SUPPLEMENTARY: END-TO-END ROBUST JOINT UNSUPERVISED IMAGE ALIGNMENT AND CLUSTERING

Xiangrui Zeng[†]
Computational Biology Department
Carnegie Mellon University
Pittsburgh, PA 15213, USA
xiangruz@andrew.cmu.edu

Gregory Howe[†]
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213, USA
gregory.s.howe@gmail.com

Min Xu[*]
Computational Biology Department
Carnegie Mellon University
Pittsburgh, PA 15213, USA
mxu1@cs.cmu.edu

Contents

[*]Corresponding author. [†]Equal contribution.

List of Figures

List of Tables

S1 Background

What is structural biology?

Structural biology is a field of molecular biology that primarily studies the structure of cellular macromolecules, particularly proteins and nucleic acids, by microscopical [47, 64], spectroscopical [26, 68], computational [37, 71, 72], or bioinformatical [50, 57] techniques. Macromolecules carry out the basic functions of cellular process. By understanding how they acquire their specific structure and how the alteration of their structure affects their function and dynamics, structural biologists are able to decipher how the function/dysfunction of macromolecules and their networks relates to the health/disease states.

What is cryo-electron tomography?

Cryo-Electron Tomography (cryo-ET) is a cell microscopy imaging technique that produces 3D views of cellular samples at nanometer resolution ($<$ 4 nm) [65]. The cellular samples are first vitrified at cryogenic temperature ($<$ -150 ^oC). The non-crystalline cryogenic condition keeps the biological structures in the sample undisrupted during the imaging. By contrast, the conventional chemical fixation or dehydration will disrupt the biological structures. After vitrification, the sample is placed in a grid will be thinned by cryo-focused-ion-beam milling [21] to carve out a 100-250 nm lamella region before imaging.

Figure S1 shows the concept of cryo-ET imaging and reconstruction. The electron beams passing through the cell sample placed under a cryo-transmission electron microscope. The electrons are then detected by an electron detector. The detection results in a projection image. The projection image is formed as electrons are less likely to pass through a thick structural region. In cryo-ET, the cell sample is tilted through a series of angles, typically at 1^o to 3^o tilt step from -60^o to $+60^o$. At each angle view, a projection image is produced.

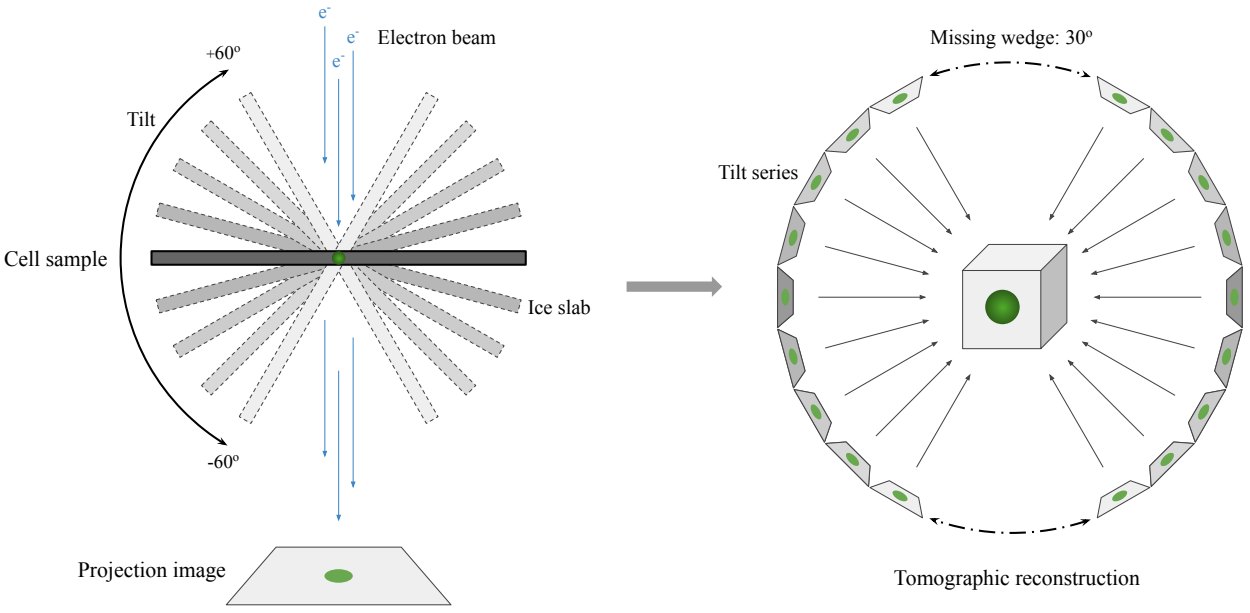

Figure S1: Illustration of cryo-ET imaging and reconstruction processes.

After acquiring a tilt-series of projection images, the 3D view of the cell sample can be reconstructed computationally [19, 45] through algorithmic steps of artifact detection and correction, alignment, back projection. The final 3D image from tomographic reconstruction is called a tomogram (Figure S2), which is a grayscale volume containing all the structural objects inside the field of view. Because a raw tomogram is very large such as of size $6000 \times 6000 \times 1500$ voxels, the computational data analysis is usually performed on the subtomogram level, of which a subtomogram (Figure S2) is a 3D cubic subvolume potentially containing a macromolecule extracted from a tomogram.

How is cryo-ET different from cryo-EM?

Cryo-electron microscopy (cryo-EM) is a closely related technique to cryo-ET. Similar to cryo-ET, cryo-EM images are acquired using a cryo-transmission electron microscope. However, the objective of cryo-EM is different from cryo-ET. Cryo-EM aims to image isolated and purified macromolecules in order to recover high-resolution structure of a known

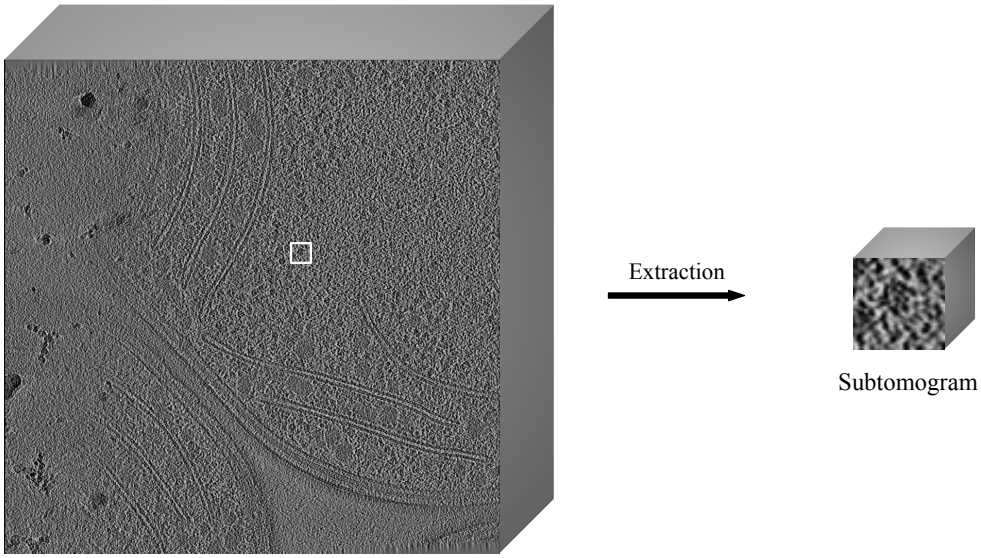

Figure S2: Illustration of extracting a subtomogram from a tomogram. The white box corresponds to the location of the extracted subtomogram.

type of macromolecule [6]. Cryo-EM only takes one projection image of the sample. Since the purified macromolecules lay in random poses on the grid, the 2D view of them can be aggregated and aligned to 3D structure [15].

In contrast, cryo-ET images the cell sample *in situ* to provide a complete structural description of the cell's native molecular landscape. *In situ* cryo-ET enables the study of both the known and unknown macromolecular structures and their spatial organization and interaction with cellular organelles, in their native cytoplasm environment [46], which is not attainable by any other imaging techniques including cryo-EM.

Biological applications of cryo-ET: Because of the unique strength of cryo-ET in visualizing wide range of 3D subcellular structures *in situ*, cryo-ET researchers have successfully revealed the native structure of large molecular complexes such as human nuclear pore complexes [44], chemoreceptor arrays [41], and chemotaxis core signaling complex [4] or organelles such as chloroplast [12], apicoplast [36], mammalian primary cilia [30]. Other than resolving native subcellular structures, cryo-ET has provided insights into important cellular functions including neural proteasome recruitment [18], dynein recruitment in intracellular trafficking network [16], and synaptic vesicle tethering [13].

