Fast and Direct Angle Inference Using Masked Projection Modelling in 2D Tomography with Unknown Views

Jiakang Chen, Renke Wang, Vincent C. H. Leung, and Pier Luigi Dragotti
Department of Electrical and Electronic Engineering, Imperial College London, United Kingdom
{jiakang.chen21, renke.wang19, vincent.leung, p.dragotti}@imperial.ac.uk

***Abstract*—Tomography with unknown views has been extensively studied due to its wide applications in various fields. The challenge of this problem stems from the difficulty of dealing with unknown projecting angles, which makes conventional ab initio algorithms like expectation–maximization very time-consuming. Moreover, these methods often prioritize the quality of reconstruction over the determination of unknown angles. Therefore, a method that can quickly approximate the angles of unknown view projections within a certain error margin would significantly streamline and enhance both the speed and quality of reconstruction. In response to this challenge, we propose a learning-based approach for direct angle inference. Training a network to take noisy projections from arbitrary shapes as input and predict their angles is exceptionally challenging. To address this, we introduce Masked Projection Modelling (MPM) as a surrogate task, combined with Probabilistic Angle Estimation (PAE) strategy, making direct angle inference feasible. We show that our system can estimate the angles of $\sim 10^4$ noisy projections, each generated from arbitrary shapes, in less than a second with reasonable errors, thereby greatly simplifying and accelerating the reconstruction process.**

***Index Terms*—Unknown view tomography, shape image, self-supervised learning, masked autoencoders, implicit neural representations**

I. Introduction

Tomography is a non-invasive imaging technique that effectively measures line integrals of a cross-section and has led to a wide range of applications including biomedical imaging [1], [2], oceanography [3] and geophysical imaging [4]. The problem of reconstructing the unknown image from a set of tomographic projections is often solved under the assumption that the projection angles are imprecisely known [5]. Nonetheless, there are many real world scenarios where such information is unattainable. For instance, in single particle cryo-electron microscopy (Cryo-EM) [6]–[8], as biomolecules are rapidly frozen in a thin layer of amorphous ice under cryogenic temperature, the acquired images contain many 2D projections of the particles at arbitrary unknown orientations. Hence, in this paper, we consider the tomographic reconstruction problem when the projection angles are unknown.

Existing tomographic reconstruction techniques allow to reconstruct with unknown angles by assuming prior knowledge about the signal, for instance, that the signal is bandlimited and the projection angle distribution is known [9], [10]. We instead consider arbitrary geometries as long as they have smooth boundaries, and assume that we have no prior knowledge of the distribution of the projection angles.

In recent years, deep learning-based tomographic reconstruction methods have displayed promising results in learning the latent representation of the projections [11]–[15]. Moreover, works such as [16] have shown that it may be possible to estimate the projection angles entirely from the projections without reconstructing the original shape. This stems naturally from the well-known data consistency conditions, which describe precisely the small redundancy that exists between projections in the form of homogeneous trigonometric functions of the projection angles [17]. Thus, in this work, we aim to learn this underlying relationship between the projections and their corresponding angles to recover the projection angles, and subsequently reconstruct multiple arbitrary shapes at the same time using deep learning techniques.

We propose a novel learning-based method that directly infers the projection angle difference between a pair of projections. Our proposed method consists of two stages: Masked Projection Modelling (MPM) and Probabilistic Angle Estimation (PAE). For the MPM stage, we first learn a masked autoencoder [18] that uses a reference projection to reconstruct a second projection from its randomly masked version. This acts as a surrogate task to allow an encoder to capture the underlying geometric relationship between projections and produces a high-dimensional latent representation, from which a lightweight decoder recovers the masked segments of the projections. The decoder is then simply replaced by a single linear layer to infer the projection angles in the PAE stage.

The rest of the paper is organised as follows: In Section II, we describe the mathematical definition and the geometry of tomographic projections. We then present our two-stage tomographic reconstruction method in Section III. In Section IV, we detail our experimental results on projection angle estimation and compare our reconstruction performance against one of the most effective existing learning-based algorithm. We conclude in Section V.

