# End-to-end robust joint unsupervised image alignment and clustering

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

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

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

## Abstract

*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 [25, 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-

*Corresponding author. †Equal contribution.

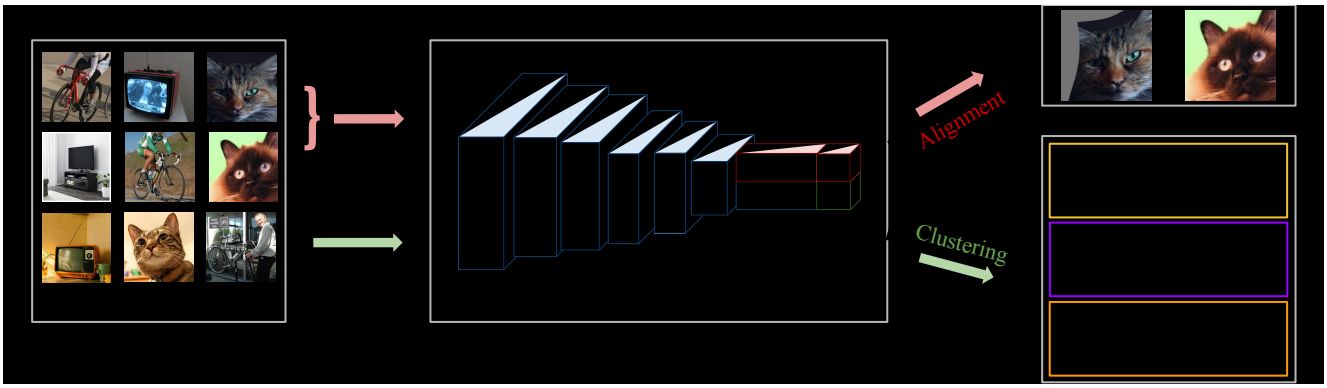


Figure 1. Illustration of Jim-Net: a multi-task learning model that can align an image pair and predict the cluster assignment of the source image. The red arrows denote the alignment branch and the green arrows denote the clustering branch.

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] (*J*oint *i*mage align*m*ent and clustering *Net*work), 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.

[1] https://github.com/xulabs/aitom

## 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.

**Semi-supervised:** [55] combines $L_2$-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 [58, 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. PICA [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, $M_{\text{AL}}$ (Section 3.1). $M_{\text{AL}}$ 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 $\mathcal{L}_{\text{AL}}$ (Section 3.1.2), and a deep feature representation of $s$, $f_{s|t} \coloneqq F(s,t)$. The deep feature representation $f_{s|t}$ is passed to a cluster assignment prediction module $M_{\text{CP}}$ that predicts which cluster $s$ was assigned to during the clustering phase. We backpropogate on a categorical-crossentropy loss $\mathcal{L}_{\text{CE}}$ for $M_{\text{CP}}$ and an alignment loss $\mathcal{L}_{\text{AL}}$ for $M_{\text{AL}}$ to jointly train $M_{\text{CP}}$, and $M_{\text{AL}}$, which share a feature extractor.

For clustering images through feature learning, Jim-Net runs $M_{\text{AL}}$ on an image $x$ paired with itself to produce deep features $f_{x|x} \coloneqq F(x,x)$. 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 $P \coloneqq \{(s,t,\ell_s) : (s,t \in X) \wedge (\ell_s = \ell_t)\}$ *i.e*. every pair has the same cluster assignment. The source and target image pairs from $P$ are used as inputs to train $M_{\text{AL}}$ 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 $\mathcal{L}_{AL}(s',t)$.

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 $\mathcal{T}^{(1)},...,\mathcal{T}^{(m)}$ such that $\mathcal{T}^{(i)} : \mathcal{I} \times \Phi^{(i)} \rightarrow \mathcal{I}$ where $\mathcal{I}$ is an image space and $\Phi^{(i)}$ is a parameter space for transformation $\mathcal{T}^{(i)}$; these transformation functions may be different for different $i$, *e.g*. $\mathcal{T}^{(1)}$ could be an affine transformation while $\mathcal{T}^{(2)}$ 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-

