CryoFormer:Continuous Reconstruction of 3D Structures from Cryo-EM Data using Transformer-based Neural Representations

BY ZHIJUN ZENG

2023年6月8日

1 Introduction:Target

- In this paper, we propose a novel approach, cryoFormer, that utilizes a transformer-based network architecture for continuous heterogeneous cryo-EM reconstruction.
- We directly reconstruct continuous conformations of 3D structures using an implicit feature volume in the 3D spatial domain to model local changes of conformations.
- A deformation transformer decoder further improves reconstruction quality and, more importantly, locates and robustly tackles flexible 3D regions caused by conformations.

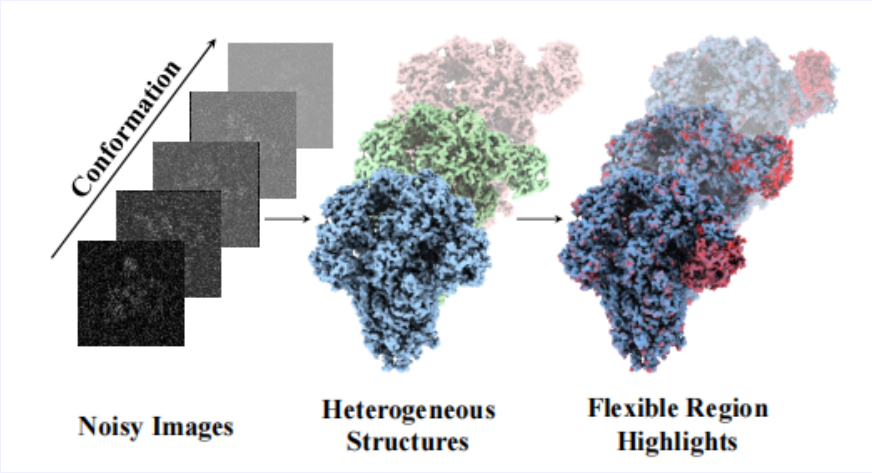
Introduction: Image Formation Model

The 3D structure is represented as a function $\sigma\colon \mathbb{R}^3 \to \mathbb{R}^+$, which expresses the Coulomb potential induced by the atoms.

The projection $\{\boldsymbol{I}_i\}_{1\leqslant i\leqslant n}$ can be expressed as

$$\boldsymbol{I}(x,y) = g * \int_{\mathbb{R}} \sigma(\boldsymbol{R}^T\boldsymbol{x} + \boldsymbol{t})\mathrm{dz} + \varepsilon, \boldsymbol{x} = (x,y,z)^T, \boldsymbol{t} = (t_x, t_y, 0)^T$$

