Computing dense pixel-to-pixel image correspondences is a fundamental task of computer vision. Often, the objective is to align image pairs from the same semantic category for manipulation or segmentation purposes. Despite achieving superior performance, existing deep learning alignment methods cannot cluster images; consequently, clustering and pairing images needed to be a separate laborious and expensive step.

Given a dataset with diverse semantic categories, we propose a multi-task model, Jim-Net, that can directly learn to cluster and align images without any pixel-level or image-level annotations. We design a pair-matching alignment unsupervised training algorithm that selectively matches and aligns image pairs from the clustering branch. Our unsupervised Jim-Net achieves comparable accuracy with state-of-the-art supervised methods on benchmark 2D image alignment dataset PF-PASCAL. Specifically, we apply Jim-Net to cryo-electron tomography, a revolutionary 3D microscopy imaging technique of native subcellular structures. After extensive evaluation on seven datasets, we demonstrate that Jim-Net enables systematic discovery and recovery of representative macromolecular structures in situ, which is essential for revealing molecular mechanisms underlying cellular functions. To our knowledge, Jim-Net is the first end-to-end model that can simultaneously align and cluster images, which significantly improves the performance as compared to performing each task alone.

1. Introduction

Image alignment that establishes dense pixel-to-pixel correspondences between images is a fundamental research area in computer vision [86]. Various algorithms have been widely applied to important fields such as face recognition [23, 105], medical diagnosis [85, 21], remote sensing [100, 60], and structural biology [74, 37]. These have successfully led to technological advances including Google street view [5], and historic scientific discoveries including the first image of a black hole [13] and atomic resolution macromolecular structures [9, 106].

Since 2017, end-to-end deep learning methods emerged for image alignment tasks and achieved superior performance in terms of both accuracy and efficiency, especially for the common scenario of image pairs with large transformation variations [91, 101]. In most cases, the objective of image alignment is to align image pairs of the same semantic category for feature localization [54, 90], image manipulation [34, 4] or image segmentation purposes [81, 94, 46]. Nevertheless, the majority of image datasets contain diverse semantic categories [35, 28, 87]. Existing supervised and unsupervised deep learning image alignment approaches only perform alignment and cannot predict the categorical class of input images. As a result, they still need to categorize the images as a separate step, either by manual grouping, which is tedious and time-consuming, or by computational methods, which may be prone to errors [44] due to the large pose variations of objects (Figure 1).

Unsupervised categorization (clustering) has been extensively studied in machine learning settings where the ground-truth categorical labels needed for supervised training are hard to acquire. The accuracy of traditional image clustering algorithms is usually low because of their limited ability in learning representations of high-dimensional visual features [31]. Recently, end-to-end deep image clustering methods have achieved similar performance to supervised classification [15, 39, 96, 72] and demonstrated their ability to extract meaningful semantic features. As mentioned above, most image datasets contain multiple semantic categories but the majority of image alignment applications require categorical information of images. Therefore, it is critical to have a model that can simultaneously cluster and align images. Such a model will enable learning from scratch the clustering and alignment of images from a completely unlabeled dataset with diverse semantic categories.

In particular, we motivate the importance of this prob-
lem for *in situ* cryo-electron tomography (cryo-ET) with the goal to subdivide heterogeneous macromolecular structures into homogeneous structural subsets and align their poses. Cryo-ET (detailed background description in Supplementary section S1) is a rapidly developing 3D variant of 2017 Nobel chemistry prize cryo-Electron Microscopy. Cryo-ET has revolutionized structural biology by enabling systematic 3D visualization of subcellular structures [83]. The data is collected *in situ*, meaning that the subcellular structures are imaged in their native cytoplasm environment. Therefore, the spatial organization of macromolecules and their interactions with organelles can be revealed, which cannot be done by any other imaging methods [11]. With this unique strength, cryo-ET has continually provided new insights into the structure and function of important biological processes including bacterial effector secretion [20, 18, 19, 17] and mammalian neural function [32, 10, 88, 7, 26].

More importantly, cryo-ET has been applied extensively [65, 98, 50, 48, 45] to characterize the structure and replication of SARS-CoV-2 in host cells by localizing individual proteins with high precision, promoting both the scientific understanding and treatment development of COVID-19. However, cryo-ET data analysis is very challenging. Because of the structural heterogeneity and high noise level, to recognize different structures, macromolecules imaged (represented as subtomograms: cubic subvolumes extracted from a tomogram) must be clustered into homogeneous groups and aligned. High-throughput methods that can simultaneously align and cluster 3D cryo-ET macromolecules will substantially benefit the discovery and recovery of higher-resolution structures *in situ* [12].

