

This CVPR 2020 paper is the Open Access version, provided by the Computer Vision Foundation. Except for this watermark, it is identical to the accepted version; the final published version of the proceedings is available on IEEE Xplore.

# Gum-Net: Unsupervised Geometric Matching for Fast and Accurate 3D Subtomogram Image Alignment and Averaging

Xiangrui Zeng Min Xu\* Computational Biology Department Carnegie Mellon University Pittsburgh, PA 15213, USA.

xiangruz@andrew.cmu.edu mxul@cs.cmu.edu

## Abstract

We propose a Geometric unsupervised matching Network (Gum-Net) for finding the geometric correspondence between two images with application to 3D subtomogram alignment and averaging. Subtomogram alignment is the most important task in cryo-electron tomography (cryo-ET), a revolutionary 3D imaging technique for visualizing the molecular organization of unperturbed cellular landscapes in single cells. However, subtomogram alignment and averaging are very challenging due to severe imaging limits such as noise and missing wedge effects. We introduce an end-to-end trainable architecture with three novel modules specifically designed for preserving feature spatial information and propagating feature matching information. The training is performed in a fully unsupervised fashion to optimize a matching metric. No ground truth transformation information nor category-level or instance-level matching supervision information is needed. After systematic assessments on six real and nine simulated datasets, we demonstrate that Gum-Net reduced the alignment error by 40 to 50% and improved the averaging resolution by 10%. Gum-Net also achieved 70 to 110 times speedup in practice with GPU acceleration compared to state-of-the-art subtomogram alignment methods. Our work is the first 3D unsupervised geometric matching method for images of strong transformation variation and high noise level. The training code, trained model, and datasets are available in our open-source software AITom<sup>1</sup>.

### 1. Introduction

Given a transformation model, geometric matching aims to estimate the geometric correspondence between related images. In two and three dimensions, geometric matching is widely applied to fields such as pattern recognition [48, 74], 3D image reconstruction [35, 29], medical image alignment and registration [19, 27], and computational chemistry [86]. Finding the global optimal parameters consistent with a geometric transformation model such as affine or rigid transformation has a fundamental bottleneck. The parametric space needs to be exhaustively searched but the computational cost is infeasible [36]. Many popular methods have been proposed that alleviate the computational cost by detecting and matching hand-crafted local features [50, 15, 73] to estimate the global geometric transformation robustly [23, 67, 47, 51].

Recently, end-to-end trainable image alignment attracts attention. There are two major advantages over traditional non-trainable methods: (1) a properly trained convolutional neural network (CNN) model can process a large amount of data in a significantly shorter time and (2) with increasing amount of data collected, the deep learning model performance can be improved progressively by better feature learning [60].

In this paper, we focus on an important geometric matching application field, cryo-electron tomography (cryo-ET). In recent years, cryo-ET emerges as a revolutionary *in situ* 3D structural biology imaging technique for studying macromolecular complexes in single cells, the nano-machines that govern cellular biological processes [44]. Cryo-ET captures the 3D native structure and spatial distribution of all macromolecular complexes together with other subcellular components without disrupting the cell [11]. Nevertheless, cryo-ET data is heavily affected by a low signal-to-noise ratio (SNR) (example input data and mathematical definition in Supplementary Section S3) due to the complex cytoplasm environment and missing wedge effects<sup>2</sup>. Therefore, the macromolecular structures in the 3D tomogram need to be detected and recovered for further biomedical interpretation.

A subtomogram from a tomogram is a small cubic sub-

<sup>\*</sup>Corresponding author

<sup>&</sup>lt;sup>1</sup>https://github.com/xulabs/aitom

<sup>&</sup>lt;sup>2</sup>Partial sampling of images due to limited tilt angle ranges (description in Supplementary Section S1)



Figure 1. Gum-Net model pipeline. The model is unsupervised and feed-forward. The model inputs two subtomograms  $s_a$  and  $s_b$  (underlying structures are shown in isosurface representation) and outputs the transformed subtomogram  $\hat{s}_b$  to geometrically match  $s_a$ , in addition to the transformation model parameters  $\phi^{\text{tr}}$  and  $\phi^{\text{rot}}$ . The dash-line denotes that the parameters are shared between the two feature extractors.

volume generally containing one macromolecular complex. Subtomogram alignment is the most critical cryo-ET data processing technique for two reasons: first, high-resolution macromolecular structures can be recovered through subtomogram averaging based on alignment. Second, the spatial distribution of a certain structure can be detected through alignment. To recover the structure, subtomograms containing the same macromolecular structure but in different poses must be iteratively aligned and averaged. Subtomogram averaging improves resolution by reducing noise and missing wedge artifacts [83]. Subtomogram alignment is a considerably more challenging geometric matching task than related tasks such as 3D deformable medical image registration from two aspects: first, there is strong transformation variation because the structure inside a subtomogram is of completely random orientation and displacement. Second, medical images are relatively clean tissue images whereas subtomograms are cellular images with a low SNR (around 0.01 to 0.1) due to the complex cytoplasm environment and the low electron dose used for imaging [16] (example input data in Supplementary Section S3).

Given the 3D rigid transformation model, subtomogram alignment computes the six parameters (three rotational and three translational). We and others have proposed methods [87, 13] to approximate the constrained correlation objective function [25] as heuristics to limit the computational time to a feasible range. However, it is possible nowadays to collect a set of tomograms in several days containing millions of subtomograms [5]. Existing state-of-the-art subtomogram alignment methods [87, 13] generally align a pair of subtomograms on the scale of several seconds, which is too slow for processing such a large amount of data. Moreover, their accuracy is limited because they are approximation methods.

We propose Gum-Net (Geometric unsupervised matching Network), a deep architecture for 3D subtomogram alignment and averaging through unsupervised rigid geometric matching. Integrating three novel modules, Gum-Net inputs two subtomograms to estimate the transformation parameters by extracting and matching convolutional features. Gum-Net achieved significant improvement in efficiency (70 to 110 times speedup) and accuracy (40 to 50 % reduction in alignment error) over two state-of-the-art subtomogram alignment methods [87, 13]. The improvements from proposed modules were demonstrated in three ablation studies.

Main contributions. Our work is the first 3D unsupervised geometric matching method for images of strong transformation variation and high noise level. We integrated three novel modules (Figure 1): (1) we observe that as the max pooling and averaging pooling operations in the standard deep feature extraction process seek to achieve local transformation invariance, it is not suitable for accurate geometric matching, because the feature spatial locations need to be preserved to a large extent during feature extraction. Therefore, we introduce a feature extraction module with spectral operations including pooling and filtering to preserve the spatial location of extracted features. (2) We propose a novel Siamese matching module that improves spatial correlation information propagation by processing two feature correlation maps in parallel. (3) We incorporate a modified spatial transformer network [37] with a differentiable missing wedge imputation strategy into the alignment module. We achieved fully unsupervised training by feeding into random pairs of subtomograms regardless of their structural class information. Therefore, in contrast to other weakly-supervised geometric matching methods [71, 70, 42, 80, 58], no supervision such as instance-level or category-level matching information is needed.

