CryoFormer: Continuous Heterogeneous Cryo-EM Reconstruction using Transformer-based Neural Representations

Xinhang Liu[1,2,4⋆†], Yan Zeng[1,2⋆], Yifan Qin[1,2], Hao Li[1,2,3‡], Jiakai Zhang[1,2], Lan Xu[1], and Jingyi Yu[1]

[1] ShanghaiTech University
[2] Cellverse
[3] iHuman Institute
[4] HKUST

Abstract. Cryo-electron microscopy (cryo-EM) has revolutionized structural biology by resolving 3D structures of biomolecules at near-atomic resolution. However, revealing the continuous conformational heterogeneity from hundreds of thousands of noisy particle images remains challenging. Recent advances in heterogeneous reconstruction, often conducted in the Fourier domain, suffer from a lack of interpretability and are limited in achieving higher resolution in locally flexible regions. To address this issue, we propose CryoFormer, a novel approach for high-resolution and continuous heterogeneous cryo-EM reconstruction. CryoFormer leverages a feature volume in the real domain to capture fine-grained local changes. We then design a novel query-based transformer architecture that incorporates deformation-aware features and region-wise spatial features using a cross-attention mechanism. Our transformer-based pipeline further supports pose refinement and can automatically highlight flexible regions by visualizing 3D attention maps. Extensive experiments show that our method achieves the best performance on five datasets (two synthetic and three experimental). We also contribute a new synthetic dataset of the PEDV spike protein for more comprehensive evaluations. Both the code and the PEDV dataset will be released for better reproducibility.

Keywords: Cryo-electron Microscopy · Neural Representation · Dynamic Reconstruction

1 Introduction

Dynamic objects as giant as planets and as minute as proteins constitute our physical world and produce nearly infinite possibilities of life forms. Accurate

⋆ Equal contribution.
† Work done while studying at ShanghaiTech University.
‡ Work done while working at Cellverse.

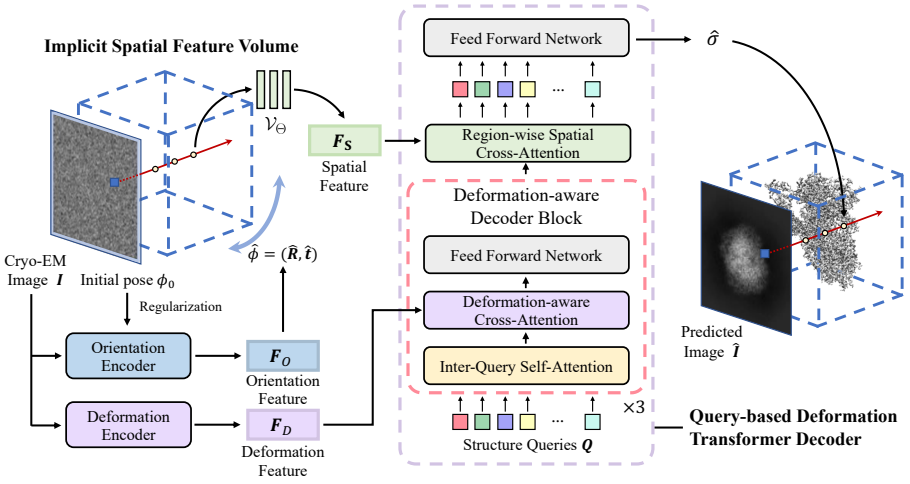

Fig. 1: Pipeline of CryoFormer. 1) Given an input image, our orientation encoder and deformation encoder first extract orientation representations and deformation features. We use pre-computed pose estimations to regularize the orientation encoder. **2)** We convert the orientation representation into a pose estimation and transformed coordinates are fed into our implicit neural spatial feature volume to produce a spatial feature. **3)** The spatial feature and the deformation image feature then interact in the deformation transformer decoder to output the density prediction.

recovery of their 3D shape, appearance, and movement helps to reflect the fundamental laws of nature. Conventional computer vision techniques combine specialized imaging apparatus such as domes or camera arrays with tailored reconstruction algorithms (SfM [56], NeRF [40], and most recently 3DGS [24]) to capture and model the fine-grained 3D dynamic entities at an object level.

Similar approaches have been adopted to recover shape and motion at a micro-scale level. In particular, to computationally determine protein structures, cryo-electron microscopy (cryo-EM) flash-freezes a purified solution with hundreds of thousands of particles of the target protein in a thin layer of vitreous ice. In a cryo-EM experiment, an electron gun generates a high-energy electron beam that interacts with the sample, and a detector captures scattered electrons during a brief duration, resulting in a 2D projection image that contains many particles. Given projection images, the single particle analysis (SPA) technique iteratively optimizes for recovering a high-resolution 3D protein structure [29, 43, 51]. Applications are numerous, ranging from revealing virus fundamental processes [71] in biodynamics to unveiling drug-protein interactions [21] in drug development.

Compared with macro-scale reconstruction, cryo-EM reconstruction presents unique challenges. First, cryo-EM images exhibit a low signal-to-noise ratio (SNR) with unknown particle orientations, leading to severe corruption of the structural signals. In addition, the flexible regions of proteins induce confor-

mational heterogeneity, further disrupting orientation estimation. Conventional software packages [49, 55] only model conformations with a small discrete set to reduce the complexity. Such approaches often yield low-resolution reconstructions of flexible regions and necessitate guidance from human experts. Recently, neural approaches exploit coordinate-based representations for heterogeneous cryo-EM reconstruction [13,27,32,75,76]. To mitigate the computational expense through the usage of the Fourier slice theorem [6], they perform reconstruction in the Fourier domain. A downside, however, is that modeling and interpreting local density variations between conformations in the Fourier domain is arduous and counter-intuitive, resulting in a reconstruction resolution falling short of practical application requirements.

Different from previous **Fourier** domain approaches [31,76], we propose *CryoFormer* (Figure 1), conducting reconstruction in the **real** domain to facilitate the modeling and interpretation of local flexible regions. Taking 2D particle images as inputs, our orientation encoder and deformation encoders extract image features related to orientation and deformation. We construct an implicit feature volume in the real domain as the core of our approach and introduce a novel query-based transformer decoder to generate a continuous, heterogeneous density volume. Specifically, our deformation-aware cross-attention mechanism embeds image deformation features into a series of structural queries. These queries then interact with spatial features through region-wise spatial cross-attention.

Our proposed transformer architecture excels at capturing fine-grained structures and refining coarse pose estimations. Through the analysis of 3D attention maps, our method further enables a novel function of highlighting spatial local changes, significantly improving interpretability.

