Mithun Das

[Company name]  [Company address]

**Denoising 3D TEM tomography via Advanced Neural Radiance Fields**

**Thesis Structure:**

1. Abstract
2. Introduction
3. Background
4. Related Works
5. Preliminary
6. Method
7. Experiments
8. Results
9. Datasets
10. Comparisons
11. Limitations and Future work
12. Discussion and Conclusions
13. Acknowledgments
14. References

Additional Baseline method details

**Introduction**

Electron tomography, often known as ET, is now the method of choice for determining the three-dimensional ultrastructure of organelles and cells at nanoscale resolutions(Neumüller, 2018). The 3D volume of the specimen can be reproduced by first collecting a tilted set of 2D transmission electron microscopy (TEM) images over a wide-angle range (usually +/- 60° to 80°), and then computationally recombining the images(Chreifi et al., 2019). However, the low electron doses that are applicable to biological samples (which are typically less than 100 e−/Å2) result in extremely poor signal-to-noise ratios (SNR) in the tomograms that are produced as a result (Frangakis, 2021). The leading causes of the noise are the stochastic character of the events involving electron scattering and the constraints imposed by electron detection (Joy, 2008). In addition, flaws in the alignment of the tilt axis, beam-induced specimen deformation, and distortions that are inherent to electron lenses all contribute to the contamination of the TEM data(Ellis & Cohen-Gould, 1927). This leads to very tiny structural features, which are essential for interpreting the sophisticated cellular processes and chemical interactions, to get confused and distorted. Therefore, reducing the amount of noise in tomograms is a critical step in the preprocessing phase that comes before extracting information that has biological significance.

In order to improve 3D ET reconstructions, a number of different denoising algorithms have been implemented. To a certain extent, straightforward linear filters like median filtering, Gaussian smoothing, and anisotropic diffusion filtration can reduce noise, but at the cost of a significant loss of high-resolution information(Fernandez, 2009; Frangakis & Hegerl, 2001). More enhanced regularization approaches such as total variation (TV) minimization (Goris et al., 2012)and sparse coding exploit image priors to preserve edges and the rigidity of an image. However, these methods frequently require extensive parameter tuning to strike a balance between the removal of noise and the absorption of detailed information. Deep learning models such as DnCNN (K. Zhang et al., 2016) have shown promise for 2D image denoising tasks. However, directly implementing such networks to tomograms slice-by-slice is unable to effectively exploit 3D contextual information and the spatial relationships between coordinates. Although several algorithms are capable of doing block-wise 3D denoising, they are restricted by computational restrictions (González-Ruiz & Fernández, 2023). Other methods involve training on simulated data, which might not translate well to tomograms taken from actual data(Martinez-Sanchez et al., 2023). In addition, the majority of currently available deep learning algorithms lack interpretability into the newly acquired features and have difficulty denoising non-uniform noise distributions, which are frequently seen in practice.

Recently, neural radiance fields (NeRF) (Mildenhall et al., 2020)have demonstrated unprecedented ability to synthesize photorealistic novel views of complex 3D scenes using a continuous volumetric Representations use a typical multilayer perceptron to learn a 5D radiance field in which each 3D coordinates (X,Y,Z) is correlated to an emission color (R,G,B) and volume density (MLP). The use of NeRF to noisy TEM tilt series data is a considerable difficulty, despite the fact that it has demonstrated impressive results for the new view synthesis of uncontaminated images. The COLMAP structure-from-motion algorithm is utilized by the standard NeRF pipeline in order to perform camera pose estimation for each input image. The high noise levels of TEM projections, on the other hand, can make it difficult for COLMAP to properly establish the viewing angles. Because of this, it is challenging to train NeRF directly on raw TEM pictures that contain noise. Within the scope of this research, I suggest alterations to the NeRF framework that, if implemented, will make it possible to obtain a more accurate camera pose estimation from noisy TEM tilt series, where COLMAP fails to provide any camera poses. Additionally, we study various training methods to properly condition the model on the noise characteristics of actual TEM data. We hope to overcome the constraints of regular NeRF when it is used directly out of the box to electron tomography volumes by customizing it to accommodate noisy inputs in this manner. Our noise aware NeRF model has the potential to open up new avenues for the high-fidelity 3D denoising and analysis of ET reconstructions.

