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

Runmin Jiang[1], Jackson Daggett[1], Shriya Pingulkar[4], Yizhou Zhao[1],
Priyanshu Dhingra[5], Daniel Brown[1], Qifeng Wu[1], Xiangrui Zeng[2,3], Xingjian Li[1,*], Min Xu[1,*]

[1] Carnegie Mellon University [2] Harvard Medical School [3] Massachusetts General Hospital
[4] K. J. Somaiya College of Engineering, [5]Rajiv Gandhi Institute of Petroleum Technology

Abstract

Subtomogram alignment is a critical task in cryo-electron tomography (cryo-ET) analysis, essential for achieving high-resolution reconstructions of macromolecular complexes. However, learning effective positional representations remains challenging due to limited labels and high noise levels inherent in cryo-ET data. In this work, we address this challenge by proposing a self-supervised learning approach that leverages intrinsic geometric transformations as implicit supervisory signals, enabling robust representation learning despite data scarcity. We introduce BOE-ViT, the first Vision Transformer (ViT) framework for 3D subtomogram alignment. Recognizing that traditional ViTs lack equivariance and are therefore suboptimal for orientation estimation, we enhance the model with two innovative modules that introduce equivariance include 1) the Polyshift module for improved shift estimation and 2) Multi-Axis Rotation Encoding (MARE) for enhanced rotation estimation. Experimental results demonstrate that BOE-ViT significantly outperforms state-of-the-art methods. Notably, at SNR 0.01 dataset, our approach achieves a 77.3% reduction in rotation estimation error and a 62.5% reduction in translation estimation error, effectively overcoming the challenges in cryo-ET subtomogram alignment.

1. Introduction

In recent years, cryo-ET has emerged as a groundbreaking *in situ* 3D structural biology imaging technique, enabling the study of macromolecular complexes—nanoscale machines that drive cellular processes—within single cells [31]. With the rapid growth in the volume of cryo-ET data available for structural and biological research, deep learning has become a valuable tool for automated cryo-ET image analysis [8, 39].

In typical cryo-ET image analysis pipelines, subtomogram alignment is a crucial task [26, 51]. [1] This importance arises from the fact that 3D cryo-ET images often suffer from low quality due to their nature of low signal-to-noise ratio (SNR) and the missing wedge issue [53]. To achieve high-resolution biological structures, it is essential to first align these identical particles in different orientations, and subsequently denoise them through averaging. Precise alignment also allows for spatial localization of specific structures within tomograms.

The goal of subtomogram alignment is to determine the six parameters of a 3D rigid transformation—three rotational and three translational. Despite its straightforward formulation, this task presents a challenging geometric matching problem, even more complex than 3D deformable medical image registration. The difficulties raise from three main factors: (1) macromolecular structures exhibit random orientations and displacements, creating high variability; (2) subtomograms have low signal-to-noise ratios (around 0.01 to 0.1), with complex cytoplasmic backgrounds and low electron doses [13]; and (3) structural heterogeneity, as different macromolecular complexes can appear vastly different, unlike the relatively consistent structures in medical images.

Traditional methods rely on exhaustive searches over six-dimensional parameter spaces (e.g., Euler angles and translations)[2, 28, 56, 57]. These geometric approaches are relatively slow due to the large search space, although several improvements have been proposed. Subsequent work explores data-driven approaches using deep CNNs [58, 59], which significantly improve both speed and accuracy. However, the error in parameter prediction is still very high, especially on low SNR data.

To better address the challenging problem of aligning highly complex and noisy 3D subtomograms, we are inspired to explore Vision Transformers (ViTs), which

*Corresponding Authors

[1] A subtomogram is defined as a subvolume with a macromolecular complex.

have demonstrated great potential in pushing performance boundaries on large-scale datasets in general computer vision domains [17, 23, 25, 36, 41, 49]. Intuitively, the attention mechanism of ViTs can effectively capture global dependencies, making it suited for understanding the entire particle structure and thereby benefiting global alignment. However, applying ViTs to the subtomogram alignment problem remains non-trivial due to their own limitations. First, ViTs require large amounts of data for effective training, but the available cryo-ET datasets are significantly smaller compared to natural image datasets. Second, the ViT architecture lacks inherent sensitivity to the rotational and translational state of input images due to the lack of inherent equivariance. Therefore, improving the capacity for accurate orientation estimation is a unique challenge not usually considered in standard computer vision tasks.

In summary, addressing this realistic problem requires the consideration of two crucial aspects:

- *How to learn representations under conditions of limited positional label and high noise in 3D cryo-ET subtomograms?*
- *How to enhance orientation estimation in vision models, which is crucial for alignment tasks?*

To address these challenges, we propose a framework termed Boosting Orientation Estimation with Equivariance in Vision Transformer (BOE-ViT) for 3D subtomogram alignment, as illustrated in Figure.1. We design a self-supervised learning task that leverages intrinsic geometric transformations as implicit supervisory signals, allowing the model to learn 3D structural relationships without explicit annotations. To improve shift-equivariance in ViT, we introduce the *Polyshift* module, which anchors on the polyphase component with the highest norm to stabilize multi-scale feature extraction across spatial transformations. Shift-equivariant positional encoding is achieved using circularly padded depth-wise convolution. To optimize the attention mechanism for rotation estimation, we propose Multi-Axis Rotation positional Encoding (*MARE*), which enhances spatial sensitivity by allowing position vectors to rotate independently along each axis. The effectiveness of BOE-ViT is validated through comprehensive experiments on various datasets, demonstrating its superiority over state-of-the-art approaches.

The main contributions of this paper are:

- **New Task:** We tackle a realistic yet underexplored challenge of multi-subtomogram alignment in conditions of label scarcity and high noise using self-supervised learning.
- **Novel Methodology:** We propose BOE-ViT as the **first** Vision Transformer approach for 3D subtomogram alignment. This framework introduces two key innovations for boosting orientation estimation with **equivariance**: (1) *Polyshift* for **shift** estimation, and (2) *MARE* for **rotation** estimation.
- **Promising Results:** Experiments across various subtomogram types and SNR conditions demonstrate the superiority of BOE-ViT over state-of-the-art methods.