For better benchmarking heterogeneous cryo-EM reconstructions, we present a novel synthetic dataset of the porcine epidemic diarrhea virus (PEDV) trimeric spike protein, a primary target for vaccine development and antigen analysis. We validate CryoFormer on the PEDV spike protein synthetic dataset and four existing public datasets. Our approach outperforms the state-of-the-art methods including popular traditional software [49] as well as recent neural approaches [27, 48, 75] on both synthetic and experimental datasets. Moreover, our experiments demonstrate that our method can identify dynamic regions within structures, thereby enabling more effective analysis of functional areas. We will release our code and PEDV spike protein dataset.

2 Related Work

Dynamic Neural 3D Representations. Neural Radiance Fields (NeRFs) [40] and their subsequent variants [25, 41] have achieved impressive results in novel view synthesis. Follow-up works have since emerged to enhance NeRFs and expand their applications [62], such as improving rendering quality [2–4, 35, 69], acceleration [17, 65, 72], and 3D scene understanding [26, 28, 36].

Numerous studies have introduced extensions of NeRF for dynamic scenes [15, 33, 34, 37, 45, 60, 70, 73, 74]. Most of these dynamic neural representations either

construct a static canonical field and use a deformation field to warp this to the arbitrary timesteps [44, 47, 63, 74], or represent the scene using a 4D space-time grid representation, often with planar decomposition or hash functions for efficiency [1, 9, 16, 58].

Conventional Cryo-EM Reconstruction. Traditional cryo-EM reconstruction involves the creation of a low-resolution initial model [30, 49] followed by the iterative refinement [19, 49, 55]. These algorithms perform reconstruction in the Fourier domain since this can reduce computational cost via Fourier slice theorem [6]. When tackling structural heterogeneity, they classify conformational states into several discrete states [39, 54]. While this paradigm is sufficient when the structure has only a small number of discrete conformations, it is nearly impossible to individually reconstruct every state of a protein with continuous conformational changes in a flexible region [46].

Neural Representations for Cryo-EM Reconstruction. Recent work has widely adopted neural representations for cryo-EM reconstruction [27, 31, 32, 59, 75]. CryoDRGN [75] first proposed a VAE architecture to encode conformational states from images and decode them by a coordinated-based MLP that represents the 3D Fourier volume. Such a design can model the continuous heterogeneity of protein and achieve higher spatial resolution compared with traditional methods. To reduce the computational cost of large MLPs, SFBP [27] uses a voxel grid representation. To enable an end-to-end reconstruction, there are some *ab*-initio neural methods [10, 31, 32, 76] directly reconstruct protein from images without requiring pre-computed poses from traditional methods. CryoFIRE [32] attempts to use an encoder to estimate poses from the input images by minimizing reconstruction loss directly, but the performance is still limited due to the ambiguity of conformation and orientation in the extremely noisy image. To model the 3D local motion, 3DFlex [48] and DynaMight [57] perform reconstruction in the real domain by using a flow or deformation field to model the structural motion, but they both require a canonical structure as input.

Transformers in 3D. Transformers have become a ubiquitous learning architecture capable of capturing long-range dependencies in sequential data, and have demonstrated remarkable success across a range of applications, including natural language processing [7, 12, 64], computer vision [14, 38], and protein structure determination [23]. Transformers have also been proven to benefit 3D reconstruction. IBRNet [67] employs a transformer to predict density from features to achieve generalizability. NeRFormer [50] utilizes attention modules to aggregate source views to construct feature volumes. GNT [66] uses transformers to render pixel color. However, these related works only apply transformers and attention mechanisms to the reconstruction of macroscopic static scenes, while we have designed deformation-aware cross-attention and region-wise spatial cross-attention to model the dynamic microstructures of biological entities.

3 Method

We propose CryoFormer, a novel approach that leverages a real domain implicit spatial feature volume coupled with a query-based transformer architecture for continuous heterogeneous cryo-EM reconstruction. In this section, we begin by laying out the cryo-EM image formation model in Section 3.1. We then introduce the procedural framework of CryoFormer (Figure 1), encompassing orientation and deformation encoders (Section 3.2), an implicit spatial feature volume $\mathcal{V}_\Theta$ (Section 3.3) and a query-based transformer decoder (Section 3.4), with the training scheme described in Section 3.5.

3.1 Cryo-EM Image Formation Model

In the cryo-EM image formation model, the 3D biological structure is represented as a function $\sigma : \mathbb{R}^3 \mapsto \mathbb{R}$, which expresses the Coulomb potential induced by the atoms. To recover the potential function, the probing electron beam interacts with the electrostatic potential, resulting in projections $\{\mathbf{I}_i\}_{1\leq i\leq n}$. Specifically, each projection can be expressed as

$$\mathbf{I}(x,y) = g \star \int_{\mathbb{R}} \sigma\left(\mathbf{R}^\top \mathbf{x} + \mathbf{t}\right) \mathrm{d}z + \epsilon, \quad \mathbf{x} = (x,y,z)^\top \tag{1}$$

where $\mathbf{R} \in SO(3)$ is an orientation representing the 3D rotation of the molecule and $\mathbf{t} = (t_x, t_y, 0)^\top$ is an in-plane translation corresponding to an offset between the center of projected particles and center of the image. The projection is, by convention, assumed to be along the z-direction after rotation. The image signal is convolved with g, a pre-estimated point spread function (PSF) for the microscope, before being corrupted with the noise ϵ and registered on a discrete grid of size $D \times D$, where D is the size of the image along one dimension. Cryo-EM reconstruction is typically performed on a per-structure basis, where reconstruction algorithms are capable of determining a structure (with motion) from cryo-EM images obtained in a single experiment. We give a more detailed formulation for cryo-EM reconstruction in the appendix.

3.2 Orientation and Deformation Image Encoding

Given an input image $\mathbf{I}$, we extract latent features for its orientation and deformation using image encoders. The former is used to optimize the initial coarse pose estimation, while the latter reflects the conformational state of $\mathbf{I}$. The final reconstructed density volume will be conditioned on the deformation feature.[‡]

Orientation Encoding. Initial pose estimations via off-the-shelf software such as RELION [55] and cryoSPARC [49] can be imprecise when structures exhibit

[‡] After training, the inference of our model does not necessarily require a specific cryo-EM image or its deformation feature. Instead, the user can sample in the latent space of the deformation feature and obtain a corresponding density volume.

