Equivariant Self-supervised Deep Pose Estimation for Cryo EM

Gabriele Cesa [1 2] **Pratik Kumar** [1] **Arash Behboodi** [1]

Abstract

Reconstructing the 3D volume of a molecule from its differently oriented 2D projections is the central problem of Cryogenic Electron Microscopy (cryo-EM), one of the main techniques for macro-molecule imaging. Because the orientations are unknown, the estimation of the images' poses is essential to solve this inverse problem. Typical methods either rely on *synchronization*, which leverages the estimated relative poses of the images to constrain their absolute ones, or *jointly estimate* the poses and the 3D density of the molecule in an iterative fashion. Unfortunately, synchronization methods don't account for the complete images' generative process and, therefore, achieve lower noise robustness. In the second case, the iterative joint optimization suffers from convergence issues and a higher computational cost, due to the 3D reconstruction steps. In this work, we directly estimate individual poses with an equivariant deep graph network trained using a self-supervised loss, which enforces agreement in Fourier domain of image pairs along the *common lines* defined by their poses. In particular, the *equivariant* design turns out essential for the proper convergence. As a result, our method can leverage the synchronization constraints - encoded by the synchronization graph structure - to improve convergence as well as the images generative process - via the common lines loss -, with no need to perform intermediate reconstructions.

1. Introduction

Cryogenic electron microscopy (Cryo-EM) is one of the major techniques in structural biology for capturing and studying the structure of macromolecules (Nogales, 2016; Henderson et al., 1990). In single particle cryo-EM, a field of the intended specimen is prepared, and the solution is frozen to cryogenic temperature. A single image is taken using projections by an electron microscope, yielding many 2D images of the intended specimen (macromolecules or proteins). The central task is to find the 3D structure of the molecule from the noisy 2D-images obtained through this process. The Nobel prize-winning cryo-EM provides many advantages compared to competing imaging techniques (Benjin & Ling, 2020). As an example, unlike X-ray crystallography, cryo-EM does not require protein crystallization, which is difficult for some molecules like membrane protein. Unfortunately, 3D reconstruction, considered as an inverse problem, includes many challenges: low signal-to-noise ratio (SNR), model mismatch with contrast transfer function (CTF) of the microscope, heterogeneity of the imaged molecules and molecule in-place translations (Singer & Sigworth, 2020) but, most importantly, *unknown molecule poses* in the 2D images. Indeed, the frozen specimens are differently oriented in the space prior to tomographic projections. Note that, assuming known poses, a 3D reconstruction can be estimated by inverting the projection step, as commonly done in general tomographic imaging.

Dealing with unknown poses remains a crucial step in the reconstruction pipeline. The class of inverse problems with similar pose ambiguities is mathematically formulated in the general framework of multi-reference alignment (Singer, 2018), and there is a plethora of techniques for cryo-EM reconstruction accompanied with software packages (Scheres, 2012; Punjani et al., 2017; Fernandez-Leiro & Scheres, 2017). Dealing with unknown orientations and the alignment problem remains an active area of research, see for example (Fan et al., 2021; Fan & Zhao, 2019a;b; Bandeira et al., 2020; 2017; Perry et al., 2018; Singer et al., 2011). Pure *synchronization* algorithms, which do not take into account the image formation model, tend to suffer in performance in the lowest SNR regimes (Singer & Sigworth, 2020). On the other hand, in Expectation-Maximization (EM) based algorithms, pose estimation and 3D reconstruction happen in an iterative fashion. For example, this is the approach followed in the popular RELION software (Scheres, 2012). Although this approach directly incorporates the data's generative process when estimating the

[*]Equal contribution [1]Qualcomm AI Research, Amsterdam. Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc. [2]QUVA Lab, AMLab, University of Amsterdam. Correspondence to: Gabriele Cesa <gcesa@qti.qualcomm.com>.

Proceedings of the 2^{nd} Annual Workshop on Topology, Algebra, and Geometry in Machine Learning (TAG-ML) at the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. 2023.

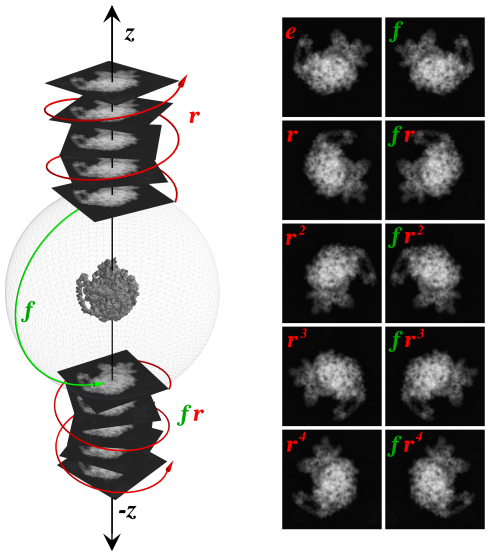


Figure 1: Formation model of Cryo-EM images. Images sharing the same ($\boldsymbol{z}$) or opposite ($-\boldsymbol{z}$) viewing axis differ respectively by a planar rotation $r \in \mathrm{SO}(2)$ or a planar reflection and rotation $fr \in \mathrm{O}(2)$.

poses, it suffers from convergence issues and the additional overhead of performing the 3D reconstruction at each iteration.

In this paper, we propose a deep learning based method able to directly infer the images' poses, while accounting for the generation model of the images. In particular, we adopt a multi-layer *equivariant* graph neural network which simultaneously process a dataset (or a subset) of projection images and predicts an initial estimation of the underlying poses. The equivariant design enables us to encode some of our prior knowledge about the geometry of the problem into the architecture: in particular, our model guarantees that the predicted poses are consistent across rotated and mirrored versions of the same image. However, we emphasise that the model's predictions are not expected to be optimal estimators of the images' poses, but rather a sufficiently good initialization for a second refinement step, which can be performed with more expensive state-of-the-art methods such as RELION. Our equivariant deep learning solution is inspired by the *group synchronization* framework (Perry et al., 2018; Cesa et al., 2022a), which helps the model avoid local optima during the initial phases of pose estimation.

Since the ground truth orientations are not available in real datasets, we follow a *self-supervised* learning approach by introducing a *common-lines* based loss to train the network. Many classical approaches in cryo-EM rely on the common-lines method (Van Heel, 1987; Goncharov, 1986; Singer & Shkolnisky, 2011), although this is less frequent now. The principle is that any two 2D images should contain a pair of central lines on which their Fourier Transforms agree, see Sec. 3 for more details. This *common line* captures two out of three angles in the relative pose of underlying molecules, and all common lines can be used for final pose estimation and reconstruction. Unfortunately, the estimation of common lines is itself expensive and sensitive to noise (Singer et al., 2010). In contrast with Singer & Shkolnisky (2011), we use the information of common lines directly and do not rely purely on relative poses: our loss enforces the consistency of image pairs along the common line defined by their estimated poses. As argued in Sec. 3.2, this allows our method to explicitly account for the generative process of the images during the training phase, thereby circumventing the limitations of pure synchronization methods. Finally, our deep learning design can amortize the cost of pose estimation over images and can be scaled up by using batches of random subsets of images at each iteration.

In summary, our method estimates the final poses using complete information available in image pairs without the overhead of 3D reconstruction, and it can potentially scale up to large number of samples by using amortized inference of poses. While our models unfortunately show *unsatisfying performance* in Sec. 5, we discuss some interesting findings in Sec. 6.

1.1. Cryo-EM image formation model

In a simplified, abstract setting, the cryo-EM image formation model can be summarized as follows.

Let $\Psi : \mathbb{R}^3 \to \mathbb{R}$ be the 3D density function of a molecule. Let $\mathrm{SO}(3)$ be the group of 3D rotations, $\mathrm{SO}(2)$ the group os 2D rotations and $\mathrm{O}(2)$ the group of 2D rotations and reflections. Let $R_i \in \mathrm{SO}(3)$ be a rotation in 3D; in particular, we write $R_i = (\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{z}_i) \in \mathbb{R}^{3\times 3}$, with $\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{z}_i \in \mathbb{R}^3$ to indicate the three orthonormal columns of the matrix R_i.

Then, an image $o_i : \mathbb{R}^2 \to \mathbb{R}$ is generated by the *tomographic* projection Π along the Z axis of the molecule Ψ, after being rotated by R_i^{-1}, i.e. $o_i = \Pi(R_i^{-1}.\Psi)$:

$$o_i(x, y) = [\Pi(R_i^{-1}.\Psi)](x, y) \tag{1}$$

$$= \int_z \Psi\left(R_i(x, y, z)^T\right) dz \tag{2}$$

$$= \int_z \Psi\left(x\boldsymbol{x}_i, y\boldsymbol{y}_i, z\boldsymbol{z}_i\right) dz \tag{3}$$

where $(x, y, z)^T \in \mathbb{R}^3$ is interpreted as a 3D vector. Then, the vector $\boldsymbol{z}_i \in \mathbb{R}^3$ is the direction along which the projection is performed. Fig. 1 provides an example of image formation: note that two images obtained by projecting the molecule along the same axis $\boldsymbol{z}$ are related by a planar rotation $r \in \mathrm{SO}(2)$, while two images obtained by projecting along the opposite axes $\boldsymbol{z}$ and $-\boldsymbol{z}$ are related by a planar rotation and reflection $fr \in \mathrm{O}(2)$.