**Contributions:** (i) We introduce the first end-to-end model, *Jim-Net* [1](Joint image alignment and clustering Network), to jointly perform image alignment and image clustering. (ii) To alleviate the requirement for annotated training data, we design loss functions and a pair-matching alignment training algorithm to make *Jim-Net* fully unsupervised. (iii) We incorporate coarse-to-fine alignment techniques to improve the robustness to large geometric transformations. This includes a novel layer for composing 3D transformations, which leads to fewer deformities and less gradient vanishing/exploding as it acts as a residual connection to the coarse module. (iv) We incorporate spectral pooling, stacked dilated convolution, and constrained cross-correlation techniques to improve the robustness to noise for both tasks. (v) Our unsupervised *Jim-Net* achieve comparable alignment accuracy with state-of-the-art weakly supervised methods on benchmark dataset PF-PASCAL. (vi) *Jim-Net* significantly improves the performance as compared to baseline methods performing each task alone on cryo-ET datasets.

2. Related Work

2.1. End-to-end image alignment

Image alignment is a challenging task [89] because it often focuses on large transformation variations which may result in large changes in objects’ appearance (Figure 1). Many traditional geometric methods have been proposed for pairwise [27, 73, 63] or batch [66, 59] alignment. However, their accuracy and efficiency are still limited [2, 77]. Earlier works used pre-trained [68, 8] or learned [102, 53, 92] feature extractors combined with traditional feature matching methods to generate the dense correspondence. In 2017, the first end-to-end model for image alignment was proposed [77]. With regard to the level of supervision, the end-to-end image alignment methods can be divided into five categories. **Supervised** alignment requires labeled transformation parameters [77, 84] or keypoints [69, 24, 36, 56, 62, 61] to train a network. However, the precise ground truth are hard and time-consuming to prepare, either by exhaustively grid searching the transformation parametric space, or by laborious manual annotation.

https://github.com/xulabs/aitom
Semi-supervised: \[53\] combines L2-norm loss for annotated data and cycle consistency loss for unannotated data. **Weakly-supervised** alignment methods only require categorically matched image pairs or foreground masks \[53, 57\] for training. \[78, 49\] integrate a differentiable transformation module with \[77, 43\] aligns image pairs by learning the forward and backward transformation to be consistent. \[80, 79\] propose neighbourhood consensus networks to learn feature correspondences from known matching and non-matching image pairs. **Self-supervised:** \[89, 84\] demonstrate that image alignment methods could be trained solely in a synthetic data augmentation fashion by randomly transforming an image and learning to regress to the transformation parameters. **Unsupervised:** \[103\] aligns 3D images by a fully unsupervised network trained by inputting arbitrary image pairs.

Yet, all of the methods only learn a single-task alignment model, which cannot predict the semantic cluster assignment of input images. For Jim-Net, we aim to learn a fully unsupervised multi-task model that aligns image pairs and predicts cluster assignments at the same time.

### 2.2. End-to-end unsupervised clustering

Image clustering is also a fundamental problem in computer vision. Unsupervised deep image clustering is substantially more challenging than supervised classification because neural networks must rely on adequate signals for backpropagation \[47\]. We focus on end-to-end models as they are mostly related to our method. JULE \[95\] proposed a recurrent framework to merge close clusters together in a hierarchical agglomerative manner. \[104\] optimizes a probabilistic assignment network to improve the image reconstruction by a mixture of autoencoders. Similarly, \[33\] jointly optimizes reconstruction loss and a clustering oriented loss based on encodings. DeepCluster \[15\] clusters learned features using K-means and uses the cluster labels to fine-tune the model to iteratively improve the feature separation and clustering accuracy. PICCA \[39\] introduces a differentiable index to maximize the global partition confidence of clustering solution, which established the state-of-the-art performance on several benchmark datasets, including over 60% accuracy on CIFAR-10 \[51\].

Unlike these single-task clustering models, Jim-Net has shared feature extractors to jointly learn two tasks, which mutually reinforce each other for better feature learning.

### 3. Method

Our end-to-end model as shown in Figure 2, Jim-Net, learns to propose cluster assignments and transformations by alternating between an alignment and cluster assignment predicting step and a clustering step. For alignment and cluster predicting, a source image \(s\) and a target image \(t\) are processed by an alignment module, \(M_{AL}\) (Section 3.1). \(M_{AL}\) has two outputs, a transformed instance of the source image \(s'\), which has been aligned with \(t\) in a way that minimizes some alignment loss \(L_{AL}\) (Section 3.1.2), and a deep feature representation of \(s\), \(f_{s|t} := F(s, t)\). The deep feature representation \(f_{s|t}\) is passed to a cluster assignment prediction module \(M_{CP}\) that predicts which cluster \(s\) was assigned to during the clustering phase. We backpropagate on a categorical-crossentropy loss \(L_{CE}\) for \(M_{CP}\) and an alignment loss \(L_{AL}\) for \(M_{AL}\) to jointly train \(M_{CP}\), and \(M_{AL}\), which share a feature extractor.