significant motion. Given an input image $\mathbf{I}$ and its associated initial pose estimation $\phi_0 = (\mathbf{R}_0, \mathbf{t}_0)$, our orientation encoding improves ϕ_0 and makes an optimized pose estimation $\hat{\phi} = (\hat{\mathbf{R}}, \hat{\mathbf{t}})$. Specifically, our orientation encoder produces an orientation feature $\boldsymbol{F}_{\mathrm{O}}$ in 8-dimensional space. This feature is represented within 6-dimensional space $\mathbb{S}^2 \times \mathbb{S}^2$, accounting for rotations, and the remaining 2 dimensions representing translations. Orientation features can uniquely determine a pose estimation through spatial transformations, effectively addressing the discontinuity issues associated with directly predicting the $SO(3)$ group via a network [77]. We regularize the orientation encoder using the term:

$$\mathcal{L}_{\text{pose}} = \sum_{i=1}^{n} \left(\frac{1}{9} \left\| \hat{\mathbf{R}}_i - \mathbf{R}_{0,i} \right\|_2 + \frac{1}{2} \left\| \hat{\mathbf{t}}_i - \mathbf{t}_{0,i} \right\|_1 \right). \tag{2}$$

During training, the orientation encoder estimates the pose of each image to transform the 3D structure representation for the minimization of the image loss (Equation (7)). In addition to the pose regularizer, gradients from the image loss are also backpropagated to the pose encoder. This allows the pose encoder to find a balance between adhering to the prior knowledge from the initial coarse estimation and minimizing the image loss, thereby optimizing the initial coarse estimation.

Deformation Encoding. To extract information related to the conformational state from a projection $\mathbf{I}$, our deformation encoder maps it into a deformation feature $\boldsymbol{F}_{\mathrm{D}}$. It subsequently interacts with 3D spatial features within the query-based deformation transformer decoder (Section 3.4) and thus the final reconstructed density volume will be conditioned on the deformation feature.

3.3 Real Domain Implicit Feature Volume

In Fourier domain reconstruction, due to the usage of the Fourier slice theorem, evaluating one pixel of a 2D Fourier image is equivalent to making one single inference on the center slice of a 3D Fourier volume. In contrast, for real domain reconstruction, evaluating one pixel of a 2D image involves a projection (numerical integral) along the z-direction and thus an order of magnitude more inference than the former. Consequently, directly using coordinate-based MLP in the style of Fourier reconstruction methods [31, 32, 75] is computationally prohibitive for real domain reconstruction.

To reduce the computational cost, we adopt multi-resolution hash grid encoding [41]. To be specific, our 3D representation involves a hash grid $\mathcal{V}_\Theta$ parameterized by Θ. For any given input coordinate $\mathbf{x} = (x, y, z)^\top$, its high-dimensional spatial feature is represented as

$$\boldsymbol{F}_{\mathrm{S}}(\mathbf{x}) = \mathcal{V}_\Theta(\mathbf{x}; \Theta). \tag{3}$$

This feature encapsulates the local structural information of the specified input location.

3.4 Query-based Deformation Transformer

We use Attention to denote the scaled dot-product attention, operating as

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{C}}\right)\boldsymbol{V}. \tag{4}$$

Previously introduced, for any particle image $\mathbf{I}$, we have extracted its deformation feature $\boldsymbol{F}_{\text{D}}$ representing its global conformational state. To recover the density value $\hat{\sigma}$ for an arbitrary coordinate $\mathbf{x}$ at this conformational state, we propose a novel query-based deformation transformer decoder to allow for the interaction between the global deformation information $\boldsymbol{F}_{\text{D}}$ and the local spatial feature $\boldsymbol{F}_{\text{S}}(\mathbf{x})$.

Structure Queries. We introduce learnable structure queries $\boldsymbol{Q} \in \mathbb{R}^{N\times C}$, where N is the number of queries and C is the number of dimensions of each query. Structure queries serve as a carrier to integrate $\boldsymbol{F}_{\text{S}}$ with $\boldsymbol{F}_{\text{D}}$. Specifically, we partition the 3D space uniformly into $N = n \times n \times n$ blocks, with each query corresponding to one of these blocks, where n is the number of blocks along each axis. We compel each query to focus exclusively on a specific region within the space via a region-wise spatial cross-attention, which we will detail below.

Deformation-aware Decoder Block. For any image, structure queries $\boldsymbol{Q}$ first interact with its deformation feature $\boldsymbol{F}_{\text{D}}$ in the deformation-aware decoder blocks to extract conformational information exclusive to this specific image. Each deformation-aware block sequentially consists of an inter-query self-attention block ($\text{Attention}(\boldsymbol{Q}, \boldsymbol{Q}, \boldsymbol{Q})$), a deformation-aware cross-attention layer, and a feed-forward network (FFN), where the deformation-aware cross-attention layer is computed as $\text{Attention}(\boldsymbol{Q}, \boldsymbol{F}_{\text{D}}, \boldsymbol{Q})$. We stack three decoder blocks to fuse deformation cues into structure queries.

Region-wise Spatial Cross-Attention. To determine the density value at a coordinate $\mathbf{x}$, we first extract its spatial feature $\boldsymbol{F}_{\text{S}}(\mathbf{x})$ from the feature volume. Given that the density value $\sigma(\mathbf{x})$ depends on the conformational state, we introduce the region-wise spatial cross-attention mechanism. It aggregates the spatial feature $\boldsymbol{F}_{\text{S}}(\mathbf{x})$ and the structural queries $\boldsymbol{Q}$. Structural queries $\boldsymbol{Q}$ are expected to incorporate conformational state information from previous blocks. Specifically, the region-wise spatial cross-attention mechanism operates as:

$$\text{Attention}(\text{GetQuery}(\mathbf{x}), \tilde{\boldsymbol{F}}_{\text{S}}(\mathbf{x}), \boldsymbol{F}_{\text{S}}(\mathbf{x})), \tag{5}$$

where $\text{GetQuery}(\mathbf{x})$ returns the specific query associated with the spatial coordinate $\mathbf{x}$, based on the region into which $\mathbf{x}$ falls. (As previously mentioned, we uniformly partition the 3D space into blocks, with each query corresponding to one of these blocks). To avoid the expensive computational cost, we downsample the $\boldsymbol{F}_{\text{S}}$ to obtain a region-wise spatial feature $\tilde{\boldsymbol{F}}_{\text{S}}$ to an affordable spatial resolution. After region-wise spatial cross-attention, an FFN projects the queries to the final density prediction $\hat{\sigma}(\mathbf{x})$.

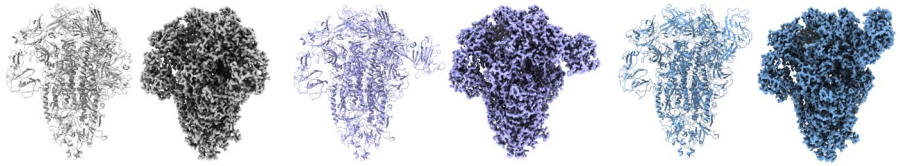

Fig. 2: Visualization of PEDV spike protein dataset. On the left in each pair are our manually modified atomic models (PDB files) in their intermediate states; on the right are their corresponding converted density fields (MRC files).

