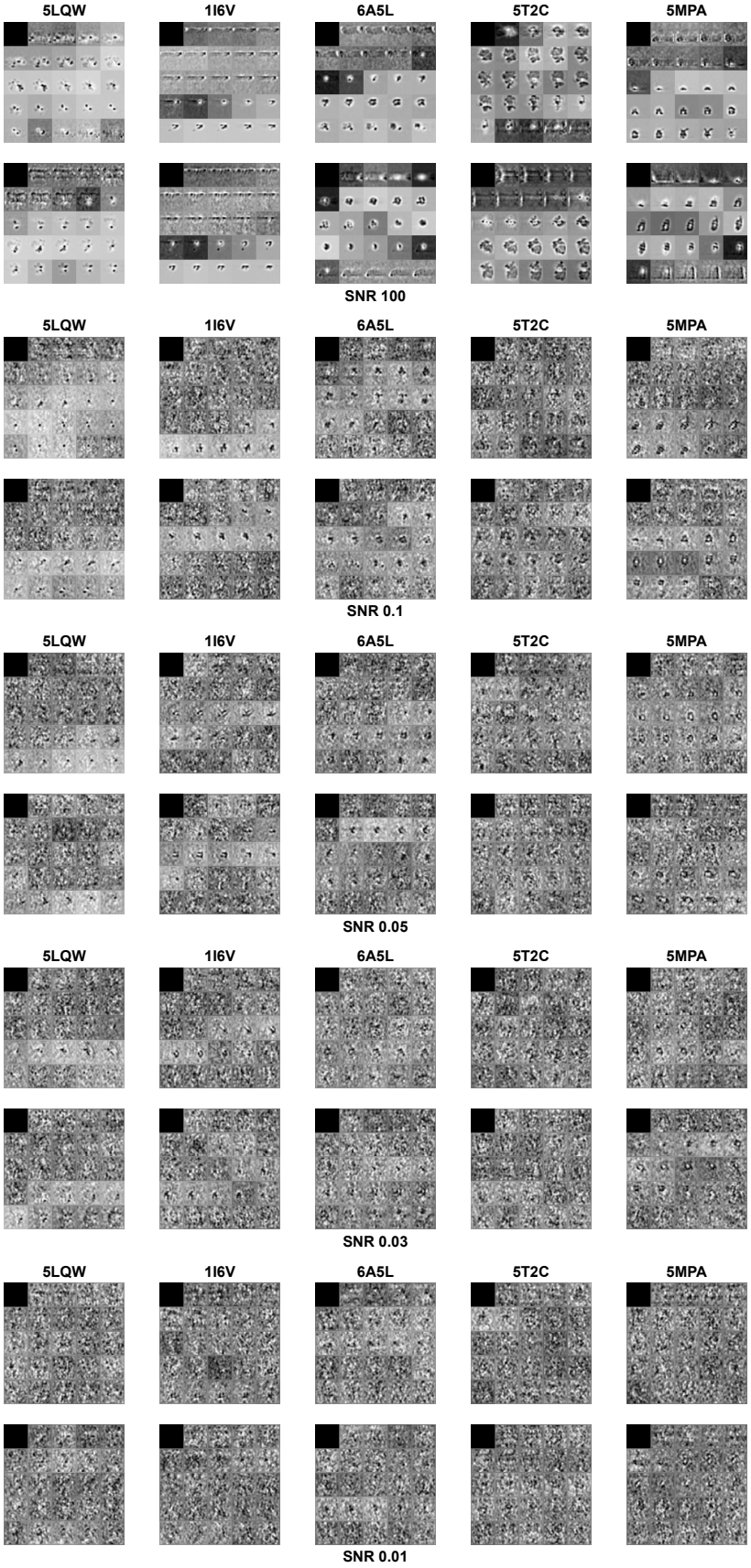

Figure 3: 2D slices representation of input subtomograms.

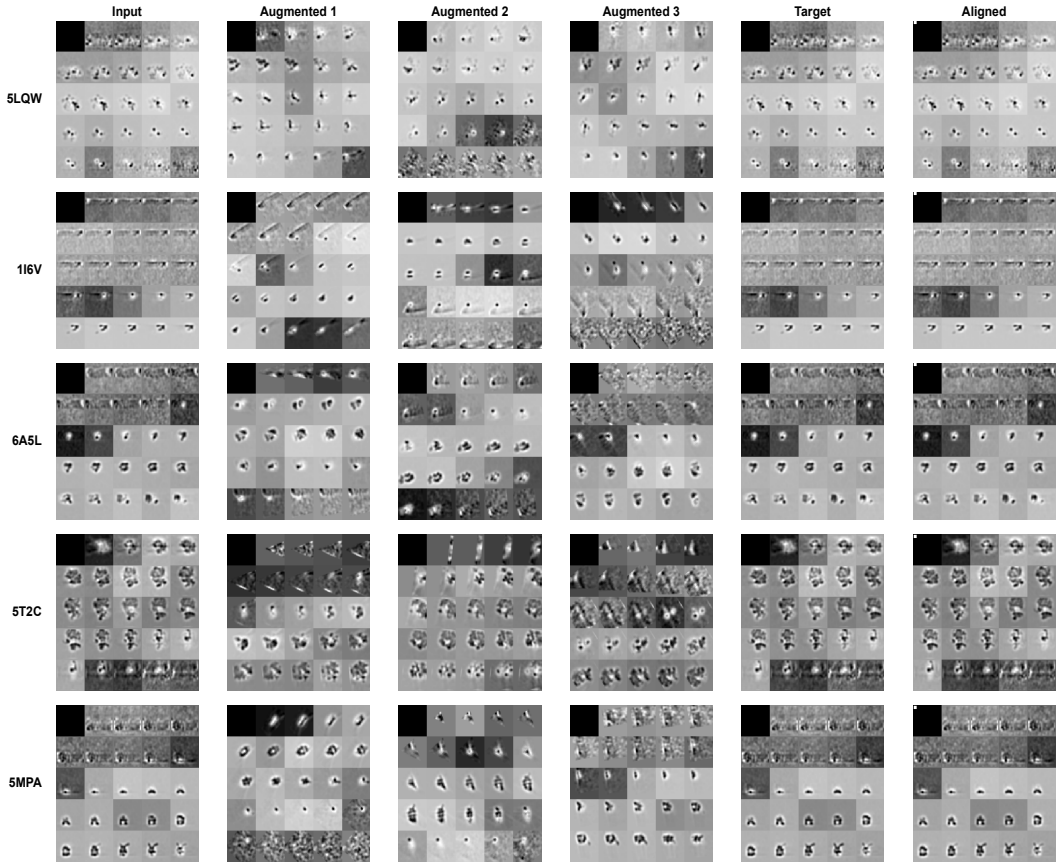

Figure 4: 2D slices representation of processed subtomograms.