For clustering images through feature learning, Jim-Net runs \(M_{AL}\) on an image \(x\) paired with itself to produce deep features \(f_{x|x} := F(x, x)\). This process is repeated for every image in the dataset \(X\). The resulting feature dataset \(f_{X|X}\) is clustered via a Gaussian Mixture Model (GMM) (Section 3.2) which is used as guidance to pair images and to train the network. Each \(x\) is given a cluster assignment \(\ell_x\). A new, paired dataset \(P := \{(s, t, \ell_s) : (s, t) \in X \land (\ell_s = \ell_t)\}\) i.e. every pair has the same cluster assignment. The source and target image pairs from \(P\) are used as inputs to train \(M_{AL}\) from the previous step. Once Jim-Net is trained, it can be deployed to simultaneously align image pairs and predict image assignment to semantically meaningful clusters.

### 3.1. Image alignment branch

The goal of image alignment is to take a source image \(s\) and a target image \(t\) and geometrically transform \(s\) into \(s'\) in a way that minimizes some dissimilarity loss \(L_{AL}(s', t)\).

Jim-Net employs a coarse-to-fine alignment architecture. Coarse-to-fine alignment architectures propose an image transformation, carry out the transformation, and use the transformed image to propose a new transformation on the already transformed image. This process stacks multiple alignment-proposing functions for progressively finer transformations. Intuitively, by dividing the transformation process into multiple stages, the architecture can propose transformations on different feature resolutions with fine-tuning transformations at later stages, which has been shown to improve the quality of the alignment \[107, 16\].

Formally, the coarse-to-fine approach can be described by defining transformation functions \(T^{(1)}, ..., T^{(m)}\), such that \(T^{(i)} : \mathcal{I} \times \Phi^{(i)} \rightarrow \mathcal{I}\) where \(\mathcal{I}\) is an image space and \(\Phi^{(i)}\) is a parameter space for transformation \(\tilde{T}^{(i)}\); these transformation functions may be different for different \(i\), e.g. \(T^{(1)}\) could be an affine transformation while \(T^{(2)}\) could be a thin plate spline transformation. We also define feature extractors \(f^{(i)} : \mathcal{I} \rightarrow \mathcal{F}^{(i)}\) where \(\mathcal{F}^{(1)}, ..., \mathcal{F}^{(m)}\) are feature spaces with possibly different resolutions. Next, we define transformation parameter proposing functions \(r^{(1)}, ..., r^{(m)}\) that operate on the feature representations such that each of the functions \(r^{(i)} : \mathcal{F}^{(i)} \times \mathcal{F}^{(i)} \rightarrow \Phi^{(i)}\), which are typically composed of a feature matching layer (Section 3.1.1) followed by a regressor. For conciseness, we de-
fine image transforming functions \( g^{(1)}, \ldots, g^{(m)} \) such that \( g^{(i)}(s, t) = T^{(i)}(s, r^{(i)}(f^{(i)}(s), f^{(i)}(t))) \) to both propose and perform the image transformations. We define the intermediate image transformation functions

\[
G^{(i)}(s, t) = \begin{cases} 
  g^{(i)}(G^{(i-1)}(s, t), t), & \text{if } 0 < i \leq m \\
  s, & \text{if } i = 0 
\end{cases} 
\tag{1}
\]

and the feature extraction

\[
F(s, t) = f^{(1)}(G^{(0)}(s, t)) ⊕ \ldots ⊕ f^{(m)}(G^{(m-1)}(s, t)), \tag{2}
\]

where \( ⊕ \) is the concatenation operation. The entire coarse-to-fine alignment module \( M_{AL} \) is defined as \( M_{AL}(s, t) = (F(s, t), G^{(m)}(s, t)) \).

By choosing differentiable functions \( T^{(1)}, \ldots, T^{(m)}, f^{(1)}, \ldots, f^{(m)} \), and \( r^{(1)}, \ldots, r^{(m)} \), the entire coarse-to-fine alignment module, \( M_{AL} \), is differentiable, and thus, we can employ gradient based optimization techniques. There are many transformation functions \( T^{(i)} \) that satisfy the differentiability requirement such as the spatial transformer layer [42]. There are many feature extracting functions \( f^{(1)}, \ldots, f^{(m)} \) and regression functions \( r^{(1)}, \ldots, r^{(m)} \) that satisfy the differentiability requirement.