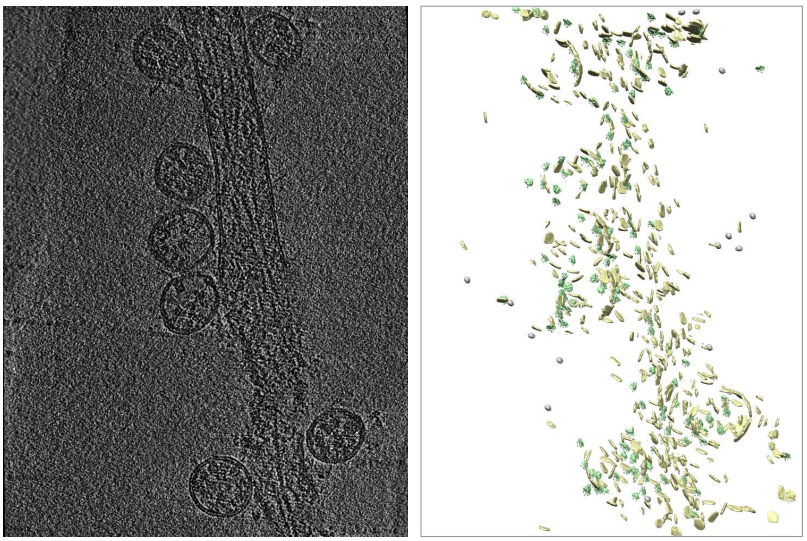

Figure S3: Sample slice and Jim-Net embedding results of a cryo-ET tomogram of SARS-CoV-2 virions released from VerE6 cells [33].

Recently, cryo-ET has been applied extensively in studying SARS-COV-2, the virus that caused the COVID-19 pandemic. Cryo-ET researchers have revealed the native structure [63] (Figure S3) and distribution of SARS-COV-2 spike proteins [29, 40] and ribonucleo-

proteins [69], and the viral replication compartment and the budding mechanism of SARS-COV-2 [33]. These critical structural insights provide meaningful clues into the drug and treatment design against SARS-COV-2 infection.

Medical applications of cryo-ET: Cryo-ET also benefits medical diagnostics to complement conventional methods. For example, the conventional electron microscopy tool can only achieve 70% accuracy in detecting primary ciliary dyskinesia, a model using cryo-ET data of human ciliary structure samples has clarified the previously unresolved primary central pair complex abnormalities by conventional EM [39]. Cryo-ET imaging of platelet samples from ovarian cancer patients has revealed alternations of the length of microtubules and the number of mitochondria compared with healthy control, which lead to diagnostic accuracy of 87.0%. Other cryo-ET researches have identified structural defects in disease states including pathogen infection [24, 51], Huntington's disease [2], Parkinson's disease [67], and Leigh syndrome [61].

What are the cause and effect of the missing wedge?

The missing wedge effect is caused by the limited tilt-angle range during the imaging process (Figure S1). The cell sample cannot be imaged at the full 180^o tilt-angle range [48] because of (1) the structure of the sample holder and (2) increasing sample thickness at high tilt-angles. $\pm 60^o$ to $\pm 70^o$ are typically used in cryo-ET. Because of the missing information at missing tilt-angles, the reconstructed tomogram will show the missing wedge effect such as elongation and blurring of the objects along the z-axis [48]. The image distortions caused by the missing wedge effect must be taken into consideration during data processing.

Why is the cryo-ET image noise high?

The high noise level of cryo-ET data (as shown in Figure S2) is mainly caused by two factors. First, unlike purified macromolecules imaged in cryo-EM, the cell sample imaged in cryo-ET is relatively much thicker with very diverse structures in the cytoplasm environment [34]. Second, because the cell sample is imaged multiple times at different tilt-angle views, the electron dose used is low [9] to prevent excessive electron beam damage to the cell sample for subsequent imaging. The relatively high sample thickness and low electron dose for imaging together result in the high noise level of cryo-ET data. Therefore, computational algorithms robust to noise are critical for cryo-ET data processing.

S2 Method

S2.1 Fine transformation regularization loss

Intuitively, coarse-to-fine alignment architectures should perform finer transformations at later layers to fine-tune the alignment, but in standard alignment losses, there is no direct incentive for the network to progressively propose finer transformations at the later layers. We craft a regularization loss $\mathcal{L}_{\text{R}}$ to penalize the network from proposing large transformations at later stages of the alignment network. By penalizing larger transformations, the space of transformation parameters that the network proposes tends to be drastically reduced, which should help stabilize and speed up training.

The regularization loss, $\mathcal{L}_{\text{R}}$, penalizes rotations according to the L_2^2 distance that each point in a unit n-sphere moves. More formally, let V be a unit n-sphere (n is 2 or 3), let s be a shift, and A be a matrix. Note, s and A together define an affine transformation.

$$\begin{aligned}
\int_V ||(Ax+s)-x||_2^2 dv &= \int_{x\in V} (Ax+s-x)^T(Ax+s-x)dx \\
&= \int_{x\in V} ((A-I)x+s)^T((A-I)x+s)dx \\
&= \int_{x\in V} x^T(A-I)^T(A-I)x + 2s^T(A-I)x + s^T s dx \\
&= \int_{x\in V} x^T(A-I)^T(A-I)x dx + 2s^T(A-I)\int_{x\in V} x dx + \int_{x\in V} s^T s dx
\end{aligned}$$

because V is symmetric and centered at 0, we have

$$\begin{aligned}
&= \int_{x\in V} x^T(A-I)^T(A-I)x dx + \int_{x\in V} s^T s dx \\
&= \int_{x\in V} x^T(A-I)^T(A-I)x dx + V s^T s \\
&= \int_{x\in V} x^T(A^TA - A - A^T + I)x dx + V s^T s \\
&= \int_{x\in V} x^T A^T A x dx - 2\int_{x\in V} x^T A x dx + \int_{x\in V} x^T x dx + V s^T s
\end{aligned}$$

because the trace of a scalar is the original scalar, we have

$$= \int_{x\in V} \text{Trace}(x^T A^T A x)dx - 2\int_{x\in V} \text{Trace}(x^T A x)dx + \int_{x\in V} x^T x dx + V s^T s$$

because of the cyclic property of traces, we have

$$= \int_{x\in V} \text{Trace}(A^T A x x^T)dx - 2\int_{x\in V} \text{Trace}(A x x^T)dx + \int_{x\in V} x^T x dx + V s^T s$$

because traces can be swapped with integrals, and A and A^TA are not functions of x, we have

$$= \text{Trace}(A^TA\int_{x\in V} xx^T dx) - 2\cdot\text{Trace}(A\int_{x\in V} xx^T dx) + \int_{x\in V} x^T x dx + V s^T s$$

because V is symmetric and centered at 0 therefore the integral over the off-diagonal entries of xx^T is 0. Additionally, because V is an n-sphere we have $\int_{x\in V} x_i^2 dx = \frac{1}{n}\int_{x\in V} x^T x dx$. Combining these two observations, we have

$$\begin{aligned}
&= \text{Trace}(A^TA\frac{1}{n}I\int_{x\in V} x^T x dx) - 2\cdot\text{Trace}(A\frac{1}{n}I\int_{x\in V} x^T x dx) + \int_{x\in V} x^T x dx + V s^T s \\
&= \text{Trace}(A^TA)\frac{1}{n}\int_{x\in V} x^T x dx - 2\cdot\text{Trace}(A)\frac{1}{n}\int_{x\in V} x^T x dx + \int_{x\in V} x^T x dx + V s^T s
\end{aligned}$$

$$= \frac{\int_{x \in V} x^T x dx}{n}(n + \text{Trace}(A^T A) - 2 \cdot \text{Trace}(A)) + V s^T s$$

assuming V is 3 dimensional, we have

$$= \frac{1}{3} \cdot \frac{4\pi}{5}(3 + \text{Trace}(A^T A) - 2 \cdot \text{Trace}(A)) + \frac{4\pi}{3} s^T s$$
$$= \frac{4\pi}{3}(\frac{1}{5}(3 + \text{Trace}(A^T A) - 2 \cdot \text{Trace}(A)) + s^T s)$$

assuming A is a rotation matrix, we have

$$= \frac{4\pi}{3}(\frac{1}{5}(3 + \text{Trace}(I_3) - 2 \cdot \text{Trace}(A)) + s^T s)$$
$$= \frac{4\pi}{3}(\frac{2}{5}(3 - \text{Trace}(A)) + s^T s)$$

Using the above derivation we define the L_2^2 regularization loss for a 3-d rigid transformation parameterized by ϕ with corresponding shift s_ϕ and rotation matrix A_ϕ as follows:

$$\mathcal{L}_{\text{R}}(\phi) = \frac{4\pi}{3}(\frac{2}{5}(3 - \text{Trace}(A_\phi)) + s_\phi^T s_\phi) \tag{1}$$

Notice that this loss is differentiable with respect to both the rotation matrix A_ϕ and the shift s_ϕ, which means it is amenable to gradient based optimization techniques.