2. For all $x \in X$,
$$\alpha(e, x) = x, \tag{2}$$
where e is the identity element of G.

Under this action, X is referred to as a (left) G-set.

Equivariance. Let X and Y be G-sets with group actions $\alpha : G \times X \to X$ and $\beta : G \times Y \to Y$, respectively. A function $f : X \to Y$ is called *equivariant* if, for all $x \in X$ and $g \in G$,
$$f(\alpha(g, x)) = \beta(g, f(x)). \tag{3}$$

If f is *invariant* under the action of G, meaning that β is the identity map, then
$$f(\alpha(g, x)) = f(x). \tag{4}$$

Equivariance of ViT module. The Vision Transformer (ViT) architecture consists of a patch embedding layer, positional encoding, transformer blocks, and MLP layers. As highlighted by **(author?)** [18], the patch embedding layer is neither shift- nor rotation-equivariant due to downsampling effects, which disrupt spatial consistency. Additionally, both absolute positional encoding [19] and relative positional encoding [20, 21] are not equivariant to shifts or rotations. While the normalization, global self-attention, and MLP layers are shift-equivariant, special design is required to achieve rotation equivariance [22, 23].

D Detailed Architecture

D.1 Feature Extraction and Patch Embedding

As shown in Fig. 5, our feature extraction and patch embedding pipeline implements the Polyshift module described in Section 4.2. The Polyphase Feature Extractor anchors input and target volumes using polyphase decomposition for shift-equivariance following Eq.(6). A group-equivariant CNN processes each polyphase component X(p,q,r) independently, extracting features Fi through multi-scale extraction with stride 2i while maintaining shift-equivariance. These spatial features then pass through a Shift-Equivariant Positional Encoder using 3D convolution with circular padding and 4×4×4 stride. Finally, a Feature Comparator analyzes relationships between input and target embeddings through parallel computations of differences, products, and concatenations, producing a combined embedding that captures both spatial and relational information.

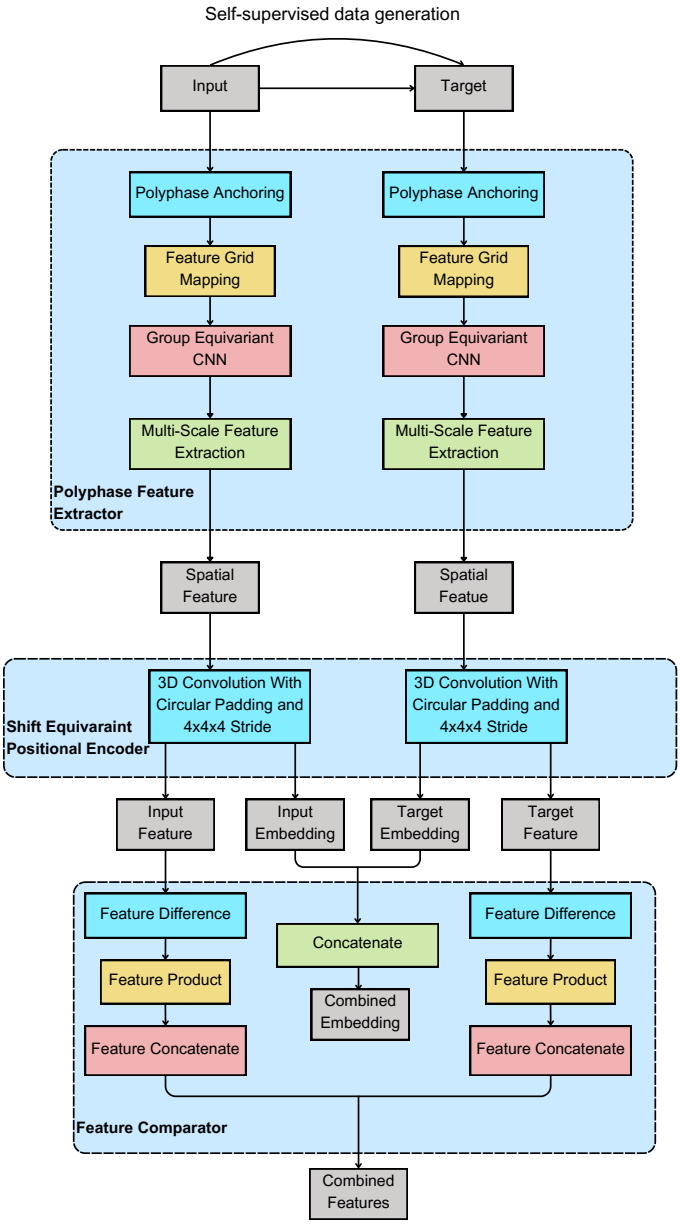

Figure 5: Feature Extraction and Patch Embedding.

D.2 Multi-Axis Rotation Positional Encoding

As illustrated in Fig. 6, the Multi-Axis Rotation positional Encoding (MARE) module enhances both rotational and translational equivariance by processing features along three axes (d,h,w) independently. Following Section 4.3, for each axis, MARE first computes rotation parameters using learnable matrices, then generates rotation matrices through skew-symmetric matrix transformation, and applies them to queries and keys while preserving original position vectors. The attention mechanism combines the rotated features with scaling factor to stabilize training, and the final attention output integrates information across all axes. This design ensures rotation equivariance through axis-specific transformations while preserving translation equivariance by maintaining relative spatial relationships in the attention computation.