**Fine transformation regularization loss:** Intuitively, later transformations should be for fine tuning and thus, should be progressively smaller. We enforce this intuition by crafting a regularization loss \( L_R \) to penalize large transformations, which should help stabilize and speed up training. For 3D rigid transformations (cryo-ET data) parameterized by rotation matrix \( R_ϕ \) and shift \( s_ϕ \), the \( L_2^2 \) regularization over a unit sphere \( V \) is derived (Supplementary Section S2) as:

\[
L_R(\hat{ϕ}) = \int_{x \in V} \| (R_ϕ x + s_ϕ) - x \|_2^2 dx \\
= \frac{4\pi}{3} \frac{2}{5} (3 - \text{Trace}(R_ϕ)) + \| s_ϕ \|_2^2 
\tag{3}
\]

### 3.1.1 Feature matching

Establishing correspondences between two feature representations allows architectures to utilize regressors on an integrated representation of the two images. Feature matching is typically accomplished globally [77, 78, 103]. Global feature matching can capture correspondences between any two positions in two feature representations but is computationally expensive and requires lots of memory, so it is usually applied at coarse resolutions in early alignment stages when images are geometrically distant. In contrast, local feature matching [40, 41, 86] can only capture correspondences between nearby positions in feature representations but is spatially much more efficient at higher feature resolutions, so it is integrated into late alignment stages when images are already geometrically close.

**Global feature matching:** The correlation layer for global feature matching [77] computes correlations between each of the indices in the source and target’s feature representation. In 2D settings, it is defined for two \( h \times w \times c \), \( L_2 \) normalized, feature representations \( f^{(s)} \) and \( f^{(t)} \) as follows:

\[
C^{(G)}(f^{(s)}, f^{(t)})_{ijkl} = \langle f^{(s)}_{ij}; f^{(t)}_{kl} \rangle 
\tag{4}
\]

where \( f^{(s)}_{ij} \) and \( f^{(t)}_{kl} \) are the feature vectors (or channels) for locations \((i, j)\) and \((k, l)\) in images \( f^{(s)} \) and \( f^{(t)} \) respectively. Notice, that the output tensor has dimension \( h \times w \times (h \times w) \), which is large for high resolutions.

**Local feature matching:** For fine alignment, the correlation layer for local feature matching computes the correlations between indices at most \( r \) away from each other under the \( L_\infty \) norm. This considerably reduces the dimension of the output and computation cost when \( r \) is small, which makes using higher feature resolutions feasible. The local correlation layer can be defined for \( L_2 \) normalized feature representations \( f^{(s)} \) and \( f^{(t)} \) and radius \( r \) as follows:

\[
C^{(L)}(f^{(s)}, f^{(t)})_{ijkl} = \langle f^{(s)}_{ij}; f^{(t)}_{(i-k)(j-l)} \rangle 
\tag{5}
\]
where \( \|(k, l)\|_\infty \leq r \). The output dimension of this map is \( h \times w \times (2r + 1)^2 \), where typically \( (2r + 1) \ll h, w \).

### 3.1.2 Image alignment loss functions

An image alignment loss function, \( \mathcal{L}_{\text{AL}} \), numerically evaluates how well two images are aligned. Unsupervised losses that do not require correspondence maps were proposed including contextual loss \([67]\) and cycle-consistency loss \([43]\). The soft inlier loss for 2D image alignment is outlined in Supplementary Section S2. For 3D cryo-ET data, we design a constrained cross-correlation loss inspired from traditional geometry-based subtomogram alignment methods.

**Constrained cross-correlation loss:** Cross-correlation and its variants are commonly used to assess the alignment between two 3D subtomograms \([23, 22, 3]\). Considering (1) the missing wedge effect introduced by the limited view-angle to reconstruct the 3D tomogram and (2) the low SNR due to the thickness of the cell sample, we implement a constrained cross-correlation measure as the loss function for subtomogram alignment. This loss function is constrained to observed regions in the Fourier space lower function for subtomogram alignment. This loss function is numerically evaluated in Supplementary Section S2. For 3D cryo-ET data, we design a constrained cross-correlation loss inspired from traditional geometry-based subtomogram alignment methods.