2. Related Work

2.1. Subtomogram and Image Alignment

Subtomogram alignment is one of the key parts of cryo-electron tomography (cryo-ET) analysis [5, 15, 42]. While sharing similar principles with general image alignment [4, 60–64], subtomogram alignment faces unique challenges due to the high noise level, missing wedge effects, and complex 3D structural variations in cellular environments [58].

Traditional approaches frequently employ exhaustive search methods [3, 52], which are computationally expensive and time-consuming for processing large-scale datasets. More efficient methods have emerged employing fast rotational matching algorithms [47, 57], drawing inspiration from modern computer vision techniques for geometric matching.

In alignment tasks, more recent methods can be roughly classified into the following:

Supervised: Such methods require datasets with ground truth key points [32–34, 38] or transformation parameter annotated [43, 45]. Therefore, preparing datasets for these methods would demand massive amount of human labor, which is considered as the limitation of this paradigm.

Self-supervised: Seo et al. [45], Truong et al. [50] demonstrate that image alignment networks can be trained solely on synthetic data with random transformations. Jiang et al. [24] used self-supervised anatomical consistency for aligning medical volumes, while Shi et al. [46] introduced a self-supervised cross-modality alignment method through iterative homography estimation.

Unsupervised: Zeng and Xu [58] adopts an unsupervised paradigm to train a network that aligns 3D images in a pairwise fashion. Followed by Zeng et al. [59], a network can handle both 2D and 3D images and align multiple images is proposed.

2.2. Group Equivariant Neural Networks

Previous studies [1] have demonstrated that modern CNNs [20, 29] lack shift equivariance due to their reliance on pooling layers. To improve shift-equivariance, G-CNNs [12] extended CNN's translation equivariance to various symmetry groups. Following this, researchers expanded the theory to continuous groups and more complex symmetries [27, 55].

Vision Transformers [17, 23, 25, 36, 41, 49] have emerged as powerful alternatives to CNNs [21, 30] in computer vision by capturing global dependencies more effectively through self-attention mechanisms. While SETrans-

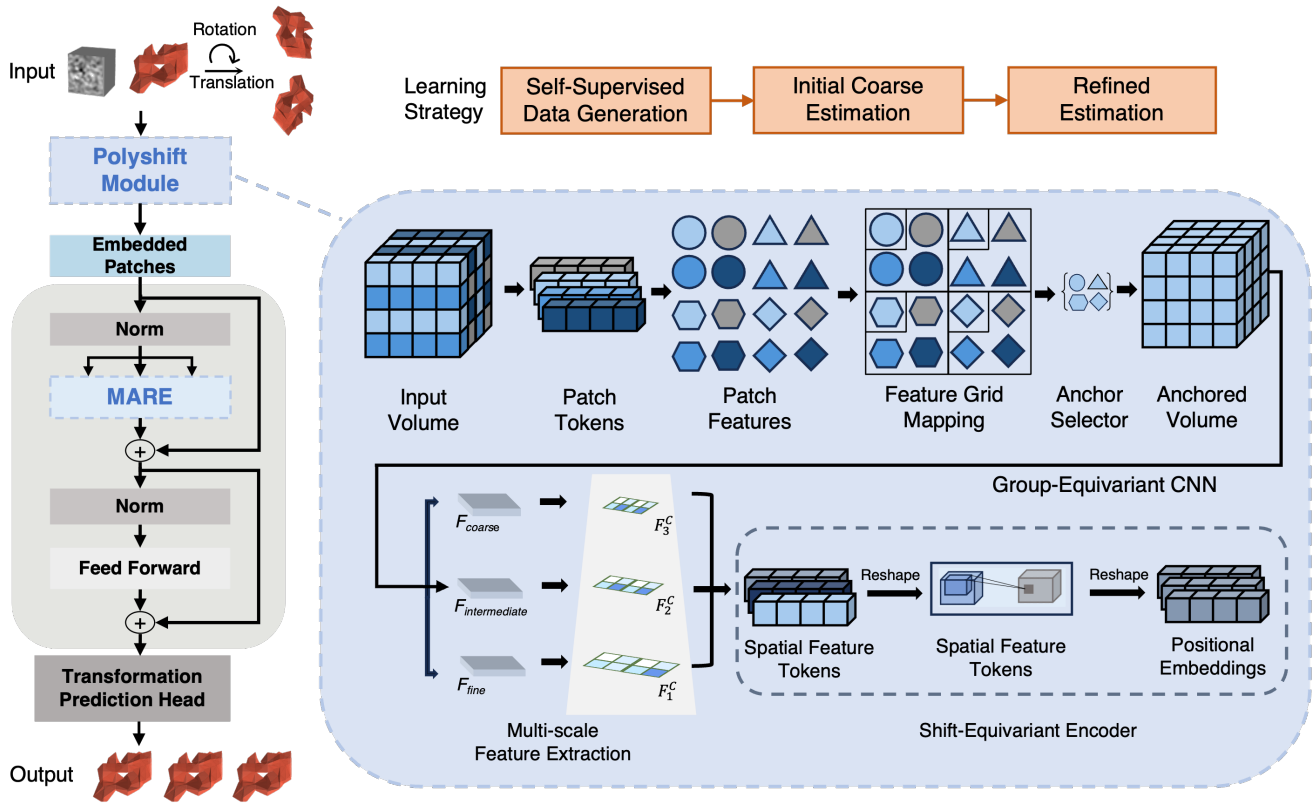

Figure 1. Overview of the BOE-ViT architecture. Subtomograms are processed through the PolyShift module for shift-equivariant feature extraction and multi-scale refinement. Positional encoding and axial attention capture spatial relationships, while the MARE module introduces rotation-equivariant encodings using axis-specific rotational matrices. The model predicts transformation parameters (θ for rotation, t for translation) to align subtomograms.

formers [7, 18, 19, 35, 54] and tensor field networks [48] have primarily been developed for point cloud data, they are less common in image-based applications due to challenges in making patch embedding and positional encoding equivariant [14]. Recently, novel adaptive designs have been introduced to enable shift-equivariance in ViTs [10, 44].