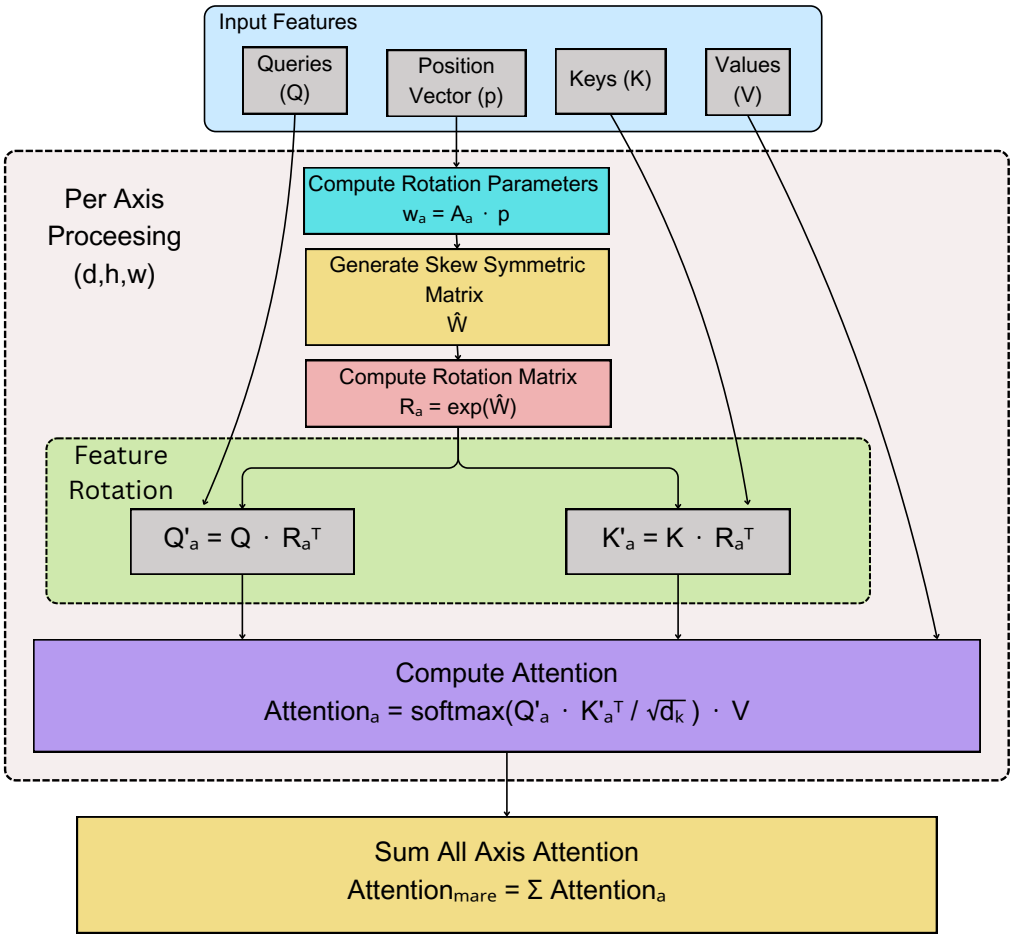

Figure 6: Workflow of MARE.

D.3 Transformer Block

As shown in Fig. 7, our BOE-ViT transformer block processes concatenated features and embeddings through three sequential stages. First, the MARE module applies axis-wise attention and rotational positional embedding to enhance spatial equivariance. Then, a standard feed-forward network with two linear layers processes the features. Finally, the transformer outputs aligned input, predicted parameters, processed features, and updated embeddings, effectively combining equivariance-enhanced attention with traditional vision transformer architecture.

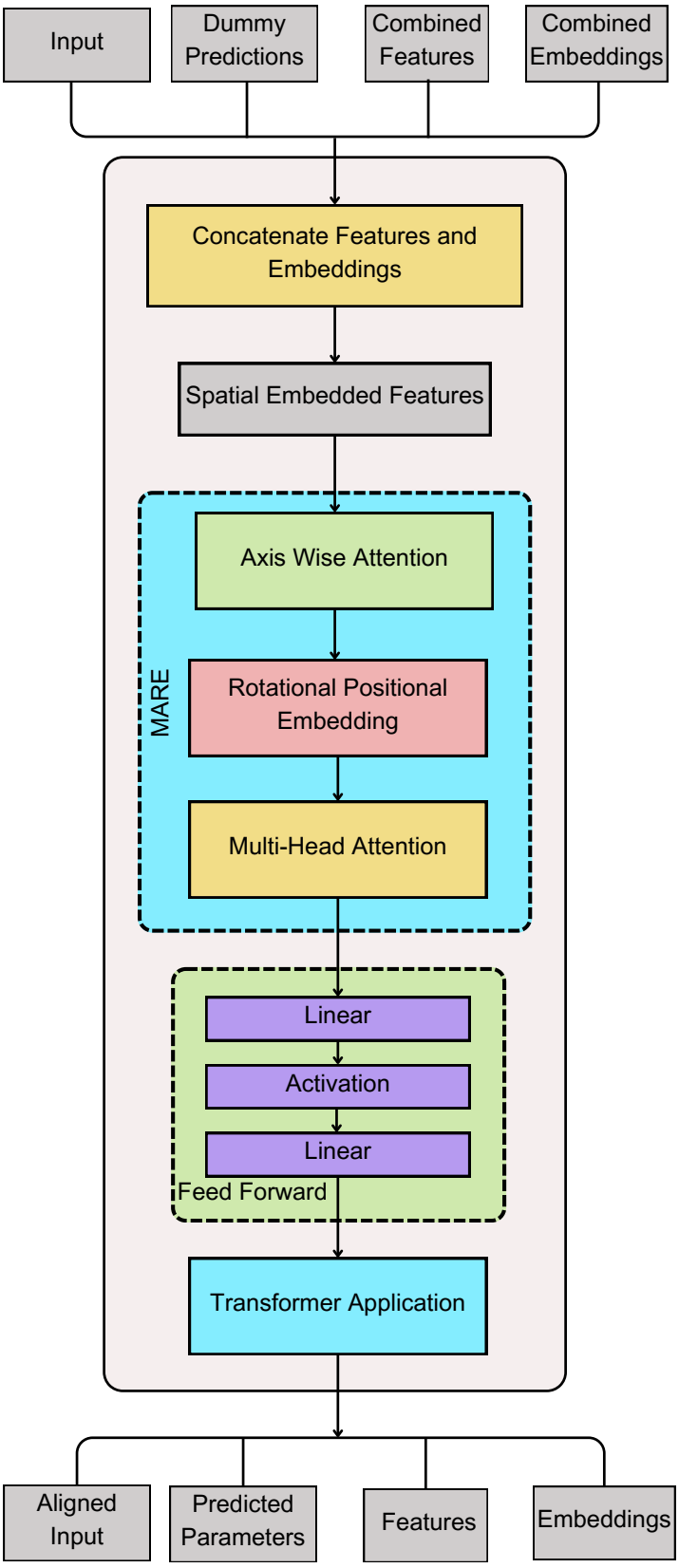

Figure 7: BOE-ViT Transformer Block.

E Mathematical Analysis of Equivariance Properties

E.1 Shift Equivariance Analysis of Polyshift Module

In this section, we provide detailed proofs for the equivariance properties of our Polyshift module. We first prove the fundamental equivariance property of the polyphase anchoring operator (Lemma 4.1), followed by an important corollary about the nature of the induced translations. We then establish the equivariance of the complete module including strided convolutions (Lemma 4.2).

Lemma E.1 (Polyphase Anchoring Equivariance)**.** *Let $\mathcal{P}$ be the polyphase anchoring operator and $\mathcal{T}_{\mathbf{g}}$ denote a translation operator that shifts $\mathbf{X}$ spatially by $\mathbf{g} = (g_D, g_H, g_W)$. Then, for multi-subtomograms $\mathbf{X} \in \mathbb{R}^{B\times C\times D\times H\times W}$, there exists a translation $\mathbf{g}' = (g'_D, g'_H, g'_W)$, corresponding to an integer multiple of the patch size $\mathbf{s} = (s_D, s_H, s_W)$, such that:*

$$\mathcal{P}(\mathcal{T}_{\mathbf{g}}\mathbf{X}) = \mathcal{T}_{\mathbf{g}'}(\mathcal{P}(\mathbf{X})) \tag{5}$$

This implies that polyphase anchoring is shift-equivariant up to a known shift $\mathbf{g}'$ dependent on the original shift $\mathbf{g}$ and the patch size $\mathbf{s}$.