\[
\mathcal{L}_{\text{CCC}}(s, t) = 1 - \frac{\sum_{i=1}^{N}(s^*_i - \hat{s}^*) (t^*_i - \hat{t}^*)}{\sqrt{\sum_{i=1}^{N}(s^*_i - \hat{s}^*)^2} \sqrt{\sum_{i=1}^{N}(t^*_i - \hat{t}^*)^2}},
\]

where \( s \) is a source image and \( t \) is a target image. \( s^* \) and \( t^* \) denote the real space subtomogram with coefficients constrained to the intersection of two missing wedge regions and the low-pass filter. \( s^* \) (similarly for \( t^* \)) is mathematically defined as: \( s^* = \mathcal{R}\{\mathcal{F}^{-1}[[\mathcal{F}s] \cdot \mathcal{H} \cdot \mathcal{T}_{R_0}(\mathcal{M}_s)]\} \), where \( \mathcal{R} \) is the real part of a complex function; \( \mathcal{F} \) is the Fourier transform operator; \( \mathcal{T}_{R_0} \) is the 3D rotation operator with parameters from the network regressor; \( \mathcal{H} \) is the low-pass filter defined by \( \mathcal{H}(\xi) = \frac{1}{|\xi|} \mathcal{H}(\xi) \), where \( \xi \) is a 3D location of the Fourier coefficient in the Fourier space, and \( t \) is the frequency threshold for the low-pass filter; \( \mathcal{M}_s \) (similarly for \( \mathcal{M}_t \)) is the binary mask indicating the observed region of \( s \) due to the missing wedge effect as defined by \( \mathcal{M}_s(\xi) = \frac{1}{|\xi|} \mathcal{H}(\xi) \), where \( \theta \) refers to the tilt-angle range \( \pm \theta \) in single-tilt cryo-ET.

### 3.2. Image clustering branch

Manually categorizing images is slow and expensive, so it is ideal to create methods that predict semantically meaningful categories (or clustering assignments) without supervision. We accomplish this by using a GMM (Supplementary Section S2) on learned features as guidance to train Jim-Net: (1) for alignment \( \mathcal{M}_{AL} \), we use image pairs that belong to the same GMM clusters as inputs; (2) for cluster assignment prediction \( \mathcal{M}_{CP} \), we use cluster labels estimated by the GMM as training labels in each iteration. We note that GMM clustering is only used in training. The trained Jim-Net predicts cluster assignment directly.

**Image pairing:** GMM clustering considers feature covariance information which is important for discriminating semantic classes by neural networks \([11, 99]\). During the clustering step, Jim-Net applies a GMM, \( \mathcal{G}_{\mu, \Sigma} \), to cluster the feature representations of the images, \( \mathcal{F}_s, \mathcal{F}_t \). Intuitively, if \( s \) is assigned the same cluster \( \ell_s := \mathcal{G}_{\mu, \Sigma}(s) \) as \( t \), then \( s, t \) should be semantically similar, so they can be aligned meaningfully. Following this intuition, Jim-Net creates a source and target pair dataset \( P = \{(s, t, \ell_s) : (s, t \in X) \land (\ell_s = \ell_t)\} \) as inputs to train Jim-Net (Algorithm 1).

**Algorithm 1:** Pair-matching alignment training

Input: Image dataset \( X \) 

1. for \( m \) iterations:
   2. Feature representations: \( \mathcal{F}_X \leftarrow \mathcal{F}(X, X) \)
   3. GMM: Fit \( \mathcal{G}_{\mu, \Sigma} \) to \( \mathcal{F}_X \)
   4. Cluster predictions: \( \ell_s \leftarrow \mathcal{G}_{\mu, \Sigma}(\mathcal{F}_X) \)
   5. Pairs: \( P \leftarrow \{(s, t, \ell_s) : (s, t \in X) \land (\ell_s = \ell_t)\} \)
   6. for \( (s, t, \ell_s) \in P \):
     7. Update model weights by loss \( \mathcal{L}(s, t, \ell_s) \)

**Joint training:** Jim-Net learns feature extractors \( f^{(1)}, ..., f^{(m)} \) by jointly optimizing \( \mathcal{L}_{\text{AL}} \) and \( \mathcal{L}_{\text{CE}} \). By jointly optimizing both \( \mathcal{L}_{\text{AL}} \) and \( \mathcal{L}_{\text{CE}} \), Jim-Net optimizes \( f^{(1)}, ..., f^{(m)} \) to output feature representations that are useful for alignment and are easily separable. Optimizing the clustering branch helps Jim-Net maintain semantically meaningful categories, which are used to propose image pairs that are amenable to reinforce the alignment branch. We define the combined loss for Jim-Net as:

\[
\mathcal{L}(s, t, \ell_s) = \mathcal{L}_{\text{CE}}(\mathcal{M}_{CP}(\mathcal{F}(s, t)), \ell_s) + \sum_{i=1}^{m} \gamma_i \mathcal{L}_{\text{AL}}(G^{(i)}(s, t), t),
\]

where coefficients \( \gamma_1, ..., \gamma_m \) determine the linear weight placed on each of the intermediate alignments. The alignment module \( \mathcal{M}_{AL} \) and the cluster predicting module \( \mathcal{M}_{CP} \) share feature extractors \( f^{(1)}, ..., f^{(m)} \), thus during backpropagation the classification loss \( \mathcal{L}_{\text{CE}} \) and the alignment losses \( \mathcal{L}_{\text{AL}} \) jointly influence the feature extractors’ update step. We keep the regularization loss \( \mathcal{L}_{\text{R}}(\theta) \) as an optional parameter.