Using an intermediate result from the above derivation we can also define a loss for arbitrary affine transformations defined by a matrix A_ϕ and a shift s_ϕ as follows:

$$\mathcal{L}_{\text{A}}(\phi) = \frac{4\pi}{3}(\frac{1}{5}(3 + \text{Trace}(A_\phi^T A_\phi) - 2 \cdot \text{Trace}(A_\phi)) + s_\phi^T s_\phi) \tag{2}$$

S2.2 Soft inlier loss

The soft inlier loss computes localized feature correlations between a target image and a transformed source image; Intuitively, if the sum of the correlations of learned features between a target and transformed source image pair is high, then the alignment is in some way semantically meaningful image patches with similar features are nearby. In 2D settings, the soft inlier loss utilizes the output from a correlation layer $C^{(G)}$, c, a transformation function $\mathcal{T}$, proposed transformation parameters ϕ, a distance metric d, and an $h \times w \times h \times w$ identity mask m^{I} defined as:

$$m^{\text{I}}_{ijkl} = \begin{cases} 1, & \text{if } d((i,j),(k,l)) < t \\ 0, & \text{otherwise} \end{cases}. \tag{3}$$

The soft inlier loss is then computed as follows:

$$\mathcal{L}_{\text{SIL}} = -\sum_{i,j,k,l} c_{ijkl} \mathcal{T}(m^{\text{I}}, \phi)_{ijkl}. \tag{4}$$

S2.3 Gaussian mixture model

We integrate Gaussian Mixture Models (GMMs) into Jim-Net's clustering step. In the clustering step, the GMM is used to assign each image a cluster. These cluster assignments are used for pairing images to be aligned by the classification module and for providing a target for training the cluster predicting module.

GMMs are one of the most used clustering models. Intuitively, GMMs make two pivotal assumptions: first, the data comes from k distinct underlying distributions; second each of the k underlying distributions is a normal distribution.

GMMs describe a dataset $X \in \Re^{n \times d}$ where n is the number of data points and d is the dimension of the input features as a probability distribution defined by the weighted sum of the probability density functions (pdfs) of k normal distributions [52]. More formally, a GMM, $\mathcal{G}_{\mu,\Sigma,w}$, defines a probability distribution with pdf:

$$p(x|\mu, \Sigma, w) = \sum_{i=1}^{k} w_i g(x|\mu_i, \Sigma_i)$$

where $g(\cdot|\mu_i, \Sigma_i)$ is the pdf of a Multivariate Gaussian Distribution with covariance matrix Σ_i and mean μ_i; we choose notation that follows [52]. Note that the weighting, w, is constrained to be non-negative and to have its entries sum to 1; this enforces that $\mathcal{G}_{\mu,\Sigma,w}$ describes a proper probability distribution. μ, Σ, and w are learned using the expectation maximization algorithm.

S2.4 Network architecture

We designed an end-to-end Convolutional Neural Network (CNN) architecture incorporating the components described above: shared extractors for feature learning $f^{(1)}, ..., f^{(m)}$, cluster predicting module M_{CP}, and coarse-to-fine alignment module M_{AL}.

Feature extractor: The shared feature extractors utilize stacked dilated convolution layers for convolution operations and spectral pooling layers for feature map size reduction alternatively. For the convolution operation, we use a layer with multiple dilated convolutions stacked in parallel. With multiple convolution operations at different dilation rates, this layer has a large receptive field and has been shown to maintain high spatial resolution for dense pixel-level matching tasks [59]. For reducing the feature map size, we use a spectral pooling layer that can resize a feature map to an arbitrary size, not limited to a certain downsize factor such as 2. We reduce the feature map size progressively by each spectral pooling layer: downsizing to 75% for the coarse alignment step and to 85% for the fine alignment step. Additionally, the spectral pooling layer has been theoretically proven to preserve better spatial information as compared to a max pooling layer [53].

Feature processing for clustering: Given a feature representation, $f_{s|t}$ we predict cluster assignment, ℓ_s, for s by passing $f_{s|t}$ to classification module, M_{CP}, composed of additional feature processing layers followed by a classifier. After each stacked dilated convolution layer, the feature maps are processed by an additional convolution layer and dimensionally reduced by a global max-pooling layer. The feature vectors are concatenated and processed by two fully connected layers to output the final cluster assignment prediction. Outputs from the second last layer is used as feature representations $f_{X|X}$ for the GMM. By utilizing feature maps from every stacked dilated convolution layer, we are able to pass both low and high-level feature information to the classifier for better prediction.

Transformation regression for alignment: For the ith alignment $G^{(i)}(s,t)$ Jim-Net creates feature representations $f^{(i)}(G^{(i-1)}(s,t))$ and $f^{(i)}(t)$ which are passed to a transformation regressor $r^{(i)}$. $r^{(i)}$ consists of a feature matching layer, $C^{(G)}$ or $C^{(L)}$, followed by a CNN regressor and a spatial transformer [27].

Jim-Net's architecture is outlined below. We define 'SDC (Stacked Dilated Convolution) Same' and 'SDC Valid' as separate figures due to their common usage in Jim-Net. We also define the feature extraction components for the coarse and fine alignment modules 'Feature Extractor 1' and 'Feature Extractor 3', respectively, as indexed in main document Figure 2. In the architecture diagram, dotted lines represent weight sharing and arrows represent the output of one layer being an input to another layer.

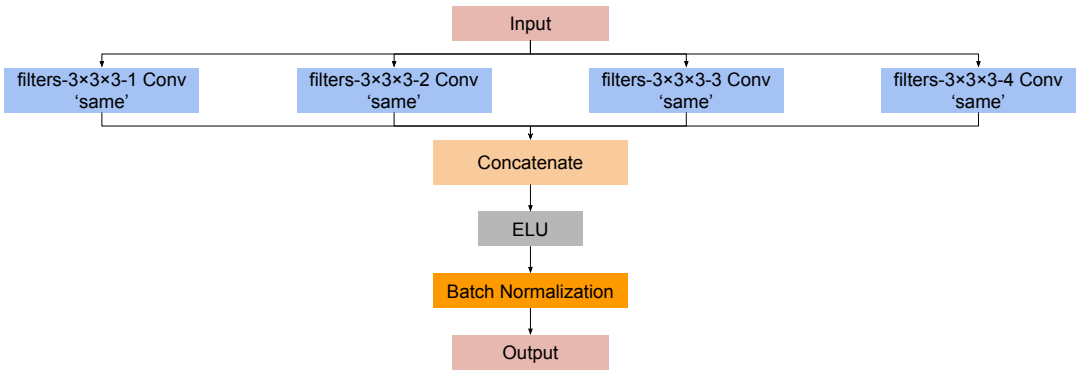

Figure S4: "SDC Same" Architecture. Each box represents a layer or composition of layers. "filters-3x3x3-d Conv 'same'" represents a 3D convolutional layer with kernel size 3, dilation rate d, number of filters set to 'filters', and 'same' padding (padding of the same size of the layer input). The concatenate layer concatenates inputs channel wise, "ELU" [8] is an ELU activation layer and "Batch Normalization" [25] is a batch normalization layer. The peach colored boxes labeled 'input' and 'output' correspond to the input and output of the "SDC Same"

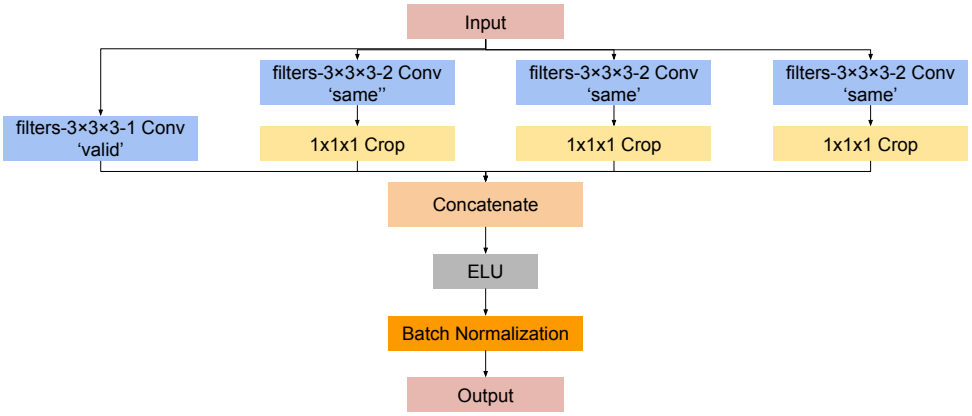

Figure S5: "SDC Valid" Architecture. "filters-3x3x3-d Conv 'valid'" represents a 3D convolutional layer with kernel size 3, dilation rate d, number of filters set to 'filters', and 'valid' padding (no padding). The "1x1x1 Crop" layer represents a 3D cropping layer that crops 1 pixel from the start and end of each dimension. All other definitions are the same as in Figure S4.