Proof. By definition of polyphase decomposition, the input tensor $\mathbf{X}$ is divided into polyphase components:

$$\mathbf{X}_{(p,q,r)} = \{\mathbf{X}_{:,:,i\cdot s_D+p,j\cdot s_H+q,k\cdot s_W+r} | i, j, k \in \mathbb{Z}_{\geq 0}\} \tag{6}$$

For each component, the norm is computed:

$$N_{(p,q,r)} = \|\mathbf{X}_{(p,q,r)}\|_p \tag{7}$$

When applying translation $\mathcal{T}_{\mathbf{g}}$, we have:

$$\mathcal{P}(\mathcal{T}_{\mathbf{g}}\mathbf{X}) = \mathcal{T}_{\Delta\hat{k}|\mathcal{T}_{\mathbf{g}}\mathbf{X}} \cdot \mathcal{T}_{\mathbf{g}} \cdot \mathbf{X} \tag{8}$$

where $\mathcal{T}_{\Delta\hat{k}|\mathcal{T}_{\mathbf{g}}\mathbf{X}}$ represents the anchoring shift determined by:

$$(\hat{p}, \hat{q}, \hat{r}) = \arg\max_{(p,q,r)} N_{(p,q,r)_{\mathcal{T}_{\mathbf{g}}\mathbf{X}}} \tag{9}$$

Similarly for the original input:

$$\mathcal{P}(\mathbf{X}) = \mathcal{T}_{\Delta\hat{k}|\mathbf{X}} \cdot \mathbf{X} \tag{10}$$

Due to the circular padding in the polyphase decomposition, the relative ordering of norms $N_{(p,q,r)}$ either remains unchanged or undergoes a cyclic permutation under translation. Therefore, there exists a translation $\mathbf{g}' \in G$ such that:

$$\mathcal{T}_{\Delta\hat{k}|\mathcal{T}_{\mathbf{g}}\mathbf{X}} \cdot \mathcal{T}_{\mathbf{g}} \cdot \mathbf{X} = \mathcal{T}_{\mathbf{g}'} \cdot \mathcal{T}_{\Delta\hat{k}|\mathbf{X}} \cdot \mathbf{X} \tag{11}$$

$$\mathcal{P}(\mathcal{T}_{\mathbf{g}}\mathbf{X}) = \mathcal{T}_{\mathbf{g}'} \cdot \mathcal{P}(\mathbf{X}) \tag{12}$$

Corollary E.1 (Integer Stride Translation)**.** *The translation $\mathbf{g}'$ in Lemma E.1 shifts $\mathcal{P}(\mathbf{X})$ by integer multiples of the patch size $\mathbf{s} = (s_D, s_H, s_W)$. Specifically:*

$$g'_D = k_D s_D, \quad g'_H = k_H s_H, \quad g'_W = k_W s_W \tag{13}$$

for some integers $k_D, k_H, k_W \in \mathbb{Z}$.

Proof. According to Algorithm 1, the polyphase component with maximum norm determines the anchoring shift. Under translation $\mathcal{T}_{\mathbf{g}}$, this maximum norm component must belong to the same equivalence class modulo $\mathbf{s}$. Therefore,

$$\mathcal{P}(\mathbf{X})_{0::s_D,0::s_H,0::s_W} = \arg\max_{\mathcal{P}(\mathbf{X})_{i::s_D,j::s_H,k::s_W}} \|\mathcal{P}(\mathbf{X})_{i::s_D,j::s_H,k::s_W}\| \tag{14}$$

where $i < s_D, j < s_H, k < s_W$.

This periodic structure ensures that $\mathbf{g}'$ must be integer multiples of the patch size in each dimension.

Lemma E.2 (Strided Convolution Equivariance)**.** *Let $\mathcal{P}$ be the polyphase anchoring operator and $*_{\mathbf{s}}$ denote strided convolution with stride $\mathbf{s}$. For any translation $\mathcal{T}_{\mathbf{g}}$ and convolution kernel $\mathbf{h}$, when the stride sizes in polyphase anchoring ($\mathbf{s}_1$) and convolution ($\mathbf{s}_2$) are equal ($\mathbf{s}_1 = \mathbf{s}_2 = \mathbf{s}$), the following equivariance property holds:*

$$\mathcal{P}(\mathcal{T}_{\mathbf{g}}\mathbf{X}) *_{\mathbf{s}} \mathbf{h} = \mathcal{T}_{\mathbf{g}'}(\mathcal{P}(\mathbf{X}) *_{\mathbf{s}} \mathbf{h}) \tag{15}$$

where $\mathbf{g}'$ is as defined in Lemma E.1 and Corollary E.1.

Proof. The strided convolution can be decomposed as:

$$\mathbf{X} *_{\mathbf{s}} \mathbf{h} = D_{\mathbf{s}}(\mathbf{X} * \mathbf{h}) \tag{16}$$

where $D_{\mathbf{s}}$ represents downsampling with stride $\mathbf{s}$.

Then:

$$\begin{aligned}
\mathcal{P}(\mathcal{T}_{\mathbf{g}}\mathbf{X}) *_{\mathbf{s}} \mathbf{h} &= D_{\mathbf{s}}(\mathcal{P}(\mathcal{T}_{\mathbf{g}}\mathbf{X}) * \mathbf{h}) && (17)\\
&= D_{\mathbf{s}}(\mathcal{T}_{\mathbf{g}'}\mathcal{P}(\mathbf{X}) * \mathbf{h}) \quad \text{(by Lemma E.1)} && (18)\\
&= D_{\mathbf{s}}(\mathcal{T}_{\mathbf{g}'}(\mathcal{P}(\mathbf{X}) * \mathbf{h})) \quad \text{(convolution equivariance)} && (19)\\
&= \mathcal{T}_{\mathbf{g}'} D_{\mathbf{s}}(\mathcal{P}(\mathbf{X}) * \mathbf{h}) \quad (\text{since } \mathbf{g}' \text{ is multiple of } \mathbf{s}) && (20)\\
&= \mathcal{T}_{\mathbf{g}'}(\mathcal{P}(\mathbf{X}) *_{\mathbf{s}} \mathbf{h}) && (21)
\end{aligned}$$

The last equality holds because $\mathbf{g}'$ is an integer multiple of the stride $\mathbf{s}$ in each dimension (by Corollary E.1), allowing the downsampling and translation operations to commute.