We design an end-to-end multi-task learning Convolutional Neural Network (CNN) architecture (detailed illustration in Supplementary Section S2) according to the proposed model and algorithm.

### 4. Experimental Validation

We systematically evaluated Jim-Net on several datasets for both tasks and analyzed the results (training and results details in Supplementary Section S3).

3858
4.1. Benchmark datasets

Simulated cryo-ET subtomograms: [103] proposed five benchmark realistically simulated cryo-ET datasets following standard procedure. They contain 3D grayscale heterogeneous structures (spliceosome, RNA polymerase, rifampicin complex, RNA polymerase II elongation complex, ribosome, and capped proteasome) at five different SNR levels (100, 0.1, 0.05, 0.03, 0.01). In addition, [103] provided the results of two popular traditional geometry-based subtomogram alignment methods on these datasets.

PF-PASCAL: This 2D RGB dataset [35] contains 1351 images pairs from 20 semantic categories with challenging object appearance variations. In our experiments, the dataset contains 2702 single images as the paired information was not used for training. We split it into training, validation, and testing datasets the same way as proposed in [35].

4.2. Real cryo-ET datasets

Air-water interface single-particle: This dataset consists of tomograms to study the macromolecule (purified through biochemical means) distribution within vitrified ice and air-water interface [71]. We manually picked 2800 subtomograms. There are in total 400 rabbit muscle aldolase (AI), 400 glutamate dehydrogenase (GD), 400 DNAB helicase, helicase loader (Hel), 400 T20S proteasome (T20S), 400 apoferitin (Ap), 400 hemagglutinin (Hem), and 400 insulin-bounded insulin receptor (Ins) subtomograms. Each subtomogram is of size $32^3$ with voxel spacing 8.75 Å, imaged with a 45° missing wedge.

Synecochystis cell: This dataset contains two cellular tomograms of the cyanobacterium Synecochystis to study the photosynthetic machinery during the biogenesis of thylakoid membranes [70]. We extract 12912 subtomograms from the two tomograms using the difference of Gaussian particle picking method. A subtomogram has size $32^3$ with voxel spacing 13.68 Å, imaged with a 30° missing wedge.

4.3. Results

Cryo-ET Benchmark: For alignment, these simulated datasets have ground truth 3D rigid-body transformation parameters available. These ground truth transformation parameters are unique as the structures chosen are asymmetric. Similar to the single-task baseline methods [23], [103], we evaluated the alignment accuracy by measuring the translation error and rotation error separately, defined as the $L_2$ distance between the estimated 3D translation, rotation matrix, and the ground truth, respectively. Table I shows the alignment accuracy on the five simulated datasets. Similar to Gum-Net, Jim-Net was instantiated with random weights. Jim-Net achieved the overall best performance, outperforming Gum-Net, the state-of-the-art for subtomogram alignment.

For image clustering, the simulated datasets have pre-specified labels for the macromolecular structures contained in each subtomogram. We compared Jim-Net with three single-task baseline methods: (1) DeepCluster [15], the most popular unsupervised image clustering method, which enabled end-to-end training of visual features on large-scale
datasets; (2) PICA [39], a recent end-to-end unsupervised image clustering method that established the state-of-the-art performance on several benchmark datasets; (3) Jim-Net (cluster), an ablation baseline with the clustering branch of Jim-Net only. DeepCluster and PICA were extended to 3D image data. We evaluated the clustering accuracy by the percentage of correctly predicted subtomograms. We used the Hungarian algorithm (linear assignment) [52] to match cluster labels with ground truth to calculate the accuracy. Jim-Net achieved the best performance on all datasets with large margins (Table 2), especially for datasets with low SNR, showing the robustness of Jim-Net. We visualized examples of Jim-Net alignment and clustering in Figure 3. We note that since the benchmark datasets contain separate training, validation, and testing datasets, this ensures that there is no over-fitting by any of the learning-based methods. Moreover, as Jim-Net is completely unsupervised, when a new dataset is available for alignment and clustering, a trained Jim-Net model can be readily applied and fine-tuned on the new dataset as no annotated training data is required.