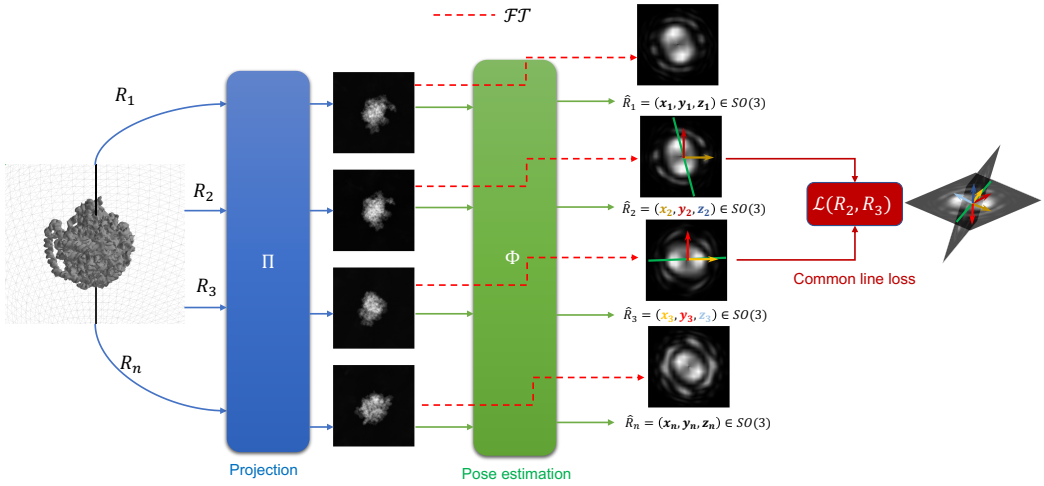


Figure 2: Visualization of the Fourier-Slice theorem together with our pose estimation differentiable model Φ and the common-line loss $\mathcal{L}$.

1.2. Proposed idea

In this work, we train a deep learning model Φ_θ which, given a set of images $\{o_i\}_{i=1}^N$ in input, outputs an estimation of their corresponding poses $\{R_i\}_i^N$ or models a posterior distribution $q_\theta(\{R_i\}_i|\{o_i\}_i)$ over them. In Sec. 4, we describe a few variations of this model; in particular, we will consider models predicting the pose of each image independently or jointly and models leveraging an equivariant design.

Since a typical cryo-EM dataset does not come with ground truth data about the poses or the 3D density, we set up a "*self-supervised*" geometric loss to train the model. For each pair of images o_i and o_j, we use a neural network Φ_θ to sample or estimate the poses R_i and R_j and train it to minimize the discrepancy between the images o_i and o_j in the Fourier domain along their corresponding *common line*, which can be estimated directly from R_i and R_j, as we will see in Sec. 3.

2. Related Work

Group Synchronization: Given a noisy observation of the relative poses between a set of images, the group synchronization problem consists in *finding an assignment of absolute poses* to the images which is *most consistent with the relative ones*. Perry et al. (2018) derived an Approximate Message Passing (AMP) algorithm (an approximation of Belief Propagation) to solve the generic group synchronization problem when all relative poses are observed. In cryo-EM, the relative pose $R_{ij} = R_j^{-1} R_i$ can be estimated by directly comparing the images o_i and o_j whenever they approximately have the same or opposite viewing direction, in which case $R_{ij} \in \mathrm{O}(2)$ is only a planar rotation or reflection. Cesa et al. (2022a) studied the synchronization when only relative poses in $\mathrm{O}(2)$ are available and considered a spectral relaxation of the problem.

Common line methods: Previous works (Pragier & Shkolnisky, 2019; Singer et al., 2010; Singer & Shkolnisky, 2011; Shkolnisky & Singer, 2012; Wang et al., 2013) studied the fact that the Fourier transforms of cryo-EM image pairs must align along a line passing through their origin. This property can be used to establish constraints on the absolute poses of each pair of images, which can be solved in various ways to estimate the final poses. For example, Bandeira et al. (2020) used Semi-Definite Programming (SDP) while Singer & Shkolnisky (2011); Shkolnisky & Singer (2012) considered a spectral relaxation of the problem, which is faster to solve. Unfortunately, the estimation of common lines itself is an expensive process that involves comparing each pair of images, and it is highly sensitive to noise; see (Singer et al., 2010).

Deep learning approaches for cryo-EM: A variety of deep learning solutions for cryo-EM have been proposed in the literature. Ullrich et al. (2019) first proposed using a variational framework to solve the reconstruction problem, assuming known poses. Rosenbaum et al. (2021) used amortized inference over the unknown poses but required prior information about the backbone of the protein. Other works leverage implicit neural functions to model the 3D molecular density (Zhong et al., 2019; 2021). More recently, (Levy et al., 2022a;b) successfully applied amortized inference for complete *ab initio* reconstruction. See also (Donnat et al., 2022; Toader et al., 2023) for recent surveys of the deep learning methods. In Sec. 4, our MLP architectures predict the pose of each image independently too: while they resemble the previous deep amortized inference approaches, we don't use a reconstruction based loss to train them. We are not aware of previous deep learning methods predicting poses using a multitude of images, rather than single ones,

like our message-passing models in Sec. 4.

SOTA methods: RELION (Scheres, 2012) and cryoSPARC (Punjani et al., 2017) are two popular softwares implementing state-of-the-art reconstruction methods based on EM-like algorithm with soft-assignment of poses.

3. The Common Lines Loss

3.1. Fourier Slice Theorem and Common Lines

The tomographic projection Π is a linear operator and corresponds to a 1D frequency-0 *Fourier Transform* $\mathcal{F}_1$ of the 3D density along the Z axis. For this reason, it is convenient to work with the images and the 3D density parameterized in the Fourier domain.

Formally, we assume both the 3D density Ψ and any 2D image o_i to be approximately *band-limited* and locally-supported. We also assume the density and the images, as well as their Fourier transform, to be square-integrable, ensuring the invertibility of the Fourier transform and the unitary action of the rotations on them. In particular, by denoting with $\mathcal{F}_{\mathrm{d}}[\cdot]$ the d-dimensional Fourier transform operator, we assume $\Psi, \mathcal{F}_3[\Psi] \in L^2(\mathbb{R}^3)$ and $o_i, \mathcal{F}_2[o_i] \in L^2(\mathbb{R}^2)$. Note also that the Fourier Transform $\mathcal{F}_{\mathrm{d}}[\cdot] : \mathbb{R}^d \to \mathbb{C}$ is a complex-valued signal.

First, define the 2D and 3D Fourier transforms as:

$$\widehat{o_i}(k_x, k_y) = \mathcal{F}_2[o_i](k_x, k_y) = \int_{\mathbb{R}^2} o_i(x, y) e^{-i2\pi(xk_x + yk_y)} dxdy \tag{4}$$

$$\widehat{\Psi}(k_x, k_y, k_z) = \mathcal{F}_3[\Psi](k_x, k_y, k_z) = \int_{\mathbb{R}^3} \Psi(x, y, z) e^{-i2\pi(xk_x + yk_y + zk_z)} dxdydz \tag{5}$$

By applying the Fourier transform on Eq. 1 and defining $\boldsymbol{k} = (k_x, k_y, 0)^T$, one can show[1]:

$$\mathcal{F}_2[o_i](k_x, k_y) = \mathcal{F}_3\left[R_i^{-1}.\Psi\right](k_x, k_y, 0) \tag{6}$$

$$= \mathcal{F}_3\left[R_i^{-1}.\Psi\right](\boldsymbol{k}) = \mathcal{F}_3[\Psi](R_i.\boldsymbol{k}) \tag{7}$$

i.e. the Fourier Transform of the image o_i corresponds to a 2D slice of the Fourier transform of the 3D density Ψ along the plane obtained by rotating the XY plane (orthogonal to the Z axis) with R_i. In particular, if $R_i = (\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{z}_i) \in \mathrm{SO}(3)$, this plane is spanned precisely by $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ and it is orthogonal to $\boldsymbol{z}_i$. This is visualized in Fig. 2, where the Fourier transforms of different images correspond to different 2D slices of the Fourier Transform of the 3D density.

[1]This equality is true up to a constant factor depending on the normalization considered in the definition of the Fourier transform. When considering an unitary discrete Fourier transform over a grid of size D, a factor $\sqrt{D}$ should be included.

Hence, it is more practical to consider a definition of the tomographic projection operator directly in the Fourier domain $\Pi : L^2(\mathbb{R}^3) \to L^2(\mathbb{R}^2)$ as

$$[\Pi\,\mathcal{F}_3[\Psi]](x, y) := \mathcal{F}_3[\Psi](x, y, 0) \tag{8}$$

To simplify the notation, we will not distinguish a density from its Fourier transform unless necessary, i.e. Ψ and o_i refer to the 3D density function and the i-th image in a basis-independent manner. Similarly, we will let Π generally operate on both density functions or their Fourier transforms, so we can generally write $o_i = \Pi R_i^{-1}\Psi$, where R_i^{-1} is interpreted as the *unitary* linear operator acting on $L^2(\mathbb{R}^3)$, the vector space of density functions, independently from the choice of basis for this space (i.e. in the spatial or the Fourier domain).

Common Lines This property leads to an important observation: the Fourier transforms $(\mathcal{F}_2[o_i], \mathcal{F}_2[o_j])$ of any two (non-coplanar) images (o_i, o_j) agree exactly along a line passing through the origin, i.e. along the intersection of the corresponding 2D slices. Geometrically, because this line belongs to both planes, it must be orthogonal to both $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$. It follows that the common line is spanned by the *cross-product* of them, that is by the vector $\boldsymbol{l}_{ij} = \frac{\boldsymbol{z}_i \times \boldsymbol{z}_j}{\|\boldsymbol{z}_i \times \boldsymbol{z}_j\|} \in \mathbb{R}^3$.

3.2. Deriving the Common-Line Loss

In this section, we derive our self-supervised loss from variational inference principles to show it encodes all information about the cryo-EM generative process.

Following the variational inference framework (Kingma & Welling, 2013), we consider the unknown poses as latent variables, a posterior (encoder) $q_\theta(\{R_i\}_i | \{x_i\}_i)$, parameterized by our neural network Φ_θ, and a generative process (decoder) $p_\Psi(\{o_i\}_i | \{R_i\}_i) = \prod_i p_\Psi(o_i | R_i)$, parameterized by the 3D molecular density Ψ. As commonly done in the literature, we assume i.i.d. Gaussian noise with variance σ^2, i.e. $p_\Psi(o_i | R_i) = \mathcal{N}(o_i | \Pi R_i^{-1}\Psi, \sigma^2 I)$. Assuming unitary Fourier transform, the i.i.d. Gaussian assumption holds in the Fourier domain, too, i.e. $p_\Psi(\mathcal{F}_2[o_i] | R_i) = \mathcal{N}(\mathcal{F}_2[o_i] | \mathcal{F}_3\left[\Pi R_i^{-1}\Psi\right], \sigma^2 I)$. From now on, we will not explicitly write the Fourier transform operator $\mathcal{F}_{\mathrm{d}}[\cdot]$.