# 2. Related Work

#### 2.1. 2D image alignment based on CNN

2D image alignment usually consists of two steps: (1) obtaining image feature descriptors and (2) matching feature descriptors according to a geometric model. Recently, some

methods have employed pre-trained [81] or trainable [41, 63] CNN-based feature extractors. Specifically, [22] proposed a hierarchical metric learning strategy to learn better feature descriptors for geometric matching. However, all the networks are combined with traditional matching methods.

In 2017, Rocco et al. proposed the first end-to-end convolutional neural network for geometric matching of 2D images [69]. This fully supervised model utilizes a pre-trained network [77] to extract features from the two images to be matched. Then a correlation layer matches the features followed by a network to regress to the known transformation parameters for supervised training. Later, they extended this model to be weakly-supervised for finding category-level [70] and instance-level correspondence [71]. Other weakly supervised methods have been proposed for similar tasks including semantic attribute matching [42], simultaneous alignment and segmentation [80], and alignment under large intra-class variations [58]. However, they still require additional training supervision such as matching image pairs on the instance level or category level.

#### 2.2. Unsupervised optical flow estimation

Optical flow estimation describes the small displacements of pixels in a sequence of 2D images using a dense or sparse vector field. Early unsupervised methods have used the gated restricted Boltzmann machine to learn image transformations [56, 57]. Recent CNN-based methods applied techniques such as frame interpolation [49], occlusion reasoning [38], and unsupervised losses in terms of brightness constancy [39] or bidirectional census [53]. Although these methods are all unsupervised, they require their input images to be highly similar with only small pixel shifts.

# 2.3. Unsupervised deformable medical image registration

3D image registration is the 3D analog to the 2D optical flow estimation. Deformable image registration has been extensively applied to 3D medical images such as brain MRI [85, 59], CT [33, 76], and cardiac images [91, 72]. Recent works present unsupervised CNN models based on spatial transformation function [18, 4, 17] or generative adversarial networks [52, 40]. Similar to optical flow estimation, these methods require the input pair of fixed and moving volumes to be similar. The information from the two volumes is integrated by stacking them as one input to the CNN models. However, simply stacking the input image pairs works poorly when there is strong transformation variation because the image similarity comparison is spatially constrained to a local neighborhood [55].

#### 2.4. Non-learning-based subtomogram alignment

Early works have used exhaustive grid search of rotations and translations with fixed intervals such as 1 voxel and  $5^{\circ}$ 

to align subtomograms [8, 24, 3]. To reduce the computational cost of searching the 6D parametric space exhaustively, high-throughput alignment proposed in [87] applied the fast rotational matching algorithm [43]. Fast and accurate alignment proposed in [13] also used the fast rotational matching algorithm and takes into account more information including amplitude and phase into their procedure. Another approach is to collaboratively align multiple subtomograms together based on nuclear-norm [46].

In this paper, we focus on pairwise subtomogram alignment and compared our method against the two most popular subtomogram alignment methods as baselines [87, 13].

# 3. Method

Our model is shown in Figure 1 (detailed architecture in Supplementary Section S2). Two subtomograms (3D grayscale cubic images)  $s_a$  and  $s_b$  are processed using feature extractors with shared weights to produce two feature maps  $v_a$  and  $v_b$ . Then a Siamese matching module computes two correlation maps  $c_{ab}$  and  $c_{ba}$ . At a specific position (i,j,k),  $c_{ab}$  contains the similarity between  $v_a$  at that position (i,j,k) and all the features of  $v_b$ , whereas  $c_{ba}$  is similarly defined.  $c_{ab}$  and  $c_{ba}$  are processed with the same network architecture and are later concatenated to estimate the transformation parameters. The 6D transformation parameters, which consist of  $\phi^{tr} = \{q_x, q_y, q_z\}$  for 3D translation and  $\phi^{\text{rot}} = \{q_{\alpha}, q_{\beta}, q_{\gamma}\}$  for 3D rotation in ZYZ convention, are feed into a differentiable spatial transformer network to compute the output, a transformed subtomogram  $\hat{s}_b = \mathcal{T}_{\phi}(s_b) = \mathcal{T}_{\phi^{\text{tr}}} \mathcal{T}_{\phi^{\text{rot}}}(s_b)$  with the missing wedge region imputed (Section 3.3). A spectral data imputation technique is integrated into the spatial transformer network to compensate for the missing wedge effects. In the training process, we do not have the ground truth transformation parameters to regress to as in [69]. Therefore, to assess the geometric matching performance, our objective is to find 3D rigid transformation parameters to maximize the crosscorrelation between  $s_a$  and  $\hat{s}_b$  in an unsupervised fashion. The cross-correlation-based loss is back-propagated to update the model weights.

#### 3.1. Feature extraction module

Feature extraction is a dimensionality reduction process to efficiently learn a compact feature vector representation of interesting parts of raw images. There are various popular feature extraction techniques such as DenseNet [34], InceptionNet [79], and ResNet [32]. Subsampling methods such as max pooling and average pooling are used in these convolutional neural networks to reduce feature map dimensionality and facilitate computation. Compared with max pooling and average pooling, spectral representation for convolutional neural networks preserves considerably more spatial information per parameter and enables flexibility in the pooling output dimensionality [68]. 2D spectral pooling layers that perform dimension reduction in the frequency domain have been proposed based on discrete Fourier transform (DFT) [68], discrete cosine transform (DCT) [78], and Hartley transform [92]. However, these methods are designed for 2D images and do not take into account image noise.

We propose a 3D DCT-based spectral layer with pooling and filtering operations. Since our inputs are 3D noisy images, the novel filtering operation is for feature map highfrequency noise reduction, and pooling operation for feature map dimension reduction. We choose the DCT because it stores only real-valued coefficients and compacts more energy in a smaller portion of the spectra compared to the DFT [84].

For an input feature map  $v \in \mathbb{R}^{L \times W \times H}$ , its 3D type-II DCT is defined as [2]:

$$\mathbb{C}(v)_{lhw} = \frac{8}{LWH} \epsilon_l \epsilon_h \epsilon_w \sum_{i=0}^{L-1} \sum_{j=0}^{H-1} \sum_{k=0}^{W-1} v_{ijk} \cos\left(\frac{l\pi (2i+1)}{2L}\right)$$
$$\cos\left(\frac{h\pi (2j+1)}{2H}\right) \cos\left(\frac{w\pi (2k+1)}{2W}\right),$$
(1)

where  $\epsilon_l = \begin{cases} \frac{1}{\sqrt{2}} & \text{for } l = 0\\ 1 & \text{otherwise} \end{cases}$ ,  $\forall l \in \{0, ..., L-1\}$ .  $\epsilon_h$  and  $\epsilon_w$ 

are similarly defined  $\forall h \in \{0, ..., H-1\}, \forall w \in \{0, ..., W-1\}$ . The inverse transform  $\mathbb{C}^{-1}$  of 3D type-II DCT is well defined as 3D type-III DCT [2]. Therefore, the pooled and filtered representation in the frequency domain can be transformed back through type-III DCT to the spatial domain as the output of the layer.



