DUIL: Deep Unsupervised Inverse Learning for *in situ* Macromolecular Morphology Identification

Mostofa Rafid Uddin[1]* **Seonghui Min**[1]*
Mahek Bhavesh Vora[1] **Qifeng Wu**[1] **Muyuan Chen**[2†] **Min Xu**[1†]
[1]Carnegie Mellon University [2]Stanford University
{mru,seonghum}@andrew.cmu.edu mxu1@cs.cmu.edu

Abstract

Emerging microscopic technologies such as cryo-electron tomography (cryo-ET) provide direct 3D visualization of macromolecules within the cell, enabling analysis of their *in situ* morphology. This morphology can be regarded as an $SE(3)$-invariant, denoised volumetric representation of subvolumes extracted from tomograms, termed as subtomograms. Morphology identification from a set of subtomograms is formulated as an inverse problem of estimating a set of template morphologies and per-subtomogram $SE(3)$ transformations with respect to any one of the templates. The existing expectation-maximization-based solution to this end often struggles with high structural heterogeneity and requires manual selection of a large number of hyperparameters. Addressing this issue, we present a novel deep unsupervised learning framework called DUIL. Given a set of subtomograms, DUIL first models their $SE(3)$-invariant morphological code using a siamese-like neural network with a multi-choice learning module. The learned morphological codes are clustered and used to generate a set of template morphologies through a generator network. The generated templates are used as references to estimate $SE(3)$ transformations for each subtomogram through pose optimization. The subtomograms with identical template morphologies are then aligned and averaged to iteratively refine the templates. Experiments on simulated and real cryo-ET datasets demonstrate clear improvements over prior methods; on real cellular data, DUIL identifies candidate morphologies not recovered by existing pipelines.

1 Introduction

Over the last decade, structural biology has shifted towards morphological characterization of large macromolecular complexes, particularly in a near-native *in situ* environment [1]. Cellular cryo-electron tomography (cryo-ET) has played a pivotal role in enabling the paradigm shift, serving as a practical tool for 3D visualization of macromolecule shape and morphology in their native states within the cell [2, 1]. In this imaging technique, a specimen of a whole cell or part of a cell is placed under an electron microscope, and images are captured across different tilt angles (typically from $-60°$ to $+60°$). To prevent radiation damage, the electron dosage is kept at a low level. The low electron dosage, along with the crowded cytoplasmic environment, makes the cryo-ET images extremely noisy. The tilt-series cryo-ET images are reconstructed into large 3D grayscale volumes, known as 3D tomograms. The tomograms are very large volumetric arrays (typically in the range of $4000 \times 4000 \times 1000$ voxels), typically containing hundreds to thousands of structurally heterogeneous macromolecular complexes in diverse orientations, along with other subcellular objects, including organelles and membranes.

*Equal contributions. †Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

The cryo-ET macromolecular structure processing workflow first extracts small subvolumes from a tomogram that potentially contains a macromolecule, known as subtomograms [3, 4]. The extracted subtomograms are further analyzed to identify the morphologies of macromolecular complexes. However, identification of macromolecular morphologies from these subtomograms is a complex process due to numerous challenges, including high noise, structural and orientational heterogeneity, and other imaging artifacts [1] present in the tomograms. Identifying macromolecular morphologies from subtomograms can be formulated as an inverse problem: determining template volumes under the assumption that each subtomogram is generated by applying an $SE(3)$ (or equivalently, $SO(3) \ltimes \mathbb{R}^3$) transformation with unknown parameters to an unknown template volume. The most traditional approach to solve this inverse problem involves performing expectation-maximization-based subtomogram classification and template generation [3]. In this approach, each subtomogram is probabilistically assigned to a class based on initial template volumes for each class, which are updated iteratively [4]. The template volume for each subtomogram denotes its morphology. To obtain a fine-grained morphology, the templates are refined to an optimal resolution in the follow-up subtomogram averaging (STA) or sub-tilt reconstruction step [3]. Nevertheless, the overall morphology is identified in the subtomogram classification and template generation step [5]. This approach, despite being the go-to method for decades, is often unable to resolve morphologies from subtomogram sets with high structural heterogeneity. Moreover, performance highly depends on manually setting appropriate values for a large number of hyperparameters [4].

In this work, we addressed this crucial open problem using a novel **D**eep **U**nsupervised **I**nverse **L**earning (DUIL) approach. Our approach is a two-stage framework (Figure 1). In the first stage, a siamese-like autoencoder network is trained on each subtomogram and an augmented view generated by a small in-plane rotation and translation that limits distortion of the missing wedge. In addition to a consistency loss traditionally used in Siamese networks, the network is trained to minimize a novel multi-choice learning-based 'winner-takes-all' loss function. In this scheme, instead of directly using the decoded output to minimize reconstruction loss with respect to the input, we present the network with multiple transformed variants of the decoded output, and the reconstruction loss is minimized against the variant most similar to the input. Such a training strategy encourages the encoder to output transformation-invariant morphological codes and the decoder to generate transformation-invariant morphology. After training, the morphological codes for each subtomogram are clustered using a Gaussian Mixture Model, and the decoded medoid of each cluster serves as an initial template. In the second stage, these templates are iteratively refined through transformation-vector optimization. Specifically, the relative transformation between each subtomogram and its corresponding template is estimated, subtomograms within the same cluster are aligned and averaged to update the template, and this process is repeated for a small number of iterations ($T = 3$ in our experiments). Our entire framework is highly automated and does not require users to adjust a large number of hyperparameters, unlike the existing methods.

We tested our approach against both idealistic and realistic simulated datasets, varying the signal-to-noise ratio (SNR) and imaging artifacts. Our method showed significantly superior performance compared to existing approaches, particularly on realistic data. Our experiments also revealed the importance of individual components of our approach, such as the multi-choice learning module and the siamese branch. Finally, we applied our method to experimental subtomograms of the thylakoid membrane region in publicly available *Chlamydomonas reinhardtii* cellular cryo-ET data, where it identified several candidate morphologies that were not recovered by existing approaches. We anticipate that our method can serve as a useful tool and an alternative or complement to existing approaches for determining the morphologies of numerous macromolecular complexes with previously unknown structures within their highly crowded and structurally heterogeneous native cellular context.

We summarize our contributions as follows:

- We developed a novel deep unsupervised learning method to address a crucial, largely unsolved inverse problem: *in situ* morphology identification of macromolecules inside the cell from structurally heterogeneous cellular cryo-ET images.

- Our method is built upon a novel multi-choice learning strategy that couples with Siamese training and enables learning $SE(3)$-transformation-invariant representation and template morphologies through an autoencoder network.

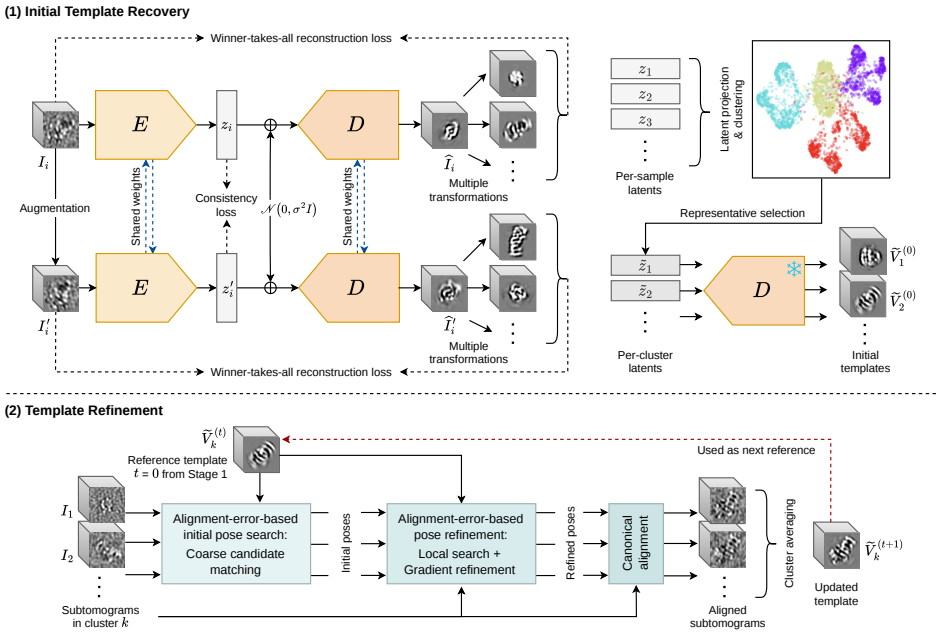

Figure 1: Schematic overview of our method. In (1) Initial Template Recovery, input subtomograms and their augmentations are encoded into consistent morphology latents, with a winner-takes-all reconstruction loss over multiple transformations and a consistency loss between augmented pairs. The learned per-sample latents are projected, clustered, and decoded through selected per-cluster latents to obtain initial templates. In (2) Template Refinement, each Stage 1 initial template is used as the reference at $t = 0$. Per-sample transformations are estimated by alignment-error-based coarse candidate matching and pose refinement, followed by canonical alignment and cluster averaging. Repeating this refinement yields the final templates.