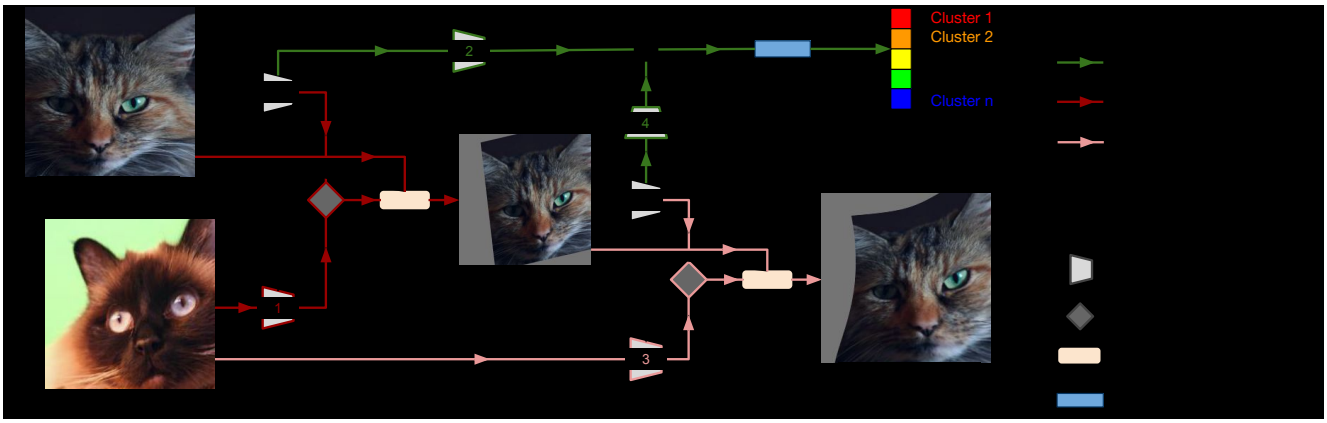


Figure 2. Jim-Net workflow. The clustering and the alignment shares the feature extractors for the source image. The alignment branch employs a coarse-to-fine alignment strategy. Different feature extractors are indexed (architecture in Supplementary Section S2).

fine image transforming functions $g^{(1)}, ..., g^{(m)}$ such that $g^{(i)}(s,t) = \mathcal{T}^{(i)}(s, r^{(i)}(f^{(i)}(s), f^{(i)}(t)))$ to both propose and perform the image transformations. We define the intermediate image transformation functions

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

and the feature extraction

$$F(s,t) = f^{(1)}(G^{(0)}(s,t)) \oplus ... \oplus f^{(m)}(G^{(m-1)}(s,t)), \quad (2)$$

where $\oplus$ is the concatenation operation. The entire coarse-to-fine alignment module $M_{\text{AL}}$ is defined as $M_{\text{AL}}(s,t) = (F(s,t), G^{(m)}(s,t))$. By choosing differentiable functions $\mathcal{T}^{(1)}, ..., \mathcal{T}^{(m)}$, $f^{(1)}, ..., f^{(m)}$, and $r^{(1)}, ..., r^{(m)}$, the entire coarse-to-fine alignment module, $M_{\text{AL}}$, is differentiable, and thus, we can employ gradient based optimization techniques. There are many transformation functions $\mathcal{T}^{(i)}$ that satisfy the differentiablity requirement such as the spatial transformer layer [42]. There are many feature extracting functions $f^{(1)}, ..., f^{(m)}$ and regression functions $r^{(1)}, ..., r^{(m)}$ 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 $\mathcal{L}_{\text{R}}$ to penalize large transformations, which should help stabilize and speed up training. For 3D rigid transformations (cryo-ET data) parameterized by rotation matrix $R_\phi$ and shift $s_\phi$, the $L_2^2$ regularization over a unit sphere $V$ is derived (Supplementary Section S2) as:

$$\begin{aligned} \mathcal{L}_{\text{R}}(\phi) &= \int_{x \in V} ||(R_\phi x + s_\phi) - x||_2^2 dx \\ &= \frac{4\pi}{3}(\frac{2}{5}(3 - \text{Trace}(R_\phi)) + ||s_\phi||_2^2) \end{aligned} \quad (3)$$

### 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, 6] 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:

$$C^{(G)}(f^{(s)}, f^{(t)})_{ijkl} = \langle f^{(s)}_{ij:}, f^{(t)}_{kl:} \rangle, \quad (4)$$

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:

$$C^{(L)}(f^{(s)}, f^{(t)})_{ijkl} = \langle f^{(s)}_{ij:}, f^{(t)}_{(i-k)(j-l):} \rangle, \quad (5)$$

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-angles 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 than a frequency cut-off threshold. This is because the high-frequency region is likely to be dominated by noise. The constrained cross-correlation loss is defined as follows:

$$\mathcal{L}_{\text{CCC}}(s,t) = 1 - \frac{\sum_{i=1}^{N}(s_i^* - \bar{s^*})(t_i^* - \bar{t^*})}{\sqrt{\sum_{i=1}^{N}(s_i^* - \bar{s^*})^2}\sqrt{\sum_{i=1}^{N}(t_i^* - \bar{t^*})^2}}, \tag{6}$$