3. Method

3.1. BOE-ViT Overview

The paper introduces Boosting Orientation estimation with Equivariance in Vision Transformer (**BOE-ViT**), a framework for self-supervised subtomogram alignment, as illustrated in Figure 1. Challenges include the scarcity of labeled data, high noise levels, and the need to adapt the framework for multi-subtomogram alignment in cryo-ET. Additionally, ViT lack equivariance, leading to significant errors in translation and rotation estimation. For preliminaries on equivariance and an analysis of ViT's equivariance properties, see Supp. Mat. Sec. C.

To address these, we formulate a self-supervised task that leverages geometric transformations within 3D subtomograms to learn structural representations without annotations. Given multi-subtomograms $\mathbf{X} \in \mathbb{R}^{B\times C\times D\times H\times W}$, we generate transformed subtomograms $\mathbf{X}'$ by applying transformation $\mathcal{T}_\theta$:

$$\mathbf{X}' = \mathcal{T}_\theta(\mathbf{X}), \tag{1}$$

The model receives $\mathbf{X}'$ as input and predicts the transformation parameters $\hat{\boldsymbol{\theta}}$ to map it back to the original subtomogram $\mathbf{X}$. This task forces the model to learn transformation-equivariant features that capture the underlying 3D geometric structures.

Our framework leverages a Vision Transformer architecture, which lacks intrinsic equivariance to rotations and translations. To overcome this, we introduce the Polyshift module (see Sec.3.2) and MARE (see Sec.3.3), which enhance shift and rotation estimation, respectively, through equivariant design. For more implementation details, please refer to Supp. Mat. Sec. D.

The Polyshift module extracts features, while MARE encodes them, enabling the attention mechanism to capture spatial relationships. We then predict transformation parameters $\hat{\boldsymbol{\theta}}$ using a prediction head.

The learning process employs a two-phase strategy:

- **Initial Phase (Coarse Estimation):** The model learns to predict large transformations, including rotations up to $\pm 180^\circ$ and translations up to 40% of the volume size. This exposure to extensive geometric variations helps the model develop a fundamental understanding of 3D structural relationships.
- **Refinement Phase (Fine Estimation):** The model fine-tunes its predictions within a constrained transformation space, with rotations up to $\pm 30^\circ$ and translations up to 30% of the volume size. This phase enhances the model's ability to capture subtle geometric relationships while retaining the robust features learned earlier.

3.2. Polyshift Module

To enhance shift-equivariance in Vision Transformers, we propose the Polyshift algorithm, which integrates polyphase anchoring, multi-scale feature extraction, and a shift-equivariant encoder. This approach ensures that feature representations are consistent under spatial transformations, making it well-suited for tasks such as orientation estimation in three-dimensional data.

Polyphase Anchoring Algorithm Inspired by adaptive polyphase sampling [6, 14], the polyphase anchoring algorithm aligns polyphase components of the input tensor to ensure shift-equivariance. By anchoring on the polyphase component with the highest norm, we stabilize feature extraction across spatial transformations.

Given multi-subtomograms $\mathbf{X}$, we define the patch size $\mathbf{s} = (s_D, s_H, s_W)$. The input is divided into polyphase components $\{\mathbf{X}_{(p,q,r)}\}$, where each component is given by:

$$\mathbf{X}_{(p,q,r)} = \{\mathbf{X}_{:,:,i\cdot s_D+p,j\cdot s_H+q,k\cdot s_W+r} \mid i,j,k \in \mathbb{Z}_{\geq 0}\}. \tag{2}$$

Polyphase Anchoring Process Algorithm 1 summarizes the polyphase anchoring process.

The shift-equivariance of the polyphase anchoring algorithm is formalized as follows:

Lemma 3.1. *Let $\mathcal{T}_\mathbf{g}$ denote a translation operator that shifts $\mathbf{X}$ spatially by $\mathbf{g} = (g_D, g_H, g_W)$. Then, there exists a translation $\mathbf{g}' = (g'_D, g'_H, g'_W)$, corresponding to an integer multiple of the patch size $\mathbf{s}$, such that:*

$$\mathcal{P}\left(\mathcal{T}_\mathbf{g}\mathbf{X}\right) = \mathcal{T}_{\mathbf{g}'}\left(\mathcal{P}(\mathbf{X})\right), \tag{3}$$

where $\mathcal{P}$ denotes the polyphase anchoring operator. 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 $\mathbf{s}$.

Algorithm 1 Polyphase Anchoring Algorithm

Input: Input tensor $\mathbf{X} \in \mathbb{R}^{B\times C\times D\times H\times W}$, patch size $\mathbf{s} = (s_D, s_H, s_W)$
Output: Shifted tensor $\hat{\mathbf{X}}$
Decompose $\mathbf{X}$ into polyphase components $\{\mathbf{X}_{(p,q,r)}\}$.
Compute norms for each polyphase component: $N_{(p,q,r)} = \|\mathbf{X}_{(p,q,r)}\|_p$.
Identify the maximum polyphase component:

$$(\hat{p}, \hat{q}, \hat{r}) = \arg\max_{(p,q,r)} N_{(p,q,r)}.$$

Circularly shift $\mathbf{X}$ to align the maximum polyphase component:

$$\hat{\mathbf{X}} = \text{Shift}\left(\mathbf{X}, -\hat{p}, -\hat{q}, -\hat{r}\right).$$

Return: $\hat{\mathbf{X}}$

Proof. When $\mathbf{X}$ is translated by $\mathcal{T}_\mathbf{g}$, the indices of the polyphase components are shifted accordingly. Due to the periodicity introduced by circular padding, the relative ordering of the norms $N_{(p,q,r)}$ remains unchanged or undergoes a cyclic permutation. The anchoring shift $\Delta_{\hat{k}}$ adjusts to align with the new maximum polyphase component, resulting in a shift of $\mathcal{P}(\mathbf{X})$ by an amount $\mathbf{g}'$, which is an integer multiple of $\mathbf{s}$.