E.2 Rotational Equivariance Analysis of MARE

Lemma E.3 (MARE Rotation Equivariance)**.** *Let $\mathcal{R}_{\boldsymbol{\phi}}$ denote a global rotation with rotation matrix $\mathbf{R}_{\boldsymbol{\phi}}$, and let $\mathbf{A}_a$ be the learnable parameter matrices for each axis $a \in \{d, h, w\}$. The MARE attention mechanism satisfies:*

$$\begin{aligned}
&\textit{MARE}(\mathbf{Q}\mathbf{R}_{\boldsymbol{\phi}}^{\top}, \mathbf{K}\mathbf{R}_{\boldsymbol{\phi}}^{\top}, \mathbf{V}, \mathbf{p}, \{\mathbf{A}_a\})\\
&\quad = \mathbf{R}_{\boldsymbol{\phi}} \cdot \textit{MARE}(\mathbf{Q}, \mathbf{K}, \mathbf{V}, \mathbf{p}, \{\mathbf{A}_a\})
\end{aligned} \tag{22}$$

Proof. For each axis $a \in \{d, h, w\}$, the rotation parameters are computed as:

$$\mathbf{w}_a = \mathbf{A}_a\mathbf{p} \tag{23}$$

When a global rotation $\mathbf{R}_{\boldsymbol{\phi}}$ is applied to the input, these parameters transform as:

$$\mathbf{w}'_a = \mathbf{A}_a(\mathbf{R}_{\boldsymbol{\phi}}\mathbf{p}) = \mathbf{R}_{\boldsymbol{\phi}}(\mathbf{A}_a\mathbf{p}) = \mathbf{R}_{\boldsymbol{\phi}}\mathbf{w}_a \tag{24}$$

The skew-symmetric matrices are constructed as:

$$\hat{\mathbf{W}}_a = \begin{bmatrix} 0 & -w_{a,z} & w_{a,y} \\ w_{a,z} & 0 & -w_{a,x} \\ -w_{a,y} & w_{a,x} & 0 \end{bmatrix} \tag{25}$$

The axis-specific rotation matrices are computed using the matrix exponential:

$$\mathbf{R}_a = \exp(\hat{\mathbf{W}}_a) \tag{26}$$

Under global rotation, these matrices transform as:

$$\mathbf{R}'_a = \mathbf{R}_{\boldsymbol{\phi}}\mathbf{R}_a\mathbf{R}_{\boldsymbol{\phi}}^{\top} \tag{27}$$

For the rotated input features:

$$\begin{aligned}
\mathbf{Q}'_a &= (\mathbf{Q}\mathbf{R}_{\boldsymbol{\phi}}^{\top})(\mathbf{R}'_a)^{\top} = \mathbf{Q}\mathbf{R}_{\boldsymbol{\phi}}^{\top}(\mathbf{R}_{\boldsymbol{\phi}}\mathbf{R}_a\mathbf{R}_{\boldsymbol{\phi}}^{\top})^{\top} && (28)\\
&= \mathbf{Q}\mathbf{R}_{\boldsymbol{\phi}}^{\top}\mathbf{R}_{\boldsymbol{\phi}}\mathbf{R}_a^{\top}\mathbf{R}_{\boldsymbol{\phi}}^{\top} = \mathbf{Q}\mathbf{R}_a^{\top}\mathbf{R}_{\boldsymbol{\phi}}^{\top} && (29)
\end{aligned}$$

Similarly for the keys:

$$\mathbf{K}'_a = \mathbf{K}\mathbf{R}_a^\top \mathbf{R}_\phi^\top \tag{30}$$

The attention logits for each axis become:

$$\mathbf{Q}'_a(\mathbf{K}'_a)^\top = \mathbf{Q}\mathbf{R}_a^\top \mathbf{R}_\phi^\top \mathbf{R}_\phi \mathbf{R}_a \mathbf{K}^\top \tag{31}$$

$$= \mathbf{Q}\mathbf{R}_a^\top \mathbf{R}_a \mathbf{K}^\top \tag{32}$$

$$= \mathbf{Q}\mathbf{K}^\top \tag{33}$$

The axis-specific attention output is:

$$\text{Attention}_a = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \tag{34}$$

The final MARE attention output transforms as:

$$\text{Attention}'_{\text{MARE}} = \sum_{a\in\{d,h,w\}} \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)(\mathbf{R}_\phi \mathbf{V}) \tag{35}$$

$$= \mathbf{R}_\phi \cdot \text{Attention}_{\text{MARE}} \tag{36}$$

This completes the proof of rotation equivariance, showing that a global rotation of the input results in a corresponding rotation of the attention output while preserving the learned axis-specific rotational patterns.

F Experiments Settings

F.1 Training Details

The model was implemented in PyTorch with CUDA for GPU acceleration and trained on a single NVIDIA V100 GPU using mixed-precision to efficiently process subtomograms of size $32 \times 32 \times 32$ from simulated and augmented SNR100 datasets.

Models were pre-trained for 100 epochs and fine-tuned for an additional 400 epochs on the concatenated training set. Testing was performed on low-SNR datasets (SNR: 0.1, 0.05, 0.03, 0.01) generated from the same complexes to simulate realistic experimental conditions, with each test dataset containing 5000 subtomograms.

The BOE-ViT architecture was initialized with 4 transformer blocks, 4 attention heads, a feed-forward hidden dimension of 256, and a transformer hidden dimension of 60. The Polyshift patch embedder used a 3D patch size of (4, 4, 4). Training was conducted with a batch size of 4 subtomograms, using the AdamW optimizer with an initial learning rate of $1 \cdot 10^{-5}$ and a weight decay of $2 \cdot 10^{-8}$.

F.2 Introduction of Baselines

Here, we provide a brief introduction to the state-of-the-art methods used as comparative baselines for cryo-ET alignment, as follows:

- **H-T align** [24]: A Fourier-based rotational alignment method designed to improve accuracy in low SNR and high tilt angle conditions.
- **F&A align** [25]: An efficient alignment algorithm leveraging spherical harmonics and Wiener-filtered corrections for reference-free subtomogram alignment.
- **Gum-Net** [15]: An unsupervised CNN-based model for 3D geometric correspondence, optimized for noisy Cryo-ET data with substantial error reduction and speedup. The three architectures of Gum-Net can be summarized as: Gum-Net MP uses max pooling for feature extraction, Gum-Net AP employs average pooling to optimize feature aggregation, and Gum-Net SC simplifies the matching module by computing only one correlation map.
- **Jim-Net** [17]: A multi-task CNN-based model that simultaneously clusters and aligns subtomograms using unsupervised pair-matching alignment.