II. Tomographic Projection

In this paper, we focus on the problem of retrieving a 2D image from its 1D projections at unknown angles.

Mathematically, the 1D parallel beam projection $\mathbf{x}_\theta$ of a given 2D image $I \subset \mathbb{R}^2$ is given by the Radon transform of I at angle $\theta \in [0, 180)$ degree:

$$\tilde{\mathbf{x}}_\theta = \mathcal{R}_\theta I + \epsilon = \mathbf{x}_\theta + \epsilon, \tag{1}$$

This work was in part supported by BBSRC under Grant BB/Y513878/1.

where $\mathcal{R}_\theta : L_2(\mathbb{R}^2) \rightarrow L_2(\mathbb{R}^1)$ is the parallel beam projection operator, and ϵ denotes the additive noise.

Within the broad range of visual images, we focus on binary shape images, as such images are widely used in computed tomography (CT) for accurate diagnosis of various organ diseases and for reliable reconstruction from severely limited projection data [5], e.g. restricted angular range by the scanning setup (electron tomography).

III. Approach

We aim to develop a system to infer directly the angle of a target noisy projection of an arbitrary 2D shape. In general, this is impossible without extra knowledge about the image. Therefore, we introduce the concept of a reference projection, that is, for a given shape image, we randomly select a projection and assume its corresponding projection angle to be zero. We then expect the system to learn the relationship between two projections, and to infer the relative angle between the target projection and the reference projection.

Learning to infer directly the angle of a target projection from only the reference projection is a highly challenging task. Therefore, as a first step, we relax the objective by considering alternatively whether a partially masked target projection can be recovered from the reference projection. This acts as a surrogate task and allows us to implicitly establish the relationship between projections with unknown angles in a self-supervised manner. We refer to this as the Masked Projection Modelling (MPM) stage. Once we have established the relationship between projections using MPM, it becomes possible to train an angle estimation system using the ground truth angle of the target projection, which we refer to as the Probabilistic Angle Estimation (PAE) stage,

Our overall two-stage system is shown in Fig. 1. To provide more effective feature representations for the angle estimation task, the encoder learned through masking strategies at the MPM stage is then used in the PAE stage. The PAE stage then leverages this powerful encoder to infer directly the projection angles with a probabilistic estimation strategy.

A. Masked Projection Modelling (MPM)

Similar to other works based on masked autoencoders [18], our study adopts the transformer architecture to construct our model. In the preprocessing stage, noise free reference projections $\mathbf{x}_{ref}$ and target noisy projections $\tilde{\mathbf{x}}_{tgt}$ are segmented into non-overlapping patches, which are then transformed into high-dimensional vectors, termed tokens, through a linear mapping process $\mathcal{L}_\alpha(\cdot)$. This mapping allows the model to learn a high-dimensional feature representation that exhibits great linear separability to capture the implicit geometric information between projections in high-dimensional space.

The tokens obtained through linear mapping are then endowed with position embedding $\mathcal{K}_\chi(\cdot)$, which helps the model to capture the positional relationship among elements in the input token sequence. This process can be represented as:

$$[\mathbf{u}_{ref}, \mathbf{u}_{tgt}] = \mathcal{K}_\chi\big(\mathcal{L}_\alpha\left(\mathbf{x}_{ref}, \tilde{\mathbf{x}}_{tgt}\right)\big), \tag{2}$$

where $\mathbf{u}_{ref}$ and $\mathbf{u}_{tgt}$ denotes the linear projected and position embedded tokens.

Next, we employ a uniform sampling strategy to randomly select a subset of tokens from the target projection for masking. This strategy forces the model to learn a more generalized feature representation, and thus avoid relying excessively on information from specific regions. This process can be represented as:

$$\mathbf{v}_{tgt} = \mathcal{M}(\mathbf{u}_{tgt}; \eta), \tag{3}$$

where $\eta \in (0, 1)$ denotes the masking ratio.

After masking, tokens from the reference projection $\mathbf{u}_{ref}$ together with tokens from the unmasked target projection $\mathbf{v}_{tgt}$ form the input to the encoder $f_\phi(\cdot)$. By forcing the model to predict the masked parts of the target projection using information only from the reference projection and the unmasked target projection, we expect the encoder to implicitly learn the geometric relationship between projections.