In a typical scenario, one would optimize both θ and Ψ by maximizing the variational lower bound:

$$\mathcal{L}(\theta, \Psi; \{o_i\}_i) = -KL\left(q_\theta(\{R_i\}_i | \{o_i\}_i) | p_\Psi(\{R_i\}_i)\right) + \mathbb{E}_{q_\theta(\{R_i\}_i | \{o_i\}_i)}\left[\log p_\Psi(\{R_i\}_i | \{o_i\}_i)\right] \tag{9}$$

which, using a uniform prior $p_\Psi(\{R_i\}_i)$ over the poses and

expanding the true posterior, equals

$$\mathcal{L}(\theta, \Psi; \{o_i\}_i) = H\left(q_\theta(\{R_i\}_i|\{o_i\}_i)\right) - \mathbb{E}_{q_\theta(\{R_i\}_i|\{o_i\}_i)}\left[\frac{1}{2\sigma^2}\sum_i \left\|\Pi(R_i^{-1}\Psi) - o_i\right\|_2^2\right] \tag{10}$$

Note that this target contains all the information about the generative process of the images and the second term resembles the quadratic loss used in Levy et al. (2022a).

However, in our case, we do not want to explicitly estimate the 3D density Ψ, yet. Instead, recall that, for a fixed assignment of the rotations $\{R_i\}_i$, maximizing over Ψ is equivalent to a *least squares solution* to the inverse linear problem given by the observed images. In other words, by denoting with $\cdot^H$ the conjugate transpose, the density Ψ which minimizes the target

$$\sum_i \left\|\Pi(R_i^{-1}\Psi) - o_i\right\|^2 = \Psi^H\left(\sum_i R_i\Pi^H\Pi R_i^{-1}\right)\Psi + \sum_i \|o_i\|^2 - 2\Psi^H\left(\sum_i R_i\Pi^H o_i\right) \tag{11}$$

is given by the Moore-Penrose pseudo-inverse[2] :

$$\Psi = \left(\sum_i R_i\Pi^H\Pi R_i^{-1}\right)^{-1}\left(\sum_i R_i\Pi^H o_i\right) \tag{12}$$

Let's study the term $\sum_i R_i\Pi^H\Pi R_i^{-1}$.

Theorem 3.1. *Let $\Pi : L^2(\mathbb{R}^3) \to L^2(\mathbb{R}^2)$ be the tomographic projection along the Z axis defined in the Fourier domain via slicing as in Eq. 8. The operator $\sum_i R_i\Pi^H\Pi R_i^{-1}$ is diagonal and counts for each frequency $\boldsymbol{k} \in \mathbb{R}^3$ the number of images containing it.*

Proof. Let $V_i \cong L^2(\mathbb{R}^2)$ be the subspace of $L^2(\mathbb{R}^3)$ of functions over the 2D plane described by the rotation $R_i = (\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{z}_i)$, i.e. orthogonal to $\boldsymbol{z}_i$. Note that $V_i \cong L^2(\mathbb{R}^2)$ is the co-image of ΠR_i^{-1} or the image of the back-projection $R_i\Pi^H$. Then, $\Pi_i := R_i\Pi^H\Pi R_i^{-1} : L^2(\mathbb{R}^3) \to L^2(\mathbb{R}^3)$ is an orthogonal *projection operator* on $V_i \subset L^2(\mathbb{R}^3)$, i.e. it acts as the identity on V_i. In other words, Π_i is the identity on the frequency $\boldsymbol{k} = (k_x, k_y, k_z)^T$ if $\boldsymbol{k}$ belongs to the plane spanned by $(\boldsymbol{x}_i, \boldsymbol{y}_i)$, i.e. if $\boldsymbol{k} \perp \boldsymbol{z}_i$, and is zero otherwise. Then, the operator $\sum_i R_i\Pi^H\Pi R_i^{-1} = \sum_i \Pi_i$ is diagonal, i.e. it acts on each frequency $\boldsymbol{k}$ independently by scaling it by a factor equal to the number of $R_i = (\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{z}_i)$ such that $\boldsymbol{k} \perp \boldsymbol{z}_i$. □

[2]We implicitly restrict our consideration to the subspace of $L^2(\mathbb{R}^3)$ corresponding to the subset of frequencies appearing in at least one image, such that $\sum_i R_i\Pi^H\Pi R_i^{-1}$ is invertible.

Then, *assuming rotations are approximately uniformly distributed*, the operator $\sum_i R_i\Pi^T\Pi R_i^T$ is approximately a scalar multiple of the identity ηI, where η is the average number of images any 3D frequency appears in.

By replacing this matrix with ηI and replacing $\Psi \approx \eta^{-1}\left(\sum_i R_i\Pi^H o_i\right)$ in the target function, we obtain

$$\sum_i \left\|\Pi(R_i^{-1}\Psi) - o_i\right\|^2 = \eta^{-2}\sum_{ijk} o_j^H\Pi R_j^{-1}R_i\Pi^H\Pi R_i^{-1}R_k\Pi^H o_k + \sum_{ij}\|o_i\|^2 - 2\eta^{-1}\sum_{ij} o_j^H\Pi R_j^{-1}R_i\Pi^H o_i \tag{13}$$

Note now that the operator $\Pi R_j R_i^{-1}\Pi^H$ projects the common line from image j to image i and, therefore, $o_j^H\Pi R_j R_i^{-1}\Pi^H o_i$ is just the inner product of the images o_i and o_j along their common line. The order-three quantity $o_j^H\Pi R_j^{-1}R_i\Pi^H\Pi R_i^{-1}R_k\Pi^H o_k$ is the inner product between o_j and o_k along those points shared between the three images i, j, k; because the intersection of three generic planes contain just the origin, this term is almost always just the average density of the molecule[3] (i.e. the frequency $\mathbf{0}$ Fourier Transform), which is a constant term.

Finally, we recognize a simple quadratic loss

$$\sum_i \left\|\Pi(R_i^{-1}\Psi) - o_i\right\|^2 \approx \sum_{ij}\|o_i\|^2 - 2\eta^{-1}\sum_{ij} o_j^H\Pi R_j^{-1}R_i\Pi^H o_i \tag{14}$$

which enforces each pair of images (o_i, o_j) to agree along the common line defined by their respective estimated poses (R_i, R_j). Hence, dropping the constant terms $\|o_i\|^2$, the final training target we want to maximize becomes

$$\mathcal{L}(\theta, \cancel{\Psi}; \{o_i\}_i) = H\left(q_\theta(\{R_i\}_i|\{o_i\}_i)\right) + \mathbb{E}_{q_\theta(\{R_i\}_i|\{o_i\}_i)}\left[\frac{1}{\eta\sigma^2}\sum_{ij}\mathcal{L}(R_i, R_j)\right] \tag{15}$$

with $\mathcal{L}(R_i, R_j) = o_j^H\Pi R_j^{-1}R_i\Pi^H o_i$, which is only a function of the parameters θ of our encoder parameterising the posteriors.

3.3. Computing the Common Line

Note that if $R_i = (\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{z}_i)^T$, the vector $\boldsymbol{z}_i$ defines the axis orthogonal to the plane spanned by $(\boldsymbol{x}_i, \boldsymbol{y}_i)$. Hence,

[3]The intersection is non-degenerate only if the common-line $\boldsymbol{l}_{ij} = \frac{\boldsymbol{z}_i \times \boldsymbol{z}_j}{\|\boldsymbol{z}_i \times \boldsymbol{z}_j\|}$ belongs to o_k too, i.e. if $\boldsymbol{z}_k^T\boldsymbol{l}_{ij} = 0$. Assuming R_i, R_j, R_k are uniformly distributed, the product $\boldsymbol{z}_k^T\boldsymbol{l}_{ij}$ is approximately uniform in $[-1, 1]$, so the case $\boldsymbol{z}_k^T\boldsymbol{l}_{ij} = 0$ is negligible.

the common line between two planes, identified by the two axes $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$, is a line orthogonal to both $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$. An orthogonal basis for this line is easily obtained via the (normalized) *cross product* $\boldsymbol{l}_{ij} = \frac{\boldsymbol{z}_i \times \boldsymbol{z}_j}{\|\boldsymbol{z}_i \times \boldsymbol{z}_j\|_2}$. Next, we need to find the equation of the line $\boldsymbol{l}_{ij}$ inside both planes, i.e. we have to express $\boldsymbol{l}_{ij}$ with respect to $\boldsymbol{x}_i, \boldsymbol{y}_i$ and with respect to $\boldsymbol{x}_j, \boldsymbol{y}_j$. Since these vectors are orthonormal, this is simply given by the projection on them, i.e.:

$$x_i = \boldsymbol{l}_{ij}^T \boldsymbol{x}_i \quad y_i = \boldsymbol{l}_{ij}^T \boldsymbol{y}_i \quad x_j = \boldsymbol{l}_{ij}^T \boldsymbol{x}_j \quad y_j = \boldsymbol{l}_{ij}^T \boldsymbol{y}_j \quad (16)$$

Note that all these computations are differentiable with respect to the predicted poses[4] R_i, R_j.

Final loss We can use this to extract the common line from two images o_i and o_j and then compute their discrepancy. The loss in Eq. 15 can be implemented by using

$$\mathcal{L}(R_i, R_j) = \int_{\mathbb{R}} \mathcal{F}_2\left[o_i\right](\lambda x_i, \lambda y_i) \cdot \overline{\mathcal{F}_2\left[o_j\right](\lambda x_j, \lambda y_j)} d\lambda \quad (17)$$

where $\bar{\cdot}$ represents complex conjugation. As argued in Sec. 3.2, this common-lines loss encodes the cryo-EM generative process and the geometry of the problem, providing all constraints needed to solve the linear inverse problem.