3.5 Training Scheme

To train our system, we first calculate the projected pixel values using the estimated density values and image poses as:

$$\hat{\mathbf{I}}(x,y) = \hat{g} \star \int_{\mathbb{R}} \hat{\sigma}\left(\hat{\mathbf{R}}^{\top}\mathbf{x} + \hat{\mathbf{t}}\right) \mathrm{d}z + \epsilon, \quad \mathbf{x} = (x,y,z)^{\top} \tag{6}$$

where $\hat{g}$ is the point spread function (PSF) of the projected image, assumed to be known from contrast transfer function (CTF) correction [52] in the image preprocessing stage. The loss function for training is to measure the squared error between the observed images $\{\mathbf{I}_i\}_{1\leq i\leq n}$ and the predicted images $\{\hat{\mathbf{I}}_i\}_{1\leq i\leq n}$:

$$\mathcal{L} = \sum_{i=1}^{n} \left\| \mathbf{I}_i - \hat{\mathbf{I}}_i \right\|_2^2. \tag{7}$$

4 Experimental Results

In this section, we evaluate the performance of CryoFormer for heterogeneous cryo-EM reconstruction on 2 synthetic and 3 experimental datasets, comparing it with the state-of-the-art approaches. We also showcase, through experimental analysis, our method's ability to locate flexible regions and refine initial pose estimation. We also validate the effectiveness of our building components through ablation studies. Please also kindly refer to our appendix and supplementary video.

4.1 PEDV Spike Protein Dataset

To evaluate CryoFormer and other heterogeneous cryo-EM reconstruction algorithms, we create a synthetic dataset of the spike protein of the *porcine epidemic diarrhea virus* (PEDV). The spike protein is a homotrimer, with each monomer containing a *domain 0* (D0) region that modulates the enteric tropism of PEDV by binding to *sialic acids* (SAs) on the surface of enterocytes [20] and can exist in both "up" and "down" states. [22] determined the atomic coordinates and deposited them in the Protein Data Bank (PDB) [5] under the accession codes *7W6M* and *7W73*.

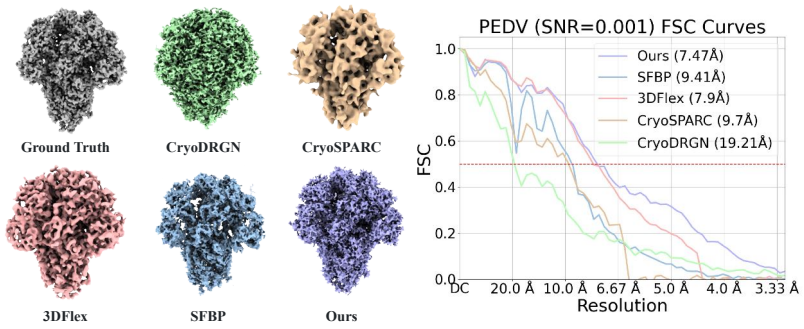

Fig. 3: Heterogenous reconstruction on PEDV spike dataset. Left: Ground truth volume and reconstructed 3D volumes with SNR = 0.001. **Right:** Curves of FSC to the ground truth volumes. Our method produces a more refined reconstruction than baselines, especially in better recovery of the flexible D0 region under severe noise. In addition, our approach yields the highest FSC curve.

We utilized *Pymol* [11] to manually supplement the reasonable process of the movement of the D0 region in the format of intermediate atomic models (Figure 2). We converted these atomic models (PDB files) to discrete potential maps (MRC files) using *pdb2mrc* module from *EMAN2* [61], which were then projected into 2D images. We then simulate the image formation model as in Equation (1) at uniformly sampled rotations and in-plane translations. On clean synthetic images, we add a zero-mean white Gaussian noise and apply the PSF. We adjust the noise scale to produce the desired SNR such as $0.1, 0.01$ and 0.001. We will make the atomic models, density maps, and simulated projections publicly available.

4.2 Experimental Setup

Implementation Details. We adopt MLPs that contain 10 hidden layers of width 128 with ReLU activations for both the orientation encoder and the deformation encoder. For the implicit spatial feature volume, we utilized a hash grid with 16 levels, where the number of features in each level is 2, the hashmap size is 2^{15}, and the base resolution is 16. This hash grid is followed by a tiny MLP with one layer and hidden dimension 64 to extract final spatial features. For the query-based deformation transformer, we adopt $N = 4096$ structure queries with $C = 8$ dimensions. For synthetic datasets, we use ground truth poses for all the methods. For real datasets, we use CryoSPARC [49] for initial pose estimation (following [75]). All experiments including training and testing have been conducted on a single NVIDIA GeForce RTX 3090 Ti GPU.

Metrics. For quantitative evaluations, we employ the Fourier Shell Correlation (FSC) curves, defined as the frequency correlation between two density maps [18]. A higher FSC curve indicates a better reconstruction result. For synthetic datasets, we compute FSC between the reconstructions and the corresponding ground truths and take the average if there are multiple conformational

Table 1: Quantitative comparison for heterogeneous reconstruction on synthetic and experimental datasets. Spatial resolution (in Å, ↓) is quantified by an FSC=0.5 threshold for synthetic datasets and 0.143 for experimental datasets. Note that for the reconstruction resolutions of Spliceosome and Integrin, some baselines achieve the highest resolution in theory, so we equally report their values.

Method	Synthetic Dataset (Å, ↓)			Experimental Dataset (Å, ↓)		
	$\text{PEDV}_{0.01}$	$\text{PEDV}_{0.001}$	1D Motion	Ribosome	Spliceosome	Integrin
CryoDRGN	6.50	19.21	3.45	3.93	**8.63**	7.43
SFBP	4.29	9.41	2.18	3.94	**8.63**	8.68
CryoSPARC	4.63	9.70	16.22	8.63	8.84	10.00
3DFlex	4.16	7.90	10.36	4.13	**8.63**	**6.46**
Ours	**4.13**	**7.47**	**2.03**	**3.80**	**8.63**	**6.46**

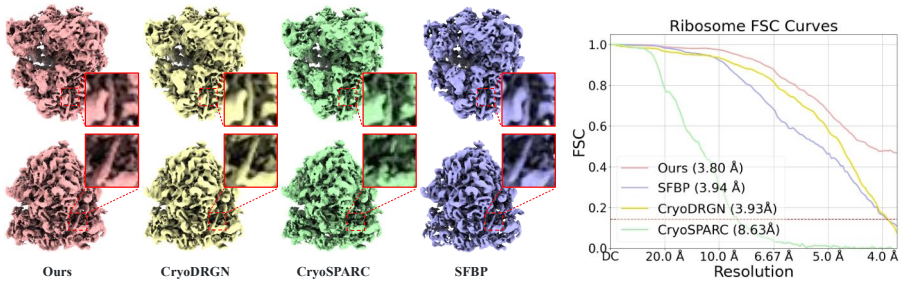

Fig. 4: Reconstruction of 80S Ribosome. Left: Reconstructed 3D volumes. **Right:** Curves of FSC between half-maps. Our method recovers secondary structures more clearly than baselines, such as the α-helices in zoom-in regions, and achieves the highest FSC curve.

states. For real experimental datasets for which we can never know what the "real" ground truth structure is, we compute FSC between two half-maps, each reconstructed from half the particle dataset. We report the spatial resolutions of the reconstructed volumes, defined as the inverse of the maximum frequency at which the FSC exceeds a threshold [53] (0.5 for synthetic datasets and 0.143 for experimental datasets).