Figure 2. Image reconstruction from the max pooling, average pooling, and DCT spectral pooling scheme at different subsampling factors. DCT spectral pooling retains substantially greater spatial information of features from the original image and offers arbitrary output map dimensionality.

We use the DCT to perform subsampling in which the input is transformed to the frequency domain and cropped there. The output with reduced dimensionality is computed by transforming the cropped spectrum back into the spatial domain. The spectral pooling operation has been shown to achieve better spatial information preservation per parameter in terms of the  $l_2$  norm as compared to the max pooling operation [68]. Figure 2 shows the image reconstruction from max pooling, average pooling, and DCT spectral pooling at different subsampling factors. Compared to other pooling operations, a major advantage of using spectral pooling & filtering layers for geometric matching tasks is that the spatial location of features in two images are significantly better preserved for accurate matching. For example, during max pooling, the maximum from the receptive field is selected to achieve local rotation and translation invariance with the intuition that the exact location of a feature does not matter to the final classification. By contrast, during the feature extraction step for geometric matching, the exact feature spatial location is critical and the information loss will lead to inaccurate downstream matching.

We implement the 3D DCT spectral pooling & filtering as differentiable layers in the feature extractor. The low-pass filtering is also performed by masking out highfrequency regions dominated by noise. The forward and back-propagation procedure of the 3D DCT spectral pooling & filtering layer is outlined in Algorithm 1 and 2.

Algorithm 1: DCT spectral pooling & filtering
<b>Input:</b> Feature map $v \in \mathbb{R}^{L \times W \times H}$
Output size $L_1 \times W_1 \times H_1$
Cropping size $L_2 \times W_2 \times H_2$
<b>Output:</b> Feature map $\hat{v} \in \mathbb{R}^{L_1 \times W_1 \times H_1}$
1 $u \leftarrow \mathbb{C}(v)$
2 $u \leftarrow \text{Crop } u \text{ to size } L_2 \times W_2 \times H_2$
$\hat{u} \leftarrow \text{ZeroPad } u \text{ to size } L_1 \times W_1 \times H_1$
$4 \ \hat{v} \leftarrow \mathbb{C}^{-1}(\hat{u})$

Algorithm 2:	DCT	spectral	pooling	&	filtering	back-
propagation						

Input: Gradient w.r.t layer output  $\frac{\partial L}{\partial \hat{v}}$ Output: Gradient w.r.t layer input  $\frac{\partial L}{\partial v}$  $y \leftarrow \mathbb{C}(\frac{\partial L}{\partial \hat{v}})$  $y \leftarrow \text{Crop } y \text{ to size } L_2 \times W_2 \times H_2$  $\hat{y} \leftarrow \text{ZeroPad } y \text{ to size } L \times W \times H$  $\frac{\partial L}{\partial v} \leftarrow \mathbb{C}^{-1}(\hat{y})$ 

The arbitrary output size of spectral pooling & filtering layers offers another major advantage for geometric matching tasks. If the output two feature maps are of size  $L \times W \times H$ with C channels, the Siamese correlation layer (Section 3.2) will create two correlation maps, each of size  $L \times W \times H$ with (LWH) channels. The output feature map size from the feature extraction module to the Siamese matching module needs to be carefully manipulated, especially for 3D images. If the output feature map is too small, such as  $3 \times 3 \times 3$ , there is too much information loss for matching. If the output feature map is too large, such as  $20 \times 20 \times 20$ , the resulting correlation maps will be of size  $20 \times 20 \times 20 \times 8000$ , which is too large to be processed. Unlike max pooling or average pooling layers which aggressively reduce each dimension to half of the size and remove 87.5 % of the information, spectral pooling & filtering layers can gradually reduce the feature map size to the desired feature extraction module output size. Therefore, no additional spatial cropping or padding layer is needed to control the feature map size.

#### 3.2. Siamese matching module

The matching of extracted features from images is usually performed as an independent post-processing step [31, 90, 75, 54, 64]. The 2D correlation layer proposed in [69] achieved the state-of-the-art for integrating the matching information from two images. It is essentially a normalized cross-correlation function  $G : \mathbb{R}^{H \times W \times C} \times \mathbb{R}^{H \times W \times C} \rightarrow \mathbb{R}^{H \times W \times (HW)}$ . One of the input feature maps  $v_a$  is first flattened into shape  $v_a \in \mathbb{R}^{N \times C}$ , where N = HW, in order to keep the output correlation map 2D. Then for each feature (pixel) in  $v_a$  and  $v_b$ , the dot product is computed over all the channels (as feature descriptors) to obtain the correlation, which is later normalized. Nevertheless, to control the dimension of the output correlation map, all axes of one input feature map are broken and later cast into the channels of the output whereas the other input feature map is preserved.

We propose a novel Siamese matching module for pairwise 3D feature matching. To better utilize and process the feature correlation information, we design a Siamese correlation layer. Different from the correlation layer in [69] which computes only  $c_{ab}$ , the Siamese correlation layer is intuitive and symmetrically designed, which computes two correlation maps  $c_{ab}$  and  $c_{ba}$ . Each of them preserves the spatial coordinates of one input feature map. The use of two correlation maps propagates more feature spatial correlation. Element at a specific position *lwhc* is defined as:

$$(c_{ab})_{lwhc} = \frac{\langle v_{an:}, v_{b_{lwh:}} \rangle}{\sqrt{\sum_{i,j,k} \langle v_{an:}, v_{b_{ijk:}} \rangle}},$$

$$(c_{ba})_{lwhc} = \frac{\langle v_{bn:}, v_{a_{lwh:}} \rangle}{\sqrt{\sum_{i,j,k} \langle v_{bn:}, v_{a_{ijk:}} \rangle}}.$$

$$(2)$$

The two correlation maps are feed into a pseudo-Siamese network consisting of convolution layers and convolved separately but later concatenated for one fully connected layer. After another fully connected layer, the Siamese matching module outputs the estimated rigid transformation parameters  $\phi^{tr}$  and  $\phi^{rot}$ . Detailed model architecture can be found in Supplementary Section S2.

#### 3.3. Unsupervised geometric alignment module

Existing subtomogram alignment methods optimize a matching metric [87, 13, 6, 3]. In practice, preparing the subtomogram alignment ground truth for training is extremely time-consuming (need to exhaustively search the 6D parametric space). Therefore, the deep model should be unsupervised for this task. To achieve this goal, we propose an unsupervised geometric alignment module utilizing the spatial transformer network [37] with spectral data imputation designed specifically for subtomogram data.