In practice, we find a mean squared error more practical and helpful to achieve convergence:

$$\mathcal{L}(R_i, R_j) = \quad (18)$$
$$- \int_{\mathbb{R}} \left|\mathcal{F}_2\left[o_i\right](\lambda x_i, \lambda y_i) - \mathcal{F}_2\left[o_j\right](\lambda x_j, \lambda y_j)\right|^2 d\lambda$$

Note that this squared error includes the inner product in Eq. 15 but also penalizes common lines which have higher norm. This is helpful to avoid local optima where the model only picks the line within an image with highest norm, regardless of its alignment with the lines in the other images; see Apx. B. Finally, the loss is implemented by sampling a discrete number L of points along the common line in both images o_i and o_j. This operation is differentiable with respect to the sampling coordinates $\{(\lambda_l x_i, \lambda_l y_i)\}_l^L$. We use `torch.nn.function.grid_sample` to sample $L = 101$ points with bilinear interpolation from each image.

Frequency Marching As commonly done in the literature (Zhong et al., 2021), we follow a frequency-marching strategy, i.e. we only use a low-resolution (heavily band-limited) version of the images in our loss function, but we gradually increase their resolution during the training process. This approach is helpful since the Fourier transforms contain most energy in the low frequencies, while higher frequencies are more affected by the i.i.d. white noise. By initially using only the lowest frequencies, it is easier for the models to avoid spurious local optima created by the noise.

Additional Regularization It is possible that a model's predictions are mostly concentrated around the same pose; in particular, this is likely to happen at initialization. In this case, our loss is not suitable, since we have assumed that all poses are different such that the common line between each pair of images is well defined. Indeed, whenever R_i and R_j share a similar or opposite viewing direction $\boldsymbol{z}_i \approx \pm \boldsymbol{z}_j$, the gradient $\frac{\partial \mathcal{L}(R_i, R_j)}{\partial \boldsymbol{z}_i}$ is particularly noisy and unstable; see Apx. A and Fig. 4. In order to prevent the model from getting stuck in these solutions, we include an additional regularization term which forces the predicted poses to spread. Specifically, we consider a linear combination of three terms. 1) The first term forces the *center* of the set of vectors $\{\boldsymbol{z}_i\}_i$ (recall that $\boldsymbol{z}_i$ is the viewing direction along which the volume is projected to generate o_i) to be close to zero $\lambda^{(1)}(\{\boldsymbol{z}_i\}) = \frac{1}{3}\left\|\frac{1}{N}\sum_i \boldsymbol{z}_i\right\|_2^2$. 2) The second term forces the covariance of the vectors $\{\boldsymbol{z}_i\}_i$ to be close to identity matrix divided by 3 (this is the covariance of a uniform distribution on the unit sphere) $\lambda^{(2)}(\{\boldsymbol{z}_i\}) = \frac{1}{9}|\operatorname{Cov}(\{\boldsymbol{z}_i\}) - \frac{1}{3}I|$. 3) The last term $\lambda^{(3)}$ is an energy function modelling repulsive forces between each pair of vectors in $\{\boldsymbol{z}_i\}_i$, defined[5] as $\lambda_{ij}^{(3)}(\boldsymbol{z}_i, \boldsymbol{z}_j) = \min(|\boldsymbol{z}_i^T \boldsymbol{z}_j|, 0.6)$. The final regularization term to minimize is given by $0.15\lambda^{(1)}(\{\boldsymbol{z}_i\}) + 0.3\lambda^{(2)}(\{\boldsymbol{z}_i\}) + \frac{1}{N^2}\sum_{i \neq j} \lambda_{ij}^{(3)}(\boldsymbol{z}_i, \boldsymbol{z}_j)$. We found these coefficients with a short hyperparameter search, comparing the final performance of our models.

4. Deep Learning and Inductive Biases

In this section, we consider a few different approaches to design the network Φ_θ used to predict the poses.

Steerable PCA features In all cases, for simplicity, we do not use the full images $\{o_i\}_i$ as inputs to our model. Instead, we leverage steerable PCA (Zhao et al., 2016; Zhao & Singer, 2013) to project the images in the dataset to *lower dimensional* feature vectors. In particular, we use the ASPIRE software to project to the top 400 principal components with angular frequency smaller than 12. Note that steerable PCA is typically used in ASPIRE (Zhao & Singer, 2014) as a preprocessing step before estimating the images with similar

[4] Here, we consider $R_i, R_j \in \mathbb{R}^{3\times 3}$ as matrices. A neural network outputs an element of $\mathrm{SO}(3)$ (e.g. using quaternions or SVD as in Sec. 4.1), which then needs to be converted into a 3×3 matrix to compute the loss. If this conversion is differentiable, the gradient can be back-propagated through the network's output.

[5] Note that the use of the absolute value $|\cdot|$ implies the forces depend on the angular distance between the axes aligned with the vectors but they are independent of the vectors' directions. min is used to ensure the regularization only includes local repulsion and does not enforce a uniform distribution too heavily.

viewing directions and their relative rotations. Fig. 3 shows some examples of denoised images using steerable PCA.

Variational Inference vs Point Estimate In Sec. 3.2 we derived our loss under the variational inference framework, but recent deep learning approaches to cryo-EM only rely on deterministic encoder architectures, which just provide point estimates of the poses, e.g. (Levy et al., 2022a). Similarly, our methods also only output point-estimates.

π-rotation augmentation As observed in Levy et al. (2022a), noisy low-resolution images are difficult to distinguish from their version rotated by π. Levy et al. (2022a) consider a "symmetrized loss" by only backpropagating through the lowest loss achieved by an image o_i or its rotated version $r_\pi.o_i$. Instead of adapting our loss[6], for each image o_i, our models output both the estimated pose $\widehat{R}_i$ and a binary distribution over $\{\widehat{R}_i, \widehat{R}_i r_\pi\}$. We use the Gumbel-Softmax trick to sample and train the models; by using an initial high-temperature, we initially sample the two rotations randomly, ensuring sufficient exploration, but we gradually decrease the temperature over the first three training epochs, allowing the models to learn to prefer only one of the two solutions.

Amortized Inference with MLP The simplest design employs the same network to predict the pose of each image independently from the others, i.e. $\widehat{R}_i = \Phi_\theta(o_i)$. Because we use the steerable PCA features instead of the raw images, this architecture is implemented by a simple MLP. This approach is a simplified version of the amortized inference used in Levy et al. (2022a), which trains a convolutional neural network (CNN) to predict the pose of each image independently, but differs by the training loss used.

Message Passing Inspired by the *group synchronization framework* (see Sec. 2), we also consider more complex architectures which predict an image's pose conditioned on all images in a batch. The intuition behind this idea is that the relative poses between pairs of images provide sufficient information to compute an approximate estimation of the absolute poses. Moreover, the synchronization problem is typically easier to solve than the complete cryo-EM inverse problem (even state-of-the-art solutions can suffer from convergence issues); for instance, Perry et al. (2018) describe an iterative message passing algorithm which provably converges to a solution, while Cesa et al. (2022a) consider a spectral relaxation of the problem which can be directly solved via eigenvalue decomposition. For this reason, we expect our models can benefit from the relative poses as well as all images' features to estimate a single image's pose. Following this design principle, in Sec. 4.2, we consider an architecture which uses attention to leverage the intermediate features of all images in a batch to estimate their poses. Instead, in Sec. 4.4, we describe an equivariant architecture which includes a message-passing module similar to Perry et al. (2018), sharing messages between nodes corresponding to images with similar viewing directions and aligning the features in the neighborhood of a node using the relative rotations on the edges.

[6]Note that, if our model is $\mathrm{O}(2)$ equivariant as in Sec. 4.3, the prediction of o_i and $r_\pi.o_i$ are guaranteed to be consistent, so there is no benefit in using the symmetrized loss in such case.

4.1. Parameterizing $\mathrm{SO}(3)$ elements

Our architectures need to output elements in $\mathrm{SO}(3)$. To do so, our models output two vectors $\boldsymbol{x}', \boldsymbol{y}' \in \mathbb{R}^3$. We map this output to $\mathrm{SO}(3)$ by projecting the matrix $R' = (\boldsymbol{x}', \boldsymbol{y}', \boldsymbol{x}' \times \boldsymbol{y}') \in \mathbb{R}^{3\times 3}$ to the closest $\mathrm{SO}(3)$ matrix via SVD[7], i.e. if $R' = U\Sigma V^T$ is the SVD of R', the matrix $R = UV^T$ is its projection to $\mathrm{SO}(3)$. Importantly, this construction is equivariant to the group $\mathrm{O}(2)$ acting on the plane spanned by $(\boldsymbol{x}', \boldsymbol{y}')$, which is important in Sec. 4.3.

4.2. Non-equivariant baselines

The first baseline is a simple 6-layer MLP. We also consider an attention-based architecture (MLP-Self-Attention), including an MLP (processing each image independently) followed by 4 self-attention layers. Self-attention is applied across the full set of images present in a mini-batch.

4.3. Local $\mathrm{O}(2)$ Equivariance

The cryo-EM problem presents a number of symmetries which can be leveraged by *equivariant neural networks* (Cohen & Welling, 2016a; Cohen et al., 2018; Kondor & Trivedi, 2018; Weiler et al., 2021). We summarize these symmetries in Tab. 1; see also (Cesa et al., 2022a) for a more detailed discussion.