In the encoding phase, we employ the self-attention mechanism to encode the tokens. The self-attention mechanism calculates the attention scores among all elements in the token sequence, enabling the model to fully exploit the underlying relationship between pairwise tokens. This mechanism allows the model to establish dynamic, context-dependent connections between unmasked reference projections and partially masked target projections, and thus, effectively capturing implicit geometric relationships between projections, even in the presence of partial information loss. The whole encoding process can be written as:

$$[\mathbf{z}_{ref}, \mathbf{z}_{tgt}] = f_\phi(\mathbf{u}_{ref}, \mathbf{v}_{tgt}), \tag{4}$$

where f_ϕ is the encoder parameterized by ϕ, $\mathbf{z}_{ref}$ denotes the encoded tokens of reference projections, and $\mathbf{z}_{tgt}$ denotes the encoded unmasked tokens of target projections.

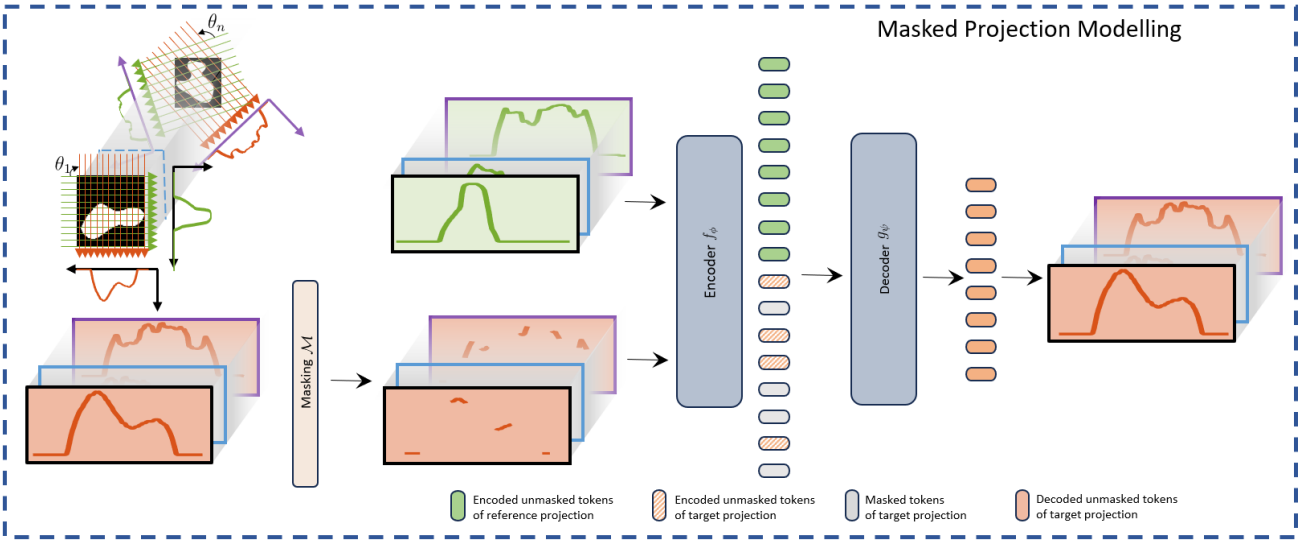
The decoder then takes encoded tokens $\mathbf{z}_{ref}$ and $\mathbf{z}_{tgt}$ as input. For indicating which parts of the target projection are missing and need to be reconstructed, the masked tokens $\mathbf{w}_{tgt}$ of the target projection are defined, which are vectors that act as placeholders. To ensure that the masked tokens can express specific positional information about the original target projection, positional embeddings $\mathcal{K}_\gamma(\cdot)$ are added to the aforementioned tokens. Similar to the encoder, the decoder g_ψ employs the self-attention mechanism for decoding. Finally, we use a linear mapping $\mathcal{L}_\beta$ to transform the decoded tokens to the reconstructed target projection, denoted as $\hat{\mathbf{x}}_{tgt}$. The overall decoding process can be summarized as follows:

$$\hat{\mathbf{x}}_{tgt} = \mathcal{L}_\beta\Big(g_\psi\big(\mathcal{K}_\gamma(\mathbf{z}_{ref}, \mathbf{z}_{tgt}, \mathbf{w}_{tgt})\big)\Big), \tag{5}$$