In a tomogram with fixed voxel spacing (around 1nm), a certain type of macromolecular structure does not scale or reflect. Therefore, we restrict ourselves to 3D rigid transformation. Denoting the transformation matrix generated by the 3D rigid transformation parameters as  $M_{\theta}$  [21] and the 3D warping as  $\mathcal{T}_{\phi} : \mathbb{R}^3 \to \mathbb{R}^3$ , we have:

$$\begin{pmatrix} x_{i}^{s} \\ y_{i}^{s} \\ z_{i}^{s} \\ 1 \end{pmatrix} = \mathcal{T}_{\phi} \begin{pmatrix} x_{i}^{t}, y_{i}^{t}, z_{i}^{t} \end{pmatrix} = M_{\theta} \begin{pmatrix} x_{i}^{t} \\ y_{i}^{t} \\ z_{i}^{t} \\ 1 \end{pmatrix}$$

$$= \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} & \theta_{14} \\ \theta_{21} & \theta_{22} & \theta_{23} & \theta_{24} \\ \theta_{31} & \theta_{32} & \theta_{33} & \theta_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} x_{i}^{t} \\ y_{i}^{t} \\ z_{i}^{t} \\ 1 \end{pmatrix},$$

$$(3)$$

where  $(x_i^t, y_i^t, z_i^t)$  is the target coordinates on the transformed output 3D image and  $(x_i^s, y_i^s, z_i^s)$  is the source coordinates on the input 3D image.  $\theta$  is an element of the transformation matrix. The 3 × 3 orthogonal rotation matrix is from  $\theta_{11}$  to  $\theta_{33}$ . The displacement along each axis is specified by  $\theta_{14}$ ,  $\theta_{24}$ , and  $\theta_{34}$ . The 3D warping is differentiable and therefore able to be trained end-to-end.

In order to compensate for the missing wedge effects and thus to decrease the bias introduced, we integrate a spectral data imputation strategy from our previous work [88] into the spatial transformer network. For a subtomogram, we use its current estimated transformation to compute the rotated missing wedge mask m, as an indicator function to represent whether the Fourier coefficients are valid or missing in certain regions, and impute the missing ones with those from its transformation target subtomogram  $s_a$ . We can form a transformed and imputed subtomogram  $\hat{s}_b$  such that:

$$(\mathcal{F}\hat{s}_b)(\xi) = \begin{cases} [\mathcal{F}s_a](\xi) & \text{if } m(\xi) = 0\\ [\mathcal{F}\mathcal{T}_{\phi}(s_b)](\xi) & \text{if } m(\xi) = 1 \end{cases},$$

$$m(\xi) = \begin{cases} 0 & \text{if the Fourier coefficient at } \xi \text{ is missing} \\ 1 & \text{if the Fourier coefficient at } \xi \text{ is valid} \end{cases},$$
(4)

where  $\mathcal{F}$  is the Fourier transform operator,  $\xi \in \mathbb{R}^3$  is a Fourier space location, and  $m(\xi)$  is the rotated missing wedge mask according to  $\phi^{\text{rot}}$ . Since the magnitude of Fourier transform is translation-invariant, we only need to rotate  $m(\xi)$  without

using  $\phi^{tr}$  [25]. The imputation operation facilitates the unsupervised geometric matching task because only when the optimal alignment is obtained, the imputed data results in the highest consistency with the transformed subtomogram.

We note that since the rotation of the missing wedge mask m is implemented along with the transformation of the input subtomogram in the differentiable spatial transformer network and the inverse discrete Fourier transformation is well defined, this spectral data imputation step is differentiable in a similar manner as Algorithm 2.

**Loss function.** Pearson' correlation and its variants are widely used for assessing the alignment between two subtomograms [25, 6, 87, 13, 3] because of its simplicity and effectiveness. We implement it as a loss function to Gum-Net:

$$\mathcal{L} = 1 - \frac{\sum_{i=1}^{N} (s_{a_i} - \bar{s}_a) \left(\hat{s}_{b_i} - \bar{s}_b\right)}{\sqrt{\sum_{i=1}^{N} (s_{a_i} - \bar{s}_a)^2} \sqrt{\sum_{i=1}^{N} \left(\hat{s}_{b_i} - \bar{s}_b\right)^2}}, \quad (5)$$

where N is the total number of voxels in an input subtomogram. Compared to existing methods [87, 13], which utilize translation-invariant upper bound to approximate the Pearson's correlation objective to reduce the computational cost, Gum-Net optimizes Pearson's correlation directly for more accurate alignment.

#### 3.4. Baseline methods

We implemented two most popular state-of-the-art subtomogram alignment methods for comparison: H-T align [87] and F&A align [13]. We performed three ablation studies with existing modules: Gum-Net Max Pooling (Gum-Net MP), Gum-Net Average Pooling (Gum-Net AP), and Gum-Net Single Correlation (Gum-Net SC). Detailed implementation can be found in Supplementary Section S2.

#### 4. Experiments

Gum-Net was evaluated on six real and nine realistically simulated datasets at different SNR. On the simulated datasets, the accuracy of subtomogram alignment was evaluated by comparing the estimated transformation parameters  $\phi^{tr}$  and  $\phi^{rot}$  to the ground truth. On the real datasets, since the transformation ground truth is not available, in practice, the optimal transformation is usually obtained by parametric space exhaustive grid search to optimize the crosscorrelation between  $s_a$  and  $\hat{s}_b$ . Therefore, we compared the cross-correlation between  $s_a$  and  $\hat{s}_b$  computed by Gum-Net and baseline methods as an indirect indicator of the alignment accuracy. The visualization of subtomograms in different datasets can be found in Supplementary Section S3.

#### 4.1. Datasets

#### 4.1.1 Real datasets

GroEL/GroES dataset: this dataset contains 786 experimental subtomograms of purified GroEL and GroEL/GroES complexes from 24 tomograms [25]. Each subtomogram is rescaled to size  $32^3$  with voxel size 0.933 nm and  $25^\circ$  missing wedge.

**Rat neuron culture dataset**: this recent dataset is a set of tomograms from rat neuron culture [28]. In total 1095 ribosome subtomograms and 1527 capped proteasome subtomograms were extracted by template matching [8] and biology expert annotation. Each subtomogram is of size  $32^3$  with voxel size 1.368 nm and  $30^\circ$  missing wedge.

*S. cerevisiae* **80S** ribosome dataset: this dataset contains 3120 subtomograms extracted from 7 tomograms of purified *S. cerevisiae* 80S ribosomes [7]. Each subtomogram is rescaled to size  $32^3$  with voxel size 1.365 nm and  $30^\circ$  missing wedge. **TMV dataset**: this dataset contains 2742 Tobacco Mosaic Virus (TMV) subtomograms, a type of helical virus [45]. Each subtomogram is binned to size  $32^3$  with voxel size 1.080 nm and  $30^\circ$  missing wedge.

Aldolase dataset: this recent dataset contains 400 purified rabbit muscle aldolase subtomograms [61]. Each subtomogram is rescaled to size  $32^3$  with voxel size 0.750 nm and  $30^\circ$  missing wedge.

**Insulin receptor dataset**: this recent dataset contains 400 purified human insulin-bounded insulin receptor subtomograms [62]. Each subtomogram is rescaled to size  $32^3$  with voxel size 0.876 nm and  $45^\circ$  missing wedge.