For shift-equivariant feature extraction, the Group Convolution CNN independently extracts features from each polyphase component $\mathbf{X}_{(p,q,r)}$, defined as:

$$\mathbf{F}_{(p,q,r)} = \phi_{\text{poly}}(\mathbf{X}_{(p,q,r)}). \tag{4}$$

To achieve multi-scale extraction, a strided convolution with stride 2^i downscales $\mathbf{X}$ at each scale i:

$$\mathbf{F}_i = \phi_{\text{scale},i}(\mathbf{X}), \tag{5}$$

preserving alignment and shift-equivariance across scales.

By applying polyphase anchoring prior to each convolution, we ensure shift-equivariance across scales. In particular, for strided convolution with stride $\mathbf{s}$, the following property holds:

Lemma 3.2. *Let $\mathcal{P}$ be the polyphase anchoring operator and $*_\mathbf{s}$ denote strided convolution with stride $\mathbf{s}$. For any translation $\mathcal{T}_\mathbf{g}$, we have:*

$$\mathcal{P}\left(\mathcal{T}_\mathbf{g}\mathbf{X}\right) *_\mathbf{s} \mathbf{h} = \mathcal{T}_{\mathbf{g}'}\left(\mathcal{P}(\mathbf{X}) *_\mathbf{s} \mathbf{h}\right), \tag{6}$$

where $\mathbf{h}$ is the convolution kernel, and $\mathbf{g}'$ is as defined in Lemma 3.1.

Proof. From Lemma 3.1, we have $\mathcal{P}\left(\mathcal{T}_\mathbf{g}\mathbf{X}\right) = \mathcal{T}_{\mathbf{g}'}\left(\mathcal{P}(\mathbf{X})\right)$. Since strided convolution samples the input at positions aligned with the stride $\mathbf{s}$, and $\mathbf{g}'$ is an integer multiple of $\mathbf{s}$, the operation $*_\mathbf{s}\mathbf{h}$ applied to $\mathcal{T}_{\mathbf{g}'}\left(\mathcal{P}(\mathbf{X})\right)$ results in a shifted version of the output $\mathcal{P}(\mathbf{X}) *_\mathbf{s} \mathbf{h}$.

Shift-Equivariant Positional Embedding Positional embeddings are critical in models like ViTs for encoding spatial information. However, traditional absolute and relative positional embeddings [16, 36, 37] disrupt shift-equivariance. To address this, we utilize a shift-equivariant positional encoder that leverages circularly padded depthwise convolution, inspired by [11], to maintain shift-equivariance across positional encoding.

Given an input tensor $\mathbf{F} \in \mathbb{R}^{B\times C\times D\times H\times W}$, which is the output from the feature extractor, we generate positional embeddings by first constructing normalized spatial coordinate grids $\mathbf{C} \in \mathbb{R}^{3\times D\times H\times W}$, where each channel corresponds to depth (z), height (y), and width (x) coordinates, respectively. Specifically, each spatial dimension is linearly spaced in the range $[-1, 1]$ to normalize coordinates. This grid is then processed through a series of circularly-padded depthwise convolutions to yield the positional embedding $\mathbf{E} \in \mathbb{R}^{B\times C\times D\times H\times W}$. This is in contrast to conventional positional embeddings, which lack shift-equivariance due to fixed absolute or relative positions.

For more implementation details, please refer to Supp. Mat. Sec. D.1. For shift equivariance Analysis, see Supp. Mat. Sec. E.1.

3.3. Multi-Axis Rotation Positional Encoding (MARE)

Inspired by rotary position encodings [22, 40], we introduce Multi-Axis Rotation positional Encoding (*MARE*), designed to enhance ViTs for precise 3D rotation estimation. MARE utilizes multi-axis rotation decomposition, allowing for independent encoding of rotational relationships along the depth, height, and width axes.

In MARE, we apply axis-specific rotations to the feature embeddings (queries and keys) based on the positional information. For each axis $a \in \{d, h, w\}$, we perform the following steps:

Rotation Parameter Computation Compute the rotation parameter vector $\mathbf{w}_a$ as a learnable linear transformation of the position vector $\mathbf{p}$:

$$\mathbf{w}_a = \mathbf{A}_a\mathbf{p}, \tag{7}$$

where $\mathbf{A}_a \in \mathbb{R}^{3\times 3}$ is a learnable parameter matrix for axis a.

Skew-Symmetric Matrix Construction Construct the skew-symmetric matrix $\hat{\mathbf{W}}_a$ from $\mathbf{w}_a$:

$$\hat{\mathbf{W}}_a = \begin{bmatrix} 0 & -w_{a,z} & w_{a,y} \\ w_{a,z} & 0 & -w_{a,x} \\ -w_{a,y} & w_{a,x} & 0 \end{bmatrix}, \tag{8}$$

where $w_{a,x}, w_{a,y}, w_{a,z}$ are the components of $\mathbf{w}_a$.

Exponential Map for Rotation Compute the rotation matrix $\mathbf{R}_a$ using the matrix exponential of $\hat{\mathbf{W}}_a$:

$$\mathbf{R}_a = \exp(\hat{\mathbf{W}}_a). \tag{9}$$

This can be approximated using Rodrigues' rotation formula.

Feature Rotation Apply the rotation matrix $\mathbf{R}_a$ to the feature embeddings (queries and keys):

$$\mathbf{Q}'_a = \mathbf{Q}\mathbf{R}_a^\top, \quad \mathbf{K}'_a = \mathbf{K}\mathbf{R}_a^\top. \tag{10}$$

Note that the rotation is applied to the features, not the positional encodings. The positions $\mathbf{p}$ remain unchanged.

Algorithm 2 provides the full MARE attention mechanism.