F.3 Mathematical Definition of Metrics

Error Metrics To quantitatively evaluate subtomogram alignment accuracy, we compute both rotational and translational errors between the estimated and ground truth parameters. For rotation matrices $\mathbf{R}_{\text{estimated}}$ and $\mathbf{R}_{\text{true}}$, the rotational error e_r (in degrees) is defined as:

$$e_{\text{rot}} = \arccos\left(\frac{\text{tr}(\mathbf{R}_{\text{est}}^T \mathbf{R}_{\text{gt}}) - 1}{2}\right) \cdot \frac{180}{\pi} \tag{37}$$

where $\mathbf{R}_{\text{est}}, \mathbf{R}_{\text{gt}} \in \text{SO}(3)$ are rotation matrices, $\text{tr}(\cdot)$ denotes the matrix trace. This formulation yields rotation errors in degrees within the range [0°, 180°]. The translational error e_t (in voxels) is calculated as the Euclidean distance between estimated and true positions:

$$e_{\text{trans}} = \|\mathbf{t}_{\text{est}} - \mathbf{t}_{\text{gt}}\|_2 \tag{38}$$

Signal-to-Noise Ratio (SNR) The Signal-to-Noise Ratio quantifies image quality through the Pearson's correlation coefficient (c) between two optimally aligned subtomograms of identical structure:

$$\text{SNR} = \frac{c}{1-c}, \quad c = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{39}$$

where x_i and y_i represent corresponding voxel intensities in the two subtomograms, and $\bar{x}$, $\bar{y}$ their respective means. In cryo-ET, SNR typically ranges from 0.01 to 0.1 due to dose fractionation across tilt series, increased sample thickness at high tilts, and complex cellular backgrounds. This inherently low SNR poses significant challenges for accurate alignment, necessitating robust computational methods.

G Further Experiments

We conduct extensive experiments to validate our BOE-ViT framework. First, we evaluate our method on five challenging cryo-ET datasets across various SNR levels, demonstrating superior alignment accuracy compared to state-of-the-art approaches in appendix G.1. Second, we explore the sensitivity of BOE-ViT to key hyperparameters including patch size, batch size, attention heads, hidden dimension, and loss parameters, in appendix G.2, providing insights for optimal model configuration. In the Tables 1-9, each cell reports the mean and standard deviation of the rotation error (first term) and translation error (second term).

G.1 Performance Across Diverse Macromolecular Structures

We evaluated BOE-ViT on five representative macromolecular complexes under varying SNR conditions. As shown in Tables 1-5, BOE-ViT demonstrates superior performance and robustness compared to existing methods, including H-T align, F&A align, four variants of Gum-Net (Gum-Net MP, Gum-Net AP, and Gum-Net SC, Gum-Net) and Jimnet.

Method	SNR 0.1	SNR 0.05	SNR 0.03	SNR 0.01
H-T align	1.67±1.06, 6.31±5.01	2.09±0.87, 7.65±4.56	2.22±0.74, 8.10±4.43	2.40±0.57, 10.93±4.97
F&A align	1.71±1.08, 6.63±4.96	2.06±0.90, 7.76±4.67	2.23±0.74, 8.48±4.62	2.37±0.56, 10.94±4.98
Gum-Net MP	1.38±0.75, 5.25±3.53	1.50±0.76, 5.70±3.65	1.59±0.76, 6.08±3.54	1.66±0.77, 7.06±3.39
Gum-Net AP	1.25±0.76, 4.75±3.37	1.39±0.76, 5.35±3.49	1.53±0.75, 5.81±3.46	1.65±0.77, 7.02±3.35
Gum-Net SC	1.26±0.77, 4.83±3.58	1.42±0.77, 5.43±3.62	1.53±0.76, 5.73±3.47	1.68±0.76, 6.96±3.52
Gum-Net	0.75±0.77, 2.99±3.17	0.87±0.76, 3.49±3.31	1.05±0.71, 3.96±2.77	1.42±0.78, 5.66±3.53
Jimnet	0.78±0.71, 3.15±3.13	1.03±0.74, 4.14±3.58	1.18±0.73, 4.68±3.34	1.60±0.75, 6.55±3.43
BOE-ViT	**0.33±0.16, 2.41±0.84**	**0.34±0.15, 2.31±0.81**	**0.34±0.16, 2.25±0.80**	**0.33±0.15, 2.26±0.78**

Table 1: RNA polymerase-rifampicin complex (PDB ID: 1I6V) subtomogram alignment accuracy.

Method	SNR 0.1	SNR 0.05	SNR 0.03	SNR 0.01
H-T align	0.94±0.95, 3.75±4.03	1.74±1.02, 6.31±4.60	2.21±0.75, 8.69±4.56	2.37±0.55, 11.58±5.02
F&A align	1.06±1.06, 4.31±4.41	1.85±0.99, 6.99±4.85	2.18±0.79, 8.69±4.55	2.39±0.58, 11.31±4.83
Gum-Net MP	1.13±0.74, 4.27±3.09	1.30±0.75, 4.80±3.11	1.45±0.76, 5.45±3.09	1.66±0.77, 6.99±3.28
Gum-Net AP	0.98±0.67, 3.72±2.74	1.20±0.72, 4.45±2.85	1.40±0.74, 5.29±3.02	1.64±0.77, 6.97±3.33
Gum-Net SC	1.07±0.73, 4.02±3.03	1.26±0.76, 4.56±3.07	1.47±0.77, 5.48±3.14	1.65±0.76, 6.89±3.33
Gum-Net	0.46±0.54, 1.80±1.90	0.71±0.63, 2.55±2.12	1.12±0.73, 3.93±2.45	1.45±0.76, 5.94±3.32
Jimnet	0.39±0.52, **1.67±2.01**	0.64±0.60, 2.42±2.33	0.99±0.72, 3.71±2.89	1.58±0.76, 6.69±3.38
BOE-ViT	**0.33±0.15**, 2.30±0.80	**0.34±0.16**, **2.27±0.81**	**0.35±0.15**, **2.27±0.75**	**0.34±0.15**, **2.24±0.78**

Table 2: RNA polymerase II elongation complex (PDB ID: 6A5L) subtomogram alignment accuracy.

Method	SNR 0.1	SNR 0.05	SNR 0.03	SNR 0.01
H-T align	0.61±0.87, 2.64±3.55	1.62±1.14, 6.08±4.92	2.15±0.88, 8.49±4.72	2.38±0.56, 11.36±5.13
F&A align	0.64±0.97, 2.96±3.99	1.68±1.16, 6.32±4.91	2.12±0.89, 8.39±4.79	2.35±0.59, 11.20±5.00
Gum-Net MP	1.02±0.70, 4.07±3.16	1.25±0.78, 4.89±3.30	1.38±0.75, 5.41±3.31	1.65±0.78, 6.79±3.08
Gum-Net AP	0.87±0.65, 3.56±2.78	1.12±0.74, 4.45±3.00	1.29±0.74, 5.07±3.09	1.60±0.81, 6.69±3.11
Gum-Net SC	0.96±0.71, 3.83±3.13	1.22±0.79, 4.76±3.28	1.38±0.76, 5.28±3.33	1.65±0.78, 6.82±3.20
Gum-Net	0.47±0.57, 1.94±2.26	0.68±0.64, 2.61±2.25	0.93±0.68, 3.62±2.32	1.38±0.78, 5.65±3.31
Jimnet	**0.30±0.47**, **1.42±2.01**	0.51±0.58, 2.30±2.36	0.74±0.62, 3.13±2.63	1.50±0.76, 6.30±3.13
BOE-ViT	0.33±0.15, 2.35±0.83	**0.34±0.16**, **2.27±0.79**	**0.34±0.15**, **2.24±0.77**	**0.34±0.16**, **2.21±0.77**

Table 3: Spliceosome (PDB ID: 5LQW) subtomogram alignment accuracy.