#### 4.1.2 Simulated datasets

The subtomogram dataset simulation utilized a standard procedure in [26, 65] which takes into account the tomographic reconstruction process with missing wedges and contrast transfer function (detailed simulation procedure in Supplementary Section S3). We chose five representative macromolecular complexes: spliceosome (PDB ID: 5LQW), RNA polymerase-rifampicin complex (116V), RNA polymerase II elongation complex (6A5L), ribosome (5T2C), and capped proteasome (5MPA). All five structures are asymmetric so that there exists only one alignment ground truth. We simulated five datasets, one relatively clean (SNR 100) and four with SNR close to the experimental conditions (0.1, 0.05, 0.05)0.03, and 0.01), each consists of 2100 subtomogram pairs of each structure (in total 10500 subtomogram pairs). 5000 subtomogram pairs from each dataset were used for training and 500 pairs for validation. The rest 5000 subtomogram pairs from each dataset are used for testing. For a pair of subtomograms, one structure is a randomly transformed copy of the other and the two structures were processed independently to obtain its tomographic image distortions. Each subtomogram is of size  $32^3$  with voxel size 1.2 nm. The  $s_b$ in each pair has a typical missing wedge  $30^{\circ}$  while  $s_a$  has no missing wedge.

For subtomogram averaging, we simulated four datasets of 500 ribosomes (PDB ID: 5T2C) in the same manner of

Method	SNR 100	SNR 0.1	SNR 0.05	SNR 0.03	SNR 0.01
H-T align	$0.30{\pm}0.68, 1.82{\pm}2.69$	1.22±1.07, 4.76±4.56	1.93±0.98, 7.26±4.77	$2.22 \pm 0.77, 8.86 \pm 4.72$	2.38±0.57, 11.33±5.02
F&A align	$0.33 \pm 0.70, 1.93 \pm 2.86$	$1.34{\pm}1.13, 5.39{\pm}4.90$	$1.95 \pm 0.98, 7.54 \pm 4.94$	$2.22{\pm}0.77, 8.99{\pm}4.81$	$2.38 \pm 0.57, 11.32 \pm 4.92$
Gum-Net MP	$0.90 \pm 0.87, 3.34 \pm 3.41$	$1.30 \pm 0.79, 4.93 \pm 3.36$	$1.44{\pm}0.79, 5.46{\pm}3.38$	$1.53 \pm 0.78, 5.96 \pm 3.34$	$1.67 \pm 0.77, 7.28 \pm 3.38$
Gum-Net AP	$0.60 {\pm} 0.71, 2.32 {\pm} 2.71$	$1.09 \pm 0.73, 4.20 \pm 2.96$	$1.30 \pm 0.77, 5.00 \pm 3.15$	$1.45 \pm 0.77, 5.70 \pm 3.25$	$1.65 \pm 0.78, 7.18 \pm 3.35$
Gum-Net SC	$0.70 {\pm} 0.75, 2.63 {\pm} 2.86$	$1.16 \pm 0.77, 4.41 \pm 3.23$	$1.36 \pm 0.79, 5.13 \pm 3.34$	$1.48 \pm 0.78, 5.75 \pm 3.34$	$1.67 \pm 0.77, 7.24 \pm 3.46$
Gum-Net	0.41±0.70, <b>1.59±2.63</b>	0.62±0.69, 2.41±2.61	$0.87{\pm}0.74, 3.20{\pm}2.78$	$1.13{\pm}0.75, 4.29{\pm}2.75$	$1.50{\pm}0.78, 6.78{\pm}4.22$

Table 1. Subtomogram alignment accuracy on five datasets with SNR specified. In each cell, the first term is the mean and standard deviation of the rotation error and the second term, the translation error. We highlighted Gum-Net results that are significantly better (p < 0.001) than all baselines by the paired sample t-test. More detailed results and analysis can be found in Supplementary Section S3.

SNR 0.1, 0.05, 0.03, and 0.01.

#### 4.2. Implementation

The deep models were implemented in Keras [14] with custom layers backend by Tensorflow [1]. All inputs have size  $32^3$ . We note that due to the flexibility of input and output size of the DCT spectral pooling & filtering layers, the input size can be arbitrary. Higher resolution can be achieved with larger input subtomogram sizes. Detailed implementation of Gum-Net and baselines can be found in Supplementary Section S2.

For each epoch, we randomly draw 5000 subtomogram pairs  $s_a$  and  $s_b$  from the training dataset regardless of their structural class information. Therefore, Gum-Net is fully unsupervised without instance-level or category-level matching information for weak supervision as in other geometric matching methods [71, 70, 42, 80, 58]. For a simulated dataset, there are 5000<sup>2</sup> possible image pairs. As a result, we did not observe any overfitting issue.

#### 4.3. Subtomogram alignment

Given the transformation ground truth, we measure the alignment accuracy with two metrics: (1) the translation error defined as the Euclidean distance between the translation estimation and the ground truth and (2) the rotation error defined as the Euclidean distance between the flattened rotation matrix of estimation and the ground truth.

**On simulated datasets:** Table 3.4 shows the alignment accuracy. Gum-Net achieved similar performance on the clean dataset (SNR 100). As max pooling achieves more local transformation invariance [93], Gum-Net MP performs worse than Gum-Net AP in all settings as expected. When the SNR is close to experimental condition (the real datasets have SNR around 0.01 to 0.1), CNN-based methods generally perform better than traditional methods. Specifically, Gum-Net outperformed all the baseline methods, demonstrating the improvement from the proposed modules.

In our experiments, the training, validation, and testing datasets are independent, which ensured no overfitting. However, since Gum-Net is fully unsupervised, even if the testing dataset is from a different domain source, such as collected under different imaging conditions, it is possible to fine-tune a trained model on the testing dataset (with no ground truth) for adaptation. In terms of speed, with a trained model, Gum-Net only takes 17.6 seconds to align 1000 subtomograms on a single GPU core. The training takes less than 10 hours. Since there is no available GPU-accelerated version of the traditional algorithms, H-T align and F & A align take 1916.4 seconds and 1251.2 seconds to align 1000 subtomograms on a CPU core, respectively. Therefore, in practice, this results in 70 to 110 times speedup over traditional methods.



Figure 3. Example alignment inputs and outputs at SNR 100. 2D slices representations are shown in Supplementary Section S3.

**On real datasets:** We split the GroEL/GroES dataset into a training dataset of 617 subtomograms, a validation dataset of 69 subtomograms, and a testing dataset of 100 subtomograms. There are 4950 pairs of subtomograms in the testing dataset. We align them pairwise by Gum-Net, H-T align, and F&A align and calculates the cross-correlation. Gum-Net achieved cross-correlations of  $0.0908\pm0.0204$ , significantly better (p < 0.001) than H-T align ( $0.0756\pm0.0194$ ) and F&A align ( $0.0838\pm0.0204$ ).

We split the rat neuron culture dataset into a training dataset of 2270 subtomograms, a validation dataset of 252 subtomograms, and a testing dataset of 100 ribosome and 100 capped proteasome subtomograms. There are 19900 pairs of subtomograms in the testing dataset. Gum-Net achieved cross-correlations of  $0.0615\pm0.0187$ , significantly better (p < 0.001) than H-T align ( $0.0541\pm0.0235$ ) and F&A align ( $0.0607\pm0.0199$ ). We use the pairwise correlation matrix to cluster the subtomograms by defining the pairwise distance as 1 - pairwise correlation. Applying the complete-linkage hierarchical clustering algorithm with k = 2, Gum-Net achieved an accuracy of 92%, better than F&A align



Figure 4. Illustration of alignment-based subtomogram averaging using Gum-Net. On the left are five example input subtomograms at SNR 0.1 in our experiment. On the right are subtomogram averages at different iterations and the true structure. The 2D slices representations are shown in Supplementary Section S3.