Datasets. We evaluate different approaches on two synthetic datasets:

- **1D Motion.** The synthetic dataset proposed by [75]. This dataset contains $50,000$ images with size $D = 128$ (pixel size $= 1.0$Å) and SNR $= 0.1(-10\text{dB})$ from an atomic model of a protein complex containing a 1D continuous motion [75].
- **PEDV.** Our proposed PEDV spike protein dataset containing $50,000$ image with size $D = 128$ (pixel size $= 1.6$Å), with two different levels of noise scale: SNR $= 0.01(-20\text{dB})$ and SNR $= 0.001(-30\text{dB})$.

as well as three real experimental datasets:

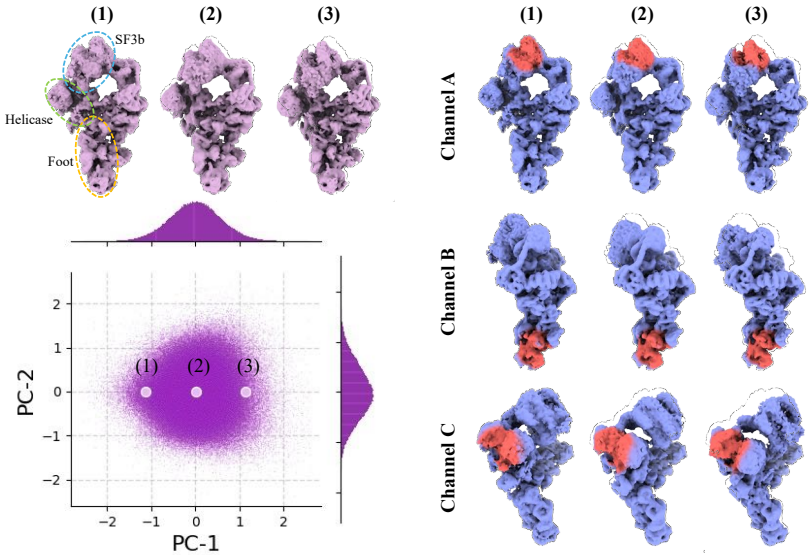

Fig. 5: Flexible Reigion Identification on Spliceosome. Left: Visualization of PCA on deformation features, as well as reconstructed volumes corresponding to three samples along one axis, exhibiting the structural motions. **Right:** Visualization of three channels of the 3D attention map by mapping attention values to the surface color of reconstructed volumes.

- **Ribosome** (EMPIAR-10028 [68]), consisting of 105,247 images of the 80S ribosome downsampled to $D = 256$ (pixel size = 1.88Å).
- **Spliceosome** (EMPIAR-10180 [46]), consisting of 327,490 images of a pre-catalytic spliceosome downsampled to $D = 128$ (pixel size = 4.2475Å).
- **Integrin** (EMPIAR-10345 [8]), consisting of 84,266 images of the asymmetric $\alpha V\beta 8$ integrin downsampled to $D = 128$ (pixel size = 3.1523Å).

Baselines. We compare CryoFormer against cryoDRGN [75] (MLPs) and Sparse Fourier Backpropagation (SFBP) [27] (voxel grids) as representatives of coordinate-based methods, as well as CryoSPARC [49] and 3DFlex [48] as representatives of conventional software.

4.3 Reconstruction Comparison

We evaluate cryoFormer compared to several state-of-the-art baselines across diverse synthetic and real datasets in terms of reconstruction resolution. As shown in Table 1, our method outperforms all baselines on every dataset. Note that our method excels on our created challenging synthetic dataset with only 0.01 SNR and surpasses most state-of-the-art approaches on experimental datasets. In conclusion, our method maintains high reconstruction resolution across varying levels of noise and motion.

PEDV Spike Protein Dataset. For qualitative comparison, we show reconstruction results in Figure 3 (left panel). When the input image noise is very

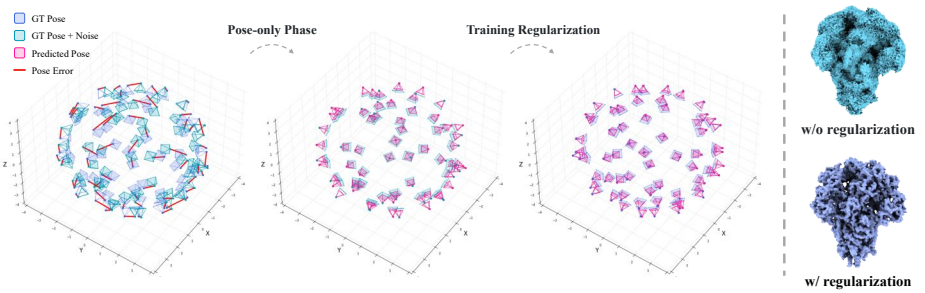

Fig. 6: Visualization of pose estimations (Left). We demonstrate the different training phases of pose refinement. Our predicted poses rapidly converge to imperfect poses during the pose-only phase and are further refined during the training regularization phase. **Qualitative comparison (Right).** With regularization of pose during the training stage, we achieve better reconstruction resolution.

high, cryoDRGN is unable to capture the contours of the protein. The reconstruction results of SFBP exhibit more voids and defects. Although cryoSPARC and 3DFlex can obtain the main structure of the protein relatively accurately, they are incapable of revealing more detailed structures. In contrast, our method can achieve more refined structures, and even accurately capture the movable D0 region. Concurrently, the FSC curves in Figure 3 (right panel) quantitatively demonstrate that the reconstruction result of our method is more refined and possesses higher resolution compared to those of other methods.

80S Ribosome. As illustrated in the left panel of Figure 4, our method manages to recover the shape and integrity of detailed structures like the α-helices (as seen in the zoom-in region) in contrast to baseline approaches. The right panel of Figure 4 shows that our FSC curve consistently surpasses those of all the baselines, quantitatively demonstrating the accuracy of our reconstructed details.