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: $s^* = \mathcal{R}\{\mathcal{F}^{-1}[(\mathcal{F}s) \cdot \mathcal{H} \cdot \mathcal{M}_t \cdot \mathcal{T}_{R_\phi}(\mathcal{M}_s)]\}$, where $\mathcal{R}$ is the real part of a complex function; $\mathcal{F}$ is the Fourier transform operator; $\mathcal{T}_{R_\phi}$ is the 3D rotation operator with parameters from the network regressor; $\mathcal{H}$ is the low-pass filter defined by $\mathcal{H}(\xi) = \mathbb{1}_{||\xi||_\infty \leq t}(\xi)$, 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 $\mathcal{M}_s(\xi) = \mathbb{1}_{|\xi_3| \leq |\xi_1| \tan(\theta)}(\xi)$, 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 $M_{\text{AL}}$, we use image pairs that belong to the same GMM clusters as inputs; (2) for cluster assignment prediction $M_{\text{CP}}$, 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 [1, 99]. During the clustering step, Jim-Net applies a GMM, $\mathcal{G}_{\mu,\Sigma}$, to cluster the feature representations of the images, $f_{X|X}$. 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 $P = \{(s,t,\ell_s) : (s,t \in X) \wedge (\ell_s = \ell_t)\}$ as inputs to train Jim-Net (Algorithm 1).

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

**Input:** Image dataset $X$
**for** m iterations:
  Feature representations: $f_{X|X} \leftarrow F(X,X)$
  GMM: Fit $\mathcal{G}_{\mu,\Sigma}$ to $f_{X|X}$
  Cluster predictions: $\ell_X \leftarrow \mathcal{G}_{\mu,\Sigma}(f_{X|X})$
  Pairs: $P \leftarrow \{(s,t,\ell_s) : (s,t \in X) \wedge (\ell_s = \ell_t)\}$
  **for** (s, t, $\ell_s$) $\in P$ :
    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:

$$\mathcal{L}(s,t,\ell_s) = \mathcal{L}_{\text{CE}}(M_{\text{CP}}(F(s,t)),\ell_s) + \sum_{i=1}^{m}\gamma_i \mathcal{L}_{\text{AL}}(G^{(i)}(s,t),t), \tag{7}$$

where coefficients $\gamma_1, ..., \gamma_m$ determine the linear weight placed on each of the intermediate alignments. The alignment module $M_{\text{AL}}$ and the cluster predicting module $M_{\text{CP}}$ share feature extractors $f^{(1)}, ..., f^{(m)}$, thus during backpropogation 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 $\mathcal{L}_{\text{R}}(\phi)$ 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).

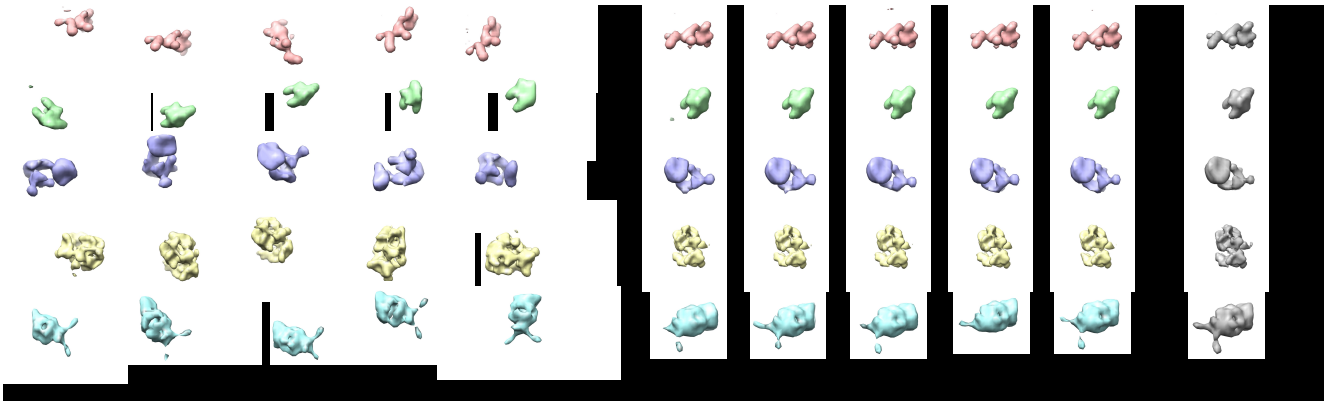

Figure 3. Example of isosurface representation of Jim-Net alignment on SNR 100 dataset. A random subtomogram from each cluster was chosen as the target subtomogram and the rest subtomograms from the same cluster were aligned to it.