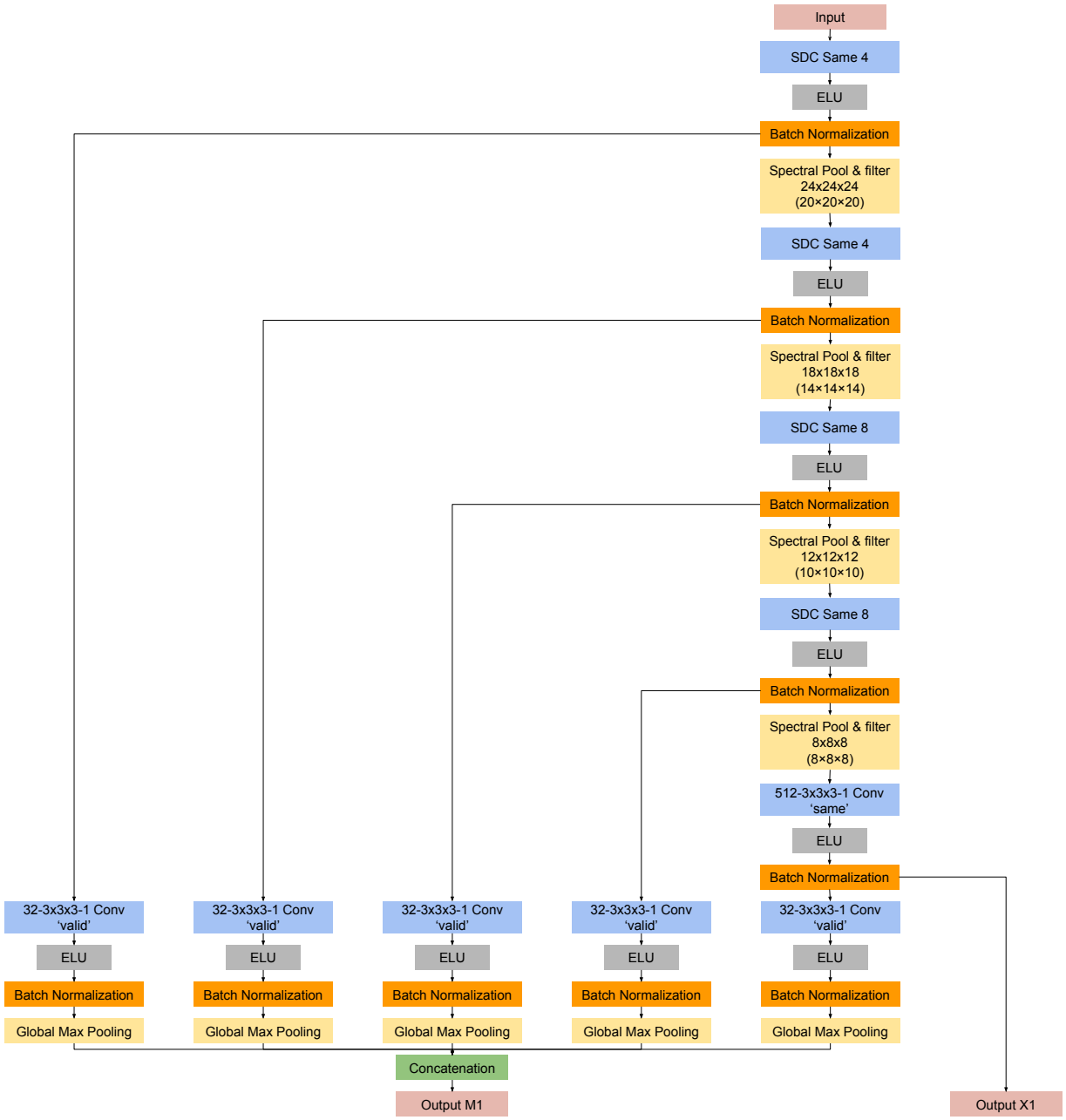

Figure S6: Feature Extractor 1 Architecture. Feature Extractor 1 is used for feature extraction during the coarse alignment phase of the Alignment Module. "Global Max Pooling" represents a global max pooling operation that pools across all indices to reduce the dimension to be the amount of channels. "Spectral Pool & filter $Z \times Z \times Z$ ($Y \times Y \times Y$)" represents a spectral pooling and filtering layer with output size $Z \times Z \times Z$ which is cropped to be of size $Y \times Y \times Y$.

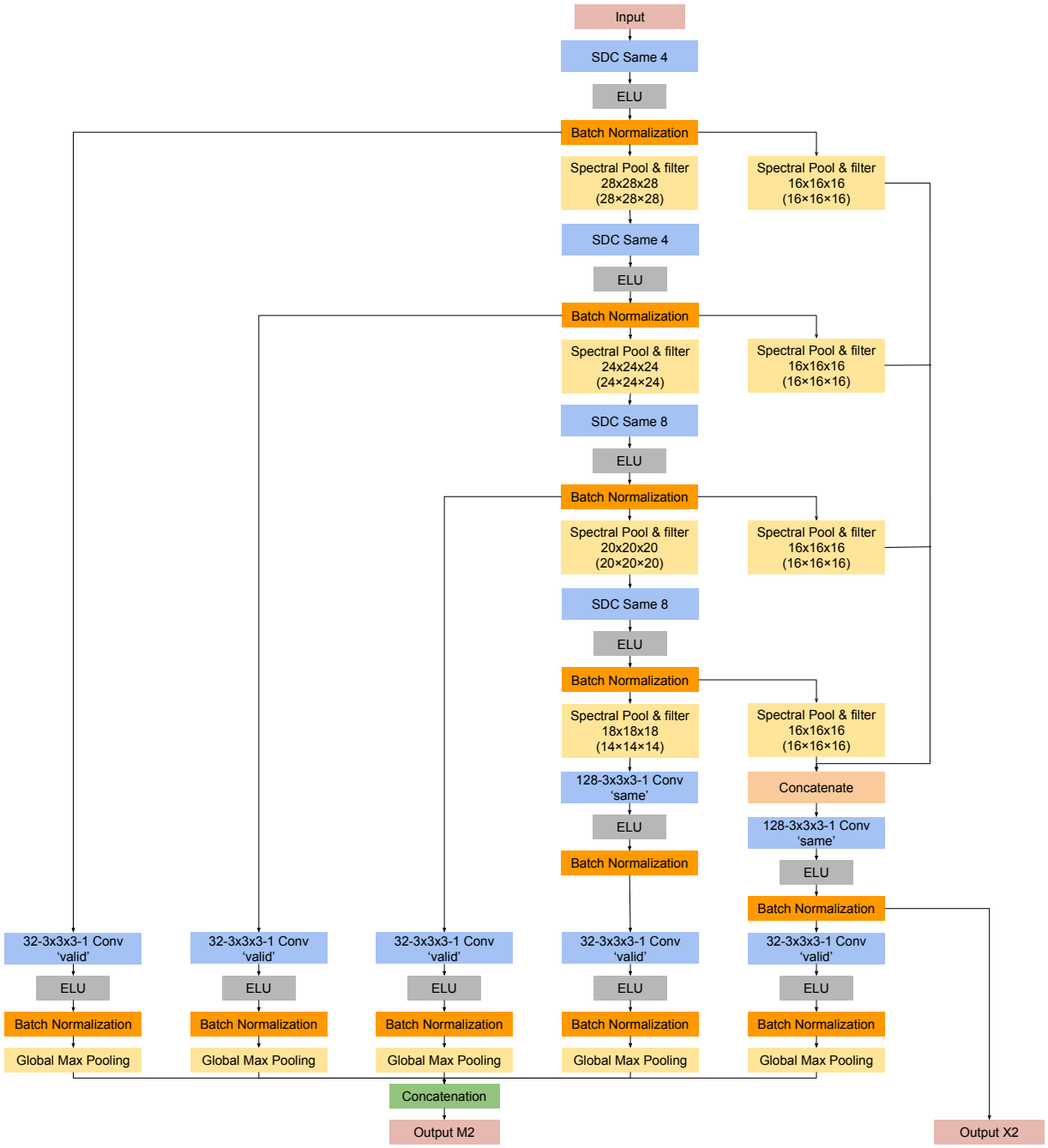

Figure S7: Feature Extractor 3 Architecture. Feature Extractor 3 is used for feature extraction during the fine alignment phase of the Alignment Module.