Indeed, note that if a single image o_i is mirrored or transformed by a planar rotation $g \in \mathrm{O}(2)$, the pose of the new image $g.o_i$ is related to the original one by a similar transformation, i.e. $R_i g^{-1}$ (rows 1 and 4 of Tab. 1 and Fig. 1). Here, the action of $g = r_\alpha f^c \in \mathrm{O}(2)$, with $\alpha \in [0, 2\pi)$ and $c \in \{0, 1\}$, on $\mathrm{SO}(3)$ in $R_i g^{-1}$ is given by

$$g : R_i \mapsto R_i g^{-1} = R_i \begin{bmatrix} -1^c & & \\ & 1 & \\ & & -1^c \end{bmatrix} \begin{bmatrix} \cos\alpha & -\sin\alpha & \\ \sin\alpha & \cos\alpha & \\ & & 1 \end{bmatrix} \tag{19}$$

This *local* $\mathrm{O}(2)$ symmetry can be encoded into a neural network via equivariance. Specifically, an $\mathrm{O}(2)$ equivariant model *satisfies the following constraint by design*:

$$\Phi_\theta(g.o_i, \{o_j\}_{j\neq i}) = \Phi_\theta(o_i, \{o_j\}_{j\neq i}) g^{-1} \quad \forall g \in \mathrm{O}(2) \ .$$

[7]Note that, since R' is constructed by using $\boldsymbol{x}' \times \boldsymbol{y}'$, if $\boldsymbol{x}' \neq \boldsymbol{y}'$, $\det R' > 0$. Since Σ has non-negative entries, it follows that $R = UV^T$ already has positive determinant.

Table 1: Summary of the Cryo-EM symmetries. 1-4 are *local* symmetries, involving only the pose of single images. 5-6 are *global* symmetries, involving a transformation of a reference density Ψ and, therefore, of all images' poses.

		description	symmetry
1	$r.o_i = \Pi(rR_i^{-1}.\Psi)$	$r \in \mathrm{SO}(2)$ rotation around Z axis	$\mathrm{SO}(2)$ equivariance
2	$o_i = \Pi(\mathrm{m}_z\, R_i^{-1}.\Psi)$	$\mathrm{m}_z \in \mathrm{O}(3)$ mirroring along Z axis	Mirroring (Z) invariance
3	$\mathrm{f}\,.o_i = \Pi(m_x R_i^{-1}.\Psi)$	$\mathrm{m}_x \in \mathrm{O}(3), \mathrm{f} \in \mathrm{O}(2)$ mirroring along X axis	Mirroring (XY) equivariance
4	$\mathrm{f}\,.o_i = \Pi(r_y R_i^{-1}.\Psi)$	$r_y = \mathrm{m}_x\, \mathrm{m}_z \in \mathrm{SO}(3)$ rotation by π around Y axis	Reflection equivariance
5	$o_i = \Pi((RR_i)^{-1}.R\Psi)$	$R \in \mathrm{SO}(3)$	pose ambiguity
6	$o_i = \Pi((R_i r_z)^{-1}.\mathrm{i}\,\Psi)$	$\mathrm{i} = -1 \cdot I \in \mathrm{O}(3)$ inversion, $r_z \in \mathrm{SO}(3)$ rotation by π around Z axis	chirality ambiguity

We construct a 6-layer equivariant MLP using the `escnn` library (Weiler & Cesa, 2019; Cesa et al., 2022b). In the design of the network, the input features already carry an action of $\mathrm{O}(2)$ since they are generated by steerable PCA. The output are two vectors $\boldsymbol{x}', \boldsymbol{y}'$ as described in Sec. 4.1; the action of $\mathrm{O}(2)$ on them is the one defined in Eq. 19, restricted on the first two columns of R_i. In the intermediate layers of the network, we use ~ 500 channels containing copies of regular representations of $\mathrm{O}(2)$, band-limited up to a maximum frequency L, decreasing with the depth of the network from $L = 12$ in input to $L = 1$ in output. We apply pointwise ELU non-linearities computed using a discretized Fourier transform as described in Cesa et al. (2022b).

4.4. The group synchronization problem and Global $\mathrm{SO}(3)$ Equivariance

As we argued earlier, the estimation of the poses of a set of images can be related to the group synchronization problem. In this section, we assume the reader has some familiarity with the framework of *steerable* and *gauge CNNs* (Cohen & Welling, 2016b; Weiler et al., 2021; Cesa et al., 2022b) and the message passing algorithm from Perry et al. (2018).

First, we note that the AMP algorithm of Perry et al. (2018) interestingly resembles the typical design of *steerable* and *gauge CNNs*, which use the following building block[8]:

$$f^{l+1}(i) = \sigma\left(W_l \sum_{j\sim i} \rho^l(g_{ij}) f^l(j) + b_l\right) \qquad (20)$$

where ρ^l is the *representation* of the equivariance group G acting on the features $f^l(j) \in \mathbb{R}^{c_l}$ and $W_l \in \mathbb{R}^{c_{l+1}\times c_l}$ is an equivariant linear map. In particular, we highlight the following facts: 1) each channel in the features $f^l(i)$ associated to a node i are (bandlimited) Fourier transforms of probability density functions over $\mathrm{SO}(3)$, representing the estimated posteriors. 2) the message-passing module parallel-transports these features along an edge $j \sim i$ by rotating them via $\rho(g_{ij})$ according to the relative rotation $g_{ij} \in G$ on the edge. 3) the algorithm alternates a message-passing step with the application of a softmax activation σ, which turns the aggregated messages into a probability distribution. A gauge CNN like de Haan et al. (2021) follows a similar pattern (possibly, replacing softmax with another activation σ).

Given this observation, we unfold the basic message passing of AMP (without Onsager correction) into a neural network. A similar, albeit more complex, strategy is typically used to train neural networks to solve compressed-sensing problems in a principled way, e.g. (Gregor & LeCun, 2010; Borgerding & Schniter, 2016). Specifically, each layer uses this message passing for each channel independently and then learns a G-equivariant linear map W_l to mix the features of each node. We replace softmax with a simpler ELU activation applied over features in the $\mathrm{SO}(3)$ regular representation band-limited up to frequency $L = 2$. Our 6-layer message-passing network is preceded by a 2-layer $\mathrm{O}(2)$ equivariant MLP encoder, which processes each image independently and initializes the message passing features.

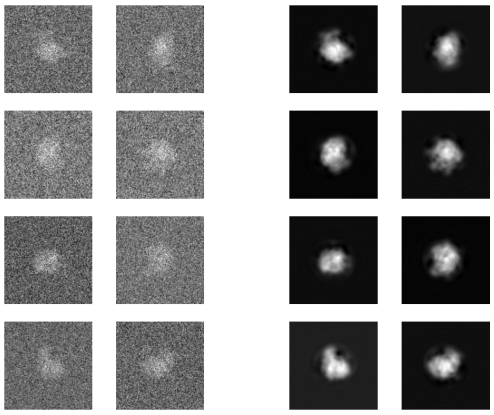
With respect to Perry et al. (2018), we don't know the full synchronization graph, with the relative poses in $\mathrm{SO}(3)$ of every pair, but we can still rely on a local version of it. As shown in Cesa et al. (2022a), these local relative poses in $G = \mathrm{O}(2)$ still sufficiently constrain the global ones. Moreover, they show that these local messages approximate a *local parallel transport operator* over the *projective plane*, which further motivates the relation with Gauge CNNs.

However, the solution to the synchronization problem is not unique, and the solution space presents a global symmetry. Indeed, note that the set of poses $\{R_i\}_i^N$ can be simultaneously transformed as $\{RR_i\}_i^N$ by a 3D rotation $R \in \mathrm{SO}(3)$ while still being valid (row 5 of Tab. 1). This symmetry is also related to the fact that cryo-EM images do not contain information about the actual pose of the molecule. In summary, the full symmetry is given by $\mathrm{SO}(3) \times \mathrm{O}(2)^{\times N}$, where $\mathrm{SO}(3)$ acts (globally) on the *left* while each $\mathrm{O}(2)$ acts (locally) on the *right* of an estimated pose. Hence, here $G = \mathrm{SO}(3) \times \mathrm{O}(2)$. Since the features $f^l(i)$ of each node i contain band-limited functions over $\mathrm{SO}(3)$, the action ρ^l of G is just induced by its action on the elements of $\mathrm{SO}(3)$.

AMP is equivariant to this full symmetry, and our neural version will be equivariant too if its linear layers are also $G = \mathrm{SO}(3)\times\mathrm{O}(2)$ equivariant. While our network does not

[8] Although messages can be weighted by non-isotropic convolution kernels $W_l(\Delta x_{ij})$ rather than a constant kernel W_l.

need $\mathrm{SO}(3)$ equivariance (the global symmetry is already broken by the $\mathrm{O}(2)$ equivariant MLP encoder preceding the message passing), we choose to evaluate both versions of the architecture in our experiments.



(a) Original images with $SNR = 0.12$.

(b) Denoised images with Steerable PCA.

Figure 3: Examples of images from the dataset and their denoised version using Steerable PCA.

5. Experiments

To validate our method, we use a synthetic dataset generated from the density of the 70S Ribosome[9] (Hentschel et al., 2017). We use 5000 images with SNR 0.12 and size 97×97; see Fig. 3. Since we consider only synthetic data with known poses, we can directly evaluate the quality of the predicted poses. During training and evaluation, we use different random batches containing subsets of the dataset; the final performance is given by the averaged predictions. Estimated and real poses are compared using the correlation[10] $\frac{1}{N\sqrt{3}} \left\| \sum_i R_i \widehat{R}_i^T \right\|_F$. During training, for each image o_i in the batch, the common-line loss is evaluated on the pair (i, j) for 200 random images o_j in the batch. A short hyperparameter search is performed to tune some design choices and standard training parameters.

Table 2: Pose correlation for different methods.