| Method | SNR 100 | SNR 0.1 | SNR 0.05 | SNR 0.03 | SNR 0.01 |
|---|---|---|---|---|---|
| H-T align [93] | 0.30±0.68, 1.82±2.69 | 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 [23] | 0.33±0.70, 1.93±2.86 | 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 [103] | 0.41±0.70, 1.59±2.63 | 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 | **0.29±0.53**, **1.28±2.10** | **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** |

Table 1. Subtomogram alignment accuracy on benchmark datasets. In each cell (best results highlighted), the first term is the mean and standard deviation of the rotation error and the second term, the translation error. Baseline results were directly taken from [103].

### 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 image 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 [36].

### 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 (Al), 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 $45^\circ$ missing wedge.

**Synechocystis cell:** This dataset contains two cellular tomograms of the cyanobacterium *Synechocystis* to study the photosynthetic machinery during the biogenesis of thylakoid membranes [76]. 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 $30^\circ$ 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 [93, 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 1 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.

| Method | SNR 100 | 0.1 | 0.05 | 0.03 | 0.01 |
|---|---|---|---|---|---|
| DeepCluster [15] | 68.7 | 48.8 | 39.4 | 34.0 | 27.2 |
| PICA [39] | 100 | 86.5 | 55.8 | 29.2 | 28.4 |
| Jim-Net (cluster) | 99.5 | 77.5 | 62.8 | 51.3 | 35.3 |
| Jim-Net | **100** | **99.7** | **96.1** | **85.5** | **48.1** |

Table 2. Subtomogram clustering accuracy (best results highlighted) on benchmark datasets.

**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.

| Method | Al | GD | Hel | T20S | Ap | Hem | Ins | Overall |
|---|---|---|---|---|---|---|---|---|
| DeepCluster [15] | 12.5 | **100** | 0 | 97.3 | 0 | **99.0** | 95.0 | 57.7 |
| PICA [39] | 69.8 | 64.0 | 26.0 | 51.5 | 47.2 | 84.5 | 80.5 | 60.5 |
| Jim-Net (cluster) | 75.0 | **100** | 35.0 | **99.7** | 98.0 | 90.2 | **100** | 85.4 |
| Jim-Net | **97.5** | 97.3 | **89.3** | 95.5 | **99.0** | 97.8 | **100** | **96.6** |
| Resolution | 35.3 | 17.8 | 31.5 | 42.5 | 17.6 | 28.6 | 17.5 | - |

Table 3. Clustering accuracy on air-water interface dataset. The last raw 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.

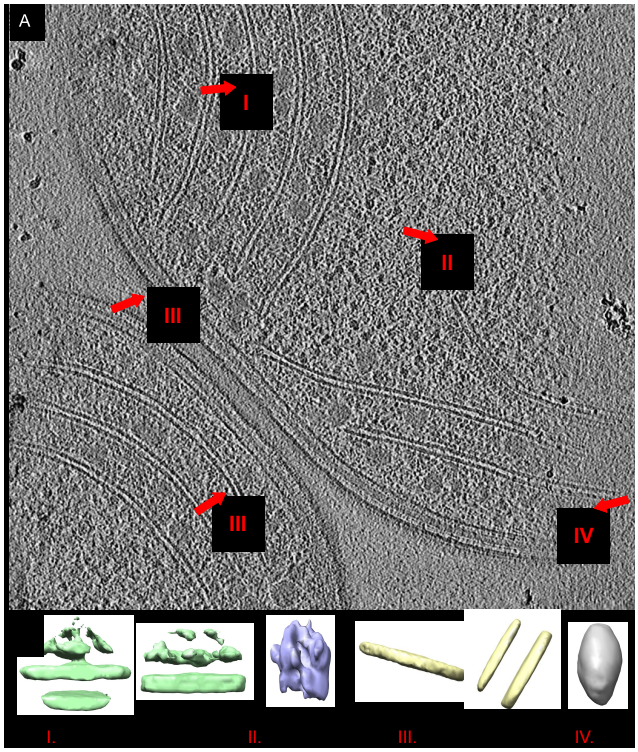


Figure 4. A. a 2D slice of the 3D synechocystis cellular tomogram. B. structures discovered by Jim-Net. C. structures embedded to the tomogram space by Jim-Net alignment branch.