Algorithm 2 Multi-Axis Rotation Positional Encoding (MARE) Attention

Procedure `MAREAttention`$(\mathbf{Q}, \mathbf{K}, \mathbf{V}, \mathbf{p}, \{\mathbf{A}_d, \mathbf{A}_h, \mathbf{A}_w\})$
Input: Queries $\mathbf{Q}$, Keys $\mathbf{K}$, Values $\mathbf{V}$; position vector $\mathbf{p}$; learnable parameter matrices $\{\mathbf{A}_d, \mathbf{A}_h, \mathbf{A}_w\}$.
for each axis $a \in \{d, h, w\}$ **do**
Compute rotation parameter vector: $\mathbf{w}_a = \mathbf{A}_a\mathbf{p}$
Compute skew-symmetric matrix $\hat{\mathbf{W}}_a$ from $\mathbf{w}_a$
Compute rotation matrix: $\mathbf{R}_a = \exp(\hat{\mathbf{W}}_a)$
Rotate queries and keys:

$$\mathbf{Q}'_a = \mathbf{Q}\mathbf{R}_a^\top, \quad \mathbf{K}'_a = \mathbf{K}\mathbf{R}_a^\top$$

Compute attention output for axis a:

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

end for
Return: Combined attention output:

$$\text{Attention}_{\text{MARE}} = \sum_{a\in\{d,h,w\}} \text{Attention}_a$$

For more implementation details, please refer to Supp. Mat. Sec. D.2. For rotational equivariance Analysis, see Supp. Mat. Sec. E.2.

4. Experiments

4.1. Experimental Settings

Datasets. We selected five asymmetric macromolecular complexes—spliceosome (5LQW), RNA polymerase-rifampicin complex (1I6V), RNA polymerase II elongation complex (6A5L), ribosome (5T2C), and capped proteasome

Method	SNR 0.1	SNR 0.05	SNR 0.03	SNR 0.01
H-T align	1.22±1.07, 4.76±4.56	1.93±0.98, 7.26±4.77	2.22±0.77, 8.86±4.72	2.38±0.57, 11.33±5.02
F&A align	1.34±1.13, 5.39±4.90	1.95±0.98, 7.54±4.94	2.22±0.77, 8.99±4.81	2.38±0.57, 11.32±4.92
Gum-Net MP	1.30±0.79, 4.93±3.36	1.44±0.79, 5.46±3.38	1.53±0.78, 5.96±3.34	1.67±0.77, 7.28±3.38
Gum-Net AP	1.09±0.73, 4.20±2.96	1.30±0.77, 5.00±3.15	1.45±0.77, 5.70±3.25	1.65±0.78, 7.18±3.35
Gum-Net SC	1.16±0.77, 4.41±3.23	1.36±0.79, 5.13±3.34	1.48±0.78, 5.75±3.34	1.67±0.77, 7.24±3.46
Gum-Net	0.62±0.69, 2.41±2.61	0.87±0.74, 3.20±2.78	1.13±0.75, 4.29±2.75	1.50±0.78, 6.78±4.22
Jim-Net	0.51±0.62, **2.12±2.47**	0.80±0.73, 3.20±3.02	1.02±0.75, 4.12±3.12	1.58±0.77, 6.78±3.44
BOE-ViT	**0.33±0.15**, 2.58±0.93	**0.34±0.15**, **2.45±0.87**	**0.34±0.15**, **2.50±0.89**	**0.34±0.15**, **2.54±0.91**

Table 1. Subtomogram alignment accuracy across various datasets with specified SNR levels. Each cell reports the mean and standard deviation of the rotation error (first term) and translation error (second term). We highlighted the BOE-ViT results which surpass state-of-the-art alignment methods.

PDB ID	Method	SNR 0.1	SNR 0.05	SNR 0.03	SNR 0.01
1I6V	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	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**
6A5L	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	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**
5LQW	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	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**
5T2C	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	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**
5MPA	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	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 2. Subtomogram alignment accuracy across various protein datasets with different SNR levels. Each cell reports the mean and standard deviation of the rotation error (first term) and translation error (second term).

(5MPA)—to ensure unique alignment ground truth. For each, we generated five simulated grayscale datasets with varying noise levels, following [58]. Training used one high-SNR (100) dataset, while testing employed four realistic low-SNR conditions (0.1–0.01). Results and subtomogram visualizations (Supp. Mat. Sec. B) are reported below.

Implementation. BOE-ViT is implemented in PyTorch with CUDA acceleration. Training proceeds in two phases: (1) pre-training with large perturbations to build broad correction capabilities, followed by (2) fine-tuning to adapt the model to subtle misalignments encountered in practical applications. All experiments were conducted on NVIDIA V100 GPUs.