### Air-water interface single-particle
This dataset has been manually categorized into 7 different structural categories of macromolecules. Therefore, we first evaluated the clustering performance of Jim-Net as compared to baseline methods (Table 3). A common problem of end-to-end unsupervised clustering methods is the degeneration of clusters. This phenomenon is also known as algorithm-agnostic trivial solutions [39], of which the model learns discriminative features between only a few clusters and results in low accuracy of other clusters. Clearly, the degeneration of clusters is only observed in baseline methods. Jim-Net obtained high accuracy across all clusters because the shared feature extractors between the alignment and clustering branches learned robust features to avoid trivial solutions.

<table>
<thead>
<tr>
<th>Method</th>
<th>Al</th>
<th>Od</th>
<th>Hcl</th>
<th>T20S</th>
<th>Ap</th>
<th>Mem</th>
<th>Ins</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepCluster</td>
<td>75.0</td>
<td>97.5</td>
<td>99.9</td>
<td>99.9</td>
<td>96.1</td>
<td>85.4</td>
<td>100</td>
<td>95.0</td>
</tr>
<tr>
<td>PICA [39]</td>
<td>70.8</td>
<td>92.0</td>
<td>99.0</td>
<td>99.0</td>
<td>92.0</td>
<td>85.0</td>
<td>100</td>
<td>95.0</td>
</tr>
<tr>
<td>Jim-Net (cluster)</td>
<td>67.2</td>
<td>97.9</td>
<td>98.0</td>
<td>98.0</td>
<td>92.0</td>
<td>90.0</td>
<td>100</td>
<td>85.4</td>
</tr>
<tr>
<td>Jim-Net</td>
<td>97.5</td>
<td>97.3</td>
<td>89.3</td>
<td>95.5</td>
<td>99.0</td>
<td>97.8</td>
<td>100</td>
<td>96.6</td>
</tr>
<tr>
<td>Resolution</td>
<td>35.3</td>
<td>17.8</td>
<td>31.8</td>
<td>42.8</td>
<td>17.6</td>
<td>28.6</td>
<td>17.3</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 3. Clustering accuracy on air-water interface dataset. The last row shows the resolution of template estimation by Jim-Net.

Because of the difficulty of manual alignment and the symmetry in some structures, there is no alignment ground truth for real cryo-ET datasets. Instead, we evaluated Jim-Net alignment by the standard alignment-based template estimation [14]. Because Jim-Net predicts cluster assignment, we are able to calculate resolution inside each cluster, which cannot be accomplished by the previous unsupervised alignment method [103]. Raw tomograms usually have a resolution worse than 70 Å [30]. By aligning and averaging multiple copies of the same type of macromolecules, the resolution can be improved to better than 40 Å to sufficiently recognize different macromolecular structures. Table 3 reports the resolution as measured by the gold-standard Fourier shell correlation [64]. Jim-Net alignment successfully attained resolution better than 40 Å for most structures.

**Synechocystis cell**: As *in situ* cryo-ET images subcellular structures in their native cytoplasm environment, 3D image alignment and clustering is critical for discovering and recovering representative and abundant structures. The 12912 synechocystis cell subtomograms were clustered and aligned by Jim-Net for template estimation. The results (Figure 4) were evaluated by the resolution measurement and interpreted by structural biology experts.

Several representative structures were discovered and recovered by alignment-based template estimation from clusters. As shown in Figure 4, we recognized structures likely to be (1) array-forming phycobilisomes that are thylakoid-
bounded light-harvesting antennas for photosystem II of photosynthetic electron flow; (2) free cytosolic 70S ribosomes that synthesize photosynthetic membrane proteins; (3) membrane structures that correspond to cell membrane single layer and thylakoid membrane carrying out the light reactions of photosynthesis; (4) ice contaminants. The structural discovery was validated by embedding them back to the tomogram space. Moreover, on the ribosome structure, a basic macromolecule in all living cells, we measured the resolution as 36.48 Å. This is the first time to recover in situ structures with resolution better than 40 Å by a completely automatic and unsupervised method.

**PF-PASCAL:** The standard alignment metric, percentage of correct keypoints (PCK), measures the number of keypoints with a transfer error below 0.1 [97]. We listed state-of-the-art self- or weakly-supervised methods that reported their PCK on PF-PASCAL in Table 3. The comparison is fair as all methods used the same ResNet-101 [38] feature extraction backbone instantiated with ImageNet [82] weights. A higher PCK indicates better alignment.