Method	SNR 0.1	SNR 0.05	SNR 0.03	SNR 0.01
H-T align	1.16±1.04, 4.43±4.21	2.13±0.84, 8.79±4.77	2.34±0.61, 10.59±4.98	2.36±0.59, 11.56±4.91
F&A align	1.54±1.12, 6.39±5.19	2.17±0.80, 9.39±5.09	2.35±0.58, 10.81±4.93	2.40±0.55, 11.81±4.89
Gum-Net MP	1.58±0.83, 5.51±3.07	1.71±0.80, 6.28±3.16	1.70±0.80, 6.72±3.13	1.70±0.78, 8.27±3.58
Gum-Net AP	1.30±0.79, 4.71±2.76	1.58±0.80, 5.94±3.05	1.63±0.81, 6.70±3.20	1.68±0.78, 8.14±3.51
Gum-Net SC	1.41±0.79, 4.90±2.94	1.63±0.79, 5.98±3.11	1.66±0.80, 6.54±3.15	1.71±0.77, 8.35±3.64
Gum-Net	0.73±0.81, 2.70±2.87	1.19±0.84, 4.23±3.01	1.43±0.79, 5.67±2.96	1.76±0.75, 10.46±5.10
Jimnet	0.49±0.70, **1.99±2.43**	1.09±0.86, 4.14±3.30	1.33±0.83, 5.19±3.28	1.65±0.78, 7.60±3.62
BOE-ViT	**0.34±0.15**, 2.28±0.81	**0.34±0.16**, **2.24±0.77**	**0.34±0.15**, **2.27±0.80**	**0.34±0.15**, **2.30±0.79**

Table 4: Ribosome (PDB ID: 5T2C) subtomogram alignment accuracy.

Method	SNR 0.1	SNR 0.05	SNR 0.03	SNR 0.01
H-T align	1.72±0.99, 6.65±4.55	2.08±0.88, 7.47±4.46	2.16±0.81, 8.42±4.47	2.38±0.58, 11.22±5.03
F&A align	1.73±1.01, 6.69±4.71	1.97±0.94, 7.26±4.67	2.24±0.79, 8.59±4.69	2.39±0.56, 11.33±4.88
Gum-Net MP	1.40±0.80, 5.52±3.60	1.43±0.78, 5.63±3.44	1.53±0.76, 6.12±3.45	1.68±0.77, 7.30±3.33
Gum-Net AP	1.05±0.69, 4.28±2.92	1.19±0.73, 4.78±3.04	1.37±0.73, 5.64±3.22	1.66±0.77, 7.10±3.27
Gum-Net SC	1.12±0.76, 4.47±3.30	1.24±0.78, 4.92±3.40	1.38±0.77, 5.71±3.43	1.66±0.78, 7.16±3.35
Gum-Net	0.68±0.64, 2.61±2.46	0.89±0.72, 3.13±2.68	1.12±0.72, 4.25±2.73	1.46±0.78, 6.22±3.38
Jimnet	0.57±0.56, 2.37±2.20	0.72±0.64, 3.10±2.71	0.88±0.66, 3.90±2.94	1.55±0.78, 6.75±3.47
BOE-ViT	**0.33±0.15**, **2.24±0.82**	**0.33±0.15**, **2.24±0.80**	**0.33±0.15**, **2.20±0.80**	**0.33±0.16**, **2.21±0.77**

Table 5: Capped proteasome (PDB ID: 5MPA) subtomogram alignment accuracy.

G.2 Parameter Exploration in BOE-ViT

Impact of Patch Size As demonstrated in Table 6, experimental analysis with varying patch sizes (4, 8, and 16) reveals that smaller patches consistently achieve superior performance, particularly in translation accuracy. While rotation errors exhibit stability across different patch sizes (approximately 0.33-0.34), translation errors increase monotonically with larger patches (from 2.38 to 4.14), indicating that finer-grained spatial partitioning is essential for preserving local geometric information.

Patch Size	SNR 0.1	SNR 0.05	SNR 0.03	SNR 0.01
4	0.335±0.153, 2.380±0.840	0.337±0.155, 2.334±0.818	0.339±0.152, 2.301±0.801	0.337±0.153, 2.320±0.803
8	0.336±0.154, 3.662±1.739	0.339±0.156, 3.580±1.728	0.340±0.153, 3.499±1.700	0.338±0.154, 3.516±1.687
16	0.336±0.153, 4.144±1.890	0.339±0.156, 4.162±1.889	0.341±0.153, 4.182±1.903	0.337±0.154, 4.125±1.907

Table 6: Impact of patch size.

Impact of Batch Size The quantitative results in Table 7 demonstrate that our model maintains consistent performance across batch sizes 4-16, with rotation errors stabilizing around 0.33-0.34. However, increasing the batch size to 32 results in significant degradation of translation accuracy, particularly at higher SNR levels. The optimal performance achieved with a batch size of 4 suggests that smaller batches facilitate more precise optimization of alignment parameters.

Batch Size	SNR 0.1	SNR 0.05	SNR 0.03	SNR 0.01
4	0.335±0.153, 2.380±0.840	0.337±0.155, 2.334±0.818	0.339±0.152, 2.301±0.801	0.337±0.153, 2.320±0.803
8	0.335±0.154, 2.568±0.943	0.338±0.154, 2.467±0.893	0.339±0.152, 2.412±0.860	0.338±0.154, 2.358±0.835
16	0.335±0.153, 2.598±0.957	0.338±0.155, 2.471±0.892	0.340±0.153, 2.399±0.875	0.338±0.154, 2.300±0.813
32	0.335±0.153, 3.125±1.223	0.338±0.155, 2.970±1.112	0.340±0.152, 2.799±1.062	0.338±0.154, 2.528±0.940

Table 7: Impact of batch size.

Impact of Loss Parameters We employ a weighted loss function $L(\hat{\theta}, \theta) = \alpha \cdot \text{MSE}(\theta, \hat{\theta}) + \beta \cdot \text{MSE}(t, \hat{t})$, where α and β balance the contributions of rotation (θ) and translation (t) errors respectively. As shown in Table 8, the configuration with α=1, β=2 achieves optimal performance, indicating that emphasizing translation error correction while maintaining rotational accuracy yields the most favorable results. This finding aligns with our observation that translation estimation presents greater challenges in the alignment task.