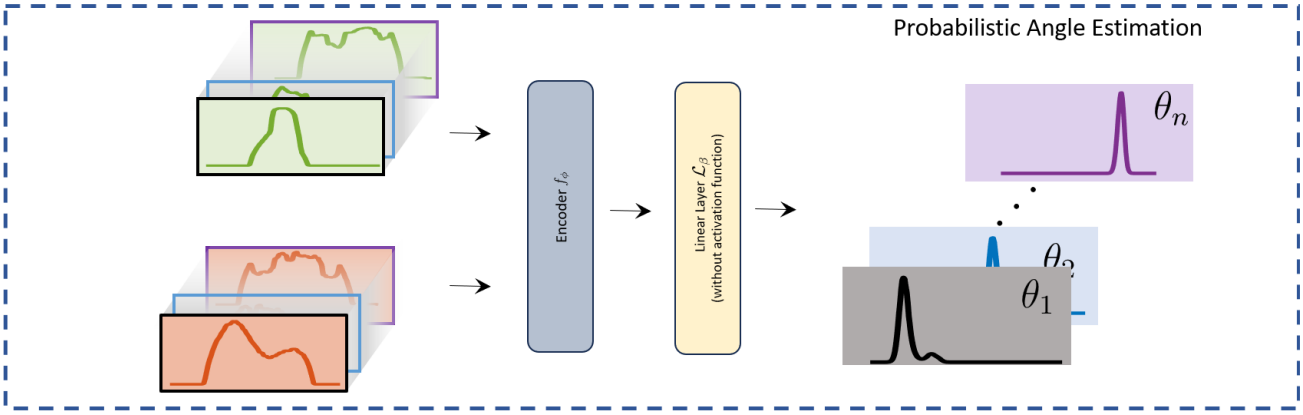
where $\mathcal{L}_\beta$ is a learnable matrix and g_ψ is the decoder parameterized by ψ. It is important to note that the decoder is only used during the MPM stage, aiming to learn a powerful encoder for the subsequent angle estimation stage. The MPM system is trained to minimize:

$$\mathcal{J} = \|\mathbf{x}_{tgt} - \hat{\mathbf{x}}_{tgt}\|_2^2, \tag{6}$$

where $\mathbf{x}_{tgt}$ is the noiseless target projection.



(a) Masked Projection Modelling (MPM): The encoder $f_\phi(\cdot)$ takes patches from an unmasked reference projection (green) and patches from a masked target projection (orange, which is a subset of its unmasked patches) as input and converts them into a high-dimensional latent representation. The Decoder $g_\psi(\cdot)$ then attempts to reconstruct the masked parts of the target projection from this latent representation.



(b) Probabilistic Angle Estimation (PAE): The encoder $f_\phi(\cdot)$ obtained from the MPM stage takes patches from an unmasked reference projection (green) and patches from an unmasked target projection (orange) as input. A simple linear layer $\mathcal{L}_\beta(\cdot)$ convert the encoded tokens to estimated angle distributions.

Fig. 1: Overview of our proposed two-stage method to infer directly the projection angles of target projections of arbitrary shapes.

B. Probabilistic Angle Estimation (PAE)

We use the encoder $f_\phi(\cdot)$ obtained from the MPM stage for the angle estimation task. This encoder $f_\phi(\cdot)$ is followed by a simple learnable linear transformation, which does not include any activation functions, as shown in Fig. 1b. To investigate the linear separability of representations obtained through the encoder at the MPM stage, we divide the implementation method for PAE into two types: (i) linear probing and (ii) end-to-end training. Specifically, approach (i) freezes the parameters of the encoder f_ϕ, only probing with a learnable linear matrix; whereas approach (ii) employs end-to-end optimization, updating both the parameters of the encoder f_ϕ and the linear matrix simultaneously.