(65%) and H-T align (53.5%).

# 4.4. Non-parametric reference-free subtomogram averaging

Structures present in multiple noisy copies (usually thousands of) in a tomogram must be averaged through geometric transformation to obtain higher resolution 3D views [83]. To eliminate potential bias, subtomogram averaging is often done without any external structural reference. One major approach of reference-free subtomogram averaging is non-parametric alignment-based averaging in which all subtomograms are iteratively aligned to their average and re-averaged for the next iteration [9]. Figure 4 illustrates such a process in which the initial average is generated by simply averaging all the subtomograms without any transformation. The structural resolution of the subtomogram average is gradually improved through the iterative process.

Method	0.1	0.05	0.03	0.01	80S	TMV	Aldolase	Insulin
H-T align	2.89	3.79	4.92	4.41	3.05	2.23	2.34	1.90
F&A align	2.78	4.36	3.81	4.53	2.77	2.52	3.13	2.18
Gum-Net	2.78	2.95	4.01	4.22	2.73	2.16	1.97	1.77

Table 2. Subtomogram averaging results in FSC resolution (nm). '0.1' denotes simulated dataset at SNR 0.1. '80S', 'TMV', 'Aldolase', and 'Insulin' denote the real datasets. The best resolution is highlighted.

The iterative alignment-based non-parametric referencefree subtomogram averaging was tested using the proposed and baseline methods. The standard resolution measurement for assessing subtomogram averaging is Fourier shell correlation (FSC) [82] (mathematical definition in Supplementary Section S3), which measures the maximal discrepant structural factors between the subtomogram average and the true structure. The smaller the value, the better the results. As shown in Table 4.4, Gum-Net achieved the overall best averaging performance and improved the resolution by around 10%.

## 5. Conclusion

Cryo-ET subtomogram alignment and averaging revolutionize the discovery of 3D native macromolecular structure details in single cells. Such information provides critical insights into the precise function/dysfunction of the cellular processes. However, with a rapidly increasing amount of cryo-ET data collected, there is an urgent need to drastically improve the efficiency of subtomogram alignment methods. We developed the first unsupervised deep learning approach for 3D subtomogram alignment and averaging. Using the three proposed modules, Gum-Net achieved fast and accurate alignment with end-to-end unsupervised learning. Gum-Net opens up the possibility for continued improvement of subtomogram alignment and averaging efficiency and accuracy with better model design and training. This work serves as an important step toward *in situ* high-throughput detection and recovery of macromolecular structures for a better understanding of the molecular machinery in cellular processes.

Gum-Net can be integrated into existing cryo-ET analvsis software in several ways. For example, EMAN2 [26] performs exhaustive 3D rotational and translational search followed by local refinement for alignment-based averaging. RELION [7] maximizes the likelihood of a model with Gaussian noise assumption by exhaustively scanning the 3D rigid transformation space for integration. Gum-Net improves the accuracy and efficiency of subtomogram alignment, especially for a large amount of cryo-ET data. Therefore, integrating Gum-Net with existing software can boost the speed of their alignment step or quickly generate initial structural models for averaging refinement. Gum-Net can also be easily extended to related tasks including tomographic tilt series alignment [30] and cryo-electron microscopy singleparticle reconstruction [94]. The proposed modules can be adapted to other geometric matching tasks for images of strong transformation variation such as face alignment under pose variations [20, 95], or of high noise level such as synthetic aperture radar imaging [89, 12] and sonar imaging [10, 66].

# Acknowledgements

This work was supported by U.S. National Science Foundation (NSF) grant DBI-1949629 and in part by U.S. National Institutes of Health (NIH) grant P41 GM103712. XZ was supported by a fellowship from Carnegie Mellon University's Center for Machine Learning and Health. We thank Hongyu Zheng, Dr. Benjamin Chidester, and Jennifer Williams at our Department for proof-reading the paper.

# References

- Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX}* Symposium on Operating Systems Design and Implementation ({OSDI} 16), pages 265–283, 2016.
- [2] O Alshibami and Said Boussakta. Fast algorithm for the 3d det. In 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221), volume 3, pages 1945–1948. IEEE, 2001.
- [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] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. An unsupervised learning model for deformable medical image registration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9252–9260, 2018.
- [5] Philip R Baldwin, Yong Zi Tan, Edward T Eng, William J Rice, Alex J Noble, Carl J Negro, Michael A Cianfrocco, Clinton S Potter, and Bridget Carragher. Big data in cryoem: automated collection, processing and accessibility of em data. *Current opinion in microbiology*, 43:1–8, 2018.
- [6] Alberto Bartesaghi, P Sprechmann, J Liu, G Randall, G Sapiro, and Sriram Subramaniam. Classification and 3d averaging with missing wedge correction in biological electron tomography. *Journal of structural biology*, 162(3):436–450, 2008.
- [7] Tanmay AM Bharat and Sjors HW Scheres. Resolving macromolecular structures from electron cryo-tomography data using subtomogram averaging in relion. *Nature protocols*, 11(11):2054, 2016.
- [8] Jochen Böhm, Achilleas S Frangakis, Reiner Hegerl, Stephan Nickell, Dieter Typke, and Wolfgang Baumeister. Toward detecting and identifying macromolecules in a cellular context: template matching applied to electron tomograms. *Proceedings of the National Academy of Sciences*, 97(26):14245– 14250, 2000.
- [9] John AG Briggs. Structural biology in situ—the potential of subtomogram averaging. *Current opinion in structural biology*, 23(2):261–267, 2013.
- [10] Cyril Chailloux, Jean-Marc Le Caillec, Didier Gueriot, and Benoit Zerr. Intensity-based block matching algorithm for mosaicing sonar images. *IEEE Journal of Oceanic Engineering*, 36(4):627–645, 2011.
- [11] Juan Chang, Xiangan Liu, Ryan H Rochat, Matthew L Baker, and Wah Chiu. Reconstructing virus structures from nanometer to near-atomic resolutions with cryo-electron microscopy and tomography. In *Viral Molecular Machines*, pages 49–90. Springer, 2012.
- [12] Min Chen, Ayman Habib, Haiqing He, Qing Zhu, and Wei Zhang. Robust feature matching method for sar and optical images by using gaussian-gamma-shaped bi-windowsbased descriptor and geometric constraint. *Remote Sensing*, 9(9):882, 2017.