- On real cellular cryo-ET subtomograms of thylakoid membranes, our method identified candidate morphologies that were not recovered by existing pipelines.

2 Related Works

Macromolecule identification in cellular Cryo-ET: Identifying macromolecules inside cellular cryo-electron tomograms has been an open and crucial challenge for several decades [1, 5]. The earliest and most traditional approach is template matching [6] with known structural templates to identify these macromolecules within the cell. However, this approach is prone to template-specific biases and cannot identify novel morphologies that lack a known template [7]. Subsequently, supervised segmentation-based approaches [8, 9] gained popularity to identify macromolecules in cellular cryo-ET data. These methods are particularly effective when many copies of a given macromolecule have already been manually annotated, since the availability of such annotations provides the training datasets required for reliable model supervision. However, the requirement of manual annotation itself poses a great challenge. It also depends on prior knowledge of structures and is limited in identifying macromolecules with unknown morphologies. Furthermore, because of the crowded and noisy nature and large size of the cellular tomograms, annotations are extremely burdensome and often impossible for macromolecules of small size.

Consequently, template-free unsupervised methods for identifying macromolecules have been adopted. We provide further details on this setup in Appendix Section A.2. Here, small subvolumes called 'subtomograms' containing peak signals are extracted from cellular tomograms, and macromolecules are identified from these subvolumes in an inverse problem manner. The most common solution to this problem is an expectation-maximization-based 3D classification and template generation approach in RELION [4]. Similar to RELION, our method also identifies macromolecules from subvolumes in a template-free, unsupervised manner. However, instead of the expectation-maximization approach, our method uses an autoencoder based framework trained with deep learning. In addition, our method

does not require users to manually define optimal values for a large number of hyperparameters similar to RELION. A learning-based unsupervised solution called DISCA [7] has been developed recently that performs subtomogram classification followed by RELION classification and refinement to identify macromolecular morphologies inside the cell. However, a large dependence on RELION remains. Unlike DISCA, our method does not depend on RELION to generate templates.

Multi-choice learning (MCL): Multi-choice learning (MCL) [10] refers to generating multiple choices or hypotheses for the model and making it learn from the most optimal choice. In this learning paradigm, a 'winner-takes-all' loss is used, where the loss is optimized through the most accurate model choice. This is different from standard mixture-of-expert setting where a weighted combination of the model choices are optimized. MCL has been used to reduce ambiguity in several machine learning prediction tasks, including image segmentation [11], human pose and shape estimation [12], motion forecasting [13], etc. Very recently, CryoSPIN [14] has used MCL to estimate pose from 2D single-particle cryo-EM images for the cryo-EM *ab initio* reconstruction task [15, 16]. The single-particle cryo-EM images are 2D projections of almost identical 3D volumes that may have subtle (conformational) heterogeneity. Unlike this work, we deal with 3D cellular cryo-ET subvolumes, which are much noisier and structurally heterogeneous than single-particle 2D cryo-EM. Moreover, CryoSPIN uses a multi-head encoder that generates 4 choices of $SO(3)$ rotations, and the model selects the most optimal one. Unlike this approach, we generate multiple choices of template volumes with varying $SE(3)$ transformations of the generator output, and let the model select the most optimal template volume. Further discussion on the difference between CryoSPIN MCL module and our proposed module is provided in Appendix Section A.5.

Single-particle cryo-EM reconstruction methods: Several deep learning based generative models, such as, cryoDRGN [17], e2GMM [18], cryoSPARC-3DVA [19], CryoAI [20], CryoSPIN [14], etc., have been developed in recent years that identify multiple conformations of macromolecules from 2D single-particle cryo-EM images. A few of them (CryoAI, CryoSPIN) do not require predefined poses and instead optimize them directly. CryoDRGN and CryoAI have been further extended for subtilt-image reconstruction in cryo-ET, named cryoDRGN-ET [16] and cryoDRGN-AI-ET [15]. However, these methods still apply to single particle cryo-ET tomograms primarily containing macromolecules of nearly homogeneous morphologies. They are not applicable for identifying highly heterogeneous morphologies from subtomogram mixtures. In Appendix Section A.1, we discuss this issue in detail. Unlike these 2D projection dependent methods, our method identifies highly heterogeneous 3D morphologies directly from 3D cellular subtomogram data.

3 Methods

3.1 Problem Formulation

Consider a set of cryo-ET subtomograms $\{I_i\}_{i=1}^N$, where $I_i \in \mathbb{R}^{d\times d\times d}$ and $d \in \mathbb{N}$ denotes the spatial dimension. Each subtomogram I_i is assumed to be generated from one of the unknown template volumes $\{V_k\}_{k=1}^K$, where $V_k \in \mathbb{R}^{d\times d\times d}$ and $K \in \mathbb{N}$ denotes the number of distinct underlying structures. Let $c_i \in \{1, \ldots, K\}$ represent the structure assignment for I_i. The observation I_i is modeled as the result of applying a rigid-body transformation T_{θ_i} to the template V_{c_i}, followed by the Cryo-ET imaging process g and additive noise. Formally,

$$I_i = g(T_{\theta_i}(V_{c_i})) + \eta_i, \qquad i = 1, \ldots, N,$$

where

- T_{θ_i} is a rigid-body transformation operator parameterized by $\theta_i = (\phi_i, \mathbf{t}_i) \in SE(3)$, with ϕ_i and $\mathbf{t}_i$ representing the rotation and translation parameters, respectively.
- g is the cryo-ET imaging operator (e.g., convolution with the point spread function).
- η_i is a noise term modeling imaging artifacts.

The problem can be formulated as an inverse problem: given only the observed subtomograms $\{I_i\}_{i=1}^N$, estimate the set of template volumes $\{V_k\}_{k=1}^K$ and the transformation parameters $\{\theta_i\}_{i=1}^N$.

3.2 Method Overview

In principle, the above formulation corresponds to a joint estimation problem over morphology and $SE(3)$ transformation. However, direct optimization is challenging due to the inherent ambiguity of both morphology and $SE(3)$ transformation under low SNR conditions. To address this, we adopt a decoupled strategy that separates representation learning from estimating transformation. Specifically, we first learn a transformation-invariant latent representation and recover initial templates from this learned representation. We subsequently perform transformation estimation and template refinement in an iterative manner.

Our method consists of two stages, as illustrated in Figure 1. **(1) Initial Template Recovery** learns latent representations using a Siamese autoencoder with a multi-choice reconstruction objective, and then clusters these representations to identify distinct structures and decode initial templates. **(2) Template Refinement** estimates per-sample transformations via coarse search and iterative refinement, followed by canonical alignment and averaging to obtain final templates.

3.3 Initial Template Recovery via Pose-Invariant Representation Learning

In Stage 1, we aim for unsupervised structural discovery by learning a pose-invariant latent representation that captures morphology while mitigating the impact of transformation variability. To this end, we adopt a multi-choice learning (MCL) objective that avoids explicit pose estimation during training, approximating marginalization over transformations via discrete candidate selection using a 'winner-takes-all loss' [10] scheme.

Let $E : \mathbb{R}^{d\times d\times d} \to \mathbb{R}^L$ denote the encoder and $D : \mathbb{R}^L \to \mathbb{R}^{d\times d\times d}$ the decoder, where $L \in \mathbb{N}$ is the latent dimensionality. Given an input subtomogram I, we construct an augmented view $I' = T_{\theta_{\text{aug}}}(I)$, where $\theta_{\text{aug}} \in SO(2) \times \mathbb{R}^3$ consists of a small in-plane rotation in the x–y plane (about the volume z axis) and a small translation. We avoid full 3D rotations because they would also rotate the missing wedge, producing wedge orientations inconsistent with the observed data, following a similar practice in TomoTwin [21]. Since a single-axis missing wedge is not rotationally symmetric, even an in-plane rotation does not preserve it exactly; we therefore restrict the rotation to a small range (Appendix A.4). The latent representations are obtained as $z = E(I)$ and $z' = E(I')$. To smooth the latent manifold, we add Gaussian perturbations $\epsilon, \epsilon' \sim \mathcal{N}(0, \sigma^2 I)$ with σ a small constant, and reconstruct the volumes as $\hat{I} = D(z + \epsilon)$ and $\hat{I}' = D(z' + \epsilon')$.