**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 4. 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.

| Method | PCK |
|---|---|
| A2-Net (S) [84] | 70.8 |
| WeakAlign (W, L) [78] | 75.8 |
| RTNs (W, L) [49] | 75.9 |
| NC-Net (W, L) [80] | 78.9 |
| SF-Net (W, M) [57, 58] | 81.9 |
| DCC-Net (W, L) [40] | 82.3 |
| Jim-Net (U) | 74.8 |

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


[1] Dinesh Acharya, Zhiwu Huang, Danda Pani Paudel, and Luc Van Gool. Covariance pooling for facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 367–374, 2018.

[2] Ebtsam Adel, Mohammed Elmogy, and Hazem Elbakry. Image stitching based on feature extraction techniques: a survey. *International Journal of Computer Applications*, 99(6):1–8, 2014.

[3] Fernando Amat, Luis R Comolli, Farshid Moussavi, John Smit, Kenneth H Downing, and Mark Horowitz. Subtomogram alignment by adaptive fourier coefficient thresholding. *Journal of structural biology*, 171(3):332–344, 2010.

[4] Brian Amberg, Andrew Blake, and Thomas Vetter. On compositional image alignment, with an application to active appearance models. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1714–1721. IEEE, 2009.

[5] Dragomir Anguelov, Carole Dulong, Daniel Filip, Christian Frueh, Stéphane Lafon, Richard Lyon, Abhijit Ogale, Luc Vincent, and Josh Weaver. Google street view: Capturing the world at street level. *Computer*, 43(6):32–38, 2010.

[6] Vasileios Argyriou. Asymmetric bilateral phase correlation for optical flow estimation in the frequency domain. In *2018 14th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pages 166–173. IEEE, 2018.

[7] Shoh Asano, Yoshiyuki Fukuda, Florian Beck, Antje Aufderheide, Friedrich Förster, Radostin Danev, and Wolfgang Baumeister. A molecular census of 26s proteasomes in intact neurons. *Science*, 347(6220):439–442, 2015.

[8] Aayush Bansal, Bryan Russell, and Abhinav Gupta. Marr revisited: 2d-3d alignment via surface normal prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5965–5974, 2016.

[9] Alberto Bartesaghi, Cecilia Aguerrebere, Veronica Falconieri, Soojay Banerjee, Lesley A Earl, Xing Zhu, Nikolaus Grigorieff, Jacqueline LS Milne, Guillermo Sapiro, Xiongwu Wu, et al. Atomic resolution cryo-em structure of $\beta$-galactosidase. *Structure*, 26(6):848–856, 2018.

[10] Felix JB Bäuerlein, Itika Saha, Archana Mishra, Maria Kalemanov, Antonio Martínez-Sánchez, Rüdiger Klein, Irina Dudanova, Mark S Hipp, F Ulrich Hartl, Wolfgang Baumeister, et al. In situ architecture and cellular interactions of polyq inclusions. *Cell*, 171(1):179–187, 2017.

[11] Martin Beck and Wolfgang Baumeister. Cryo-electron tomography: can it reveal the molecular sociology of cells in atomic detail? *Trends in cell biology*, 26(11):825–837, 2016.

[12] Jan Böhning and Tanmay AM Bharat. Towards high-throughput in situ structural biology using electron cryotomography. *Progress in Biophysics and Molecular Biology*, 2020.

[13] Katherine Bouman. Capturing the first image of a black hole. *Bulletin of the American Physical Society*, 2020.

[14] John AG Briggs. Structural biology in situ-the potential of subtomogram averaging. *Current opinion in structural biology*, 23(2):261–267, 2013.

[15] 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.

[16] Che-Han Chang, Chun-Nan Chou, and Edward Y Chang. Clkn: Cascaded lucas-kanade networks for image alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2213–2221, 2017.

[17] Yi-Wei Chang, Songye Chen, Elitza I Tocheva, Anke Treuner-Lange, Stephanie Löbach, Lotte Søgaard-Andersen, and Grant J Jensen. Correlated cryogenic photoactivated localization microscopy and cryo-electron tomography. *Nature methods*, 11(7):737, 2014.

[18] Yi-Wei Chang, Andreas Kjær, Davi R Ortega, Gabriela Kovacikova, John A Sutherland, Lee A Rettberg, Ronald K Taylor, and Grant J Jensen. Architecture of the vibrio cholerae toxin-coregulated pilus machine revealed by electron cryotomography. *Nature microbiology*, 2(4):16269, 2017.

[19] Yi-Wei Chang, Andreas Kjær, Davi R Ortega, Gabriela Kovacikova, John A Sutherland, Lee A Rettberg, Ronald K Taylor, and Grant J Jensen. Architecture of the vibrio cholerae toxin-coregulated pilus machine revealed by electron cryotomography. *Nature microbiology*, 2(4):16269, 2017.