- [13] Yuxiang Chen, Stefan Pfeffer, Thomas Hrabe, Jan Michael Schuller, and Friedrich Förster. Fast and accurate referencefree alignment of subtomograms. *Journal of structural biol*ogy, 182(3):235–245, 2013.
- [14] François Chollet et al. Keras (2015), 2017.
- [15] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *international Conference on computer vision & Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE Computer Society, 2005.
- [16] Radostin Danev, Shuji Kanamaru, Michael Marko, and Kuniaki Nagayama. Zernike phase contrast cryo-electron tomography. *Journal of structural biology*, 171(2):174–181, 2010.
- [17] Bob D de Vos, Floris F Berendsen, Max A Viergever, Hessam Sokooti, Marius Staring, and Ivana Išgum. A deep learning framework for unsupervised affine and deformable image registration. *Medical image analysis*, 52:128–143, 2019.
- [18] Bob D de Vos, Floris F Berendsen, Max A Viergever, Marius Staring, and Ivana Išgum. End-to-end unsupervised deformable image registration with a convolutional neural network. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 204–212. Springer, 2017.
- [19] Jérôme Declerck, Jacques Feldmar, Fabienne Betting, and Michael L Goris. Automatic registration and alignment on a template of cardiac stress and rest spect images. In Proceedings of the Workshop on Mathematical Methods in Biomedical Image Analysis, pages 212–221. IEEE, 1996.
- [20] Hassen Drira, Boulbaba Ben Amor, Anuj Srivastava, Mohamed Daoudi, and Rim Slama. 3d face recognition under expressions, occlusions, and pose variations. *IEEE Transactions* on Pattern Analysis and Machine Intelligence, 35(9):2270– 2283, 2013.
- [21] David W Eggert, Adele Lorusso, and Robert B Fisher. Estimating 3-d rigid body transformations: a comparison of four major algorithms. *Machine vision and applications*, 9(5-6):272–290, 1997.
- [22] Mohammed E Fathy, Quoc-Huy Tran, M Zeeshan Zia, Paul Vernaza, and Manmohan Chandraker. Hierarchical metric learning and matching for 2d and 3d geometric correspondences. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 803–819, 2018.
- [23] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications* of the ACM, 24(6):381–395, 1981.
- [24] Friedrich Förster and Reiner Hegerl. Structure determination in situ by averaging of tomograms. *Methods in cell biology*, 79:741–767, 2007.
- [25] Friedrich Förster, Sabine Pruggnaller, Anja Seybert, and Achilleas S Frangakis. Classification of cryo-electron subtomograms using constrained correlation. *Journal of structural biology*, 161(3):276–286, 2008.
- [26] Jesús G Galaz-Montoya, John Flanagan, Michael F Schmid, and Steven J Ludtke. Single particle tomography in eman2. *Journal of structural biology*, 190(3):279–290, 2015.

- [27] André P Guéziec, Xavier Pennec, and Nicholas Ayache. Medical image registration using geometric hashing. *IEEE Computational Science and Engineering*, 4(4):29–41, 1997.
- [28] 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.
- [29] Renmin Han, Xiaohua Wan, Zihao Wang, Yu Hao, Jingrong Zhang, Yu Chen, Xin Gao, Zhiyong Liu, Fei Ren, Fei Sun, et al. Autom: a novel automatic platform for electron tomography reconstruction. *Journal of structural biology*, 199(3):196– 208, 2017.
- [30] Renmin Han, Liansan Wang, Zhiyong Liu, Fei Sun, and Fa Zhang. A novel fully automatic scheme for fiducial markerbased alignment in electron tomography. *Journal of structural biology*, 192(3):403–417, 2015.
- [31] Xufeng Han, Thomas Leung, Yangqing Jia, Rahul Sukthankar, and Alexander C Berg. Matchnet: Unifying feature and metric learning for patch-based matching. In *Proceedings* of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3279–3286, 2015.
- [32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceed-ings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [33] Jidong Hou, Mariana Guerrero, Wenjuan Chen, and Warren D D'Souza. Deformable planning ct to cone-beam ct image registration in head-and-neck cancer. *Medical physics*, 38(4):2088–2094, 2011.
- [34] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4700–4708, 2017.
- [35] Qi-Xing Huang, Simon Flöry, Natasha Gelfand, Michael Hofer, and Helmut Pottmann. Reassembling fractured objects by geometric matching. ACM Transactions on Graphics (TOG), 25(3):569–578, 2006.
- [36] Piotr Indyk, Rajeev Motwani, and Suresh Venkatasubramanian. Geometric matching under noise: Combinatorial bounds and algorithms. In SODA, pages 457–465, 1999.
- [37] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In Advances in neural information processing systems, pages 2017–2025, 2015.
- [38] Joel Janai, Fatma Guney, Anurag Ranjan, Michael Black, and Andreas Geiger. Unsupervised learning of multi-frame optical flow with occlusions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 690–706, 2018.
- [39] J Yu Jason, Adam W Harley, and Konstantinos G Derpanis. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In *European Conference on Computer Vision*, pages 3–10. Springer, 2016.
- [40] Boah Kim, Jieun Kim, June-Goo Lee, Dong Hwan Kim, Seong Ho Park, and Jong Chul Ye. Unsupervised deformable image registration using cycle-consistent cnn. In *International Conference on Medical Image Computing and Computer*-*Assisted Intervention*, pages 166–174. Springer, 2019.

- [41] Seungryong Kim, Dongbo Min, Bumsub Ham, Sangryul Jeon, Stephen Lin, and Kwanghoon Sohn. Fcss: Fully convolutional self-similarity for dense semantic correspondence. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6560–6569, 2017.
- [42] Seungryong Kim, Dongbo Min, Somi Jeong, Sunok Kim, Sangryul Jeon, and Kwanghoon Sohn. Semantic attribute matching networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 12339– 12348, 2019.
- [43] Julio A Kovacs and Willy Wriggers. Fast rotational matching. Acta Crystallographica Section D: Biological Crystallography, 58(8):1282–1286, 2002.
- [44] Werner Kühlbrandt. The resolution revolution. *Science*, 343(6178):1443–1444, 2014.
- [45] Michael Kunz, Zhou Yu, and Achilleas S Frangakis. M-free: Mask-independent scoring of the reference bias. *Journal of structural biology*, 192(2):307–311, 2015.
- [46] Oleg Kuybeda, Gabriel A Frank, Alberto Bartesaghi, Mario Borgnia, Sriram Subramaniam, and Guillermo Sapiro. A collaborative framework for 3d alignment and classification of heterogeneous subvolumes in cryo-electron tomography. *Journal of structural biology*, 181(2):116–127, 2013.
- [47] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), volume 2, pages 2169–2178. IEEE, 2006.
- [48] 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.
- [49] Gucan Long, Laurent Kneip, Jose M Alvarez, Hongdong Li, Xiaohu Zhang, and Qifeng Yu. Learning image matching by simply watching video. In *European Conference on Computer Vision*, pages 434–450. Springer, 2016.
- [50] David G Lowe. Distinctive image features from scaleinvariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.
- [51] Jiayi Ma, Huabing Zhou, Ji Zhao, Yuan Gao, Junjun Jiang, and Jinwen Tian. Robust feature matching for remote sensing image registration via locally linear transforming. *IEEE Transactions on Geoscience and Remote Sensing*, 53(12):6469– 6481, 2015.
- [52] Dwarikanath Mahapatra, Bhavna Antony, Suman Sedai, and Rahil Garnavi. Deformable medical image registration using generative adversarial networks. In 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pages 1449–1453. IEEE, 2018.
- [53] Simon Meister, Junhwa Hur, and Stefan Roth. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [54] 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.