To account for pose variability, we define a finite set of candidate transformations $\Omega = \{\theta_m\}_{m=1}^M \subset SE(3)$, where each $\theta_m = (\phi_m, \mathbf{t}_m)$ is randomly sampled at each training iteration. The reconstruction loss is defined using a winner-takes-all objective over Ω:

$$\mathcal{L}_{\text{recon}} = \min_{\theta\in\Omega} \|I - T_\theta(\hat{I})\|_2^2 + \min_{\theta\in\Omega} \|I' - T_\theta(\hat{I}')\|_2^2,$$

where $\|\cdot\|_2^2$ denotes the voxel-wise squared error. To enforce consistency between augmented views, we introduce a latent alignment term $\mathcal{L}_{\text{embed}} = \|z - z'\|_1$. The overall objective is given by

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \lambda \mathcal{L}_{\text{embed}},$$

where λ controls the trade-off between reconstruction and latent consistency. This objective encourages the encoder to learn pose-invariant representations, as the reconstruction loss is minimized over transformed candidates, allowing the model to ignore pose variations while focusing on structural consistency.

Once the model is trained, we perform structure identification to obtain initial estimates of the underlying template volumes $\{V_k\}_{k=1}^K$ using the learned representations. For each subtomogram $I_i \in \{I_i\}_{i=1}^N$, we extract the latent representation $z_i = E(I_i)$. We then apply Uniform Manifold Approximation and Projection (UMAP) [22] to obtain low-dimensional embeddings $\zeta_i = \text{UMAP}(z_i) \in \mathbb{R}^2$. We fit a Gaussian Mixture Model (GMM) with K components to these embeddings and assign each sample to a cluster $\hat{c}_i = \arg\max_{c\in\{1,\dots,K\}} p(c \mid \zeta_i)$. Let $C_k = \{\, i \mid \hat{c}_i = k \,\}$ denote the index set of subtomograms assigned to cluster k. For each cluster, we select a representative latent vector using the medoid, $\tilde{z}_k = \arg\min_{j\in C_k} \sum_{i\in C_k} \|z_j - z_i\|_2$, and obtain an initial template estimate $\tilde{V}_k^{(0)} = D(\tilde{z}_k)$.

3.4 Template Refinement through Iterative Pose Optimization

While Stage 1 provides reliable initial template estimates $\{\tilde{V}_k^{(0)}\}_{k=1}^K$, the resulting volumes inherently lack high-frequency details, as the pose-invariant latent representation prioritizes global morphology over fine structural variations. To recover these details, Stage 2 refines the templates over T iterations by alternating between per-sample pose estimation and canonical averaging of the aligned subtomograms. For each cluster $k \in \{1, \ldots, K\}$, each subtomogram I_i with $i \in C_k$, and iteration $t \in \{0, \ldots, T-1\}$, we estimate the optimal pose $\theta_i^{(t)}$ relative to the current template $\tilde{V}_k^{(t)}$ by minimizing the alignment error:

$$\mathcal{E}_i^{(t)}(\theta) = \left\| I_i - T_\theta\left(\tilde{V}_k^{(t)}\right) \right\|_2^2.$$

To solve this, we employ a coarse-to-fine strategy. An initial pose $\theta_{i,\text{init}} = \arg\min_{\theta \in \Omega_{\text{coarse}}} \mathcal{E}_i^{(t)}(\theta)$ is obtained by discrete search within a finite candidate set $\Omega_{\text{coarse}} = \{\theta_m\}_{m=1}^{M_c} \subset SE(3)$. Starting from $\theta_i^{(t,0)} = \theta_{i,\text{init}}$, the pose is optimized over S inner steps. In each step $s \in \{0, \ldots, S-1\}$, we apply a gradient-based update to refine the pose estimate. To prevent the optimization from being trapped in local minima, this process is periodically augmented with a local discrete search $\delta_i^{(t,s)} = \arg\min_{\delta \in \Delta} \mathcal{E}_i^{(t)}(\theta_i^{(t,s)} + \delta)$ over a fine-grained set $\Delta = \{\delta_m\}_{m=1}^{M_\Delta} \subset \mathbb{R}^6$, where δ represents a small perturbation in the transformation parameter space. Specifically, the pose is updated as:

$$\theta_i^{(t,s+1)} = \mathcal{G}\left(\theta_i^{(t,s)} + \delta_i^{(t,s)}\right),$$

where $\mathcal{G}(\theta) = \theta - \alpha_s \nabla_\theta \mathcal{E}_i^{(t)}(\theta)$ denotes the gradient descent operator with step size α_s, and we set $\delta_i^{(t,s)} = 0$ for steps where the discrete search is not performed. Following these S refinements, the pose for the current iteration is set to $\theta_i^{(t)} = \theta_i^{(t,S)}$. Given the estimated poses $\{\theta_i^{(t)}\}_{i \in C_k}$, the template is updated by canonical averaging:

$$\tilde{V}_k^{(t+1)} = \frac{1}{|C_k|} \sum_{i \in C_k} T_{\theta_i^{(t)}}^{-1}(I_i),$$

where T_θ^{-1} denotes the inverse rigid-body transformation, computed in closed-form. After T iterations, the procedure yields the final refined template $\tilde{V}_k^{(T)}$. This stage complements the initial recovery in Stage 1 by explicitly estimating per-sample poses to restore high-frequency structural details.

4 Experiments

Simulated datasets: For benchmarking, we created several simulated subtomogram datasets containing mixtures of macromolecules. To this end, we selected four macromolecular complexes that co-exist in yeast cells and fit within the same subtomogram box size: the 80S ribosome (PDB ID: 4V7R), the 26S proteasome (PDB ID: 3JCP), fatty acid synthase (PDB ID: 2UV8), and TRiC (PDB ID: 7YLU). We filter the PDB structures to 15 Å with a pixel size of 7.5 Å. We created 1,000 subtomograms from each of the filtered PDB structures. To create the subtomograms, we first applied a random $SO(3)$ rotation and a translation sampled independently along each axis from $\mathcal{U}[-1.2, 1.2)$ voxels to the filtered PDB structures. Then we apply CTF and noise to match the desired SNR value of the subtomograms. We created two subtomogram sets: 1) an idealistic dataset with SNR 0.1 and no missing wedge effect, 2) a realistic dataset with SNR 0.01 and missing wedge effect with 30° missing wedge angle, which is commonly found in experimental subtomograms.

Experimental dataset: As experimental dataset, we used cellular cryo-ET tomograms of *Chlamydomonas reinhardtii* algae of EMPIAR-11830 (CC0 license). In particular, we studied the structurally heterogeneous and biologically significant region of the thylakoid membrane (Figure 3A). The thylakoid membrane region contains membrane proteins and several membrane-bound complexes, all of which are morphologically distinct and can be identified in the resolution range of a cryo-ET tomogram (Figure 3B). Given the relatively high signal in the membrane regions, automated picking [23] worked reasonably well. We used automated particle picking [23] to extract 55,118 subtomograms

Table 1: Quantitative comparison of our method and the baselines on the simulated macromolecule mixture datasets. (↑) indicates that higher is better. For each experimental setup, we performed three runs with different random seeds and report the mean values. For our full method on the realistic dataset, the standard deviations over the three runs are 0.055 for ARI, 2.3 for accuracy (%), and at most 0.004 for AUC-FSC across all four structures.