[20] Yi-Wei Chang, Lee A Rettberg, Anke Treuner-Lange, Janet Iwasa, Lotte Søgaard-Andersen, and Grant J Jensen. Architecture of the type iva pilus machine. *Science*, 351(6278):aad2001, 2016.

[21] Cheng Chen, Qi Dou, Hao Chen, Jing Qin, and Pheng Ann Heng. Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation. *IEEE Transactions on Medical Imaging*, 2020.

[22] Yuxiang Chen, Stefan Pfeffer, José Jesús Fernández, Carlos Oscar S Sorzano, and Friedrich Förster. Autofocused 3d classification of cryoelectron subtomograms. *Structure*, 22(10):1528–1537, 2014.

[23] 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.

[24] Christopher B Choy, JunYoung Gwak, Silvio Savarese, and Manmohan Chandraker. Universal correspondence network. In *Advances in Neural Information Processing Systems*, pages 2414–2422, 2016.

[25] Zhen Cui, Shiguang Shan, Haihong Zhang, Shihong Lao, and Xilin Chen. Image sets alignment for video-based face recognition. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2626–2633. IEEE, 2012.

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

dria. *Proceedings of the National Academy of Sciences*, 108(34):14121–14126, 2011.

[27] Georgios D Evangelidis and Emmanouil Z Psarakis. Parametric image alignment using enhanced correlation coefficient maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(10):1858–1865, 2008.

[28] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006.

[29] Alejandro F Frangi, Wiro J Niessen, Daniel Rueckert, and Julia A Schnabel. Automatic 3d asm construction via atlas-based landmarking and volumetric elastic registration. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 78–91. Springer, 2001.

[30] Jesus G Galaz-Montoya and Steven J Ludtke. The advent of structural biology in situ by single particle cryo-electron tomography. *Biophysics reports*, 3(1-3):17–35, 2017.

[31] Ashwini Gulhane, Prashant L Paikrao, and DS Chaudhari. A review of image data clustering techniques. *International Journal of Soft Computing and Engineering*, 2(1):212–215, 2012.

[32] 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.

[33] Xifeng Guo, Xinwang Liu, En Zhu, and Jianping Yin. Deep clustering with convolutional autoencoders. In *International conference on neural information processing*, pages 373–382. Springer, 2017.

[34] Yoav HaCohen, Eli Shechtman, Dan B Goldman, and Dani Lischinski. Non-rigid dense correspondence with applications for image enhancement. *ACM transactions on graphics (TOG)*, 30(4):1–10, 2011.

[35] Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal flow: Semantic correspondences from object proposals. *IEEE transactions on pattern analysis and machine intelligence*, 40(7):1711–1725, 2017.

[36] Kai Han, Rafael S Rezende, Bumsub Ham, Kwan-Yee K Wong, Minsu Cho, Cordelia Schmid, and Jean Ponce. Scnet: Learning semantic correspondence. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1831–1840, 2017.

[37] Renmin Han, Liansan Wang, Zhiyong Liu, Fei Sun, and Fa Zhang. A novel fully automatic scheme for fiducial marker-based alignment in electron tomography. *Journal of structural biology*, 192(3):403–417, 2015.

[38] 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.

[39] 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.

[40] Shuaiyi Huang, Qiuyue Wang, Songyang Zhang, Shipeng Yan, and Xuming He. Dynamic context correspondence network for semantic alignment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2010–2019, 2019.

[41] Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. Liteflownet: A lightweight convolutional neural network for optical flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8981–8989, 2018.

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

[43] Sangryul Jeon, Seungryong Kim, Dongbo Min, and Kwanghoon Sohn. Parn: Pyramidal affine regression networks for dense semantic correspondence. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 351–366, 2018.

[44] Xu Ji, João F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9865–9874, 2019.

[45] Edris Joonaki, Aliakbar Hassanpouryouzband, Caryn L Heldt, and Oluwatoyin Areo. Surface chemistry can unlock drivers of surface stability of sars-cov-2 in variety of environmental conditions. *Chem*, 2020.

[46] Zhaojie Ju, Yuehui Wang, Wei Zeng, Shengyong Chen, and Honghai Liu. Depth and rgb image alignment for hand gesture segmentation using kinect. In *2013 International Conference on Machine Learning and Cybernetics*, volume 2, pages 913–919. IEEE, 2013.

[47] Artúr István Károly, Róbert Fullér, and Péter Galambos. Unsupervised clustering for deep learning: A tutorial survey. *Acta Polytechnica Hungarica*, 15(8):29–53, 2018.

[48] 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.

[49] Seungryong Kim, Stephen Lin, Sang Ryul Jeon, Dongbo Min, and Kwanghoon Sohn. Recurrent transformer networks for semantic correspondence. In *Advances in Neural Information Processing Systems*, pages 6126–6136, 2018.