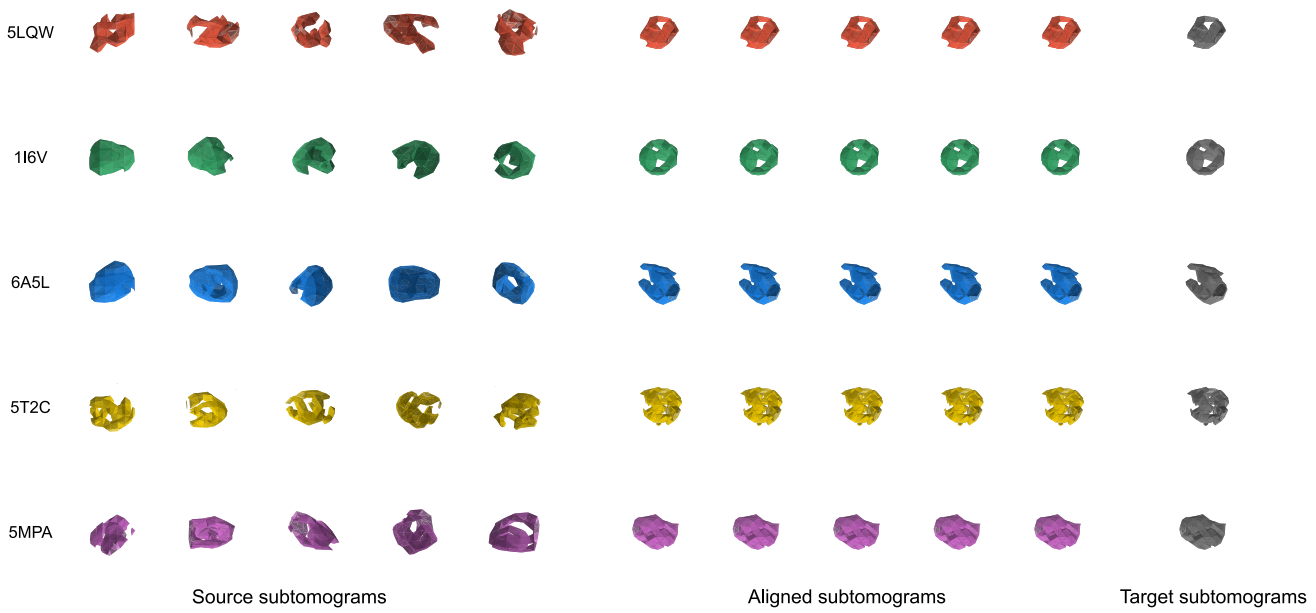

Figure 2. Comparison of source, aligned, and target subtomograms for five macromolecular complexes (5LQW, 1I6V, 6A5L, 5T2C, and 5MPA). The aligned subtomograms demonstrate the effectiveness of BOE-ViT in predicting accurate transformations.

Figure 3. Iterative refinement of subtomogram alignment during BOE-ViT training, starting from the input subtomograms (Iteration 0) and progressively aligning them closer to the true structure.

Baselines. We compare BOE-ViT with two traditional methods (H-T align [57] and F&A align [9]), and two state-of-the-art CNN-based methods, Gum-Net [58] and Jim-Net [59]. A brief introduction to experiment settings including training details, baseline methods and metrics is provided in the Supp. Mat. Sec. F.

4.2. Experimental Results

Comparison with SOTA methods. Results in Table 1 show that BOE-ViT significantly outperforms both traditional methods (H-T Align, F&A Align) and state-of-the-art CNN-based approaches, reducing rotation error by 77.3% and translation error by 62.5% at SNR 0.01. Notably, BOE-ViT exhibits exceptional robustness to noise, with its performance advantage becoming more pronounced as noise levels increase. While our method shows greater improvement in rotation estimation than translation prediction—likely due to the specific architecture choices of Polyshift and MARE—it consistently delivers superior overall performance across all noise conditions.

Diverse Macromolecular Performance. As shown in Table 2, we validated our approach on structurally diverse macromolecules, including spliceosome, RNA polymerase, ribosome (5T2C), and capped proteasome. The experimental results demonstrate that BOE-ViT, trained on different protein classes, effectively generalizes to diverse subtomograms. Further details are provided in Supp. Mat. Sec. G.1.

Ablation Study. The ablation study in Table 3 confirms the necessity of the attention module, Polyshift, and MARE for ViT-based alignment. While vanilla ViTs exhibit significantly higher translational errors than CNN-based methods across attention types, they show marginal rotational error improvements in low-SNR groups, with gains still substantially smaller than BOE-ViT's advantage over CNNs.

Regarding the ViT architecture, we observe that self-attention is more compatible with our Polyshift and MARE modules than cross-attention. Polyshift enhances alignment accuracy through shift-equivariance, demonstrating superior performance in self-attention by preserving spatial relationships throughout the network. MARE efficiently reduces both rotation and translation errors by decomposing

Model	Method	SNR 0.1	SNR 0.05	SNR 0.03	SNR 0.01
CNN-based	Gum-Net	0.62±0.69, 2.41±2.61	0.87±0.74, 3.20±2.78	1.13±0.75, 4.29±2.75	1.50±0.78, 6.78±4.22
	Jimnet	0.51±0.62, 2.12±2.47	0.80±0.73, 3.20±3.02	1.02±0.75, 4.12±3.12	1.58±0.77, 6.78±3.44
Cross-attention	CrossViT	0.66±0.30, 6.14±1.78	0.66±0.30, 6.14±1.76	0.65±0.30, 6.15±1.76	0.66±0.30, 6.20±1.78
	w/o Polyshift	0.33±0.15, 6.14±1.79	0.33±0.16, 6.14±1.77	0.33±0.15, 6.16±1.77	0.33±0.15, 6.20±1.78
	w/o MARE	0.35±0.15, 2.51±0.94	0.35±0.16, 2.52±0.94	0.35±0.15, 2.56±0.95	0.35±0.15, 2.55±0.95
	BOE-CrossViT	0.33±0.15, 2.79±1.10	0.33±0.16, 2.96±1.24	0.33±0.15, 3.10±1.36	0.33±0.16, 3.63±1.78
Self-attention	SelfViT	0.60±0.32, 7.13±1.83	0.62±0.31, 6.75±1.78	0.61±0.32, 6.59±1.83	0.61±0.31, 6.40±1.85
	w/o Polyshift	0.33±0.15, 6.14±1.80	0.33±0.16, 6.14±1.77	0.33±0.15, 6.16±1.77	0.33±0.15, 6.20±1.78
	w/o MARE	0.34±0.15, 2.63±0.87	0.34±0.15, 2.62±0.87	0.35±0.15, 2.62±0.88	0.35±0.15, 2.62±0.91
	BOE-ViT	**0.33±0.15**, **2.38±0.84**	**0.33±0.14**, **2.33±0.81**	**0.33±0.15**, **2.30±0.80**	**0.33±0.15**, **2.32±0.80**

Table 3. Ablation Study of attention, Polyshift, MARE. Each cell reports the mean and standard deviation of the rotation error (first term) and translation error (second term).

complex 3D transformations while delivering unexpected improvements in translation precision. Together, these components significantly improve alignment accuracy.

What's more, we do parameter exploration including patch size, batch size, loss parameters and attention heads, more details can be found in Supp. Mat. Sec. G.2.