- [55] Iaroslav Melekhov, Aleksei Tiulpin, Torsten Sattler, Marc Pollefeys, Esa Rahtu, and Juho Kannala. Dgc-net: Dense geometric correspondence network. In 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 1034–1042. IEEE, 2019.
- [56] Roland Memisevic and Geoffrey Hinton. Unsupervised learning of image transformations. In 2007 IEEE Conference on Computer Vision and Pattern Recognition, pages 1–8. IEEE, 2007.
- [57] Roland Memisevic and Geoffrey E Hinton. Learning to represent spatial transformations with factored higher-order boltzmann machines. *Neural computation*, 22(6):1473–1492, 2010.
- [58] 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.
- [59] Ashraf Mohamed, Evangelia I Zacharaki, Dinggang Shen, and Christos Davatzikos. Deformable registration of brain tumor images via a statistical model of tumor-induced deformation. *Medical image analysis*, 10(5):752–763, 2006.
- [60] Maryam M Najafabadi, Flavio Villanustre, Taghi M Khoshgoftaar, Naeem Seliya, Randall Wald, and Edin Muharemagic. Deep learning applications and challenges in big data analytics. *Journal of Big Data*, 2(1):1, 2015.
- [61] 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.
- [62] Alex J Noble, Hui Wei, Venkata P Dandey, Zhening Zhang, Yong Zi Tan, Clinton S Potter, and Bridget Carragher. Reducing effects of particle adsorption to the air–water interface in cryo-em. *Nature methods*, 15(10):793, 2018.
- [63] David Novotny, Diane Larlus, and Andrea Vedaldi. Anchornet: A weakly supervised network to learn geometry-sensitive features for semantic matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5277–5286, 2017.
- [64] Hyun Oh Song, Stefanie Jegelka, Vivek Rathod, and Kevin Murphy. Deep metric learning via facility location. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5382–5390, 2017.
- [65] Long Pei, Min Xu, Zachary Frazier, and Frank Alber. Simulating cryo electron tomograms of crowded cell cytoplasm for assessment of automated particle picking. *BMC bioinformatics*, 17(1):405, 2016.
- [66] Minh Tân Pham and Didier Gueriot. Guided block-matching for sonar image registration using unsupervised kohonen neural networks. In 2013 OCEANS-San Diego, pages 1–5. IEEE, 2013.
- [67] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In 2007 IEEE Conference on Computer Vision and Pattern Recognition, pages 1–8. IEEE, 2007.
- [68] Oren Rippel, Jasper Snoek, and Ryan P Adams. Spectral representations for convolutional neural networks. In *Advances*

in neural information processing systems, pages 2449–2457, 2015.

- [69] 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.
- [70] Ignacio Rocco, Relja Arandjelović, and Josef Sivic. End-toend weakly-supervised semantic alignment. In *Proceedings* of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6917–6925, 2018.
- [71] 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.
- [72] Marc-Michel Rohé, Manasi Datar, Tobias Heimann, Maxime Sermesant, and Xavier Pennec. Svf-net: Learning deformable image registration using shape matching. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 266–274. Springer, 2017.
- [73] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary R Bradski. Orb: An efficient alternative to sift or surf. In *ICCV*, volume 11, page 2. Citeseer, 2011.
- [74] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, et al. Benchmarking 6dof outdoor visual localization in changing conditions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 8601–8610, 2018.
- [75] Tanner Schmidt, Richard Newcombe, and Dieter Fox. Selfsupervised visual descriptor learning for dense correspondence. *IEEE Robotics and Automation Letters*, 2(2):420–427, 2016.
- [76] Eduard Schreibmann, Jonathon A Nye, David M Schuster, Diego R Martin, John Votaw, and Tim Fox. Mr-based attenuation correction for hybrid pet-mr brain imaging systems using deformable image registration. *Medical physics*, 37(5):2101– 2109, 2010.
- [77] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.
- [78] James S Smith and Bogdan M Wilamowski. Discrete cosine transform spectral pooling layers for convolutional neural networks. In *International Conference on Artificial Intelligence and Soft Computing*, pages 235–246. Springer, 2018.
- [79] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.
- [80] Nikolai Ufer, Kam To Lui, Katja Schwarz, Paul Warkentin, and Björn Ommer. Weakly supervised learning of dense semantic correspondences and segmentation. In *German Conference on Pattern Recognition*, pages 456–470. Springer, 2019.

- [81] Nikolai Ufer and Bjorn Ommer. Deep semantic feature matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6914–6923, 2017.
- [82] Marin Van Heel and Michael Schatz. Fourier shell correlation threshold criteria. *Journal of structural biology*, 151(3):250– 262, 2005.
- [83] W Wan and JAG Briggs. Cryo-electron tomography and subtomogram averaging. In *Methods in enzymology*, volume 579, pages 329–367. Elsevier, 2016.
- [84] Andrew B Watson. Image compression using the discrete cosine transform. *Mathematica journal*, 4(1):81, 1994.
- [85] Adam Wittek, Karol Miller, Ron Kikinis, and Simon K Warfield. Patient-specific model of brain deformation: Application to medical image registration. *Journal of biomechanics*, 40(4):919–929, 2007.
- [86] Gerhard Wolber, Thomas Seidel, Fabian Bendix, and Thierry Langer. Molecule-pharmacophore superpositioning and pattern matching in computational drug design. *Drug discovery today*, 13(1-2):23–29, 2008.
- [87] 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.
- [88] Min Xu, Jitin Singla, Elitza I Tocheva, Yi-Wei Chang, Raymond C Stevens, Grant J Jensen, and Frank Alber. De novo structural pattern mining in cellular electron cryotomograms. *Structure*, 2019.
- [89] Yuanxin Ye and Li Shen. Hopc: A novel similarity metric based on geometric structural properties for multi-modal remote sensing image matching. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 3:9, 2016.
- [90] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *European Conference on Computer Vision*, pages 467–483. Springer, 2016.
- [91] Vladimir Zagrodsky, Vivek Walimbe, Carlos R Castro-Pareja, Jian Xin Qin, Jong-Min Song, and Raj Shekhar. Registrationassisted segmentation of real-time 3-d echocardiographic data using deformable models. *IEEE Transactions on Medical Imaging*, 24(9):1089–1099, 2005.
- [92] Hao Zhang and Jianwei Ma. Hartley spectral pooling for deep learning. arXiv preprint arXiv:1810.04028, 2018.
- [93] Jiahuan Zhou, Weiqi Xu, and Ryad Chellali. Analysing the effects of pooling combinations on invariance to position and deformation in convolutional neural networks. In 2017 IEEE International Conference on Cyborg and Bionic Systems (CBS), pages 226–230. IEEE, 2017.
- [94] Z Hong Zhou. Towards atomic resolution structural determination by single-particle cryo-electron microscopy. *Current opinion in structural biology*, 18(2):218–228, 2008.
- [95] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 146–155, 2016.