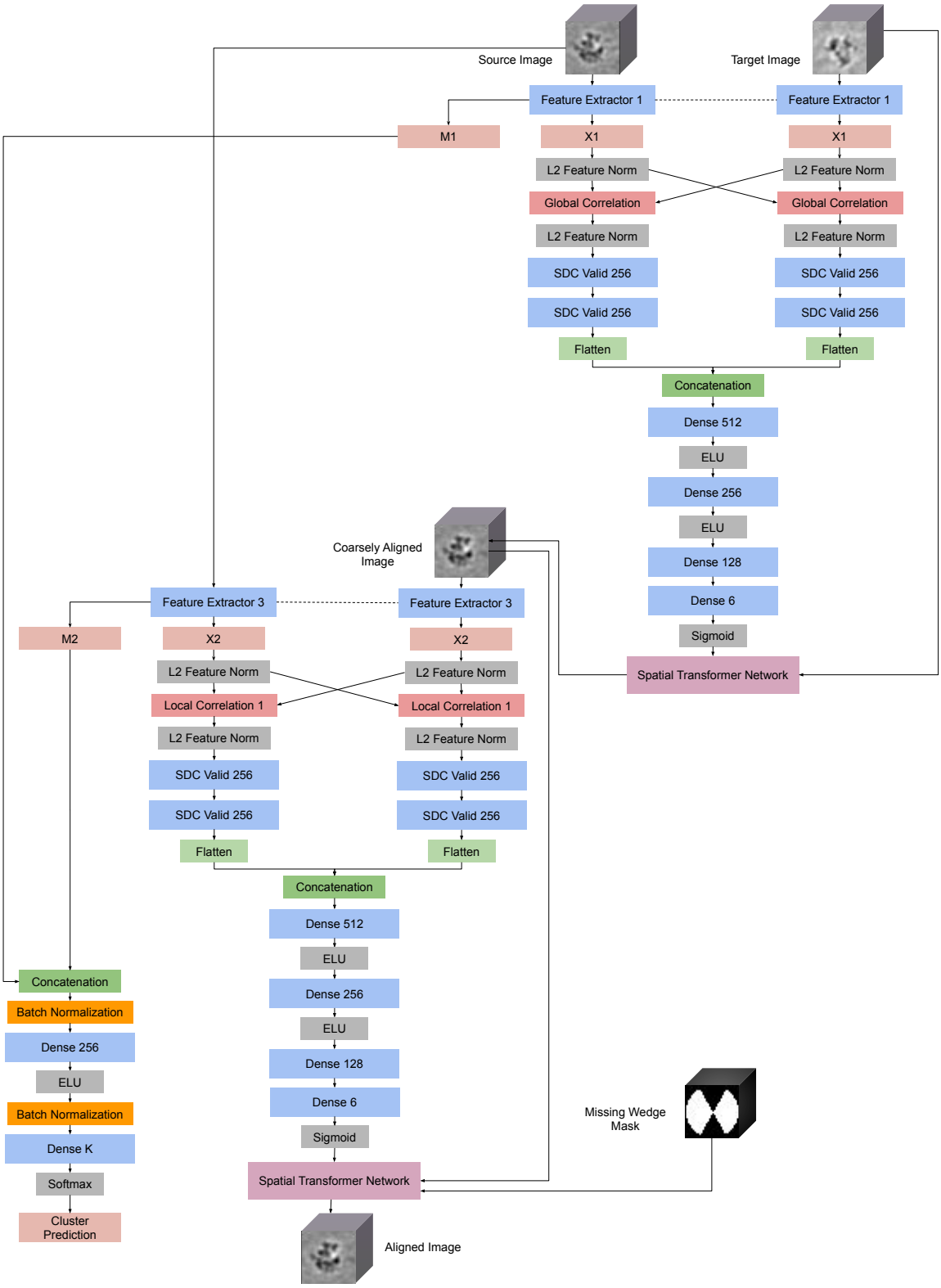

Figure S8: Jim-Net architecture. “Flatten" represents a layer that flattens the input tensor, “Dense X" is a linear layer with X nodes, “Sigmoid" is a sigmoid activation layer, and “Spatial Transformer Network" is a spatial transformer layer [27] constrained to proposing 3D rigid transformations. “Global Correlation" denotes a global correlation layer [54] and Local Correlation 1 is a local correlation layer with radius 1.

S3 Experimental validation

S3.1 Training details

We implemented the models in Pytorch [49] and Keras [7] backend by Tensorflow 2.0 [1]. We trained all models on a computer with 4 NVIDIA GeForce Titan X Pascal GPUs and 48 CPU cores. For cryo-ET simulated dataset at SNR 100, we trained Jim-Net with a learning rate of $1 \cdot 10^{-5}$ for 100 epochs using the Nadam optimizer [62] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1 \cdot 10^{-7}$. Then, similar to the baseline method Gum-Net [70], we fine-tuned the trained model from SNR 100 dataset on other datasets for 20 epochs with a learning rate of $1 \cdot 10^{-6}$. Since the baseline clustering methods DeepCluster [3] and PICA [22] had not been tested on cryo-ET data, we extended them to 3D versions by trying different architectures, from simple ones to complex ones, and kept the one with the best performance. Then, we trained the extended 3D network models in the same way as Jim-Net with the same optimizer and learning rate till convergence. For the PF-PASCAL dataset, all the baseline methods were initialized with the ResNet-101 [20] feature extraction backbone with its ImageNet [58] weights and fine-tuned on the PF-PASCAL dataset. We directly took their reported per-class PCK accuracy on the PF-PASCAL dataset (Table S11). For a fair comparison, Jim-Net's alignment module was trained with the same feature extraction backbone initialization. More specifically, as one of the baseline methods, WeakAlign [56], the feature extractor was cropped after the conv4-23 layer of ResNet-101 and additionally, as in WeakAlign, the image alignment module was initialized with the weights from [55], which was trained in a self-supervised fashion without access to annotations or class labels. We trained the model for 30 epochs using the Adam optimizer [32] with a learning rate of $1 \cdot 10^{-7}$, a weight decay of 0, and momentum terms of $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1 \cdot 10^{-7}$. Early stopping was used to choose the model with the highest validation PCK score. During each epoch, the clustering branch was warm started for 10 epochs with a learning rate of $1 \cdot 10^{-4}$. Additionally, the clustering branch had a separate learning rate of $\epsilon = 5 \cdot 10^{-4}$.

For the alignment modules, we used a spatial transformer layer [27] with a 2D affine transformation for coarse alignment and a thin plate spline transformation for fine alignment on the PF-PASCAL dataset. We used 3D rigid-body transformations for both the coarse and fine transformations on cryo-ET datasets. This is because a macromolecule is of fixed size with chirality under electron microscopy.

For GMM clustering, the learned features from the source image extracted from the feature layer were dimension reduced by TSNE [43] to speed up the clustering process. We used full covariance for fitting the GMM. Both the TSNE and GMM were implemented using the python package *scikit-learn*. For each training iteration, we matched the cluster indices from the previous iteration by the Hungarian method [35] to stabilize the clustering branch training.

S3.2 Results details

In Table S1-S5, we report the alignment accuracy by different macromolecules on the simulated benchmark cryo-ET datasets. Jim-Net achieved the overall best performance on four of the five macromolecules. We note that our model Jim-Net converges faster than Gum-Net as Gum-Net trained the initial model for 500 epochs. This is because Jim-Net learns to pair semantically similar images for more meaningful alignment training.

Method	SNR 100	SNR 0.1	SNR 0.05	SNR 0.03	SNR 0.01
H-T align	**0.06±0.02**, 1.03±0.63	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	0.08±0.13, 1.09±1.14	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	0.27±0.54, 1.13±2.03	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**
Jim-Net	0.16±0.47, **0.82±1.93**	**0.30±0.47**, **1.42±2.01**	**0.51±0.58**, **2.20±2.36**	**0.74±0.62**, **3.13±2.63**	1.50±0.76, 6.30±3.13

Table S1: Spliceosome (5LQW) subtomogram alignment accuracy on five datasets with SNR specified. In each cell, the first term is the mean and standard deviation of the rotation error and the second term, the translation error. Best results across all methods are highlighted.

Method	SNR 100	SNR 0.1	SNR 0.05	SNR 0.03	SNR 0.01
H-T align	0.63±0.99, 3.15±4.27	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	0.67±1.00, 3.22±4.24	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	0.56±0.78, 2.22±3.05	**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**
Jim-Net	**0.46±0.56**, **1.98±2.46**	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

Table S2: RNA polymerase-rifampicin complex (1I6V) subtomogram alignment accuracy.

Method	SNR 100	SNR 0.1	SNR 0.05	SNR 0.03	SNR 0.01
H-T align	**0.09±0.10**, 1.11±0.82	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	0.16±0.34, 1.31±1.62	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	0.30±0.55, 1.08±1.71	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**
Jim-Net	0.22±0.47, **0.98±1.66**	**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

Table S3: RNA polymerase II elongation complex (6A5L) subtomogram alignment accuracy.