Dataset	**Method**	**ARI**	**Acc (%)**	**AUC-FSC** (↑)			
		(↑)	(↑)	FAS	Proteasome	Ribosome	TRiC
SNR 0.1 MWA 0	RELION	0.959	98.4	0.537	0.514	0.544	0.535
	cryoDRGN-AI-ET	-	-	0.093	0.084	0.109	0.104
(idealistic)	DISCA + RELION refine	0.960	97.0	0.537	**0.517**	0.544	0.535
	Our method (w/o MCL)	0.859	94.6	0.249	0.255	0.285	0.287
	Our method (w/o Siamese)	0.945	97.9	0.435	0.430	0.430	0.458
	Our method (w/o Stage 2)	**0.992**	**99.7**	0.422	0.385	0.417	0.436
	Our method	**0.992**	**99.7**	**0.554**	0.472	**0.546**	**0.541**
SNR 0.01 MWA 30	RELION	0.343	59.3	0.469	0.111	0.120	0.140
	cryoDRGN-AI-ET	-	-	0.083	0.038	0.099	0.094
(realistic)	DISCA + RELION refine	0.490	62.0	0.415	0.430	0.151	0.400
	Our method (w/o MCL)	0.668	86.9	0.209	0.165	0.138	0.186
	Our method (w/o Siamese)	0.786	91.6	0.191	0.162	0.158	0.269
	Our method (w/o Stage 2)	**0.867**	**94.9**	0.134	0.240	0.226	0.305
	Our method	**0.867**	**94.9**	**0.485**	**0.483**	**0.478**	**0.499**

in the thylakoid membrane regions of *Chlamydomonas reinhardtii*. Unlike simulated datasets, the experimental data set does not contain 'ground truth' morphology classes or templates. Consequently, the evaluation on this dataset could be qualitative only.

Training details: In stage 1 of our method, we used a convolutional neural network without pooling layers to implement the encoder and decoder of our model. For both stage 1 and stage 2, we apply gradient descent using the Adam optimizer with a constant learning rate (0.0001 for stage 1 and 0.04 for stage 2). We used NVIDIA A5000 GPUs for training our model against the datasets. For the simulated datasets, we trained using a single GPU. For the thylakoid membrane dataset, we trained our model in a distributed manner using four GPUs. We used the PyTorch Accelerate package for the distributed training. Before training, the 3D subtomograms are usually preprocessed (preprocessing details in the Appendix). To implement the MCL module, we use $M = 96$ in our experiments. Further details of our method implementation and training is provided in the Appendix.

Evaluation Metrics: Given that we have ground truth structure labels and templates in the simulated dataset, we can perform a quantitative evaluation on the simulated dataset. To evaluate clustering performance, we use the adjusted rand index (ARI) and accuracy. For calculating accuracy, we first aligned the predicted unsupervised labels using our approach with the ground-truth labels via Hungarian matching, and then calculated the accuracy between the aligned labels and the ground-truth labels. To assess the quality of the decoder output, we calculated the Area under the Fourier Shell Correlation (FSC) curve [24] with respect to the corresponding templates we used for data simulation. Further details on these metrics and their calculation are provided in Appendix Section A.7.

5 Results

5.1 Results in simulated cellular subtomogram data

We begin our experiments with the simulated datasets. We use RELION 3D classification, DISCA, cryoDRGN-AI-ET [15], the 1st stage of our method (without MCL, without Siamese branch, and with all components), and our complete framework. For all the relevant methods, we set the number of morphology classes, $K = 4$. Among these methods, only cryoDRGN-AI-ET operates on subtilt images rather than 3D subtomograms. We include it because, like our method, it does not require particle poses as input, whereas related subtilt-based methods such as tomoDRGN [25] and cryoDRGN-ET [16] rely on poses from upstream subtomogram averaging. We applied cryoDRGN-AI-ET to the subtilt images corresponding to our subtomograms, using 61 subtilt images with a 2° interval for each subtomogram, and applied the other methods directly to the subtomograms. We

quantitatively assessed the classification performance and the template generation performance (Table 1). We investigated the distinct morphology templates obtained by each method in the datasets. We show the templates obtained for the realistic simulated dataset with SNR 0.01 and MWA 30° in Figure 2.

Table 1 shows that our method consistently shows the best classification performance (ARI and accuracy) in the simulated datasets. However, the improvement of our method over the baselines in the idealistic dataset is less pronounced than in the realistic dataset. A similar phenomenon is observed for template quality as measured by the AUC-FSC scores. This suggests the necessity of our method, particularly for real subtomogram datasets, where the performance of other methods is not satisfactory. The results in the Table 1 also underscore the necessity of each individual component (MCL loss, siamese training, pose refinement in stage 2) of our approach. A controlled test, in which rotation, translation, and missing-wedge severity are varied separately, further shows that the full model preserves cluster assignments more consistently than the ablated variants (Appendix A.8.3).

Figure 2 shows the ground truth templates (used to calculate AUC-FSC scores) and the templates recovered by our method and the alternatives for the realistic dataset. It can be observed that RELION 3D classification could only somewhat recover the 'ground truth' FAS template on the realistic simulated dataset. In fact, RELION is highly effective in recovering FAS morphology in all the simulated datasets (Table 1), largely due to the strong symmetric signal present in FAS subtomograms. However, RELION failed to correctly identify other macromolecular complexes, including the ribosome, proteasome, and TRiC, particularly in realistic simulated data, even after performing downstream template refinement. This further underscores the necessity of the proposed MCL module for realistic subtomogram data analysis. CryoDRGN-AI-ET [15], the method for heterogeneous reconstruction from sub-tilt images, could not identify any of the morphologies, as expected, which is consistent with its primary focus on structural variability within a single macromolecular species rather than mixtures of distinct complexes. While the subtomogram classification method, DISCA [7] was able to recover several macromolecular morphologies after RELION refinement, its performance was limited, and considerable improvements were necessary. Finally, with our complete method, we achieved markedly improved results, successfully recovering the coarse morphology of all 'ground truth' templates in the realistic dataset. In addition, our method requires much less time and computing resources compared to the alternative methods (Appendix Section A.8).

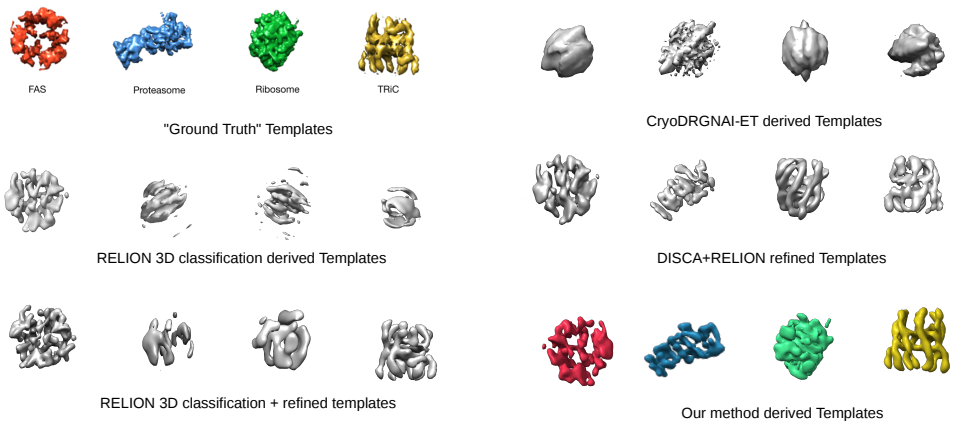

Figure 2: Results on realistic simulated data (SNR 0.01 and 30° Missing Wedge Angle).

5.2 Results in experimental cellular subtomograms data

We first used the RELION 3D classification and DISCA to identify the morphologies present in the experimental dataset. For the 3D classification, we used $K = 8$ expecting that the dataset should not have more than 8 morphologically distinct classes. The RELION 3D classification provided 8 3D structure models. We investigated the structure models with isosurface visualization (Figure 3C). However, we did not observe any structure densities other than membranes. A similar phenomenon was observed with DISCA classification followed by RELION refinement. Then we applied our method against the subtomogram dataset. After training our unsupervised method on the dataset,