4.4 Flexible Region Identification

Our approach enables flexible region location through the analysis of 3D attention maps. Specifically, after reconstruction, we can reshape structure queries into high-dimensional attention volumes, since it is designed with correspondence to uniform spatial partitions. The distribution of values across different channels in this high-dimensional volume has physical interpretations. Certain channels exhibit higher values in specific areas compared to other regions. These areas often correspond to flexible regions with local motion. Channels of a certain structure query after the spatial cross-attention encode the local deformation information.

We showcase this capability on the pre-catalytic spliceosome dataset in Figure 5. After reconstructing and performing Principal Component Analysis (PCA) on deformation features, we obtain three reconstructed volumes corresponding

Table 2: Quantitative comparison of different pose estimation strategies. Our method consistently achieves the best performance in terms of rotation error, translation error, and resolution.

Method	Rot. Med/MSE (rad,↓)	Trans. Med/MSE (px,↓)	Res.(Å,↓)
CryoDRGN-BNB	1.47/1.55	11.18/11.24	4.26
CryoFIRE	5.12/11.26	5.13/6.70	6.23
Ours	**0.12**/**0.12**	**3.95**/**4.16**	**4.10**

to three samples along one axis, each situated in different conformational states. Through mapping to the surface color of these reconstructed volumes, Channel A corresponds to SF3b, Channel B to Helicase, and Channel C to Foot, as previously defined in [42]. Thus, we achieve precise localization of interesting flexible regions.

4.5 Pose Estimation Refinement

To evaluate our approach's capability of optimizing initial pose estimations through pose regularization, we generate a dataset with 50,000 projections of PEDV spike protein with SNR = 0.1 and evaluate our approach on it. We sample particle rotations uniformly from $SO(3)$ space and particle in-plane translations uniformly from $[-10\text{pix.}, 10\text{pix.}]^2$ space and simulate imperfect pre-computed poses by perturbing the ground truth rotations using additive noise ($\mathcal{N}(\mathbf{0}, 0.1\mathbf{I})$), and the translations using another uniform distribution $[-5\text{pix.}, 5\text{pix.}]^2$. We start the training process with only the pose loss (no image loss) to optimize the orientation encoder to initialize the orientation encoder using initial estimations. Subsequently, we train the entire model by optimizing the orientation encoder through both the image loss and pose regularization term, to mitigate the impact of inaccurate initial estimation and refine pose estimation. Figure 6 visualizes pose estimations at different stages of the training process. After the pose-only phase, the pose estimation is consistent with the initial estimation, which is erroneous compared to the ground truth. However, at the end of the major training phase, the pose estimation is very close to the ground truth.

To demonstrate the impact of pose refinement on the quality of reconstruction, the right side of Figure 6 qualitatively compares the reconstructed volumes with and without pose regularization. Without regularization to eliminate the interference of inaccurate poses, the reconstructed volume has poor resolution, whereas regularization can significantly improve both pose estimation and the final quality of the reconstruction. We quantitatively compare different strategies of pose estimation. Our baselines include CryoDRGN-BNB [75] which uses initial pose estimations for initialization and employs the branch-and-bound algorithm to refine them, as well as CryoFIRE [32] which uses a deep network to predict poses without utilizing initial estimations. As is shown in Table 2, our method surpasses the baseline methods in terms of reconstruction resolution and pose error (both rotation and translation error), demonstrating that with the same coarse initialization, our pose encoder prediction outperforms the search strategy of CryoDRGN-BNB. Although CryoFIRE similarly uses a network to

Table 3: Quantitative ablation study. We explore our key design choices, and our complete model achieves the best reconstruction performance in terms of reconstruction resolution.

Domain	DDB	RSCA	N	Resolution (Å,↓)
Fourier	✓	✓	512	7.52
Real		✓	512	10.36
Real	✓		512	8.10
Real	✓	✓	8	5.22
Real	✓	✓	64	4.87
Real	✓	✓	512	**4.13**

predict poses, its inability to utilize initial estimations by nature leads to larger pose errors and poor reconstruction resolution. Please refer to the appendix for more studies about pose refinement.

4.6 Ablation Studies

To validate CryoFormer's key architecture designs, we conduct the following evaluations on our synthetic PEDV spike protein dataset with SNR = 0.01. Specifically, we ablate on the reconstruction domain, the deformation-aware decoder blocks (**DDB**), the region-wise spatial cross-attention (**RSCA**), and the number of structure queries (N). As shown in Table 3, CryoFormer's reconstruction performance is reduced in the Fourier domain compared to the real domain as query-based transformer architecture is harder to capture globally changing frequencies in the Fourier domain than capturing local changes in the real domain. Without deformation-aware decoder blocks or region-wise spatial cross-attention, simple concatenation for feature aggregation causes structural queries cannot effectively be fused with deformation or spatial features, thereby degrading resolution. The quality of reconstruction improves with the increase in the number of structure queries, with $512(16 \times 16 \times 16)$, the maximum number our computational resources can afford, achieving the best reconstruction resolution.

5 Conclusion

We have introduced CryoFormer for high-resolution continuous heterogeneous cryo-EM reconstruction. Our approach builds an implicit feature volume directly in the real domain as the 3D representation to facilitate the modeling of local flexible regions. Furthermore, we propose a novel query-based deformation transformer decoder to enhance the quality of reconstruction. Our approach can refine pre-computed pose estimations and locate flexible regions. Quantitative and qualitative experiment results show that our approach outperforms traditional methods and recent neural methods on both synthetic datasets and real datasets. In the future, we believe our method can serve as a solid work in high-resolution and interpretable continuous heterogeneous reconstruction in cryo-EM.

References

1. Attal, B., Huang, J.B., Richardt, C., Zollhoefer, M., Kopf, J., O'Toole, M., Kim, C.: Hyperreel: High-fidelity 6-dof video with ray-conditioned sampling. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16610–16620 (2023)
2. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5855–5864 (2021)
3. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5470–5479 (2022)
4. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Zip-nerf: Anti-aliased grid-based neural radiance fields. IEEE/CVF International Conference on Computer Vision (ICCV)(2023)
5. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The protein data bank. Nucleic acids research **28**(1), 235–242 (2000)
6. Bracewell, R.N.: Strip integration in radio astronomy. Australian Journal of Physics **9**(2), 198–217 (1956)
7. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in Neural Information Processing Systems (NeurIPS)**33**, 1877–1901 (2020)
8. Campbell, M.G., Cormier, A., Ito, S., Seed, R.I., Nishimura, S.L.: Cryo-em reveals integrin-mediated tgf-β activation without release from latent tgf-β. Cell **180**, 490–501.e16 (2020), `https://api.semanticscholar.org/CorpusID:210214101`
9. Cao, A., Johnson, J.: Hexplane: A fast representation for dynamic scenes. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 130–141 (2023)
10. Chen, W., Yao, L., Xia, Z., Wang, Y.: Ace-hetem for ab initio heterogenous cryo-em 3d reconstruction. arXiv preprint arXiv:2308.04956 (2023)
11. DeLano, W.L., et al.: Pymol: An open-source molecular graphics tool. CCP4 Newsl. Protein Crystallogr **40**(1), 82–92 (2002)
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
13. Donnat, C., Levy, A., Poitevin, F., Zhong, E.D., Miolane, N.: Deep generative modeling for volume reconstruction in cryo-electron microscopy. Journal of Structural Biology p. 107920 (2022)
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations (ICLR)(2020)
15. Fang, J., Yi, T., Wang, X., Xie, L., Zhang, X., Liu, W., Nießner, M., Tian, Q.: Fast dynamic radiance fields with time-aware neural voxels. In: SIGGRAPH Asia 2022 Conference Papers. pp. 1–9 (2022)
16. Fridovich-Keil, S., Meanti, G., Warburg, F.R., Recht, B., Kanazawa, A.: K-planes: Explicit radiance fields in space, time, and appearance. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12479–12488 (2023)