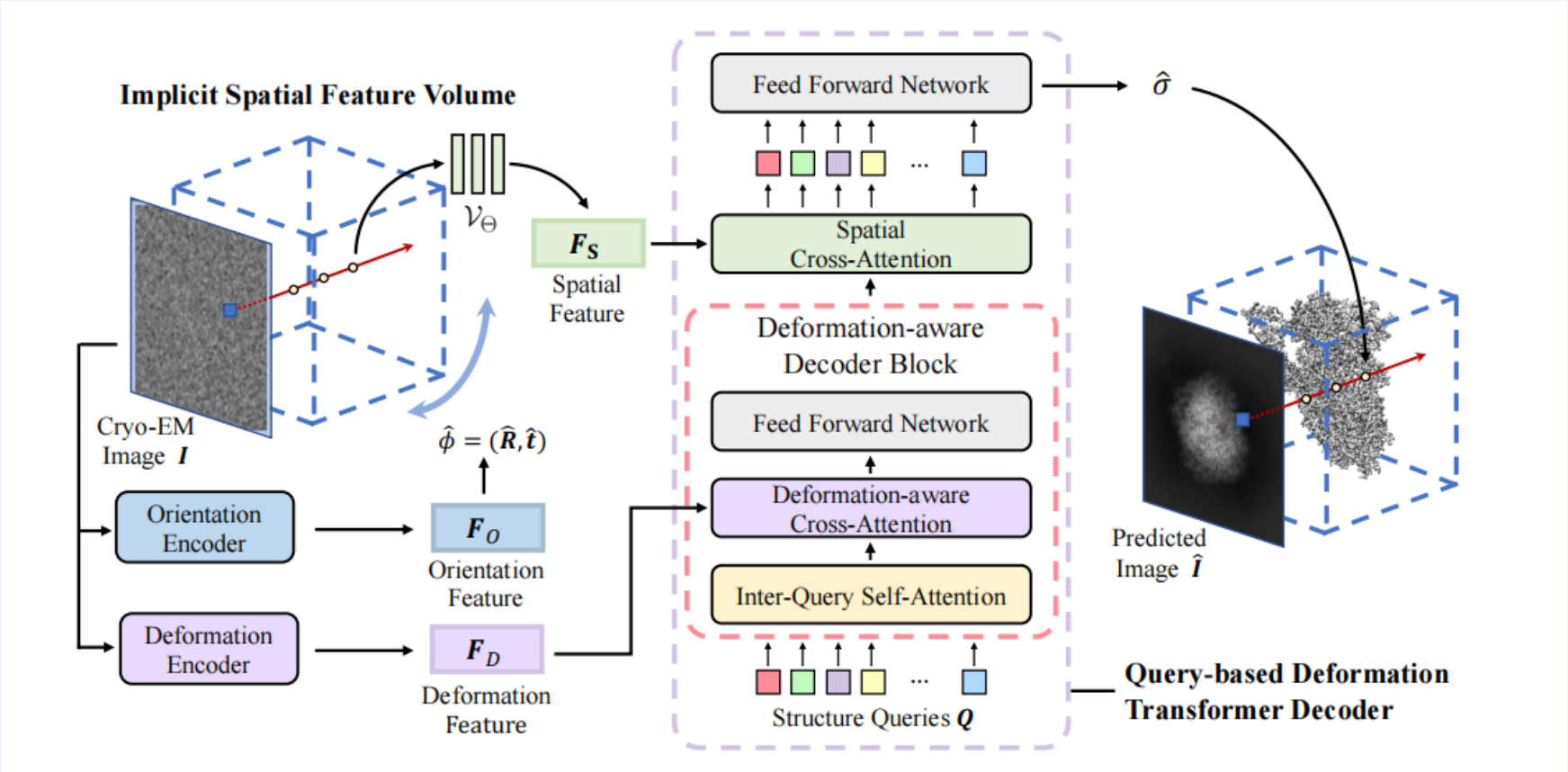
The image signal is convolved with g, a pre-estimated point spread function with noise and registered on a discrete grid of size $D \times D$.



2 Overview Of CryoFormer

- To extract conformation and pose information of image, an imput projection $\boldsymbol{I}$ is fed into image encoders that output Orientation feature F_O and a deformation feature F_D. Here F_O is then converted to pose $\phi = (\boldsymbol{R}, \boldsymbol{t})$, $\boldsymbol{R}$ is a rotation matrix to rotate a grid with D^3 size in spation domain.
- The rotated coordinates are then fed into implicit neural spatial feature volumn ν_θ which maps coordinates in $\mathbb{R}^3$ to spatial feature in $\mathbb{R}^{N \times C}$.Given a 3D point in spatial domain, INSFV output s a spatial feature F_S.
- F_S is then fed into deformation transformer decoder. The F_S spatial feature interact with structure queries Q and deformation image feature F_D to output the density prediction σ.

Pipeline of CryoFormer



2.1 Implicit Spatial Feature Volumn

We directly reconstruct continuous conformations of 3D structures using an implicit feature volume, denoted $\nu_\theta(x, y, z; \theta)$: $\mathbb{R}^3 \rightarrow \mathbb{R}^{N \times C}$, we use a similar multi-resolution hash grid encoding augmented by a single-layer MLP to decode a high dimensional spatial feature

$$\nu_\theta(x, y, z; \theta) = F_s \in \mathbb{R}^{N \times C}$$

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 16.

2.2 Orientation and Deformation Image Encoders

For both encoders, we adopt MLPs containing 10 hidden layers of width 128 with ReLU activations, following the image encoder design of cryoDRGN.

Orientation Encoder:

The spatial feature is transformed to rotation and translation $\phi = (\boldsymbol{R}, \boldsymbol{t})$, a 6-dimension space denoted as $\mathbb{S}^2 \times \mathbb{S}^2$ and converted to matrix format.The orientation encoder can be removed when an accurate pose is pre-computed for each image.

Deformation Encoder:

The projection $\boldsymbol{I}_i$ spatial feature is transformed to a deformation feature representation $F_D \in \mathbb{R}^{N \times C}$. The output density is condition on F_D.

2.3 Query-based Deformation Transformer Decoder

The transformer decoder uses a cross-attention mechanism to fuse information from different sources or modalities

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

where $Q, K, V \in \mathbb{R}^{N\times C}$.N is token number and C is hidden dimension.Note that $Q = K = V$ then we call it self attention.

Structure Query Prototypes:

We denote randomly initialized learnable structure queries as $Q \in \mathbb{R}^{N\times C}$.