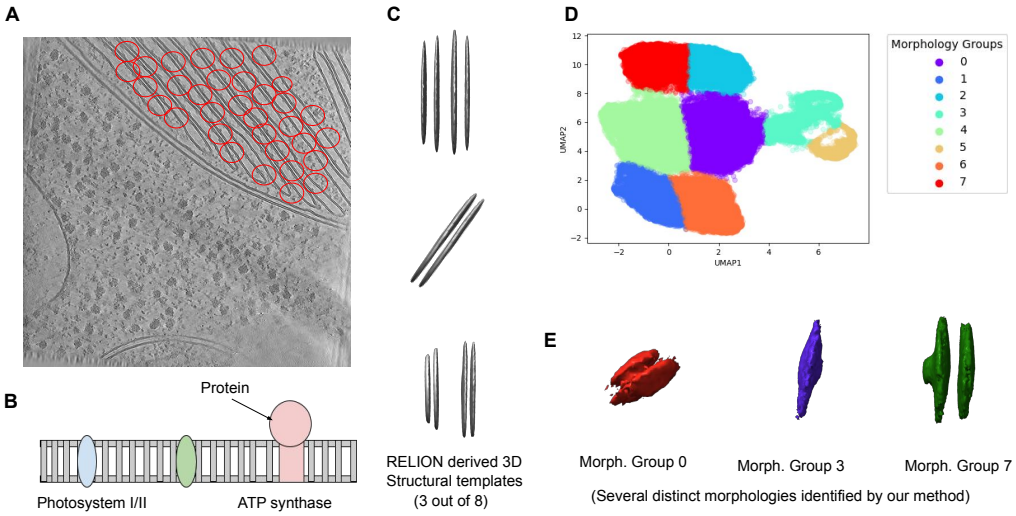

Figure 3: Our method recognizes the morphology of membrane proteins and several membrane-bound enzyme complexes in thylakoid membrane region of *Chlamydomonas reinhardtii*. **A.** A central slab (slice across depth axis) of *Chlamydomonas reinhardtii* tomogram, where subtomograms are extracted from thylakoid membrane region. **B.** Schematic representation of membrane-bound protein complexes involved in photosynthesis in the thylakoid membrane. **C.** A few of the morphologies (initial 3D models) generated by RELION and DISCA. **D.** The UMAP visualization of the morphology latent factor by our method along with the morphology groups obtained with GMM (K=8). **E.** A few of the morphologies (decoded outputs of the medoids of certain morphology classes) obtained by our method.

we inferred the semantic factors for each subtomogram and calculated the UMAP. We clustered the UMAP with Gaussian Mixture Model (GMM) with $K = 8$ (similar to the RELION experiment). We decoded a representative sample for each cluster. The decoded outputs show that morphology group 0 from our results represents the protein densities present in the thylakoid membrane region (Figure 3E). We also obtained two morphological groups: single-membrane or double-membrane with some portion of protein densities (decoder outputs are shown in Figure 3E). The remaining groups repeat these morphologies, suggesting that overclustering beyond $K = 8$ may not reveal additional heterogeneity.

Thus, our experiments with subtomograms extracted from the thylakoid membrane region of *Chlamydomonas reinhardtii* tomograms demonstrate the ability of our method to identify macromolecular-scale morphology groups that were not recovered by the compared approaches.

6 Discussion & Limitations

Identifying the *in situ* morphology of macromolecules from cellular cryo-ET images is extremely difficult given the small size of macromolecules compared to the large cryo-ET images and several other challenges discussed before. The cryo-ET community mostly used manual identification or maximum-likelihood based RELION 3D classification [4] by manually setting a large number of parameters to identify the macromolecular morphologies. Despite the manual efforts, these approaches would overlook many rare but important macromolecular morphologies. Our work pioneers as a highly automated, practical solution to identify macromolecular morphologies inside the cell, and in our experiments it recovered morphology groups that the compared approaches did not.

However, like any unsupervised classification approach [4, 26, 7], our method requires an estimate of the number of classes K to be provided by the user that it uses during the GMM-based latent space clustering step. While applying the method to experimental cryo-ET datasets, providing a high value for K is recommended to ensure that all heterogeneous morphologies are captured

(see Appendix A.8.2 for a sensitivity analysis). While this may introduce duplicate clusters, such redundancy is acceptable for a comprehensive morphology analysis. Moreover, the convergence of our method can be further improved with learned or probabilistic SO(3) priors, which we have left to future endeavor.

Our evaluation on experimental data is qualitative, as ground-truth classes and templates are unavailable, and the identified morphology groups are candidates that can be further validated, for example through subtomogram averaging or fitting to known structures. Broader evaluation is currently limited by the scarcity of public cellular cryo-ET datasets with high compositional heterogeneity, as many available datasets consist of single-particle or bacterial tomograms; we plan to benchmark our method on more heterogeneous datasets as they become available.

7 Conclusion

In this paper, we developed a novel unsupervised deep inverse learning method that enables morphology identification of macromolecular complexes from cellular cryo-ET subtomograms. Our method is specifically tailored to solve the inverse problem of macromolecular morphology identification with a subtomogram-specific method design and a novel multi-choice learning loss. Unlike the existing decade-long expectation-maximization-based solution, our method is highly automated and more effective. Our extensive experiments on simulated and experimental cellular cryo-ET subtomogram data validate this claim. We anticipate that our morphology analysis method, being coupled with the downstream subtomogram averaging or subtilt-reconstruction step can determine previously unknown structures with a higher resolution than before. Given the remarkable growth of cellular cryo-ET data collections recently [27], we foresee our method enabling the study of macromolecular morphology across cell populations, discovering novel biological insights on disease mechanisms and drug response.

Acknowledgments and Disclosure of Funding

This work was supported in part by U.S. NIH grant R35GM158094. M.C. were partially supported by NIH R01GM150905. The authors declare no competing interests.

References

[1] Martin Turk and Wolfgang Baumeister. The promise and the challenges of cryo-electron tomography. *FEBS letters*, 594(20):3243–3261, 2020.

[2] Allison Doerr. Cryo-electron tomography. *Nature Methods*, 14(1):34–34, 2017.

[3] Muyuan Chen, James M Bell, Xiaodong Shi, Stella Y Sun, Zhao Wang, and Steven J Ludtke. A complete data processing workflow for cryo-ET and subtomogram averaging. *Nature methods*, 16(11):1161–1168, 2019.

[4] Sjors HW Scheres. RELION: implementation of a bayesian approach to cryo-em structure determination. *Journal of structural biology*, 180(3):519–530, 2012.

[5] Mostofa Rafid Uddin, Sajib Acharjee Dip, Rajat Aayush Jha, Xiangrui Zeng, and Min Xu. Feature detection in cryo-electron tomography image analysis. In *Cryo-electron Tomography*, pages 173–215. Elsevier, 2025.

[6] 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.

[7] Xiangrui Zeng, Anson Kahng, Liang Xue, Julia Mahamid, Yi-Wei Chang, and Min Xu. High-throughput cryo-ET structural pattern mining by unsupervised deep iterative subtomogram clustering. *Proceedings of the National Academy of Sciences*, 120(15):e2213149120, 2023.

[8] Emmanuel Moebel, Antonio Martinez-Sanchez, Lorenz Lamm, Ricardo D Righetto, Wojciech Wietrzynski, Sahradha Albert, Damien Larivière, Eric Fourmentin, Stefan Pfeffer, Julio Ortiz, et al. Deep learning improves macromolecule identification in 3D cellular cryo-electron tomograms. *Nature methods*, 18(11):1386–1394, 2021.

[9] Guole Liu, Tongxin Niu, Mengxuan Qiu, Yun Zhu, Fei Sun, and Ge Yang. DeepETPicker: Fast and accurate 3D particle picking for cryo-electron tomography using weakly supervised deep learning. *Nature Communications*, 15(1):2090, 2024.

[10] Abner Guzman-Rivera, Dhruv Batra, and Pushmeet Kohli. Multiple choice learning: Learning to produce multiple structured outputs. *Advances in neural information processing systems*, 25, 2012.

[11] Simon Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R Ledsam, Klaus Maier-Hein, SM Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic U-Net for segmentation of ambiguous images. *Advances in neural information processing systems*, 31, 2018.

[12] Benjamin Biggs, David Novotny, Sebastien Ehrhardt, Hanbyul Joo, Ben Graham, and Andrea Vedaldi. 3D multi-bodies: Fitting sets of plausible 3D human models to ambiguous image data. *Advances in neural information processing systems*, 33:20496–20507, 2020.

[13] Ye Yuan and Kris M Kitani. Diverse trajectory forecasting with determinantal point processes. In *International Conference on Learning Representations*, 2020.

[14] Shayan Shekarforoush, David B Lindell, Marcus A Brubaker, and David J Fleet. CryoSPIN: improving ab-initio cryo-EM reconstruction with semi-amortized pose inference. *Advances in Neural Information Processing Systems*, 37:55785–55809, 2024.

[15] Axel Levy, Rishwanth Raghu, J Ryan Feathers, Michal Grzadkowski, Frédéric Poitevin, Jake D Johnston, Francesca Vallese, Oliver Biggs Clarke, Gordon Wetzstein, and Ellen D Zhong. CryoDRGN-AI: neural ab initio reconstruction of challenging cryo-EM and cryo-ET datasets. *Nature Methods*, 22(7):1486–1494, 2025.

[16] Ramya Rangan, Ryan Feathers, Sagar Khavnekar, Adam Lerer, Jake D Johnston, Ron Kelley, Martin Obr, Abhay Kotecha, and Ellen D Zhong. CryoDRGN-ET: deep reconstructing generative networks for visualizing dynamic biomolecules inside cells. *Nature Methods*, 21(8): 1537–1545, 2024.