17. Fridovich-Keil, S., Yu, A., Tancik, M., Chen, Q., Recht, B., Kanazawa, A.: Plenoxels: Radiance fields without neural networks. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5501–5510 (2022)
18. Harauz, G., van Heel, M.: Exact filters for general geometry three dimensional reconstruction. Optik. **73**(4), 146–156 (1986)
19. Hohn, M., Tang, G., Goodyear, G., Baldwin, P.R., Huang, Z., Penczek, P.A., Yang, C., Glaeser, R.M., Adams, P.D., Ludtke, S.J.: Sparx, a new environment for cryo-em image processing. Journal of structural biology **157**(1), 47–55 (2007)
20. Hou, Y., Lin, C.M., Yokoyama, M., Yount, B.L., Marthaler, D., Douglas, A.L., Ghimire, S., Qin, Y., Baric, R.S., Saif, L.J., et al.: Deletion of a 197-amino-acid region in the n-terminal domain of spike protein attenuates porcine epidemic diarrhea virus in piglets. Journal of virology **91**(14), e00227–17 (2017)
21. Hua, T., Li, X., Wu, L., Iliopoulos-Tsoutsouvas, C., Wang, Y., Wu, M., Shen, L., Brust, C.A., Nikas, S.P., Song, F., et al.: Activation and signaling mechanism revealed by cannabinoid receptor-gi complex structures. Cell **180**(4), 655–665 (2020)
22. Huang, C.Y., Draczkowski, P., Wang, Y.S., Chang, C.Y., Chien, Y.C., Cheng, Y.H., Wu, Y.M., Wang, C.H., Chang, Y.C., Chang, Y.C., et al.: In situ structure and dynamics of an alphacoronavirus spike protein by cryo-et and cryo-em. Nature communications **13**(1), 4877 (2022)
23. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al.: Highly accurate protein structure prediction with alphafold. Nature **596**(7873), 583–589 (2021)
24. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics (TOG)**42**(4) (2023)
25. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics (TOG)**42**(4), 1–14 (2023)
26. Kerr, J., Kim, C.M., Goldberg, K., Kanazawa, A., Tancik, M.: Lerf: Language embedded radiance fields. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 19729–19739 (2023)
27. Kimanius, D., Jamali, K., Scheres, S.H.: Sparse fourier backpropagation in cryo-em reconstruction. In: Advances in Neural Information Processing Systems (NeurIPS)(2022)
28. Kobayashi, S., Matsumoto, E., Sitzmann, V.: Decomposing nerf for editing via feature field distillation. In: Advances in Neural Information Processing Systems (NeurIPS)(2022)
29. Kühlbrandt, W.: The resolution revolution. Science **343**(6178), 1443–1444 (2014)
30. Leschziner, A.E., Nogales, E.: The orthogonal tilt reconstruction method: an approach to generating single-class volumes with no missing cone for ab initio reconstruction of asymmetric particles. Journal of structural biology **153**(3), 284–299 (2006)
31. Levy, A., Poitevin, F., Martel, J., Nashed, Y., Peck, A., Miolane, N., Ratner, D., Dunne, M., Wetzstein, G.: CryoAI: Amortized inference of poses for ab initio reconstruction of 3D molecular volumes from real cryo-EM images. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)(2022)
32. Levy, A., Wetzstein, G., Martel, J., Poitevin, F., Zhong, E.D.: Amortized inference for heterogeneous reconstruction in cryo-em. Advances in Neural Information Processing Systems (NeurIPS)(2022)

33. Li, T., Slavcheva, M., Zollhoefer, M., Green, S., Lassner, C., Kim, C., Schmidt, T., Lovegrove, S., Goesele, M., Newcombe, R., et al.: Neural 3d video synthesis from multi-view video. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5521–5531 (2022)
34. Li, Z., Niklaus, S., Snavely, N., Wang, O.: Neural scene flow fields for space-time view synthesis of dynamic scenes. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6498–6508 (2021)
35. Liu, X., Chen, J., Kao, S.h., Tai, Y.W., Tang, C.K.: Deceptive-nerf/3dgs: Diffusion-generated pseudo-observations for high-quality sparse-view reconstruction. European Conference on Computer Vision (ECCV)(2024)
36. Liu, X., Chen, J., Yu, H., Tai, Y.W., Tang, C.K.: Unsupervised multi-view object segmentation using radiance field propagation. Advances in Neural Information Processing Systems (NeurIPS)**35**, 17730–17743 (2022)
37. Liu, X., Tai, Y.W., Tang, C.K., Miraldo, P., Lohit, S., Chatterjee, M.: Gear-nerf: Free-viewpoint rendering and tracking with motion-aware spatio-temporal sampling. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19667–19679 (2024)
38. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10012–10022 (2021)
39. Lyumkis, D., Brilot, A.F., Theobald, D.L., Grigorieff, N.: Likelihood-based classification of cryo-em images using frealign. Journal of Structural Biology **183**(3), 377–388 (2013)
40. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: European Conference on Computer Vision (ECCV)(2020)
41. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics (TOG)**41**(4), 102:1–102:15 (Jul 2022). `https://doi.org/10.1145/3528223.3530127`, `https://doi.org/10.1145/3528223.3530127`
42. Nakane, T., Kimanius, D., Lindahl, E., Scheres, S.H.: Characterisation of molecular motions in cryo-em single-particle data by multi-body refinement in relion. elife **7**, e36861 (2018)
43. Nogales, E.: The development of cryo-em into a mainstream structural biology technique. Nature methods **13**(1), 24–27 (2016)
44. Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: Nerfies: Deformable neural radiance fields. IEEE/CVF International Conference on Computer Vision (ICCV)(2021)
45. Park, K., Sinha, U., Hedman, P., Barron, J.T., Bouaziz, S., Goldman, D.B., Martin-Brualla, R., Seitz, S.M.: Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. ACM Trans. Graph. **40**(6) (dec 2021)
46. Plaschka, C., Lin, P.C., Nagai, K.: Structure of a pre-catalytic spliceosome. Nature **546**(7660), 617–621 (2017)
47. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-nerf: Neural radiance fields for dynamic scenes. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10318–10327 (2021)
48. Punjani, A., Fleet, D.J.: 3d flexible refinement: Structure and motion of flexible proteins from cryo-em. BioRxiv pp. 2021–04 (2021)
49. Punjani, A., Rubinstein, J.L., Fleet, D.J., Brubaker, M.A.: cryosparc: algorithms for rapid unsupervised cryo-em structure determination. Nature methods **14**(3), 290–296 (2017)