Typically, learning based angle estimation methods directly perform regression on angles. However, this is often challenging due to the discontinuous nature of angles. Therefore, instead of directly predicting the projection angles, we opt to predict a probability distribution. Specifically, during training, we parameterize the ground truth angles as Gaussian distributions as follows:

$$Q_i = \frac{\exp\left(-\frac{1}{2}\left(\frac{\frac{180i}{N}-\theta}{\sigma}\right)^2\right)}{\sum_{j=1}^{N}\exp\left(-\frac{1}{2}\left(\frac{\frac{180j}{N}-\theta}{\sigma}\right)^2\right)}, \tag{7}$$

where θ is the ground truth angle, N is the length of bins and σ is the standard deviation. This approach allows us to directly optimize the Kullback-Leibler (KL) divergence $D_{\mathrm{KL}}(P\|Q)$ between the predicted distribution P and the

TABLE I: Average standard deviation of the error in angle estimation using linear probing and end-to-end training for various SNR levels.

SNR (dB)	5	10	20	30
Linear Probing (Degree)	14.56	9.21	5.62	4.14
End-to-End Training (Degree)	14.15	8.83	5.31	3.92

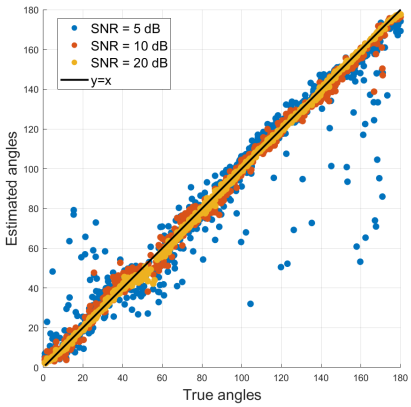


Fig. 2: Scatter plot of the estimated projection angles of the top shape in Fig. 4 using linear probing against the true angles under various SNR levels. The solid line represents a perfect agreement between the estimated and true angles. The standard deviation of the estimation is 17.36, 9.682 and 2.490 degrees for SNR level 5 dB, 10 dB and 20 dB respectively.

ground truth distribution Q during the training stage of PAE. By transforming the task of angle prediction into predicting a probability distribution of possible angles, our method can more naturally handle the discontinuities. This probabilistic approach reduces abrupt changes and uncertainties during the prediction process, thereby aiding in the model's convergence and generalization ability.

IV. Results

A. Implementation Details

We performed self-supervised Masked Projection Modelling (MPM) on the training dataset, and then used the ground truth angles for evaluating the encoded representations with (i) linear probing for angle estimation or (ii) end-to-end training for angle estimation. All our experiments are implemented in PyTorch [19] with AdamW [20] optimiser. We selected a masking rate of $\eta = 75\%$, as a higher masking rate helps us to model the implicit geometric relationships between projections more deeply and enhances the generalization performance of the model. In terms of data processing, we constructed a dataset containing 100,000 binary images with a resolution of 128×128, and randomly divided them into training, validation, and test sets in an 8:1:1 ratio. We also assumed that the shape of the images is confined within a circle with a radius of 128, thereby ensuring that the size of the generated projections is 128. We generate 100 random projections per image where the angle is random chosen within the range $[0, 180]$ degrees. Regarding the model architecture, we divided projections of

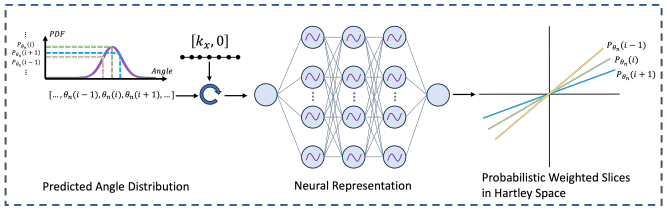


Fig. 3: 2D Reconstruction using implicit neural representation.

length 128 into non-overlapping patches of length 2, and linearly projected each patch into a high-dimensional vector. We adopted an asymmetric encoder-decoder structure: the encoder contains 16 self-attention modules, while the decoder includes 4 self-attention modules. These self-attention modules are adopted from the standard architecture of vision transformers [21].