Method	Pose Correlation
VDM $\mathrm{O}(2)$ $L = 1$ (Cesa et al., 2022b)	82.0
MFVDM $\mathrm{O}(2)$ $L = 5$ (Cesa et al., 2022b)	97.5
MLP	43.2 ± 0.1
MLP-Self-Attn	39.9 ± 0.005
MLP $\mathrm{O}(2)$	92.2
GNN $G = \mathrm{O}(2)$ $L = 2$	93.8
GNN $G = \mathrm{SO}(3) \times \mathrm{O}(2)$ $L = 2$	95.4

Using the full dataset, we build the (frequency $L = 1$) $\mathrm{O}(2)$ Vector Diffusion Map (VDM) matrix (Cesa et al., 2022b) and save its top eigenvectors. For a random batch, the local synchronization graph is efficiently estimated by using these eigenvectors[11], i.e. we use the MFVDM denoising from Cesa et al. (2022b) with $L = 1$. We also use the full MFVDM algorithm as a baseline; it denoises the nearest neighbors and the relative rotations via MFVDM up to frequency $L = 1$ or 5 and performs the final synchronization via spectral relaxation. To evaluate the appropriateness of the proposed self-supervised loss and compare it with our non-trainable baseline, we only consider the task of overfitting the training data. Tab. 2 reports our results; unfortunately, our methods *do not improve* over the baseline.

6. Discussion of Results and Conclusion

In this work, we proposed two novel ideas: 1) a common-lines based loss to train a deep pose estimator network and replace the expensive intermediate 3D reconstructions and 2) the integration of group synchronization to improve the convergence of the pose estimation. Specifically, we expected group synchronization to help avoiding local optima and the common-line loss to enable accurate estimations by explicitly encoding the generative process. Unfortunately, our preliminary experimental results demonstrated an unsatisfying performance of our methods. Still, these results yield a few interesting observations.

First, we observe that an $\mathrm{O}(2)$-*equivariant design* is not only useful but actually necessary for the models to converge to reasonable estimates. Moreover, by including a message passing component and by further enforcing $\mathrm{SO}(3)$-equivariance to more closely imitate the AMP algorithm, while not necessary, our model's performance improves. This illustrates the importance of including the right inductive biases to solve the problem in a principled way. Hence, future works could explore new strategies to combine this approach with the deep amortized inference from Levy et al. (2022a), which has been proven successful so far, possibly replacing steerable PCA features with raw images.

Second, the common-line loss might suffer from more local optima: since it compares pairs of poses which variate at each iteration, the training target quickly changes and can be unstable. Instead, the loss in Levy et al. (2022a) relies on a single reconstruction which is smoothly adapted, providing a more stable target for optimization. See also Apx. B for a simple study of the common-line loss landscape.

Finally, we suspect our model's performance is limited by the quality of the estimated noisy synchronization graph. Future works could study better ways to estimate or learn relative poses without relying on imprecise PCA features.

[9] Structure 5o60 from the Protein Data Bank database.

[10] Note its invariance to a global rotation of the estimated poses.

[11] If $\psi(i) \in \mathbb{R}^{d \times 2}$ are the top d eigenvectors at node i, the matrix $\psi(i)^T \psi(j)$ and its determinant approximate R_{ij} and $\boldsymbol{z}_i^T \boldsymbol{z}_j$.

References

Bandeira, A. S., Blum-Smith, B., Kileel, J., Perry, A., Weed, J., and Wein, A. S. Estimation under group actions: recovering orbits from invariants. *arXiv preprint arXiv:1712.10163*, 2017.

Bandeira, A. S., Chen, Y., Lederman, R. R., and Singer, A. Non-unique games over compact groups and orientation estimation in cryo-em. *Inverse Problems*, 36(6):064002, 2020.

Benjin, X. and Ling, L. Developments, applications, and prospects of cryo-electron microscopy. *Protein Science*, 29(4):872–882, 2020.

Borgerding, M. and Schniter, P. Onsager-corrected deep learning for sparse linear inverse problems. In *2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 227–231. IEEE, 2016.

Cesa, G., Behboodi, A., Cohen, T. S., and Welling, M. On the symmetries of the synchronization problem in cryo-em: Multi-frequency vector diffusion maps on the projective plane. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 5446–5458. Curran Associates, Inc., 2022a.

Cesa, G., Lang, L., and Weiler, M. A program to build E(N)-equivariant steerable CNNs. In *International Conference on Learning Representations*, 2022b.

Cohen, T. S. and Welling, M. Group Equivariant Convolutional Networks. *arXiv:1602.07576 [cs, stat]*, February 2016a. arXiv: 1602.07576.

Cohen, T. S. and Welling, M. Steerable CNNs. In *ICLR 2017*, November 2016b.

Cohen, T. S., Geiger, M., and Weiler, M. A general theory of equivariant CNNs on homogeneous spaces. *arXiv preprint arXiv:1811.02017*, 2018.

de Haan, P., Weiler, M., Cohen, T., and Welling, M. Gauge equivariant mesh cnns: Anisotropic convolutions on geometric graphs. In *International Conference on Learning Representations*, 2021.

Donnat, C., Levy, A., Poitevin, F., Zhong, E. D., and Miolane, N. Deep generative modeling for volume reconstruction in cryo-electron microscopy. *Journal of Structural Biology*, pp. 107920, 2022.

Fan, Y. and Zhao, Z. Cryo-electron microscopy image analysis using multi-frequency vector diffusion maps. *arXiv preprint arXiv:1904.07772*, 2019a.

Fan, Y. and Zhao, Z. Multi-frequency vector diffusion maps. In *International Conference on Machine Learning*, pp. 1843–1852. PMLR, 2019b.

Fan, Y., Gao, T., and Zhao, Z. Representation theoretic patterns in multi-frequency class averaging for three-dimensional cryo-electron microscopy. *Information and Inference: A Journal of the IMA*, May 2021.

Fernandez-Leiro, R. and Scheres, S. H. A pipeline approach to single-particle processing in RELION. *Acta Crystallographica Section D: Structural Biology*, 73(6):496–502, 2017. Publisher: International Union of Crystallography.

Goncharov, V. B. Determination of the spatial orientation of arbitrarily arranged identical particles of unknown structure from their projections. In *Dokl. Akad. Nauk SSSR*, volume 287, pp. 1131–1134, 1986.

Gregor, K. and LeCun, Y. Learning fast approximations of sparse coding. In *Proceedings of the 27th international conference on international conference on machine learning*, pp. 399–406, 2010.

Henderson, R., Baldwin, J. M., Ceska, T. A., Zemlin, F., Beckmann, E., and Downing, K. H. Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. *Journal of Molecular Biology*, 213 (4):899–929, June 1990. ISSN 0022-2836.

Hentschel, J., Burnside, C., Mignot, I., Leibundgut, M., Boehringer, D., and Ban, N. The complete structure of the mycobacterium smegmatis 70s ribosome. *Cell Reports*, 20(1):149–160, 2017. ISSN 2211-1247.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kondor, R. and Trivedi, S. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *International Conference on Machine Learning (ICML)*, 2018.

Levy, A., Poitevin, F., Martel, J., Nashed, Y., Peck, A., Miolane, N., Ratner, D., Dunne, M., and Wetzstein, G. Cryoai: Amortized inference of poses for ab initio reconstruction of 3d molecular volumes from real cryo-em images. *arXiv preprint arXiv:2203.08138*, 2022a.

Levy, A., Wetzstein, G., Martel, J. N., Poitevin, F., and Zhong, E. Amortized inference for heterogeneous reconstruction in cryo-em. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 13038–13049. Curran Associates, Inc., 2022b.

Nogales, E. The development of cryo-EM into a mainstream structural biology technique. *Nature methods*, 13(1):24–27, 2016. Publisher: Nature Publishing Group.

Perry, A., Wein, A. S., Bandeira, A. S., and Moitra, A. Message-passing algorithms for synchronization problems over compact groups. *Communications on Pure and Applied Mathematics*, 71(11):2275–2322, 2018.

Pragier, G. and Shkolnisky, Y. A common lines approach for ab initio modeling of cyclically symmetric molecules. *Inverse Problems*, 35(12):124005, 2019. Publisher: IOP Publishing.

Punjani, A., Rubinstein, J. L., Fleet, D. J., and Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nature Methods*, 14(3): 290–296, February 2017.

Rosenbaum, D., Garnelo, M., Zielinski, M., Beattie, C., Clancy, E., Huber, A., Kohli, P., Senior, A. W., Jumper, J., Doersch, C., et al. Inferring a continuous distribution of atom coordinates from cryo-em images using vaes. *arXiv preprint arXiv:2106.14108*, 2021.

Scheres, S. H. Relion: Implementation of a bayesian approach to cryo-em structure determination. *Journal of Structural Biology*, 180(3):519–530, 2012. ISSN 1047-8477.

Shkolnisky, Y. and Singer, A. Viewing direction estimation in cryo-EM using synchronization. *SIAM journal on imaging sciences*, 5(3):1088–1110, 2012. Publisher: SIAM.

Singer, A. Mathematics for cryo-electron microscopy. In *Proceedings of the International Congress of Mathematicians: Rio de Janeiro 2018*, pp. 3995–4014. World Scientific, 2018.

Singer, A. and Shkolnisky, Y. Three-dimensional structure determination from common lines in cryo-EM by eigenvectors and semidefinite programming. *SIAM Journal on Imaging Sciences*, 4(2):543–572, January 2011.

Singer, A. and Sigworth, F. J. Computational methods for single-particle electron cryomicroscopy. *Annual Review of Biomedical Data Science*, 13:163–190, 2020.

Singer, A., Coifman, R. R., Sigworth, F. J., Chester, D. W., and Shkolnisky, Y. Detecting consistent common lines in cryo-EM by voting. *Journal of structural biology*, 169 (3):312–322, 2010. Publisher: Elsevier.

Singer, A., Zhao, Z., Shkolnisky, Y., and Hadani, R. Viewing angle classification of cryo-electron microscopy images using eigenvectors. *SIAM J. Img. Sci.*, 4(2):723–759, June 2011.

Toader, B., Sigworth, F. J., and Lederman, R. R. Methods for cryo-em single particle reconstruction of macromolecules having continuous heterogeneity. *Journal of Molecular Biology*, 435(9):168020, 2023.

Ullrich, K., Berg, R. v. d., Brubaker, M., Fleet, D., and Welling, M. Differentiable probabilistic models of scientific imaging with the fourier slice theorem. In *proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2019.