[17] Ellen D Zhong, Tristan Bepler, Bonnie Berger, and Joseph H Davis. CryoDRGN: reconstruction of heterogeneous cryo-EM structures using neural networks. *Nature methods*, 18(2):176–185, 2021.

[18] Muyuan Chen and Steven J Ludtke. Deep learning-based mixed-dimensional Gaussian mixture model for characterizing variability in cryo-EM. *Nature methods*, 18(8):930–936, 2021.

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

[20] Axel Levy, Frédéric Poitevin, Julien Martel, Youssef Nashed, Ariana Peck, Nina Miolane, Daniel Ratner, Mike Dunne, and Gordon Wetzstein. CryoAI: Amortized inference of poses for ab initio reconstruction of 3D molecular volumes from real cryo-EM images. In *European Conference on Computer Vision*, pages 540–557. Springer, 2022.

[21] Gavin Rice, Thorsten Wagner, Markus Stabrin, Oleg Sitsel, Daniel Prumbaum, and Stefan Raunser. TomoTwin: generalized 3D localization of macromolecules in cryo-electron tomograms with structural data mining. *Nature methods*, 20(6):871–880, 2023.

[22] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29), 2018.

[23] Guang Tang, Liwei Peng, Philip R Baldwin, Deepinder S Mann, Wen Jiang, Ian Rees, and Steven J Ludtke. EMAN2: an extensible image processing suite for electron microscopy. *Journal of structural biology*, 157(1):38–46, 2007.

[24] Minkyu Jeon, Rishwanth Raghu, Miro Astore, Geoffrey Woollard, J Feathers, Alkin Kaz, Sonya Hanson, Pilar Cossio, and Ellen Zhong. CryoBench: Diverse and challenging datasets for the heterogeneity problem in cryo-EM. *Advances in Neural Information Processing Systems*, 37: 89468–89512, 2024.

[25] Barrett M Powell and Joseph H Davis. Learning structural heterogeneity from cryo-electron sub-tomograms with tomoDRGN. *Nature Methods*, 21(8):1525–1536, 2024.

[26] Jasenko Zivanov, Joaquín Otón, Zunlong Ke, Andriko von Kügelgen, Euan Pyle, Kun Qu, Dustin Morado, Daniel Castano-Díez, Giulia Zanetti, Tanmay AM Bharat, et al. A Bayesian approach to single-particle electron cryo-tomography in RELION-4.0. *elife*, 11:e83724, 2022.

[27] Mart GF Last, Lenard M Voortman, and Thomas H Sharp. Scaling data analyses in cellular cryoET using comprehensive segmentation. *bioRxiv*, 2025.

[28] Mostofa Rafid Uddin, Ajmain Yasar Ahmed, HM Shadman Tabib, Md Toki Tahmid, Md Zarif Ul Alam, Zachary Freyberg, and Min Xu. Localization of macromolecules in crowded cellular cryo-electron tomograms from extremely sparse labels. *Briefings in Bioinformatics*, 26(6): bbaf630, 2025.

A Appendix

A.1 Single-particle cryo-EM vs cellular cryo-ET

Single-particle cryo-EM collects 2D projection images of purified and isolated macromolecules that are randomly oriented and well separated in a thin layer of vitreous ice. Because many, nearly identical particles are imaged, the resulting micrographs contain relatively high contrast, more uniform backgrounds, and numerous projections of particles with nearly identical morphologies. With 'particle picking' or macromolecule localization, thousands of 2D projection images, each capturing a unique or slightly heterogeneous (different conformations) macromolecular structure with unknown camera angles, are extracted. The projection images are reconstructed into either a single consensus homogeneous structure or a series of structurally heterogeneous conformations. The latter is often referred to as 'heterogeneous reconstruction' in single-particle cryo-EM. The relatively higher SNR and the absence of surrounding cellular material enable near-atomic-resolution reconstructions. CryoDRGN [17], cryoSPARC [19], CryoAI [20], CryoSPIN [14], etc., all perform homogeneous or heterogeneous reconstruction of 3D structures from 2D single-particle cryo-EM.

On the other hand, cellular cryo-ET collects a tilt series of 2D images through thick, crowded cellular specimens, where each projection contains overlapping densities from membranes, cytoskeleton, organelles, and macromolecular complexes. The 2D tilt-series images are reconstructed into a large 3D grayscale volume, known as a tomogram. The 3D tomograms exhibit extremely low SNR, dramatic contrast attenuation at high tilt angles, and structural clutter that makes macromolecule identification and alignment substantially harder. Unlike single-particle EM micrographs, tomograms also suffer from the missing wedge, an angular region of uncollected data that leads to anisotropic resolution and elongation artifacts (discussed in detail in the following sections). Moreover, while single-particle EM images a homogeneous population of particles that may differ only in conformation, cryo-ET reveals a highly heterogeneous molecular landscape, with both high degrees of compositional and conformational heterogeneity.

The high degrees of compositional heterogeneity present in cryo-ET images make the identification of macromolecular morphology extremely difficult, and often impractical to do directly from 2D projection or tilt-series images. In Figure 4, we provide a fundamental example describing why 2D projection can be misleading to identify 3D object morphology. If a cylinder and a sphere are imaged from the top, both would appear as circles in their corresponding projection images. Thus it is impractical to distinguish between the 3D morphologies just based on the projection image. Hence, cryoDRGN-ET [16] series of models that reconstruct 3D structures from sub tilt-images are not suitable for identifying compositionally heterogeneous 3D morphologies *in situ*. Instead of projection images, it is practical to classify the 3D images to identify the 3D morphologies. Consequently, the morphologies are identified from 3D subvolumes (often called subtomograms) extracted from the 3D cryo-ET tomograms instead of the 2D tilt-series projection images.

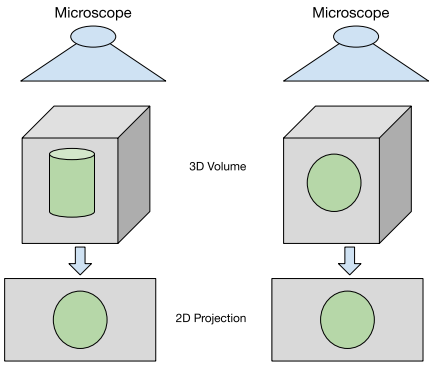

Figure 4: The image shows why projection-image-based reconstruction methods (cryoDRGN [17] and its variants) are not suitable for highly heterogeneous 3D structure identification.

A.2 Cryo-ET image analysis pipeline

In cryo-ET, a cellular sample or portion of a cellular sample is imaged with an electron microscope. The sample is tilted up to a certain range at both directions (typically -60° to $+60^\circ$) and an image is captured at each tilted position [1]. The tilt series images are then backprojected and reconstructed into a 3D voxel image, which is called a tomogram. These tomograms contain *in situ* visualizations of macromolecules and organelles inside a cell and their native spatial organization. However, this unique aspect of tomograms comes with several costs. To maintain the native context of the sample specimen, the electron dosage needs to be kept very low. Due to this low electron dose and also because of the complex cytoplasmic environment, the tomograms become very noisy. Tomograms are also usually very large (e.g., $4000 \times 6000 \times 1000$ voxels) and can not be processed as a whole. Even after binning 4 times across each axis, a tomogram is still large (e.g., $1000 \times 1500 \times 250$ voxels). Each tomogram contains hundreds to thousands of macromolecules, each occupying a minuscule portion of the tomogram. Consequently, the process for macromolecular morphology identification from tomograms occurs at the subtomogram level, where a subtomogram is a small subvolume of a tomogram that potentially contains a single macromolecule. Subtomograms are extracted from 3D tomograms using automated particle picking methods [28, 9, 23] or by manual picking. The extracted subtomograms are then classified and initial coarse 3D templates are generated. RELION 3D classification and our method perform this step. DISCA [7] only classifies the subtomograms, and depends on RELION refinement to obtain the coarse templates. The coarse templates are further refined with subtomogram averaging or subtilt reconstruction to obtain fine-grained and higher-resolution 3D template structures or morphologies. The whole pipeline and the positioning of our method relative to this pipeline are depicted in Figure 5. The figure also contains a schematic diagram visualizing the whole process.