B. Projection Angle Estimation

In Table I, we present the performance of angle estimation when target projection under different SNRs for both linear probing (fixed encoder) and end-to-end training scenarios. For each predicted angle distribution, we select the value with the highest probability. We calculated all the projections for each of the 10,000 shapes in the test dataset, then obtained the above average standard deviations in degrees. We observe that linear probing is only slightly outperformed by end-to-end training across all SNR levels. This indicates that MPM effectively captures the implicit geometric relationships between projections. Moreover, the encoded high-dimensional representations exhibit a certain degree of linear separability.

In Fig. 2, we show a scatter plot of 1000 estimated angles using linear probing and true angles under SNR levels of $5, 10, 20$ dB respectively. The noisy projections are generated from the top ground truth shape in Fig. 4.

It is also worth mentioning that our system can estimate the angles of 10,000 projections, each of which generated from different arbitrary shapes, in less than a second with a single NVIDIA RTX 4090.

C. 2D Reconstruction using Implicit Neural Representation

Given the estimated projection angles and the noisy projections, various 2D reconstruction schemes can be employed. One of the candidates is CryoFIRE [15] which uses the physics-based implicit neural representation to reconstruct the image in the Hartley space. This approach is particularly suited for tomographic reconstruction as it can learn complex structural information from sparse and severely noise corrupted data. To reconstruct the 2D image, we append our implementation of implicit neural representation to the output of our linear layer $\mathcal{L}_\beta$ in the PAE stage as shown in Fig. 3. Diverging from CryoFIRE, our implementation selects multiple angles around the maximum value from the predicted distribution to generate multiple sets of rotated coordinates, rather than using a single set of rotated coordinates as input. This method is able to leverage the probabilistic nature of our

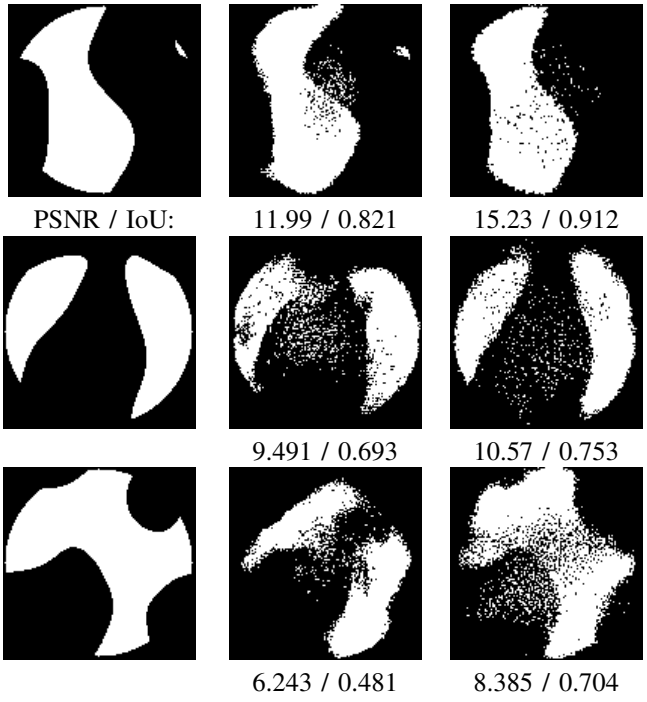


Fig. 4: Visual comparison of reconstructed images by our proposed method and CryoFIRE method from 1000 projections at SNR = 10 dB.

angle predictions by optimising over a sum of Hartley slices weighted by their corresponding predicted probabilities. For a fair comparison, we froze the parameters of encoder learned in the MPM stage and only updated the parameters of the implicit neural representation. Fig. 4 visualises the ground truth and the reconstructed images of CryoFIRE and our proposed method. It can be seen that our method has a better reconstruction result in terms of PSNR and Intersection over Union (IoU) across various example shapes.