Deformation-aware Decoder Block

Given an image with its deformation feature F_D,the structure queries first interact with F_D in deformation-aware Decoder block.

Each deformation-aware block sequentially consists of an inter-query self-attention block$\mathrm{Attn}(Q, Q, Q)$, a deformation-aware cross-attention layer$\mathrm{Attn}(Q, F_D, Q)$, and a feed-forward network (FFN).We stack three decoder blocks for fusing deformation cues into structure queries.

Spatial Density Estimation

To estimate the density value at a specifific coordinate, the queries then interact with its spatial feature F_S by spatial cross attention:

$$\mathrm{Attn}(Q, F_S, Q)$$

And finally, an FNN maps the queries to the esmitamed density $\hat{\sigma}$.

2.4 Training Scheme

To train our comprehensive system, we fifirst calculate the projected pixel value of the predicted image from the estimated density volume corresponding to the input image

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

where $\hat{g}$ is the point spread function (PSF) which is computed in pre-processing step.

The loss function is to measure the squared error between the observed images $\{I_i\}_{1 \leqslant i \leqslant n}$ and predicted $\{\hat{I}_i\}_{1 \leqslant i \leqslant n}$

$$\hat{L} = \sum_{i=1}^{n} \|I_i - \hat{I}_i\|^2$$

3 Result

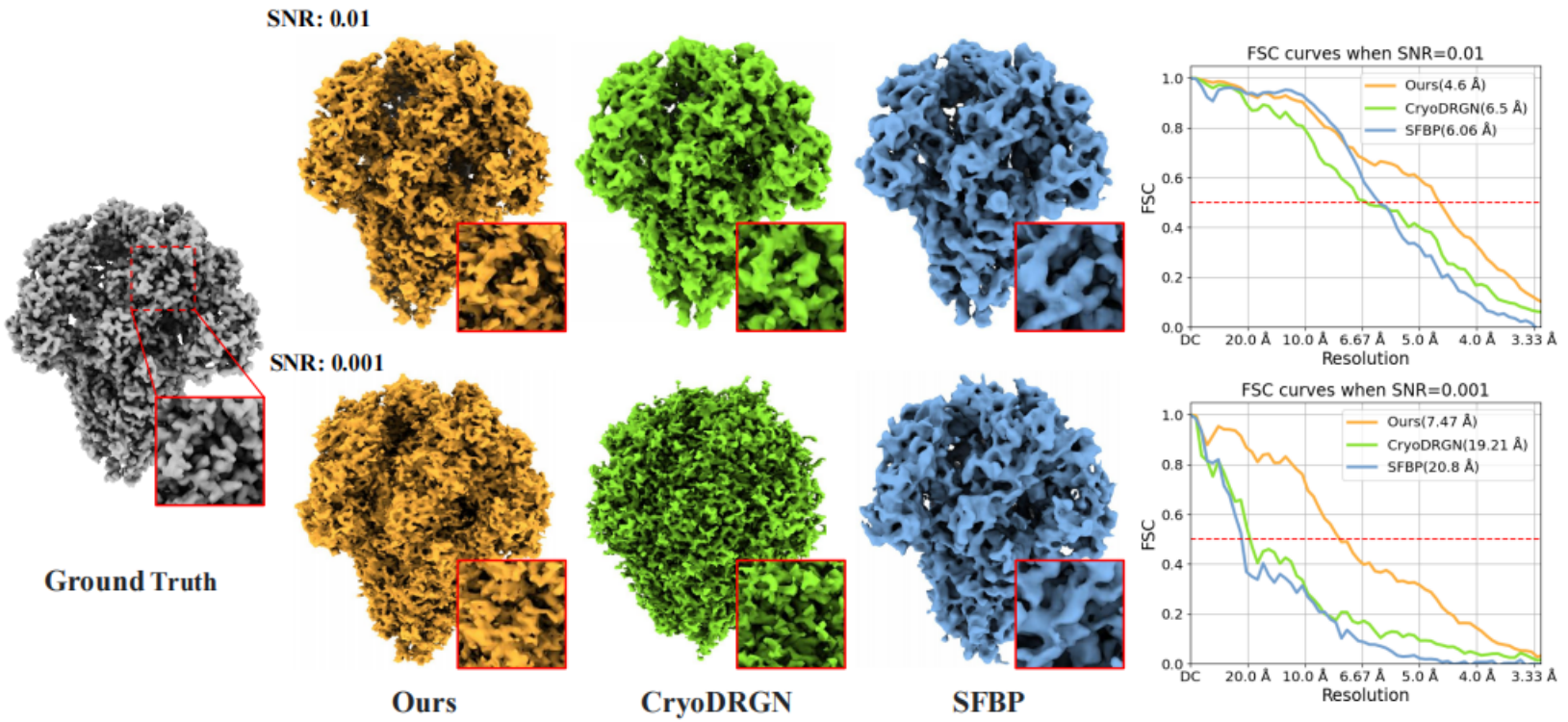