[50] 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. *Nature communications*, 11(1):1–10, 2020.

[51] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

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

[53] Amit Kumar, Rajeev Ranjan, Vishal Patel, and Rama Chellappa. Face alignment by local deep descriptor regression. *arXiv preprint arXiv:1601.07950*, 2016.

[54] Shang-Hong Lai and Ming Fang. A hybrid image alignment system for fast and precise pattern localization. *Real-Time Imaging*, 8(1):23–33, 2002.

[55] Zakaria Laskar and Juho Kannala. Semi-supervised semantic matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[56] Zakaria Laskar, Hamed Rezazadegan Tavakoli, and Juho Kannala. Semantic matching by weakly supervised 2d point set registration. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1061–1069. IEEE, 2019.

[57] Junghyup Lee, Dohyung Kim, Wonkyung Lee, Jean Ponce, and Bumsub Ham. Learning semantic correspondence exploiting an object-level prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[58] Junghyup Lee, Dohyung Kim, Jean Ponce, and Bumsub Ham. Sfnet: Learning object-aware semantic correspondence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2278–2287, 2019.

[59] Qi Li, Zhenan Sun, Ran He, and Tieniu Tan. Joint alignment and clustering via low-rank representation. In *2013 2nd IAPR Asian Conference on Pattern Recognition*, pages 591–595. IEEE, 2013.

[60] Qiaoliang Li, Guoyou Wang, Jianguo Liu, and Shaobo Chen. Robust scale-invariant feature matching for remote sensing image registration. *IEEE Geoscience and Remote Sensing Letters*, 6(2):287–291, 2009.

[61] 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*, pages 10196–10205, 2020.

[62] Xinghui Li, Kai Han, Shuda Li, and Victor Prisacariu. Dual-resolution correspondence networks. *Advances in Neural Information Processing Systems*, 33, 2020.

[63] Xinchao Li, Martha Larson, and Alan Hanjalic. Pairwise geometric matching for large-scale object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5153–5161, 2015.

[64] Hstau Y Liao and Joachim Frank. Definition and estimation of resolution in single-particle reconstructions. *Structure*, 18(7):768–775, 2010.

[65] 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.

[66] Xiaoming Liu, Yan Tong, and Frederick W Wheeler. Simultaneous alignment and clustering for an image ensemble. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1327–1334. IEEE, 2009.

[67] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 768–783, 2018.

[68] Iaroslav Melekhov, Juho Kannala, and Esa Rahtu. Image patch matching using convolutional descriptors with euclidean distance. In *Asian Conference on Computer Vision*, pages 638–653. Springer, 2016.

[69] 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*, pages 3395–3404, 2019.

[70] Takanori Nakane, Dari Kimanius, Erik Lindahl, and Sjors HW Scheres. Characterisation of molecular motions in cryo-em single-particle data by multi-body refinement in relion. *Elife*, 7:e36861, 2018.

[71] Alex J Noble, Venkata P Dandey, Hui Wei, Julia Brasch, Jillian Chase, Priyamvada Acharya, Yong Zi Tan, Zhening Zhang, Laura Y Kim, Giovanna Scapin, et al. Routine single particle cryoem sample and grid characterization by tomography. *eLife*, 7:e34257, 2018.

[72] Xi Peng, Hongyuan Zhu, Jiashi Feng, Chunhua Shen, Haixian Zhang, and Joey Tianyi Zhou. Deep clustering with sample-assignment invariance prior. *IEEE Transactions on Neural Networks and Learning Systems*, 2019.

[73] Giovanni Puglisi and Sebastiano Battiato. A robust image alignment algorithm for video stabilization purposes. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(10):1390–1400, 2011.

[74] Ali Punjani, John L Rubinstein, David J Fleet, and Marcus A Brubaker. cryosparc: algorithms for rapid unsupervised cryo-em structure determination. *Nature methods*, 14(3):290–296, 2017.

[75] B Ramsay, Tharindu De Silva, Runze Han, M Ketcha, Ali Uneri, Joseph Goerres, Niral Sheth, Matt Jacobson, Sebastian Vogt, Gerhard Kleinszig, et al. Clustered iterative sub-atlas registration for improved deformable registration using statistical shape models. In *Medical Imaging 2018: Image-Guided Procedures, Robotic Interventions, and Modeling*, volume 10576, page 105760C. International Society for Optics and Photonics, 2018.