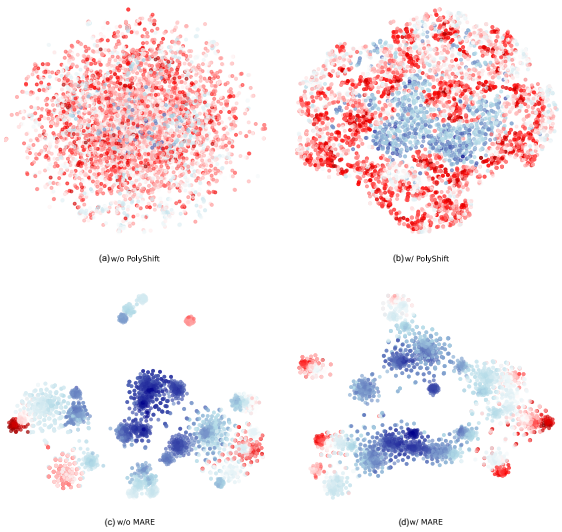

Figure 4. t-SNE visualizations of features with and without Polyshift and MARE. Polyshift enhances shift-equivariant feature extraction, resulting in more distinct clusters (b), while MARE improves rotational equivariance, yielding tighter and more spatially coherent groupings (d).

Alignment Visualization. Figure 2 compares source, aligned, and target subtomograms across five macromolecular complexes, demonstrating BOE-ViT's alignment accuracy. Figure 3 illustrates the iterative alignment process, showing how BOE-ViT progressively refines the structure from an initial average to closely match the true structure. These visualizations underscore BOE-ViT's robustness in accurately capturing complex spatial transformations. More visualization can be found in Supp. Mat. Sec. B.

t-SNE Visualization. The t-SNE visualizations (Figure 4) reveal Polyshift and MARE's impact on BOE-ViT feature organization. Without Polyshift (Figure 4a), SNR 0.01 dataset features appear dispersed, while with Polyshift (Figure 4b), features form distinct, well-separated clusters due to shift-equivariant, multi-scale extraction. Similarly, features without MARE (Figure 4c) show loose groupings, whereas with MARE (Figure 4d), they form tighter, more coherent clusters.

5. Conclusion

In this work, we address the challenging and underexplored task of 3D subtomogram alignment in cryo-ET under conditions of limited labels and high noise. We propose BOE-ViT, the first Vision Transformer framework for this task, which incorporates two innovative modules, Polyshift and MARE, to enhance shift and rotation equivariance. Through a self-supervised learning approach, our method achieves robust representation learning and significantly outperforms state-of-the-art methods, demonstrating its effectiveness in addressing key challenges in cryo-ET analysis.

Broader Impacts and Limitation. We believe that this work not only lays the foundation for applying powerful vision models to specific tasks in structural biology but also provides insights into designing equivariant Vision Transformers (ViTs) for general computer vision tasks. However, the Polyshift module increases memory usage and extends training times to over two days on a single V100 GPU.

6. Acknowledgement

We thank Mostofa Rafid Uddin for meaningful discussions. This work was supported in part by U.S. NSF grant IIS-2211597.

References

[1] Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *Journal of Machine Learning Research*, 20(184):1–25, 2019. 2

[2] Alberto Bartesaghi, P Sprechmann, J Liu, G Randall, G Sapiro, and Sriram Subramaniam. Classification and 3d averaging with missing wedge correction in biological electron tomography. *Journal of structural biology*, 162(3):436–450, 2008. 1

[3] Raphael André Bauer, Kristian Rother, Peter Moor, Knut Reinert, and Thomas Steinke. Fast structural alignment of biomolecules using a hash table, n-grams and string descriptors. *Algorithms and Molecular Sciences*, 2009. 2

[4] Lisa G Brown. A survey of image registration techniques. *ACM Computing Surveys*, 1992. 2

[5] Daniel Castaño-Díez and Giulia Zanetti. In situ structure determination by subtomogram averaging. *Current Opinion in Structural Biology*, 2019. 2

[6] Anadi Chaman and Ivan Dokmanic. Truly shift-invariant convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3773–3783, 2021. 4

[7] Evangelos Chatzipantazis, Stefanos Pertigkiozoglou, Edgar Dobriban, and Kostas Daniilidis. Se (3)-equivariant attention networks for shape reconstruction in function space. *arXiv preprint arXiv:2204.02394*, 2022. 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. 1

[9] 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. 7

[10] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, and Chunhua Shen. Conditional positional encodings for vision transformers. In *International Conference on Learning Representations*, 2021. 3

[11] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, and Chunhua Shen. Conditional positional encodings for vision transformers, 2023. 5

[12] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *ICML*, 2016. 2

[13] Radostin Danev, Shuji Kanamaru, Michael Marko, and Kuniaki Nagayama. Zernike phase contrast cryo-electron tomography. *Journal of structural biology*, 171(2):174–181, 2010. 1

[14] Peijian Ding, Davit Soselia, Thomas Armstrong, Jiahao Su, and Furong Huang. Reviving shift equivariance in vision transformers. *arXiv preprint arXiv:2306.07470*, 2023. 3, 4

[15] Terje Dokland. Back to the basics: The fundamentals of cryo-electron microscopy. *Microscopy and Microanalysis*, 2009. 2

[16] 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. 5

[17] 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 *ICLR*, 2021. 2

[18] Fabian B Fuchs, Daniel E Worrall, Volker Fischer, and Max Welling. Se(3)-transformers: 3d roto-translation equivariant attention networks. *NeurIPS*, 2020. 3

[19] Chongkai Gao, Zhengrong Xue, Shuying Deng, Tianhai Liang, Siqi Yang, Lin Shao, and Huazhe Xu. Riemann: Near real-time se (3)-equivariant robot manipulation without point cloud segmentation. *arXiv preprint arXiv:2403.19460*, 2024. 3

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. 2

[22] Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *European Conference on Computer Vision*, pages 289–305. Springer, 2025. 5

[23] Khawar Islam. Recent advances in vision transformer: A survey and outlook of recent work. *arXiv preprint arXiv:2203.01536*, 2022. 2

[24] Yankai Jiang, Mingze Sun, Heng Guo, Xiaoyu Bai, Ke Yan, Le Lu, and Minfeng Xu. Anatomical invariance modeling and semantic alignment for self-supervised learning in 3d medical image analysis. In *ICCV*, 2023. 2

[25] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM Comput. Surv.*, 2022. 2

[26] 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. 1

[27] Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. *ICML*, 2018. 2

[28] Julio A Kovacs and Willy Wriggers. Fast rotational matching. *Acta Crystallographica Section D: Biological Crystallography*, 58(8):1282–1286, 2002. 1