50. Reizenstein, J., Shapovalov, R., Henzler, P., Sbordone, L., Labatut, P., Novotny, D.: Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10901–10911 (2021)
51. Renaud, J.P., Chari, A., Ciferri, C., Liu, W.t., Rémigy, H.W., Stark, H., Wiesmann, C.: Cryo-em in drug discovery: achievements, limitations and prospects. Nature reviews Drug discovery **17**(7), 471–492 (2018)
52. Rohou, A., Grigorieff, N.: Ctffind4: Fast and accurate defocus estimation from electron micrographs. Journal of structural biology **192**(2), 216–221 (2015)
53. Rosenthal, P.B., Henderson, R.: An objective criterion for resolution assessment in single-particle electron microscopy. Journal of molecular biology **333**(4), 743–745 (2003), (Appendix to: Rosenthal, P.B., Henderson, R., 2003. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. J. Mol. Biol. 333, 721–745)
54. Scheres, S.H.: Maximum-likelihood methods in cryo-em. part ii: application to experimental data. Methods in enzymology **482**, 295 (2010)
55. Scheres, S.H.: Relion: implementation of a bayesian approach to cryo-em structure determination. Journal of Structural Biology **180**(3), 519–530 (2012)
56. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4104–4113 (2016)
57. Schwab, J., Kimanius, D., Burt, A., Dendooven, T., Scheres, S.H.W.: Dynamight: estimating molecular motions with improved reconstruction from cryo-em images. bioRxiv (2023), `https://api.semanticscholar.org/CorpusID:264378895`
58. Shao, R., Zheng, Z., Tu, H., Liu, B., Zhang, H., Liu, Y.: Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16632–16642 (2023)
59. Shekarforoush, S., Lindell, D.B., Fleet, D.J., Brubaker, M.A.: Residual multiplicative filter networks for multiscale reconstruction. In: Advances in Neural Information Processing Systems (NeurIPS)(2022)
60. Song, L., Chen, A., Li, Z., Chen, Z., Chen, L., Yuan, J., Xu, Y., Geiger, A.: Nerfplayer: A streamable dynamic scene representation with decomposed neural radiance fields. IEEE Transactions on Visualization and Computer Graphics (TVCG)**29**(5), 2732–2742 (2023)
61. Tang, G., Peng, L., Baldwin, P.R., Mann, D.S., Jiang, W., Rees, I., Ludtke, S.J.: Eman2: an extensible image processing suite for electron microscopy. Journal of structural biology **157**(1), 38–46 (2007)
62. Tewari, A., Thies, J., Mildenhall, B., Srinivasan, P., Tretschk, E., Yifan, W., Lassner, C., Sitzmann, V., Martin-Brualla, R., Lombardi, S., et al.: Advances in neural rendering. In: Computer Graphics Forum. pp. 703–735. Wiley Online Library (2022)
63. Tretschk, E., Tewari, A., Golyanik, V., Zollhöfer, M., Lassner, C., Theobalt, C.: Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12959–12970 (2021)
64. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems (NeurIPS)**30** (2017)

65. Wang, L., Zhang, J., Liu, X., Zhao, F., Zhang, Y., Zhang, Y., Wu, M., Yu, J., Xu, L.: Fourier plenoctrees for dynamic radiance field rendering in real-time. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13524–13534 (2022)
66. Wang, P., Chen, X., Chen, T., Venugopalan, S., Wang, Z., et al.: Is attention all nerf needs? International Conference on Learning Representations (ICLR)(2022)
67. Wang, Q., Wang, Z., Genova, K., Srinivasan, P.P., Zhou, H., Barron, J.T., Martin-Brualla, R., Snavely, N., Funkhouser, T.: Ibrnet: Learning multi-view image-based rendering. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4690–4699 (2021)
68. Wong, W., Bai, X.c., Brown, A., Fernandez, I.S., Hanssen, E., Condron, M., Tan, Y.H., Baum, J., Scheres, S.H.: Cryo-em structure of the plasmodium falciparum 80s ribosome bound to the anti-protozoan drug emetine. Elife **3**, e03080 (2014)
69. Wu, R., Mildenhall, B., Henzler, P., Park, K., Gao, R., Watson, D., Srinivasan, P.P., Verbin, D., Barron, J.T., Poole, B., et al.: Reconfusion: 3d reconstruction with diffusion priors. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21551–21561 (2024)
70. Xian, W., Huang, J.B., Kopf, J., Kim, C.: Space-time neural irradiance fields for free-viewpoint video. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9421–9431 (2021)
71. Yao, H., Song, Y., Chen, Y., Wu, N., Xu, J., Sun, C., Zhang, J., Weng, T., Zhang, Z., Wu, Z., et al.: Molecular architecture of the sars-cov-2 virus. Cell **183**(3), 730–738 (2020)
72. Yu, A., Li, R., Tancik, M., Li, H., Ng, R., Kanazawa, A.: Plenoctrees for real-time rendering of neural radiance fields. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5752–5761 (2021)
73. Yuan, W., Lv, Z., Schmidt, T., Lovegrove, S.: Star: Self-supervised tracking and reconstruction of rigid objects in motion with neural rendering. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13144–13152 (2021)
74. Zhang, J., Liu, X., Ye, X., Zhao, F., Zhang, Y., Wu, M., Zhang, Y., Xu, L., Yu, J.: Editable free-viewpoint video using a layered neural representation. ACM Transactions on Graphics (TOG)**40**(4), 1–18 (2021)
75. Zhong, E.D., Bepler, T., Berger, B., Davis, J.H.: Cryodrgn: reconstruction of heterogeneous cryo-EM structures using neural networks. Nature Methods **18**(2), 176–185 (2021)
76. Zhong, E.D., Lerer, A., Davis, J.H., Berger, B.: CryoDRGN2: Ab initio neural reconstruction of 3D protein structures from real cryo-EM images. In: IEEE/CVF International Conference on Computer Vision (ICCV)(2021)
77. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5745–5753 (2019)