Figure 5. **Qualitative and quantitative comparison of different methods on PEDV spike sataset.** We compare cryoFormer with cryoDRGN and SFBP on PEDV spike dataset, with SNR $= 0.01$ and SNR $= 0.001$. **Left:** Reconstructed 3D volumes. Under different levels of the noise intensity, the volumes reconstructed by cryoFormer exhibit better restoration of details than the baseline. **Right:** The average FSC to each ground truth of for cryoFormer , cryoDRGN, and SFBP, where a higher curve indicates better reconstruction.

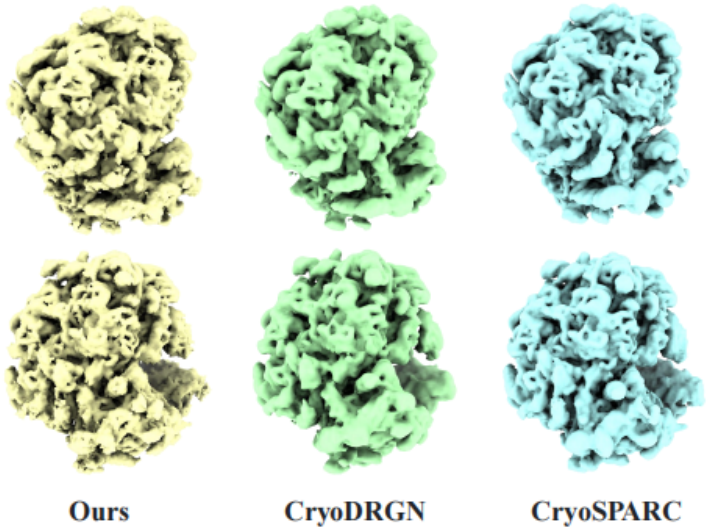


Figure 6. **Qualitative comparison for homogeneous cryo-EM reconstruction on EMPIAR-10028.** Our method matches or even outperforms baselines in terms of reconstructing fine details of biological structures. The two rows show the same reconstructed structure viewed from different angles.

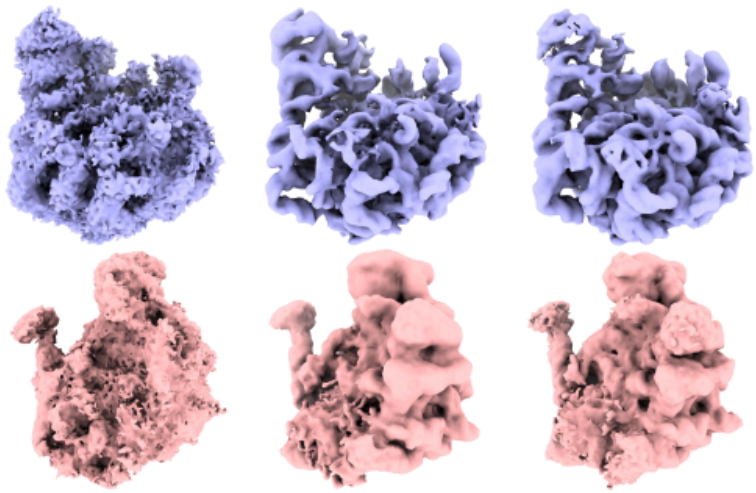


Figure 7. **Qualitative comparison for heterogeneous cryo-EM reconstruction on EMPIAR-10076.** Our method outperforms cryoDRGN and cryoSPARC in terms of the quality of heterogeneous reconstruction and is able to model the dynamic structures with finer details.

best second-best

Method	Resolution (↓)	Time (↓)
positional encoding	6.18	8.3h
Fourier domain	9.99	2.11h
w/o transformer	8.9	0.91h
Ours ($N = 32, C = 32$)	5.34	1.4h
Ours ($N = 64, C = 32$)	4.83	1.33h
Ours ($N = 32, C = 64$)	4.93	1.21h
Ours ($\mathbf{N = 64, C = 64}$)	4.74	1.39h

Table 1. **Quantitative evaluation on our design choices.** We evaluate the core design of our implicit spatial feature volume and deformation transformer decoder. When $N = 64, C = 64$, our full model in the spatial domain achieves the highest performance of spatial resolution with little sacrifice of training speed.