[29] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60:84 – 90, 2012. 2

[30] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 2012. 2

[31] Werner Kühlbrandt. The resolution revolution. *Science*, 343 (6178):1443–1444, 2014. 1

[32] Zakaria Laskar, Hamed Rezazadegan Tavakoli, and Juho Kannala. Semantic matching by weakly supervised 2d point set registration. In *IEEE Winter Conference on Applications of Computer Vision*, 2019. 2

[33] Shuda Li, Kai Han, Theo W Costain, Henry Howard-Jenkins, and Victor Prisacariu. Correspondence networks with adaptive neighbourhood consensus. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[34] Xinghui Li, Kai Han, Shuda Li, and Victor Prisacariu. Dual-resolution correspondence networks. In *Advances in Neural Information Processing Systems*, 2020. 2

[35] Xiaolong Li, Yijia Weng, Li Yi, Leonidas J Guibas, A Abbott, Shuran Song, and He Wang. Leveraging se (3) equivariance for self-supervised category-level object pose estimation from point clouds. *Advances in neural information processing systems*, 2021. 3

[36] 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. 2, 5

[37] 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. 5

[38] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Hyperpixel flow: Semantic correspondence with multi-layer neural features. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 2

[39] Emmanuel Moebel, Antonio Martinez-Sanchez, Lorenz Lamm, Ricardo D Righetto, Wojciech Wietrzynski, Sahradha Albert, Damien Larivière, Eric Fourmentin, Stefan Pfeffer, Julio Ortiz, et al. Deep learning improves macromolecule identification in 3d cellular cryo-electron tomograms. *Nature methods*, 18(11):1386–1394, 2021. 1

[40] Sophie Ostmeier, Brian Axelrod, Michael E. Moseley, Akshay Chaudhari, and Curtis Langlotz. Liere: Generalizing rotary position encodings. *arXiv preprint arXiv:2406.10322*, 2024. 5

[41] Arshi Parvaiz, Muhammad Anwaar Khalid, Rukhsana Zafar, Huma Ameer, Muhammad Ali, and Muhammad Moazam Fraz. Vision transformers in medical computer vision—a contemplative retrospection. *Engineering Applications of Artificial Intelligence*, 2023. 2

[42] Stefan Pfeffer and Julia Mahamid. Unravelling molecular complexity in structural cell biology. *Current Opinion in Structural Biology*, 2018. 2

[43] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional neural network architecture for geometric matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[44] Renán Rojas-Gómez, Teck-Yian Lim, Minh Do, and Raymond Yeh. Making vision transformers truly shift-equivariant. In *CVPR*, 2023. 3

[45] Paul Hongsuck Seo, Jongmin Lee, Deunsol Jung, Bohyung Han, and Minsu Cho. Attentive semantic alignment with offset-aware correlation kernels. In *European Conference on Computer Vision*, 2018. 2

[46] Feng Shi, Paul Marchwica, Juan Camilo Gamboa Higuera, Michael Jamieson, Mehrsan Javan, and Parthipan Siva. Self-supervised shape alignment for sports field registration. In *WACV*, 2022. 2

[47] David Střelák, Jiří Filipovič, Amaya Jiménez-Moreno, Jose María Carazo, and Carlos Óscar Sánchez Sorzano. Flexalign: An accurate and fast algorithm for movie alignment in cryo-electron microscopy. *Electronics*, 2020. 2

[48] Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219*, 2018. 3

[49] Hugo Touvron, Matthieu Cord, Piotr Bojanowski, Alaaeldin El-Nouby, Mathilde Caron, Baptiste Gall, Matthijs Douze, Hervé Jégou, and Armand Joulin. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 2

[50] Prune Truong, Martin Danelljan, and Radu Timofte. Glunet: Global-local universal network for dense flow and correspondences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2

[51] Martin Turk and Wolfgang Baumeister. The promise and the challenges of cryo-electron tomography. *FEBS letters*, 594 (20):3243–3261, 2020. 1

[52] Niels Volkmann, Dorit Hanein, Gabriel Oseroff, Howard White, and Phil Horowitz. Rotational biomolecule matching. *Biophysical Journal*, 2000. 2

[53] W Wan and John AG Briggs. Cryo-electron tomography and subtomogram averaging. *Methods in enzymology*, 579:329–367, 2016. 1

[54] Ziming Wang and Rebecka Jörnsten. Se (3)-bi-equivariant transformers for point cloud assembly. *arXiv preprint arXiv:2407.09167*, 2024. 3

[55] Maurice Weiler, Fred A Hamprecht, and Martin Storath. Learning steerable filters for rotation equivariant cnns. In *CVPR*, 2018. 2

[56] 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. 1

[57] 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. 1, 2, 7

[58] Xiangrui Zeng and Min Xu. Gum-net: Unsupervised geometric matching for fast and accurate 3d subtomogram alignment and averaging. In *CVPR*, 2020. 1, 2, 6, 7

[59] Xiangrui Zeng, Gregory Howe, and Min Xu. End-to-end robust joint unsupervised image alignment and clustering. In *ICCV*, 2021. 1, 2, 7

[60] Shengyu Zhao, Yue Dong, Eric Chang, and Yan Xu. Recursive cascaded networks for unsupervised medical image registration. In *ICCV*, 2019. 2

[61] Shengyu Zhao, Tingfung Lau, Ji Luo, Eric I-Chao Chang, and Yan Xu. Unsupervised 3d end-to-end medical image registration with volume tweening network. *IEEE Journal of Biomedical and Health Informatics*, 2020.

[62] Yizhou Zhao, Xun Guo, and Yan Lu. Semantic-aligned fusion transformer for one-shot object detection. In *CVPR*, 2022.

[63] Yizhou Zhao, Zhenyang Li, Xun Guo, and Yan Lu. Alignment-guided temporal attention for video action recognition. *Advances in Neural Information Processing Systems*, 2022.

[64] Barbara Zitova and Jan Flusser. Image registration methods: a survey. In *Image and vision computing*, 2003. 2