Loss Parameters	SNR 0.1	SNR 0.05	SNR 0.03	SNR 0.01
α=1, β=1	0.334±0.153, 3.101±1.208	0.337±0.155, 2.971±1.134	0.339±0.152, 2.865±1.081	0.337±0.154, 2.619±1.020
α=1, β=2	0.334±0.153, 2.452±0.938	0.337±0.155, 2.381±0.888	0.339±0.152, 2.308±0.864	0.338±0.154, 2.232±0.812
α=1, β=3	0.334±0.153, 2.415±0.850	0.338±0.155, 2.365±0.818	0.339±0.153, 2.363±0.806	0.336±0.154, 2.387±0.825
α=2, β=3	0.335±0.153, 2.616±1.005	0.337±0.155, 2.510±0.953	0.339±0.152, 2.410±0.916	0.337±0.153, 2.256±0.851
α=3, β=2	0.335±0.153, 4.055±1.533	0.338±0.155, 3.916±1.481	0.339±0.152, 3.747±1.417	0.338±0.153, 3.480±1.355

Table 8: Impact of loss parameters.

Impact of Attention Heads Systematic evaluation of different attention head configurations in Table 9 identifies an optimal architecture with 4 heads, achieving minimal rotation (0.334-0.339) and translation errors (2.24-2.32). This empirical finding suggests that 4 attention heads provide an optimal balance between model capacity and computational efficiency, effectively capturing spatial relationships without introducing redundant complexity.

Attention Heads	SNR 0.1	SNR 0.05	SNR 0.03	SNR 0.01
2	0.335±0.153, 2.545±0.925	0.337±0.155, 2.451±0.892	0.340±0.152, 2.384±0.860	0.337±0.153, 2.307±0.838
4	0.334±0.153, 2.318±0.821	0.338±0.155, 2.265±0.795	0.339±0.152, 2.245±0.782	0.337±0.154, 2.245±0.776
5	0.333±0.153, 2.341±0.847	0.337±0.155, 2.304±0.822	0.339±0.152, 2.267±0.797	0.337±0.154, 2.283±0.820
10	0.334±0.153, 2.363±0.829	0.337±0.155, 2.353±0.825	0.339±0.152, 2.365±0.816	0.337±0.154, 2.399±0.838

Table 9: Impact of attention heads.

References

[1] Fernández-Busnadiego R. Wagner J, Schaffer M. Cryo-electron tomography-the cell biology that came in from the cold. febs lett. 2017 sep;591(17):2520-2533. doi: 10.1002/1873-3468.12757. epub 2017 aug 2. pmid: 28726246. *FEBS Letters*, 591(17):2520–2533, 2017. 3

[2] Werner Kühlbrandt. The resolution revolution. *Science*, 343(6178):1443–1444, 2014. 3

[3] Susanne Pfeffer and Julia Mahamid. Unravelling molecular complexity in structural cell biology. *Current Opinion in Structural Biology*, 52:111–118, Oct 2018. 3

[4] Terje Dokland. Back to the basics: The fundamentals of cryo-electron microscopy. *Microscopy and Microanalysis*, 15:1538–, Jul 2009. 3

[5] Daniel Castaño-Díez and Giulia Zanetti. In situ structure determination by subtomogram averaging. *Current Opinion in Structural Biology*, 2019. 3

[6] Ruobing Han, Xuelong Wan, Zhenzhu Wang, Ying Hao, Jing Zhang, Yiming Chen, Xiaoxia Gao, Zhirong Liu, Fei Ren, Fei Sun, and Fan Zhang. Autom: A novel automatic platform for electron tomography reconstruction. *Journal of Structural Biology*, 199(3):196–208, September 2017. Epub 2017 Jul 26. 3

[7] David N. Mastronarde and Susannah R. Held. Automated tilt series alignment and tomographic reconstruction in imod. *Journal of Structural Biology*, 197(2):102–113, 2017. Electron Tomography. 3

[8] Muyuan Chen, James M Bell, Xiaodong Shi, Stella Y Sun, Zhao Wang, and Steven J Ludtke. A complete data processing workflow for cryo-et and subtomogram averaging. *Nature methods*, 16(11):1161–1168, 2019. 3

[9] Janina Böhning and Tanmay A. M. Bharat. Towards high-throughput in situ structural biology using electron cryotomography. *Progress in Biophysics and Molecular Biology*, 160:97–103, Mar 2021. 3

[10] Daniel Castaño-Díez and Giulia Zanetti. In situ structure determination by subtomogram averaging. *Current Opinion in Structural Biology*, 58:68–75, Oct 2019. 3

[11] Hannah Hyun-Sook Kim, Mostofa Rafid Uddin, Min Xu, and Yi-Wei Chang. Computational methods toward unbiased pattern mining and structure determination in cryo-electron tomography data. *Journal of molecular biology*, 435(9):168068, 2023. 3

[12] Martin Turk and Wolfgang Baumeister. The promise and the challenges of cryo-electron tomography. *FEBS letters*, 594(20):3243–3261, 2020. 3

[13] Julio A Kovacs and Willy Wriggers. Fast rotational matching. *Acta Crystallographica Section D: Biological Crystallography*, 58(8):1282–1286, 2002. 3

[14] Radostin Danev, Shuji Kanamaru, Michael Marko, and Kuniaki Nagayama. Zernike phase contrast cryo-electron tomography. *Journal of structural biology*, 171(2):174–181, 2010. 3

[15] Xiangrui Zeng and Min Xu. Gum-net: Unsupervised geometric matching for fast and accurate 3d subtomogram alignment and averaging. In *CVPR*, 2020. 3, 12

[16] Min Xu, Martin Beck, and Frank Alber. High-throughput subtomogram alignment and classification by fourier space constrained fast volumetric matching. *Journal of structural biology*, 178(2):152–164, 2012. 3

[17] Xiangrui Zeng, Gregory Howe, and Min Xu. End-to-end robust joint unsupervised image alignment and clustering. In *ICCV*, 2021. 3, 12

[18] Peijian Ding, Davit Soselia, Thomas Armstrong, Jiahao Su, and Furong Huang. Reviving shift equivariance in vision transformers. In *The Second Workshop on Spurious Correlations, Invariance and Stability (SCIS), ICML*, 2023. 6

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 6

[20] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 6

[21] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution, 2022. 6

[22] Serge Assaad, Carlton Downey, Rami Al-Rfou, Nigamaa Nayakanti, and Ben Sapp. Vn-transformer: Rotation-equivariant attention for vector neurons. *arXiv preprint arXiv:2206.04176*, 2022. 6

[23] Soumyabrata Kundu and Risi Kondor. Steerable transformers. *arXiv preprint arXiv:2405.15932*, 2024. 6

[24] Min Xu, Martin Beck, and Frank Alber. High-throughput subtomogram alignment and classification by fourier space constrained fast volumetric matching. *Journal of structural biology*, 2012. 12

[25] Yuxiang Chen, Stefan Pfeffer, Thomas Hrabe, Jan Michael Schuller, and Friedrich Förster. Fast and accurate reference-free alignment of subtomograms. *Journal of structural biology*, 182(3):235–245, 2013. 12