Van Heel, M. Angular reconstitution: a posteriori assignment of projection directions for 3D reconstruction. *Ultramicroscopy*, 21(2):111–123, 1987. Publisher: Elsevier.

Wang, L., Singer, A., and Wen, Z. Orientation determination of cryo-EM images using least unsquared deviations. *SIAM journal on imaging sciences*, 6(4):2450–2483, 2013. Publisher: SIAM.

Weiler, M. and Cesa, G. General E(2)-Equivariant Steerable CNNs. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

Weiler, M., Forré, P., Verlinde, E., and Welling, M. Coordinate independent convolutional networks–isometry and gauge equivariant convolutions on riemannian manifolds. *arXiv preprint arXiv:2106.06020*, 2021.

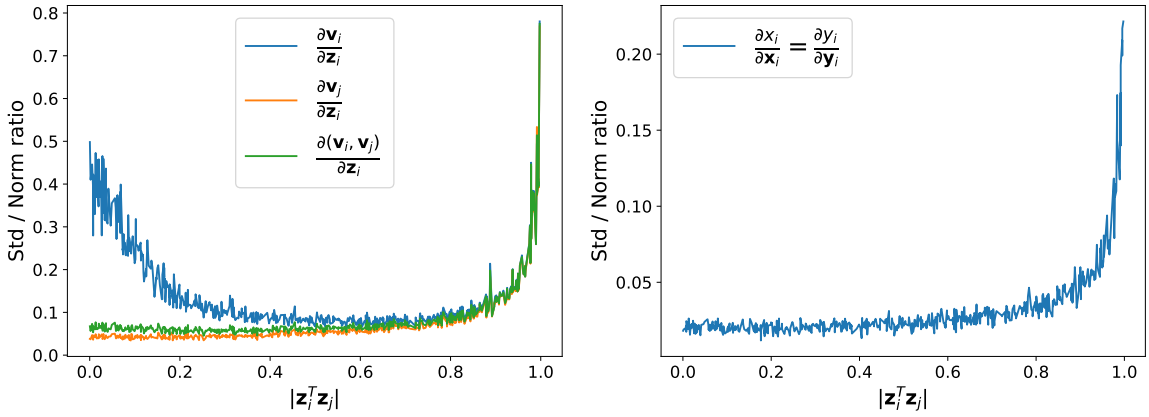
Zhao, Z. and Singer, A. Fourier–bessel rotational invariant eigenimages. *JOSA A*, 30(5):871–877, 2013.

Zhao, Z. and Singer, A. Rotationally invariant image representation for viewing direction classification in cryo-em. *Journal of structural biology*, 186(1):153–166, 2014.

Zhao, Z., Shkolnisky, Y., and Singer, A. Fast steerable principal component analysis. *IEEE transactions on computational imaging*, 2(1):1–12, 2016.

Zhong, E. D., Bepler, T., Davis, J. H., and Berger, B. Reconstructing continuous distributions of 3d protein structure from cryo-em images. *arXiv preprint arXiv:1909.05215*, 2019.

Zhong, E. D., Lerer, A., Davis, J. H., and Berger, B. Cryodrgn2: Ab initio neural reconstruction of 3d protein structures from real cryo-em images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4066–4075, October 2021.



(a) Gradient with respect to $\boldsymbol{z}_i$.

(b) Gradient with respect to $\boldsymbol{x}_i$ (or $\boldsymbol{y}_i$).

Figure 4: Ratio between the variance of the gradient and its norm as a function of the similarity between the viewing directions $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$.

A. Gradient of Common-Line loss at co-planar poses

In this section, we show how the presence of images with co-planar poses affect our common-lines based loss and, more specifically, its gradient.

Consider two images o_i and o_j and their predicted poses $R_i = (\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{z}_i) \in \mathbb{R}^{3\times 3}$ and $R_j = (\boldsymbol{x}_j, \boldsymbol{y}_j, \boldsymbol{z}_j) \in \mathbb{R}^{3\times 3}$. Note we assume $R_* \in \mathbb{R}^{3\times 3}$ rather than $R_* \in \mathrm{SO}(3)$ since our loss depends on R_i, R_j parameterized as matrices in $\mathbb{R}^{3\times 3}$ and the following arguments are independent of how our neural networks predict elements of $\mathrm{SO}(3)$ (although the particular choice of parameterization can affect how this gradient is back-propagated through the model).

Note that the loss $\mathcal{L}(R_i, R_j)$ from Eq. 17 or Eq. 18 only depends on the coordinates in Eq. 16, i.e.:

$$x_i = \boldsymbol{l}_{ij}^T \boldsymbol{x}_i \quad y_i = \boldsymbol{l}_{ij}^T \boldsymbol{y}_i \quad x_j = \boldsymbol{l}_{ij}^T \boldsymbol{x}_j \quad y_j = \boldsymbol{l}_{ij}^T \boldsymbol{y}_j$$

For convenience, define $\boldsymbol{v}_i = (x_i, y_i)^T \in \mathbb{R}^2$ and $\boldsymbol{v}_j = (x_j, y_j)^T \in \mathbb{R}^2$ and write $\mathcal{L}_{ij} = \mathcal{L}(R_i, R_j)$. Then, using the chain rule:

$$\frac{\partial \mathcal{L}_{ij}}{\partial R_i} = \frac{\partial \boldsymbol{v}_i}{\partial R_i}\frac{\partial \mathcal{L}_{ij}}{\partial \boldsymbol{v}_i} + \frac{\partial \boldsymbol{v}_j}{\partial R_i}\frac{\partial \mathcal{L}_{ij}}{\partial \boldsymbol{v}_j} \tag{21}$$

Hence, in this section, we can focus only on the partial derivatives $\frac{\partial \boldsymbol{v}_i}{\partial R_i}$ and $\frac{\partial \boldsymbol{v}_j}{\partial R_i}$. We will study the partial derivatives with respect to $\boldsymbol{x}_i, \boldsymbol{y}_i$ and $\boldsymbol{z}_i$ independently.

First, note that:

$$\frac{\partial x_i}{\partial \boldsymbol{x}_i} = \frac{\partial y_i}{\partial \boldsymbol{y}_i} = \boldsymbol{l}_{ij} \tag{22}$$

$$\frac{\partial y_i}{\partial \boldsymbol{x}_i} = \frac{\partial x_i}{\partial \boldsymbol{y}_i} = 0 \tag{23}$$

and

$$\frac{\partial x_j}{\partial \boldsymbol{x}_i} = \frac{\partial y_j}{\partial \boldsymbol{x}_i} = \frac{\partial x_j}{\partial \boldsymbol{y}_i} = \frac{\partial y_j}{\partial \boldsymbol{y}_i} = 0 \tag{24}$$

Hence, we only need to consider the following quantities: $\frac{\partial x_i}{\partial \boldsymbol{x}_i}$, $\frac{\partial \boldsymbol{v}_i}{\partial \boldsymbol{z}_i} = (\frac{\partial x_i}{\partial \boldsymbol{z}_i}, \frac{\partial y_i}{\partial \boldsymbol{z}_i})$ and $\frac{\partial \boldsymbol{v}_j}{\partial \boldsymbol{z}_i} = (\frac{\partial x_j}{\partial \boldsymbol{z}_i}, \frac{\partial y_j}{\partial \boldsymbol{z}_i})$.

To understand how the similarity of $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$, we study the variance of the gradients when R_j is perturbed by a small amount of noise, as a function of the similarity $|\boldsymbol{z}_i^T \boldsymbol{z}_j|$.

To do so, we sample 300 random pairs of rotations $R_i, R_j \in \mathrm{SO}(3)$ and compute their similarity $|\boldsymbol{z}_i^T \boldsymbol{z}_j|$. Then, for each pair, we generate 50 variations of R_j by perturbing it with a small Gaussian noise with standard deviation $\sigma = 0.04$ in the quaternion space, and compute each gradient $\frac{\partial x_i}{\partial \boldsymbol{x}_i}$, $\frac{\partial \boldsymbol{v}_i}{\partial \boldsymbol{z}_i}$ and $\frac{\partial \boldsymbol{v}_j}{\partial \boldsymbol{z}_i}$.

For each pair, we compute the average norm of the gradients (Frobenious norm for the Jacobians) and the standard deviation (over the 50 samples) of each partial derivative, which we average to obtain a single number. In Fig. 4, we plot the ratio between the standard deviation and the average norm for each pair, as a function of the similarity $|\boldsymbol{z}_i^T \boldsymbol{z}_j|$.

Whenever $\boldsymbol{z}_i$ is close to $\pm \boldsymbol{z}_j$, the variance of the gradient is very close to its average norm (the ratio approaches 1); this is particularly true for the gradients of $\boldsymbol{z}_i$, see Fig. 4a but less severe for $\boldsymbol{z}_i$ and $\boldsymbol{y}_i$. That result suggests that the training process can be particularly unstable in this setting, especially since the gradient on $\boldsymbol{z}_i$ is necessary to leave this situation but it is also the most affected by that.

B. Common-Line Loss Landscape

In this section, we provide a simple study of the common-line loss landscape. To do so, we compute the common line loss between two random images o_i and o_j, respectively at the poses $R_i R_{\theta_1}$ and $R_j R_{\theta_2}$.

$R_i = (\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{z}_i)$ and $R_j = (\boldsymbol{x}_j, \boldsymbol{y}_j, \boldsymbol{z}_j)$ are the ground-truth poses of o_i and o_j.

$R_\theta = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}$ is a 2D rotation by $\theta \in [0, 2\pi)$.

Hence, $R_i R_\theta$ is the pose of o_i rotated by θ around its projection axis $\boldsymbol{z}_i$ and simply corresponds to rotating the common line $\boldsymbol{v}_i = (\boldsymbol{x}_i^T \boldsymbol{l}_{ij}, \boldsymbol{y}_i^T \boldsymbol{l}_{ij})^T \in \mathbb{R}^2$ in the image o_i by θ. Note that if $\theta_1 = \theta_2 = \pi$, the loss is unchanged since the common line predicted is the same, only reflected.