Method	SNR 100	SNR 0.1	SNR 0.05	SNR 0.03	SNR 0.01
H-T align	0.06±0.02, 0.99±0.60	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	**0.05±0.03**, 0.98±0.61	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	0.43±0.87, 1.67±3.31	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
Jim-Net	0.16±0.50, **0.81±1.88**	**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**

Table S4: Ribosome (5T2C) subtomogram alignment accuracy.

Method	SNR 100	SNR 0.1	SNR 0.05	SNR 0.03	SNR 0.01
H-T align	0.65±0.95, 2.81±3.44	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	0.69±0.97, 3.02±3.72	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	0.48±0.67, 1.86±2.53	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**
Jim-Net	**0.45±0.52**, **1.82±2.16**	**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

Table S5: Capped proteasome (5MPA) subtomogram alignment accuracy.

In Table S6-S10, we report the clustering accuracy by different macromolecules on the simulated benchmark cryo-ET datasets. By comparing the per macromolecule clustering accuracy of different methods, we can see that the accuracy of Jim-Net is relatively consistent among all macromolecules, whereas the baseline methods have the problem of the degeneration of clusters. As discussed in the main document, Jim-Net has shared feature extractors for alignment and clustering to learn robust features and therefore avoids the degeneration of clusters.

Method	SNR 100	0.1	0.05	0.03	0.01
DeepCluster	86.7	40.8	30.0	26.6	28.4
PICA	**100**	**100**	55.1	24.7	33.5
Jim-Net (cluster)	99.8	97.8	87.5	67.8	44.6
Jim-Net	**100**	**100**	**96.1**	**87.7**	**47.3**

Table S6: Spliceosome (5LQW) subtomogram clustering accuracy on five datasets with SNR specified. Best results across all methods are highlighted.

Method	SNR 100	0.1	0.05	0.03	0.01
DeepCluster	53.9	24.2	21.2	18.7	16.5
PICA	**100**	**99.9**	40.4	39.7	23.6
Jim-Net (cluster)	99.8	78.1	37.2	29.9	25.3
Jim-Net	**100**	**99.9**	**94.9**	**83.1**	**42.2**

Table S7: RNA polymerase-rifampicin complex (1I6V) subtomogram clustering accuracy.

Method	SNR 100	0.1	0.05	0.03	0.01
DeepCluster	58.8	61.1	36.5	29.8	43.4
PICA	**100**	74.7	57.7	29.5	17.0
Jim-Net (cluster)	98.3	55.8	59.8	48.4	29.1
Jim-Net	**100**	**98.8**	**94.8**	**73.8**	**36.0**

Table S8: RNA polymerase II elongation complex (6A5L) subtomogram clustering accuracy.

Method	SNR 100	0.1	0.05	0.03	0.01
DeepCluster	**100**	84.8	78.2	69.8	22.9
PICA	**100**	**100**	97.5	32.7	33.9
Jim-Net (cluster)	**100**	99.7	88.8	76.5	41.4
Jim-Net	**100**	**100**	**99.4**	**98.5**	**58.3**

Table S9: Ribosome (5T2C) subtomogram clustering accuracy.

Method	SNR 100	0.1	0.05	0.03	0.01
DeepCluster	44.1	33.0	31.4	25.0	24.9
PICA	**100**	57.7	28.4	19.3	33.9
Jim-Net (cluster)	99.7	56.0	40.5	33.9	36.2
Jim-Net	**100**	**99.7**	**95.1**	**84.4**	**56.6**

Table S10: Capped proteasome (5MPA) subtomogram clustering accuracy.

In Table S11, we report the per-class PCK on each of the baseline methods. Among the weakly-supervised baseline methods, SF-Net achieved the overall best performance. This is mainly because SF-Net requires stronger supervision in the form of segmented foreground masks than other methods that only require image labels. We note that although our unsupervised Jim-Net did not perform better than the methods that require weak supervision (foreground masks or image labels), Jim-Net beat the only self-supervised method, A2-Net [60], overall 74.8% vs 70.8%. A2-Net is the most directly comparable method to Jim-Net due to A2-Net being the only other method in the table that does not require some form of annotated data.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	d.table	dog	house	moto	person	plant	sheep	sofa	train	tv	all
A2-Net	83.2	82.8	83.8	44.4	57.8	81.3	89.4	86.1	40.1	91.7	21.4	73.2	33.8	76.3	74.3	63.3	**100.0**	45.5	45.3	60.0	70.8
WeakAlign	83.7	88.0	83.4	58.3	68.8	90.3	92.3	83.7	47.4	91.7	28.1	76.3	**77.0**	76.0	71.4	**76.2**	80.0	59.5	62.3	63.9	75.8
RTNs	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	75.9
NC-Net	86.8	86.7	**86.7**	55.6	82.8	88.6	93.8	87.1	54.3	87.5	43.2	82.0	64.1	79.2	71.1	71.0	60.0	54.2	75.0	**82.8**	78.9
SF-Net	**89.5**	**89.2**	83.1	73.6	**85.9**	**92.6**	**95.0**	83.7	65.6	**93.8**	53.6	81.3	71.6	80.6	72.3	71.0	**100.0**	**69.3**	80.0	79.5	81.9
DCC-Net	87.3	88.6	82.0	66.7	84.4	89.6	94.0	**90.5**	64.4	91.7	51.6	**84.2**	74.3	**83.5**	72.5	72.9	60.0	68.3	81.8	81.1	**82.3**
Jim-Net	73.6	76.3	59.2	**82.6**	64.4	74.4	73.3	84.8	**71.9**	88.4	**74.9**	77.3	69.8	71.0	**81.1**	51.2	**100.0**	68.0	**82.5**	75.1	74.8

Table S11: Per-class PCK on the PF-PASCAL benchmark. Baseline results were directly taken from corresponding papers. RTNs did not report their per-class PCK.

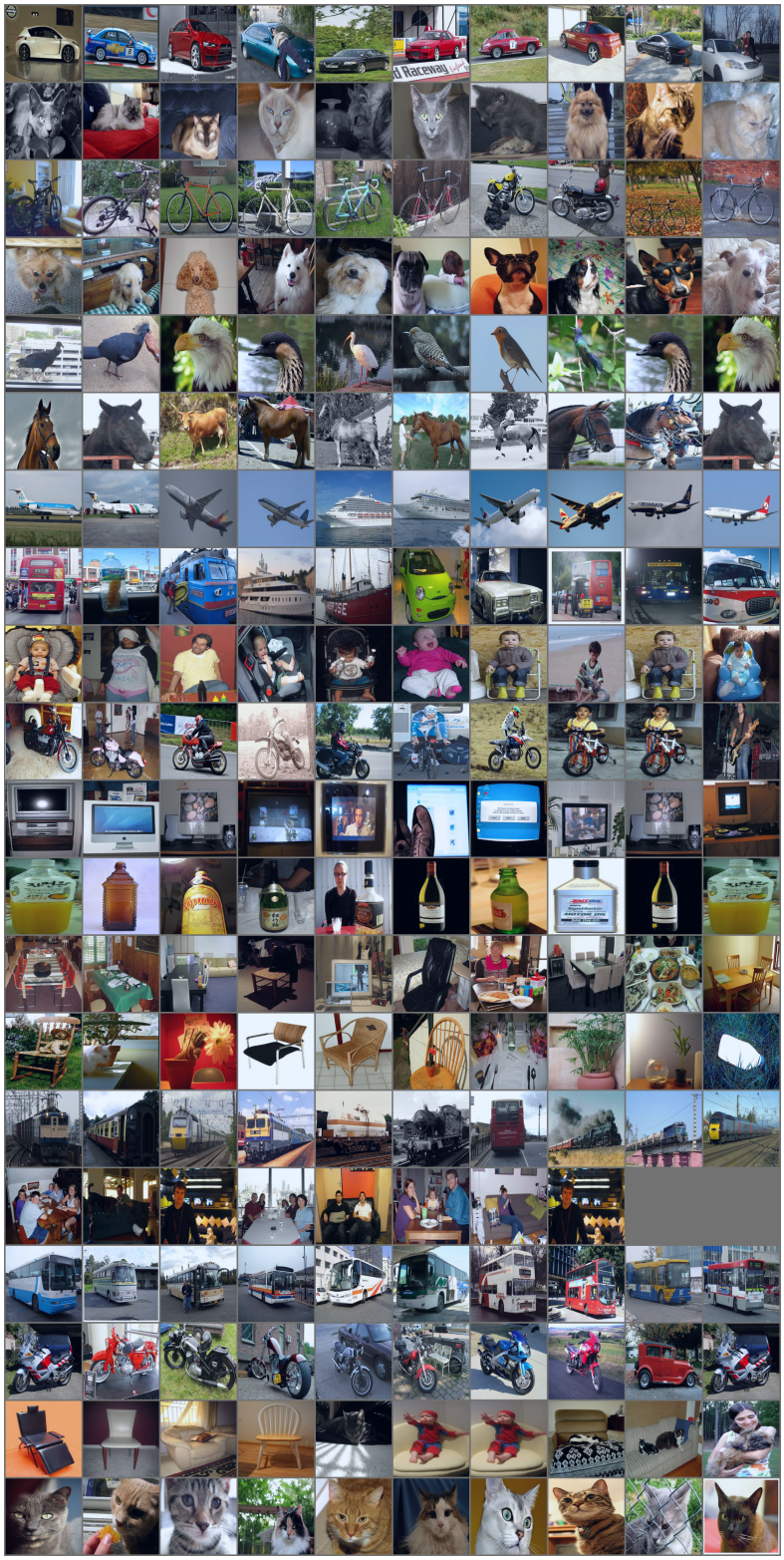

Figure S9: Sample clustering results of Jim-Net applied on PF-PASCAL. Each row corresponds to images that Jim-Net's cluster predicting branch assigned the same cluster label.