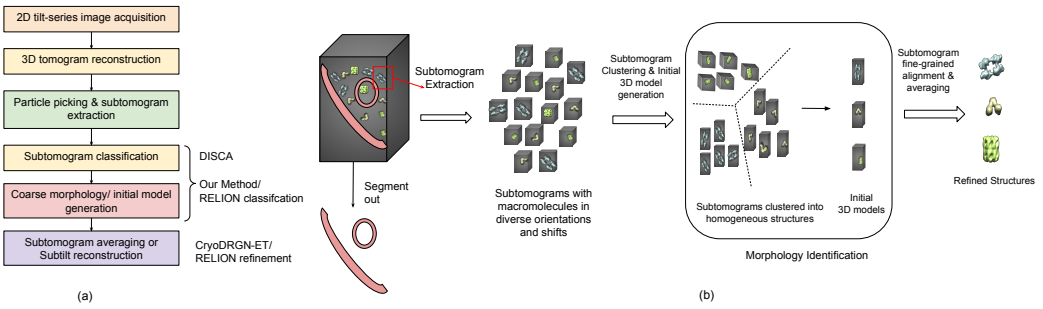

Figure 5: (a) The existing cryo-ET image processing pipeline and our method's positioning, (b) Schematic diagram of identifying refined macromolecular templates from 3D cellular cryo-ET tomogram

A.3 Missing wedge effect in Cryo-ET subtomograms

In cryo-electron tomography (cryo-ET), a cellular specimen is imaged by tilting it incrementally under the electron beam to acquire a series of 2D tilt-series images. However, due to physical and technical limitations of the microscope stage, the tilt range is restricted—typically to about ± 60–70°—instead of the full $\pm 90^\circ$. Once the 2D tilt-series images are reconstructed into a 3D tomogram, the incomplete angular coverage during the image acquisition process leaves a wedge-shaped region in the Fourier space of the tomogram unmeasured, known as the missing wedge effect.

Subtomograms extracted from the reconstructed tomogram also carry the missing wedge effect of the tomograms. Due to the missing wedge effect, subtomograms exhibit anisotropic resolution, with features elongated or distorted along the beam (Z) axis. This elongation affects both structural interpretation and subsequent computational analyses such as alignment, averaging, and classification.

A.4 Additional Experimental Details

Encoder-decoder network implementation

Our encoder comprises four 3D convolutional layers with exponentially linear unit (ELU) activations. The feature maps are progressively downsampled by strided convolutions (kernel size 4, stride 2 for the first three layers; stride 1 for the final layer), followed by two fully connected layers. The output layer produces the latent embedding vector; per-subtomogram poses are estimated in Stage 2 (Section 3). During training, a dropout layer ($p = 0.2$) is applied to improve generalization. The decoder reconstructs 3D volumes from the latent embedding using a fully connected layer followed by four transposed 3D convolution layers with ELU activations for the first three layers. This sequence progressively upsamples the latent representation back to the original ($48 \times 48 \times 48$) voxel resolution. Dropout ($p = 0.2$) is applied to the fully connected layer during training.

Preprocessing subtomograms

For preprocessing the subtomograms, we first low-pass-filter them to 15 Å with EMAN2. Balancing the trade-off between computational requirement and resolution, we use a box size of 48. To reshape the low-pass-filtered subtomograms to a box size of 48, we performed Fourier space cropping. We used these filtered and reshaped subtomograms, each of size $48 \times 48 \times 48$, to train our model. Before training, we also standardize the intensity of each subtomogram to a mean of 0 and a standard deviation of 1. Upon standardizing the intensities, we applied a soft-edged spherical mask to the subtomograms. The mask is centered within the $48 \times 48 \times 48$ volume with a radius of 24.

Siamese augmentation

The augmented view I' is generated from the same subtomogram I by an in-plane rotation in the x–y plane (about the volume z axis), with the angle sampled uniformly from the integers $\{-6^\circ, -5^\circ, \ldots, 6^\circ\}$, followed by a translation sampled independently along each axis from $\mathcal{U}[-1, 1)$ voxels.

$SE(3)$ candidate sampling for MCL:

For implementing the MCL loss, we sample from the $SO(3)$ grid. For ease of implementation, we sample 3D axis angles uniformly within a fixed range and then convert them to $SO(3)$. For initial epochs (first 80), we uniformly sample axis angles in the range $[-60^\circ, 60^\circ]$ to sample from most regions of the SO(3) grid. For the next several epochs (from 81 to 140), we uniformly sample axis angles in the range $[-45^\circ, 45^\circ]$. For the last epochs (after 140), we uniformly sample axis angles in the range $[-30^\circ, 30^\circ]$ to ensure they sample close to the identity matrix in the SO(3) grid. The translation component of each candidate is sampled independently along each axis from $\mathcal{U}[-0.5, 0.5)$ voxels. Stage 1 does not estimate poses: these candidates only approximate marginalization over small translations, while the Siamese augmentation encourages robustness to additional small shifts. Explicit pose estimation is performed in Stage 2, whose coarse-search domain spans $[-1.2, 1.2)$ voxels along each axis, matching the per-axis translation range used in the simulation (Appendix A.6).

A.5 Our MCL module vs CryoSPIN MCL module

In Figure 6, we demonstrate the difference between the MCL module in CryoSPIN [14] and the MCL module in our method. In CryoSPIN, the encoder generates four SO(3) candidates, all of which are used to transform the Fourier decoded volume. The projection images of the transformed Fourier volumes are compared with the input 2D image to compute the MCL loss. The SO(3) candidates are produced by four different encoder heads, and backpropagation is propagated through all the heads. In our work, the candidate $SE(3)$ transformations applied to the decoded volume are not generated by the encoder; rather, they are sampled independently of the encoder (Appendix A.4). We do not backpropagate through the sampling process; instead, we optimize the network with the MCL loss. Furthermore, in terms of method architecture, ours is significantly different from CryoSPIN.

We also experimented by integrating the MCL module of Shekarforoush et al. with our framework for our realistic simulated dataset. We observe that it further degrades the performance. This is partly due to the additional complexity introduced by using four additional heads to the 3D encoder of our

framework. In addition, the input to our encoder is extremely noisy 3D volumes with missing wedge artifacts; it is difficult to extract the right SO(3) candidates for transforming the decoder volume with heads attached to this encoder and backpropagating through it.

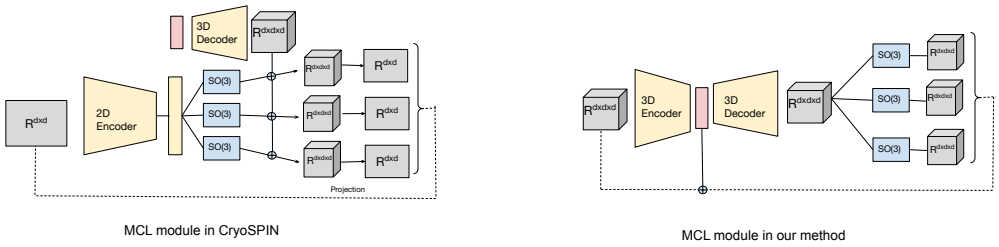

Figure 6: Difference between CryoSPIN MCL module and our MCL module.

A.6 Template refinement process details

Stage 2 refines the initial templates by iterative per-sample $SE(3)$ transformation optimization while keeping the Stage 1 network fixed. We use $T = 3$ outer refinement iterations. For the coarse search, Ω_{coarse} contains $M_c = 4096$ candidate transformations, with rotation parameters and translations sampled uniformly from $[-180^\circ, 180^\circ]^3$ and $[-1.2, 1.2)^3$ voxels, respectively. Coarse candidate evaluation uses a sample batch size of 256 and a candidate chunk size of 4096. Starting from $\theta_{i,\text{init}}$, we perform $S = 30$ gradient refinement steps using Adam with learning rate 0.04 and a batch size of 48. Every five refinement steps, we additionally perform a local discrete search over Δ, where $|\Delta| = 13$ and the candidates consist of the current pose, six $\pm 5^\circ$ rotation perturbations, and six ± 0.1 translation perturbations.

A.7 Evaluation metric details

ARI: To measure the clustering performance, we used Adjusted Rand Index (ARI). The ARI is commonly used to measure the similarity between two data clusterings, correcting for chance. Given a contingency table where:

- n_{ij} is the number of objects in both cluster i of the ground truth and cluster j of the predicted labels,
- $a_i = \sum_j n_{ij}$ is the sum over row i,
- $b_j = \sum_i n_{ij}$ is the sum over column j,
- n is the total number of data points.

The ARI is defined as:

$$\text{ARI} = \frac{\sum_{ij} \binom{n_{ij}}{2} - \frac{\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}}{\binom{n}{2}}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \frac{\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}}{\binom{n}{2}}}$$

To calculate ARI, we used the ADJUSTED_RAND_SCORE function from SKLEARN.METRICS.