In Fig. 5, we plot the common-line loss in Eq. 18 as a function of θ_1 and θ_2, for different random pairs (i, j). Note that any possible predicted pair of common lines corresponds to a point in the figure (multiple choices of $R_i, R_j \in \mathrm{SO}(3)$ lead to the same common lines). This enables us to study the complete loss landscape for the simple case of $N = 2$ images. The right column of Fig. 5 highlights the global minima of the loss.

First, we note the expected periodicity of the loss by $\theta_1 = \theta_2 = \pi$ in all images.

In the first pair (first row), we also observe spurious global minima at $\theta_1 = 0$ and $\theta_2 = \pi$ (and the opposite), which corresponds to a reflection of the correct common line in only one of the two images. This is likely related to the spurious planar symmetry described in Levy et al. (2022a) and in Sec. 4, which motivated the use of a "symmetrized loss" in Levy et al. (2022a).

We also note that the landscape can vary a lot over different pairs. While the first pair has a smooth landscape with two clear global minima at expected locations $(0, 0)$ and (π, π), other pairs show multiple global optima. In some cases, like the last row, the two locations $(0, 0)$ and (π, π) are close but not exactly global optima.

We also compare the original formulation of the loss in Eq. 17 with the modified version in Eq. 18 which we use in our experiments. Fig. 6 shows similar plots obtained using Eq. 17 (pairs are randomly sampled and don't necessary match those in Fig. 5). When using the original loss in Eq. 17, the global optima often do not include the ground-truth $(0, 0)$ and (π, π).

Finally, we emphasise that this study is limited to the case $N = 2$. However, during training, the loss is averaged over multiple pairs.

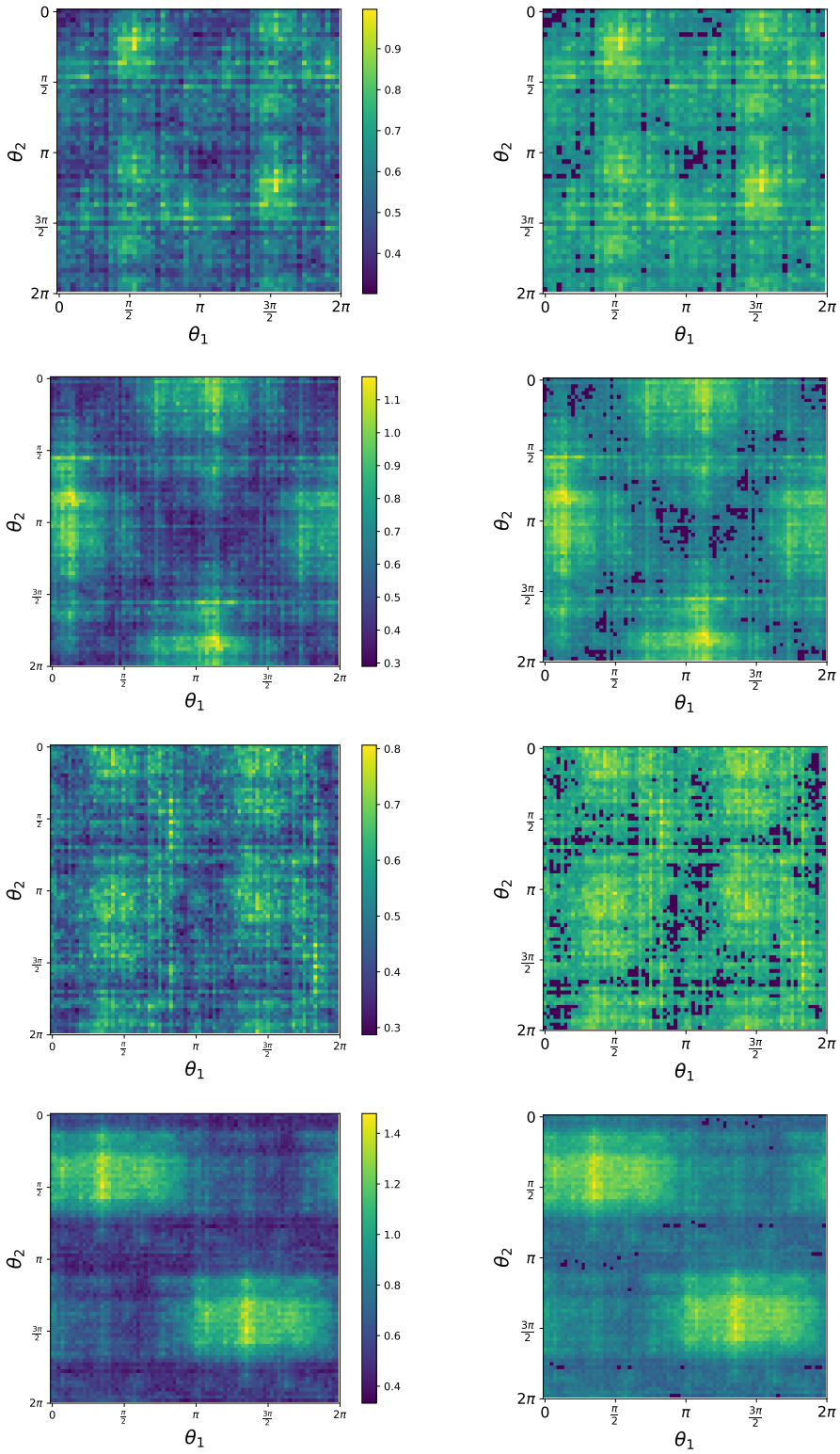


Figure 5: Common line loss using Eq. 18 between two random images o_i and o_j using the poses $R_i R_{\theta_1}$ and $R_j R_{\theta_2}$, with $\theta_1, \theta_2 \in [0, 2\pi)$. Each row is a different random pair (i, j). In the right column, areas where $\mathcal{L} < 0.4$ are highlighted with a darker color.

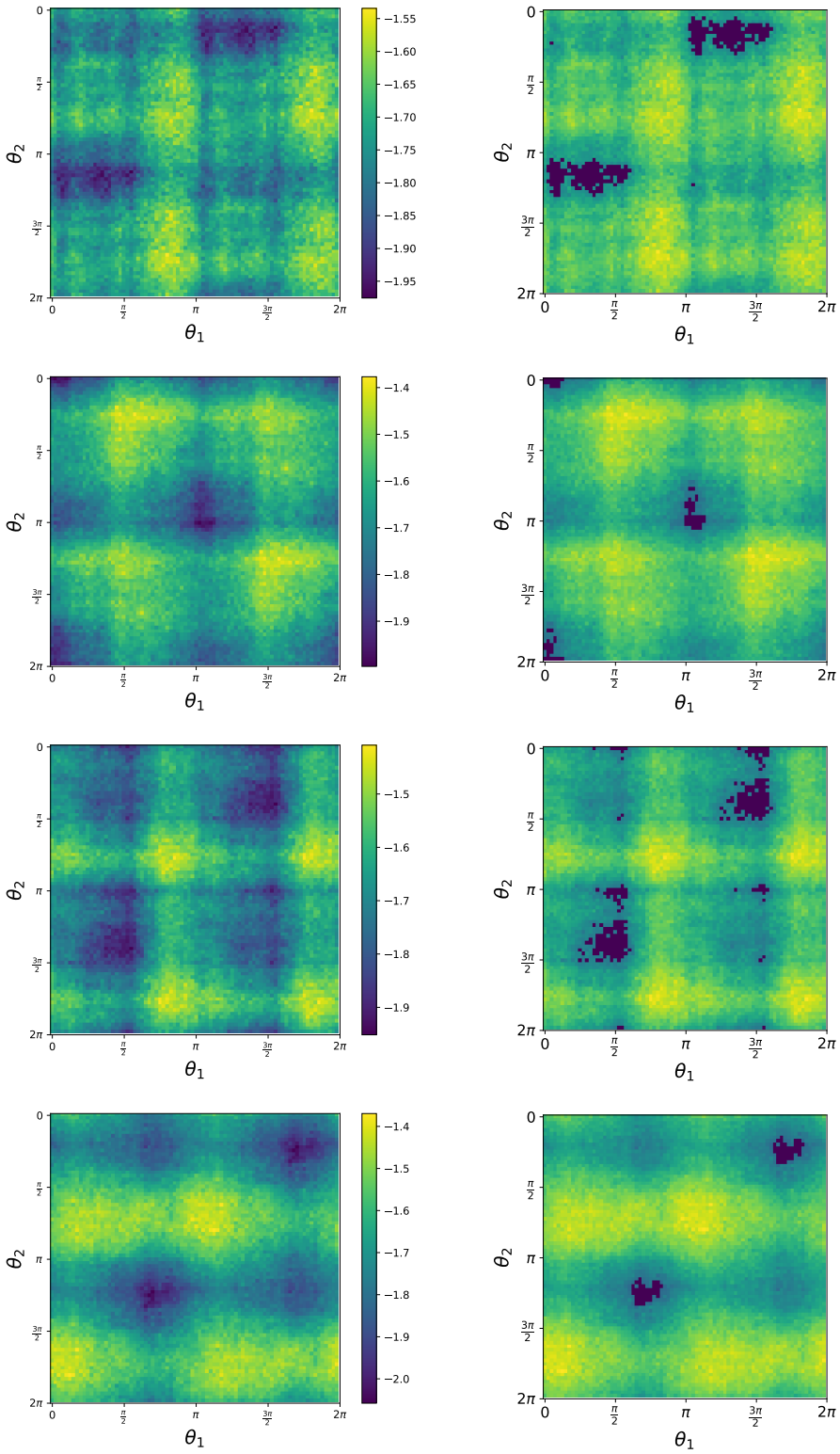


Figure 6: Common line loss using Eq. 17 between two random images o_i and o_j using the poses $R_i R_{\theta_1}$ and $R_j R_{\theta_2}$, with $\theta_1, \theta_2 \in [0, 2\pi)$. Each row is a different random pair (i, j). In the right column, the points closer to global optima are highlighted with a darker color.