Figure S10: Sample source to target alignments results of Jim-Net applied on PF-PASCAL. The images are arranged into groups of four. The first image being the source image; the second image being the source image after an affine transformation; the third image being the source image after an affine and spline transformation, and the last image being the target image.

S4 Additional materials

We put some contents originally in the main document due to page limits.

Related Work

S4.1 Optical flow/Image registration

Optical flow describes the motion pattern of objects in a visual scene by a dense or sparse vector field. [5, 14] provide good surveys on traditional approaches. Flownet [11, 23] is the first end-to-end model for optical flow estimation trained in a supervised fashion. Later, self-supervised [42, 75] and unsupervised [17, 66] optical flow models have been proposed. Deformable image registration, a closely related concept, locally registers a set of source images to a reference image, which is often applied to 3D medical images. Unsupervised deformable image registration models have been successfully applied to CT scans [28, 31, 73], MRI [10, 74], and PET images [38].

As opposed to our image alignment objective of dealing with large transformation variations, both the 2D optical flow and the 3D image registration are restricted to small local displacements between the source and target pairs.

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283, 2016.

[2] FJB Bäuerlein, A Mishra, I Dudanova, M Hipp, R Klein, FU Hartl, W Baumeister, R Fernández-Busnadiego, et al. Structural characterization of mutant huntingtin inclusion bodies by cryo-electron tomography. *Microscopy and Microanalysis*, 22(S3):80–81, 2016.

[3] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018.

[4] C Keith Cassidy, Benjamin A Himes, Dapeng Sun, Jun Ma, Gongpu Zhao, John S Parkinson, Phillip J Stansfeld, Zaida Luthey-Schulten, and Peijun Zhang. Structure and dynamics of the e. coli chemotaxis core signaling complex by cryo-electron tomography and molecular simulations. *Communications biology*, 3(1):1–10, 2020.

[5] Haiyang Chao, Yu Gu, and Marcello Napolitano. A survey of optical flow techniques for robotics navigation applications. *Journal of Intelligent & Robotic Systems*, 73(1-4):361–372, 2014.

[6] Yifan Cheng, Nikolaus Grigorieff, Pawel A Penczek, and Thomas Walz. A primer to single-particle cryo-electron microscopy. *Cell*, 161(3):438–449, 2015.

[7] François Chollet et al. Keras (2015), 2017.

[8] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus), 2016.

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

[10] Bob D de Vos, Floris F Berendsen, Max A Viergever, Hessam Sokooti, Marius Staring, and Ivana Išgum. A deep learning framework for unsupervised affine and deformable image registration. *Medical image analysis*, 52:128–143, 2019.

[11] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.

[12] Benjamin D Engel, Miroslava Schaffer, Luis Kuhn Cuellar, Elizabeth Villa, Jürgen M Plitzko, and Wolfgang Baumeister. Native architecture of the chlamydomonas chloroplast revealed by in situ cryo-electron tomography. *Elife*, 4:e04889, 2015.

[13] Rubén Fernández-Busnadiego, Shoh Asano, Ana-Maria Oprisoreanu, Eri Sakata, Michael Doengi, Zdravko Kochovski, Magdalena Zürner, Valentin Stein, Susanne Schoch, Wolfgang Baumeister, et al. Cryo–electron tomography reveals a critical role of rim1α in synaptic vesicle tethering. *Journal of Cell Biology*, 201(5):725–740, 2013.

[14] Denis Fortun, Patrick Bouthemy, and Charles Kervrann. Optical flow modeling and computation: a survey. *Computer Vision and Image Understanding*, 134:1–21, 2015.

[15] Joachim Frank, Michael Radermacher, Pawel Penczek, Jun Zhu, Yanhong Li, Mahieddine Ladjadj, and Ardean Leith. Spider and web: processing and visualization of images in 3d electron microscopy and related fields. *Journal of structural biology*, 116(1):190–199, 1996.

[16] Danielle A Grotjahn, Saikat Chowdhury, Yiru Xu, Richard J McKenney, Trina A Schroer, and Gabriel C Lander. Cryo-electron tomography reveals that dynactin recruits a team of dyneins for processive motility. *Nature structural & molecular biology*, 25(3):203–207, 2018.

[17] Shuosen Guan, Haoxin Li, and Wei-Shi Zheng. Unsupervised learning for optical flow estimation using pyramid convolution lstm. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 181–186. IEEE, 2019.

[18] Qiang Guo, Carina Lehmer, Antonio Martínez-Sánchez, Till Rudack, Florian Beck, Hannelore Hartmann, Manuela Pérez-Berlanga, Frédéric Frottin, Mark S Hipp, F Ulrich Hartl, et al. In situ structure of neuronal c9orf72 poly-ga aggregates reveals proteasome recruitment. *Cell*, 172(4):696–705, 2018.

[19] Renmin Han, Xiaohua Wan, Zihao Wang, Yu Hao, Jingrong Zhang, Yu Chen, Xin Gao, Zhiyong Liu, Fei Ren, Fei Sun, et al. Autom: a novel automatic platform for electron tomography reconstruction. *Journal of structural biology*, 199(3):196–208, 2017.

[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.

[21] Chyongere Hsieh, Thomas Schmelzer, Gregory Kishchenko, Terence Wagenknecht, and Michael Marko. Practical workflow for cryo focused-ion-beam milling of tissues and cells for cryo-tem tomography. *Journal of structural biology*, 185(1):32–41, 2014.

[22] Jiabo Huang, Shaogang Gong, and Xiatian Zhu. Deep semantic clustering by partition confidence maximisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8849–8858, 2020.

[23] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017.

[24] Simon Imhof, Jiayan Zhang, Hui Wang, Khanh Huy Bui, Hoangkim Nguyen, Ivo Atanasov, Wong H Hui, Shun Kai Yang, Z Hong Zhou, and Kent L Hill. Cryo electron tomography with volta phase plate reveals novel structural foundations of the 96-nm axonemal repeat in the pathogen trypanosoma brucei. *Elife*, 8:e52058, 2019.

[25] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015.

[26] Neil E Jacobsen. *NMR spectroscopy explained: simplified theory, applications and examples for organic chemistry and structural biology*. John Wiley & Sons, 2007.

[27] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.

[28] Zhuoran Jiang, Fang-Fang Yin, Yun Ge, and Lei Ren. A multi-scale framework with unsupervised joint training of convolutional neural networks for pulmonary deformable image registration. *Physics in Medicine & Biology*, 65(1):015011, 2020.

[29] Zunlong Ke, Joaquin Oton, Kun Qu, Mirko Cortese, Vojtech Zila, Lesley McKeane, Takanori Nakane, Jasenko Zivanov, Christopher J Neufeldt, Berati Cerikan, et al. Structures and distributions of sars-cov-2 spike proteins on intact virions. *Nature*, pages 1–7, 2020.

[30] Petra Kiesel, Gonzalo Alvarez Viar, Nikolai Tsoy, Riccardo Maraspini, Peter Gorilak, Vladimir Varga, Alf Honigmann, and Gaia Pigino. The molecular structure of mammalian primary cilia revealed by cryo-electron tomography. *Nature Structural & Molecular Biology*, pages 1–10, 2020.

[31] Boah Kim, Jieun Kim, June-Goo Lee, Dong Hwan Kim, Seong Ho Park, and Jong Chul Ye. Unsupervised deformable image registration using cycle-consistent cnn. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 166–174. Springer, 2019.