<table>
<thead>
<tr>
<th>Method</th>
<th>PCK</th>
</tr>
</thead>
<tbody>
<tr>
<td>A2-Net (S)</td>
<td>70.8</td>
</tr>
<tr>
<td>WeakAlign (W, L)</td>
<td>75.8</td>
</tr>
<tr>
<td>RTNs (W, L)</td>
<td>75.9</td>
</tr>
<tr>
<td>NC-Net (W, L)</td>
<td>78.9</td>
</tr>
<tr>
<td>SF-Net (W, M)</td>
<td>81.9</td>
</tr>
<tr>
<td>DCC-Net (W, L)</td>
<td>82.3</td>
</tr>
<tr>
<td>Jim-Net (U)</td>
<td>74.8</td>
</tr>
</tbody>
</table>

Table 4. PCK on the PF-PASCAL benchmark. Baseline results were directly taken from corresponding papers. S: self-supervised. W: weakly-supervised with: L: image labels; M: foreground masks. U: unsupervised.

Compared with the self-supervised method A2-Net [84] (essentially unsupervised), Jim-Net achieved better performance. Table S11 showed that Jim-Net achieved the highest PCK on 6 of the 20 classes compared to four SOTA weakly-supervised baselines. Without using categorical labels or foreground masks, unsupervised Jim-Net achieved close performance to those weakly-supervised methods. Note that PF-PASCAL is a highly imbalanced dataset (e.g., 140 bus pairs vs. 6 sheep pairs). If the number of samples from each class changes, the overall PCK would change for different methods. Therefore, the number of class-wise highest PCK is a better evaluation criterion.

We provide visualizations of Jim-Net alignment and clustering results for validation in Supplementary Section S3. For (1) no clustering methods have been reported on PF-PASCAL, an alignment evaluation dataset, and (2) fair comparison as Jim-Net feature extraction backbone has been pre-trained with the same weight as all baseline methods. The pre-trained weights may already contain semantic information. Therefore, we did not compare 2D clustering baselines. However, the initial clustering accuracy was 64% and steadily improved to 73% on the testing dataset, demonstrating the learning ability and unique strength of the clustering branch. We note that on the main cryo-ET application, 3D Jim-Net was initialized randomly and fairly compared with clustering baselines.

5. Discussion

Jointly learning alignment and clustering is significant as the two tasks are related for three major reasons: (1) joint alignment and clustering is the most essential and challenging cryo-ET analysis task. Such end-to-end joint learning can be easily extended to related biomedical fields including CT and MRI data sub-atlas registration [29, 75], and reconstruction of multiple conformations in single-particle cryo-EM [70]. (2) For alignment, the objective in most cases is to align image pairs of the same semantic category. Without clustering, the direct alignment is ill-posed as it requires prior categorical information. Joint learning not only alleviates this requirement but also improves the alignment accuracy, as demonstrated by the experimental comparison with Gum-Net baseline [103]. (3) For clustering, joint learning substantially improved the performance and prevented a common issue, the degeneration of clusters. When deployed, each Jim-Net branch can be executed concurrently or can be singled out for one particular task.

6. Conclusion

In this paper, we propose an unsupervised multi-task learning model to tackle an important problem in computer vision: joint learning of image alignment and image clustering. As the first end-to-end model for this problem, Jim-Net is thoroughly evaluated on both 3D cryo-ET and 2D natural image benchmark datasets. While attaining similar performance, Jim-Net alleviates the need for weak-supervision of state-of-the-art image alignment methods. Moreover, unsupervised alignment and clustering on in situ cryo-ET data specifically enable ‘visual proteomics’ that provides systematic recognition and recovery of macromolecular structures and distributions, which critically facilitates the understanding of molecular machinery of cellular processes.

Acknowledgment

This work was supported in part by U.S. NIH grants R01GM134020 and P41GM103712, NSF grants DBI-1949629 and IIS-2007595, and Mark Foundation for Cancer Research 19-044-ASP. We thank the computational resources support from AMD COVID-19 HPC Fund and Dr. Zachary Freyberg’s lab. X.Z. was supported in part by a fellowship from CMU CMLH. We thank Dr. Anson Kahng, Hongyu Zheng, Conor Igoe, and Samarth Malhotra for critical comments on the manuscript.
References


[26] Karen M Davies, Mike Strauss, Bertram Daum, Jan H Kief, Heinz D Osiewacz, Adriana Rykovska, Volker Zickermann, and Werner Kühlbrandt. Macromolecular organization of atp synthase and complex i in whole mitochon-


Michael Rubinstein, Armand Joulin, Johannes Kopf, and Ce Liu. Unsupervised joint object discovery and segmen-