[76] Anna Rast, Miroslava Schaffer, Sahradha Albert, William Wan, Stefan Pfeffer, Florian Beck, Jürgen M Plitzko, Jörg Nickelsen, and Benjamin D Engel. Biogenic regions of cyanobacterial thylakoids form contact sites with the plasma membrane. *Nature plants*, 5(4):436–446, 2019.

[77] 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.

[78] 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.

[79] Ignacio Rocco, Relja Arandjelović, and Josef Sivic. Efficient neighbourhood consensus networks via submanifold sparse convolutions. *arXiv preprint arXiv:2004.10566*, 2020.

[80] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. In *Advances in Neural Information Processing Systems*, pages 1651–1662, 2018.

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

tation in internet images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1939–1946, 2013.

[82] 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.

[83] Florian KM Schur. Toward high-resolution in situ structural biology with cryo-electron tomography and subtomogram averaging. *Current opinion in structural biology*, 58:1–9, 2019.

[84] 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.

[85] Colin Studholme, Derek LG Hill, and David J Hawkes. An overlap invariant entropy measure of 3d medical image alignment. *Pattern recognition*, 32(1):71–86, 1999.

[86] Richard Szeliski. Image alignment and stitching: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 2(1):1–104, 2006.

[87] Tatsunori Taniai, Sudipta N Sinha, and Yoichi Sato. Joint recovery of dense correspondence and cosegmentation in two images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4246–4255, 2016.

[88] Chang-Lu Tao, Yun-Tao Liu, Rong Sun, Bin Zhang, Lei Qi, Sakar Shivakoti, Chong-Li Tian, Peijun Zhang, Pak-Ming Lau, Z Hong Zhou, et al. Differentiation and characterization of excitatory and inhibitory synapses by cryo-electron tomography and correlative microscopy. *Journal of Neuroscience*, pages 1548–17, 2018.

[89] 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*, pages 6258–6268, 2020.

[90] Gerald J Van Dalen, Daniel P Magree, and Eric N Johnson. Absolute localization using image alignment and particle filtering. In *AIAA Guidance, Navigation, and Control Conference*, page 0647, 2016.

[91] Rui Wang and Zhi Xu. A coarse-to-fine strategy for the registration of the multi-wavelength high-resolution solar images. *Research in Astronomy and Astrophysics*, 20(7):103, 2020.

[92] Lingyu Wei, Qixing Huang, Duygu Ceylan, Etienne Vouga, and Hao Li. Dense human body correspondences using convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1544–1553, 2016.

[93] 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.

[94] Zhong Xue, Dinggang Shen, and Christos Davatzikos. Classic: consistent longitudinal alignment and segmentation for serial image computing. *NeuroImage*, 30(2):388–399, 2006.

[95] Jianwei Yang, Devi Parikh, and Dhruv Batra. Joint unsupervised learning of deep representations and image clusters. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5147–5156, 2016.

[96] Xu Yang, Cheng Deng, Feng Zheng, Junchi Yan, and Wei Liu. Deep spectral clustering using dual autoencoder network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4066–4075, 2019.

[97] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2878–2890, 2012.

[98] 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.

[99] Kaicheng Yu and Mathieu Salzmann. Second-order convolutional neural networks. *arXiv preprint arXiv:1703.06817*, 2017.

[100] Armand Zampieri, Guillaume Charpiat, Nicolas Girard, and Yuliya Tarabalka. Multimodal image alignment through a multiscale chain of neural networks with application to remote sensing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 657–673, 2018.

[101] Andrei Zanfir and Cristian Sminchisescu. Large displacement 3d scene flow with occlusion reasoning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4417–4425, 2015.

[102] Andy Zeng, Shuran Song, Matthias Nießner, Matthew Fisher, Jianxiong Xiao, and Thomas Funkhouser. 3dmatch: Learning local geometric descriptors from rgb-d reconstructions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1802–1811, 2017.

[103] 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.

[104] Dejiao Zhang, Yifan Sun, Brian Eriksson, and Laura Balzano. Deep unsupervised clustering using mixture of autoencoders. *arXiv preprint arXiv:1712.07788*, 2017.

[105] Yuanyi Zhong, Jiansheng Chen, and Bo Huang. Toward end-to-end face recognition through alignment learning. *IEEE signal processing letters*, 24(8):1213–1217, 2017.

[106] Bing-Rui Zhou, KN Sathish Yadav, Mario Borgnia, Jingjun Hong, Baohua Cao, Ada L Olins, Donald E Olins, Yawen Bai, and Ping Zhang. Atomic resolution cryo-em structure of a native-like cenp-a nucleosome aided by an antibody fragment. *Nature communications*, 10(1):1–7, 2019.

[107] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. Face alignment by coarse-to-fine shape searching.

In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4998–5006, 2015.