[32] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[33] Steffen Klein, Mirko Cortese, Sophie L Winter, Moritz Wachsmuth-Melm, Christopher J Neufeldt, Berati Cerikan, Megan L Stanifer, Steeve Boulant, Ralf Bartenschlager, and Petr Chlanda. Sars-cov-2 structure and replication characterized by in situ cryo-electron tomography. *BioRxiv*, 2020.

[34] Roman I Koning, Abraham J Koster, and Thomas H Sharp. Advances in cryo-electron tomography for biology and medicine. *Annals of Anatomy-Anatomischer Anzeiger*, 217:82–96, 2018.

[35] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[36] Leandro Lemgruber, Mikhail Kudryashev, Chaitali Dekiwadia, David T Riglar, Jake Baum, Henning Stahlberg, Stuart A Ralph, and Friedrich Frischknecht. Cryo-electron tomography reveals four-membrane architecture of the plasmodium apicoplast. *Malaria journal*, 12(1):25, 2013.

[37] Michael Levitt. The birth of computational structural biology. *Nature structural biology*, 8(5):392–393, 2001.

[38] Tiantian Li, Mengxi Zhang, Wenyuan Qi, Evren Asma, and Jinyi Qi. Motion correction of respiratory-gated pet images using deep learning based image registration framework. *Physics in Medicine & Biology*, 2020.

[39] Jianfeng Lin, Weining Yin, Maria C Smith, Kangkang Song, Margaret W Leigh, Maimoona A Zariwala, Michael R Knowles, Lawrence E Ostrowski, and Daniela Nicastro. Cryo-electron tomography reveals ciliary defects underlying human rsph1 primary ciliary dyskinesia. *Nature communications*, 5:5727, 2014.

[40] Chuang Liu, Luiza Mendonça, Yang Yang, Yuanzhu Gao, Chenguang Shen, Jiwei Liu, Tao Ni, Bin Ju, Congcong Liu, Xian Tang, et al. The architecture of inactivated sars-cov-2 with postfusion spikes revealed by cryo-em and cryo-et. *Structure*, 2020.

[41] Jun Liu, Bo Hu, Dustin R Morado, Sneha Jani, Michael D Manson, and William Margolin. Molecular architecture of chemoreceptor arrays revealed by cryoelectron tomography of escherichia coli minicells. *Proceedings of the National Academy of Sciences*, 109(23):E1481–E1488, 2012.

[42] Pengpeng Liu, Michael Lyu, Irwin King, and Jia Xu. Selflow: Self-supervised learning of optical flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4571–4580, 2019.

[43] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[44] Tal Maimon, Nadav Elad, Idit Dahan, and Ohad Medalia. The human nuclear pore complex as revealed by cryo-electron tomography. *Structure*, 20(6):998–1006, 2012.

[45] 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.

[46] Catherine M Oikonomou and Grant J Jensen. Cellular electron cryotomography: toward structural biology in situ. *Annual review of biochemistry*, 86:873–896, 2017.

[47] Igor Orlov, Alexander G Myasnikov, Leonid Andronov, S Kundhavai Natchiar, Heena Khatter, Brice Beinsteiner, Jean-François Ménétret, Isabelle Hazemann, Kareem Mohideen, Karima Tazibt, et al. The integrative role of cryo electron microscopy in molecular and cellular structural biology. *Biology of the Cell*, 109(2):81–93, 2017.

[48] Lassi Paavolainen, Erman Acar, Uygar Tuna, Sari Peltonen, Toshio Moriya, Pan Soonsawad, Varpu Marjomäki, R Holland Cheng, and Ulla Ruotsalainen. Compensation of missing wedge effects with sequential statistical reconstruction in electron tomography. *PloS one*, 9(10):e108978, 2014.

[49] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037, 2019.

[50] Bhumi Patel, Vijai Singh, and Dhaval Patel. Structural bioinformatics. In *Essentials of Bioinformatics, Volume I*, pages 169–199. Springer, 2019.

[51] Zhuan Qin, Jiagang Tu, Tao Lin, Steven J Norris, Chunhao Li, Md A Motaleb, and Jun Liu. Cryo-electron tomography of periplasmic flagella in borrelia burgdorferi reveals a distinct cytoplasmic atpase complex. *PLoS biology*, 16(11):e3000050, 2018.

[52] Douglas Reynolds. *Gaussian Mixture Models*, pages 659–663. Springer US, Boston, MA, 2009.

[53] Oren Rippel, Jasper Snoek, and Ryan P Adams. Spectral representations for convolutional neural networks. In *Advances in neural information processing systems*, pages 2449–2457, 2015.

[54] 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*, pages 6148–6157, 2017.

[55] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional neural network architecture for geometric matching. *CoRR*, abs/1703.05593, 2017.

[56] Ignacio Rocco, Relja Arandjelović, and Josef Sivic. End-to-end weakly-supervised semantic alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6917–6925, 2018.

[57] Xiongtao Ruan and Robert F Murphy. Evaluation of methods for generative modeling of cell and nuclear shape. *Bioinformatics*, 35(14):2475–2485, 2019.

[58] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[59] René Schuster, Oliver Wasenmuller, Christian Unger, and Didier Stricker. Sdc-stacked dilated convolution: A unified descriptor network for dense matching tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2556–2565, 2019.

[60] 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*, pages 367–383. Springer, 2018.

[61] Stephanie E Siegmund, Robert Grassucci, Stephen D Carter, Emanuele Barca, Zachary J Farino, Martí Juanola-Falgarona, Peijun Zhang, Kurenai Tanji, Michio Hirano, Eric A Schon, et al. Three-dimensional analysis of mitochondrial crista ultrastructure in a patient with leigh syndrome by in situ cryoelectron tomography. *iScience*, 6:83–91, 2018.

[62] Ange Tato and Roger Nkambou. Improving adam optimizer. 2018.

[63] Beata Turoňová, Mateusz Sikora, Christoph Schürmann, Wim JH Hagen, Sonja Welsch, Florian EC Blanc, Sören von Bülow, Michael Gecht, Katrin Bagola, Cindy Hörner, et al. In situ structural analysis of sars-cov-2 spike reveals flexibility mediated by three hinges. *Science*, 370(6513):203–208, 2020.

[64] Catherine Vénien-Bryan, Zhuolun Li, Laurent Vuillard, and Jean Albert Boutin. Cryo-electron microscopy and x-ray crystallography: complementary approaches to structural biology and drug discovery. *Acta Crystallographica Section F: Structural Biology Communications*, 73(4):174–183, 2017.

[65] Jonathan Wagner, Miroslava Schaffer, and Rubén Fernández-Busnadiego. Cryo-electron tomography—the cell biology that came in from the cold. *FEBS letters*, 591(17):2520–2533, 2017.

[66] Yang Wang, Peng Wang, Zhenheng Yang, Chenxu Luo, Yi Yang, and Wei Xu. Unos: Unified unsupervised optical-flow and stereo-depth estimation by watching videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8071–8081, 2019.

[67] Reika Watanabe, Robert Buschauer, Jan Böhning, Martina Audagnotto, Keren Lasker, Tsan-Wen Lu, Daniela Boassa, Susan Taylor, and Elizabeth Villa. The in situ structure of parkinson's disease-linked lrrk2. *Cell*, 182(6):1508–1518, 2020.

[68] Shimon Weiss. Measuring conformational dynamics of biomolecules by single molecule fluorescence spectroscopy. *Nature structural biology*, 7(9):724–729, 2000.

[69] Hangping Yao, Yutong Song, Yong Chen, Nanping Wu, Jialu Xu, Chujie Sun, Jiaxing Zhang, Tianhao Weng, Zheyuan Zhang, Zhigang Wu, et al. Molecular architecture of the sars-cov-2 virus. *Cell*, 2020.

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

[71] She Zhang, Fangyuan Chen, and Ivet Bahar. Differences in the intrinsic spatial dynamics of the chromatin contribute to cell differentiation. *Nucleic acids research*, 48(3):1131–1145, 2020.

[72] Yan Zhang, Pemra Doruker, Burak Kaynak, She Zhang, James Krieger, Hongchun Li, and Ivet Bahar. Intrinsic dynamics is evolutionarily optimized to enable allosteric behavior. *Current Opinion in Structural Biology*, 62:14–21, 2020.

[73] Shengyu Zhao, Yue Dong, Eric I Chang, Yan Xu, et al. Recursive cascaded networks for unsupervised medical image registration. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10600–10610, 2019.

[74] Shengyu Zhao, Tingfung Lau, Ji Luo, I Eric, Chao Chang, and Yan Xu. Unsupervised 3d end-to-end medical image registration with volume tweening network. *IEEE journal of biomedical and health informatics*, 24(5):1394–1404, 2019.

[75] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Ev-flownet: Self-supervised optical flow estimation for event-based cameras. *arXiv preprint arXiv:1802.06898*, 2018.