V. Conclusion

In this paper, we presented a learning-based method that is capable of fast and accurate estimation of the projection angles of noisy tomographic projections. We demonstrated the importance of establishing implicit geometric relationships between projections by reconstructing missing parts of the projection for accurate angle estimation. Simulations under various SNRs are conducted to validate the proposed method. In future, we aim to extend our approach to more complex 3D scenarios.

References

[1] J. Frank, *Three-Dimensional Electron Microscopy of Macromolecular Assemblies*. Burlington: Academic Press, Jan. 1996.

[2] J. Drenth, *Principles of Protein X-ray Crystallography*. New York, NY: Springer, 2007.

[3] A. Levis, Y. Y. Schechner, and R. Talmon, "Statistical tomography of microscopic life," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2018, pp. 6411–6420.

[4] G. Nolet, Ed., *Seismic Tomography: With Applications in Global Seismology and Exploration Geophysics*. Dordrecht: Springer Netherlands, 1987.

[5] R. Wang, T. Blu, and P. L. Dragotti, "Reconstruction of images with finite rate of innovation from noisy tomographic projections," in *Proceedings of European Signal Processing Conference (EUSIPCO)*, Sep. 2023, pp. 1953–1957.

[6] M. van Heel *et al.*, "Single-particle electron cryo-microscopy: Towards atomic resolution," *Quarterly Reviews of Biophysics*, vol. 33, no. 4, pp. 307–369, Nov. 2000.

[7] T. Bendory, A. Bartesaghi, and A. Singer, "Single-particle cryo-electron microscopy: Mathematical theory, computational challenges, and opportunities," *IEEE signal processing magazine*, vol. 37, no. 2, pp. 58–76, Mar. 2020.

[8] S. H. Scheres, "RELION: Implementation of a bayesian approach to cryo-EM structure determination," *Journal of Structural Biology*, vol. 180, no. 3, pp. 519–530, Dec. 2012.

[9] A. Kumar, "On bandlimited signal reconstruction from the distribution of unknown sampling locations," *IEEE Transactions on Signal Processing*, vol. 63, no. 5, pp. 1259–1267, 2015.

[10] S. Shah, K. S. Gurumoorthy, and A. Rajwade, "Analysis of tomographic reconstruction of 2D images using the distribution of unknown projection angles," Apr. 2023.

[11] J.-G. Wu *et al.*, "Machine learning for structure determination in single-particle cryo-electron microscopy: A systematic review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 2, pp. 452–472, Feb. 2022.

[12] E. D. Zhong *et al.*, "CryoDRGN2: Ab initio neural reconstruction of 3D protein structures from real cryo-EM images," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*. Montreal, QC, Canada: IEEE, Oct. 2021, pp. 4046–4055.

[13] H. Gupta *et al.*, "CryoGAN: A new reconstruction paradigm for single-particle cryo-EM via deep adversarial learning," *IEEE Transactions on Computational Imaging*, vol. 7, pp. 759–774, 2021.

[14] A. Levy *et al.*, "CryoAI: Amortized inference of poses for ab initio reconstruction of 3D molecular volumes from real cryo-EM images," Aug. 2022.

[15] A. Levy *et al.*, "Amortized inference for heterogeneous reconstruction in cryo-EM," *Advances in neural information processing systems*, vol. 35, pp. 13 038–13 049, Dec. 2022.

[16] M. S. Phan *et al.*, "Moment-based angular difference estimation between two tomographic projections in 2D and 3D," *Journal of Mathematical Imaging and Vision*, vol. 57, no. 2, pp. 164–182, Feb. 2017.

[17] S. Basu and Y. Bresler, "Uniqueness of tomography with unknown view angles," *IEEE Transactions on Image Processing*, vol. 9, no. 6, pp. 1094–1106, Jun. 2000.

[18] K. He *et al.*, "Masked autoencoders are scalable vision learners," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA, USA: IEEE, Jun. 2022, pp. 15 979–15 988.

[19] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proceedings of Advances in Neural Information Processing Systems*, vol. 32, 2019, pp. 8024–8035.

[20] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proceedings of International Conference on Learning Representations (ICLR)*, Jan. 2019.

[21] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2021.