AUC-FSC: To evaluate the quality of template morphologies obtained with RELION or our method, we used AUC-FSC (Area Under the Curve of Fourier Shell Correlation), which has been adopted by the community to measure structure recovery performance in heterogeneous cryo-EM/ET datasets [24]. It is a scalar metric used to summarize the overall agreement between two 3D volumes in frequency space.

The **Fourier Shell Correlation (FSC)** measures the correlation between two 3D volumes in Fourier space as a function of spatial frequency s. It is defined as:

$$\text{FSC}(s) = \frac{\sum_{i \in s} F_1(i) \cdot F_2^*(i)}{\sqrt{\sum_{i \in s} |F_1(i)|^2 \cdot \sum_{i \in s} |F_2(i)|^2}}$$

where:

- $F_1(i)$ and $F_2(i)$ are the complex Fourier coefficients of the two volumes,
- $F_2^*(i)$ is the complex conjugate of $F_2(i)$,
- $i \in s$ denotes the voxels in the shell corresponding to spatial frequency s.

The **AUC-FSC** summarizes the FSC curve over the full frequency range $[0, 1]$ and is defined as:

$$\text{AUC-FSC} = \int_0^1 \text{FSC}(s)\, ds$$

A.8 Results

A.8.1 Time and memory requirements

In Table 2, we provide the average time per epoch or iteration and the memory requirements to execute our method and the related methods on our benchmark simulated datasets. (↓) indicates that lower is better.

Table 2: Time and memory requirement of our method and the related methods

Method	**Time (GPU hours)** (↓)	**GPU Memory (GB)** (↓)
RELION	0.15	20
cryoDRGN-AI-ET	8.25	41
DISCA	0.60	18
Our method	**0.05**	**16**

A.8.2 Sensitivity to the number of clusters K

We evaluated our method on the realistic simulated dataset (SNR 0.01, missing wedge angle $30°$) with different values of K in the GMM clustering step. In addition to ARI, we report homogeneity (whether each predicted cluster contains a single true morphology), completeness (whether each true morphology is assigned to a single predicted cluster), and purity (the fraction of samples matching the majority true morphology of their assigned cluster). As shown in Table 3, ARI is highest at the true value $K = 4$. When K exceeds the true number of morphologies, completeness decreases as true morphologies are split across clusters, while homogeneity and purity remain high. In contrast, when K is smaller than the true value, distinct morphologies are merged, and homogeneity and purity drop substantially. This supports choosing a value of K larger than the expected number of morphologies for experimental data.

A.8.3 Controlled test of transformation and missing-wedge robustness

To directly examine whether the learned representation is driven by pose or missing-wedge artifacts, we constructed a controlled set of 672 volumes from the four morphologies, combining 8 sampled 3D rotations, 3 missing wedge angles ($0°$, $15°$, $30°$), and 7 translations (0 and ± 0.5 voxels along each of the x, y, and z axes). We kept the Stage 1 encoder and the UMAP/GMM models fitted on the realistic simulated dataset fixed, and only applied UMAP transformation and GMM prediction to the controlled volumes without refitting. For each factor, we varied that factor while holding the other two fixed, and defined agreement as the fraction of condition pairs within a matched family that receive the same GMM cluster label, averaged over all matched families. This metric does not require matching predicted clusters to ground-truth labels. As shown in Table 4, our method largely preserves cluster

Table 3: Sensitivity to the number of clusters K on the realistic simulated dataset. The true number of morphologies is 4. (↑) indicates that higher is better.

K	ARI (↑)	Homogeneity (↑)	Completeness (↑)	Purity (↑)
2	0.375	0.358	0.724	0.495
3	0.528	0.544	0.725	0.704
4	**0.867**	**0.838**	**0.841**	**0.949**
5	0.775	0.821	0.732	0.939
6	0.682	0.818	0.660	0.938
8	0.467	0.800	0.546	0.906

assignments across all three factors, with higher agreement than both ablated variants. Agreement is lower for rotations than for translations and missing-wedge severity, as rotations change the full orientation of the structure.

Table 4: Cluster-assignment agreement (↑) under controlled changes in rotation, translation, and missing-wedge severity. (↑) indicates that higher is better.

Controlled change	Our method	w/o MCL	w/o Siamese
8 sampled rotations	**0.865**	0.738	0.794
7 small translations	**0.968**	0.898	0.891
3 missing wedge angles	**0.967**	0.908	0.917

A.8.4 Additional template morphologies

The template morphologies generated by RELION, the ablation experiments (our method without various components) and our complete method on realistic subtomogram dataset with SNR 0.01 and missing wedge angle 30° are provided in Figure 2. In this Appendix, we further provide the template morphologies obtained by these methods for the other idealistic simulated datasets. The obtained template morphologies for SNR 0.1 and missing wedge angle 0° are provided in Figure 7.

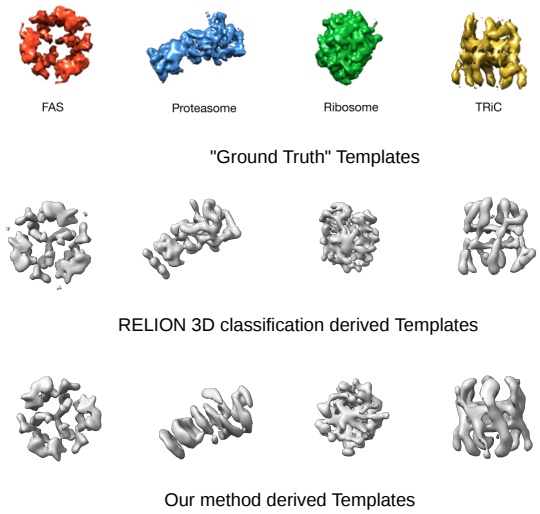

Figure 7: The morphology templates obtained by RELION and our method on idealistic SNR 0.1 missing wedge angle 0° simulated dataset

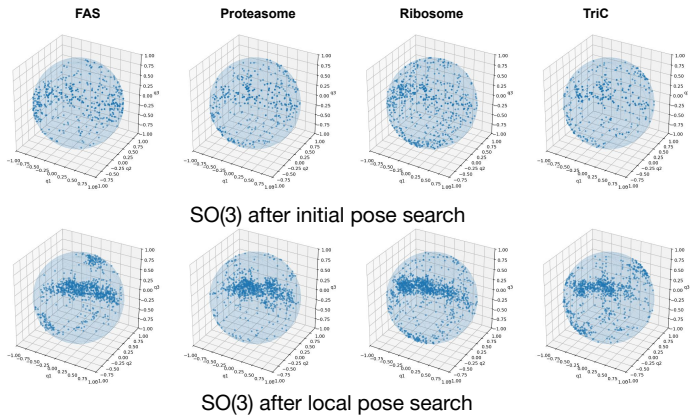

Figure 8: SO(3) distribution of subtomograms for each morphology class for our realistic dataset, both after initial pose search and after local search in stage 2 of our pipeline

A.8.5 Pose inspection in stage 2

We inspected the SO(3) poses obtained in stage 2 of our method, both after initial pose search (coarse candidate matching) and after local search (pose refinement). In Figure 8, we provide the obtained SO(3) distributions for 1000 subtomograms of each morphology class in our realistic simulated dataset with SNR 0.01 and missing wedge angle 30°. It shows that during initial pose search, the distribution is highly sparse as expected. However, after refinement, the pose distribution for each subtomogram becomes more concentrated. Furthermore, we do not observe any morphology class-specific pose distribution, which shows that our pose refinement strategy is not biased or affected by specific morphology of macromolecules.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims are supported by the results in Section 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in Section 6.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The architecture is fully discussed in Section 3. The experimental details are discussed in Section 4 and Appendix Section A.4. The code, a sample dataset and a trained model can be found in the supplementary material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code, a sample dataset, and a trained model, together with instructions to run them, are provided in the supplementary material.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Specified in Section 4 and Appendix Sections A.4 and A.6.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Results in Table 1 are means over three runs with different random seeds; the standard deviations for the full method on the realistic dataset are reported in the table caption.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Provided in Section 4 and Table 2 in Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research does not violate any of the ethics code. It does not involve any human subjects or participants, does not face data privacy issue and does not have any potential harmful social impact.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The potential positive impacts are discussed in Section 7. We believe it does not have any potential negative societal impact.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper uses existing dataset available under CC0 license. It is specified in Section 4.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The released code and sample data are documented with usage instructions in the supplementary material.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.