**Background Information**

**Neural radiance field:** A fully connected neural network called a neural radiance field (NeRF) may provide inventive renderings of intricate 3D scenes from a sparse collection of 2D photos. It has been trained to replicate input views of a scene using a rendering loss. It functions by interpolating between input photos of a scene to create a single rendered scene. NeRF is a very efficient method for creating images from synthetic data (Mildenhall et al., 2020).

To render new views, a NeRF (*Neural Radiance Field (NeRF): A Gentle Introduction*, n.d.)network is trained to directly map from viewing direction and spatial location (5D input) to opacity and color (4D output). NeRF is a computationally demanding technique, and it might take hours or even days to process complex scenes. New algorithms, nevertheless, are readily available and significantly boost performance.

**Camera Parameters:**

**Denoising:** (Mildenhall et al., 2021; Pearl et al., 2022)

**Noise Modeling:** (Kniesel et al., n.d.)

**Implicit reconstruction:** (Kniesel et al., n.d.)

**NeRF Math:** (Bian et al., 2022)

**Joint Optimization of Poses and NeRF:** (Bian et al., 2022)

**View synthesis and image-based rendering:** (Mildenhall et al., 2020)

**TEM**

A beam of electrons is used in transmission electron microscopy (TEM), which generates images of specimens with a resolution that is far higher than that of optical microscopes(Egerton et al., 2004; Tang & Yang, 2017). In transmission electron microscopy, electrons are emitted by a tungsten filament or field emission source and then accelerated under high voltage (typically 100-300 kV) (Gault et al., 2008). Electromagnetic lenses concentrate the electron beam such that it is directed toward the extremely thin sample. Electrons, when they go through the sample, have a variety of interactions with the sample, depending on the density and the thickness of the material. This produces an electron diffraction pattern, which may be interpreted to reveal information about the structure of the material(Tang & Yang, 2017)

Additional lenses concentrate the transmitted electrons so that they may be captured as an image on a detector or camera(Gault et al., 2008). The transmission electron microscope (TEM) may provide magnifications of up to 2 million times (Gault et al., 2008), which enables the viewing of structures and details on a scale as tiny as a nanometer or an angstrom (Egerton et al., 2004). Because of this, it is an extremely useful instrument for study in the fields of materials science, cell biology, molecular structure analysis, and semiconductors(Egerton et al., 2004).

Imaging mode and diffraction mode are the major modes of operation for the transmission electron microscope (TEM) (Adrian et al., 1984). The image that is created by the transmitted electrons is used by the imaging mode. It is possible to examine either the diffraction pattern or the image depending on how the magnetic lenses are adjusted. The electron diffraction patterns are the primary focus of the diffraction mode, which focuses on the crystal structure(Adrian et al., 1984).

The preparation of samples is an essential part of TEM. To facilitate electron transmission, specimens must have a thickness of between 50 and 100 nanometers (nm)(Adrian et al., 1984). Staining with substantial amounts of heavy metal salts is required for biological and polymer materials to produce contrast(Adrian et al., 1984). Imaging of hydrated materials is possible because to specialized methods such as cryo-TEM, which vitrifies the samples(Adrian et al., 1984).There is a possibility that radiation will destroy sensitive specimens, which is one of the TEM's limitations(Egerton et al., 2004). Imaging of living biological samples is likewise not possible due to the vacuum environment(Egerton et al., 2004). Nevertheless, transmission electron microscopy continues to be an essential instrument for high-resolution structural characterization in both the physical and biological sciences(Egerton et al., 2004).

In this study, transmission electron microscopy (TEM) was used to examine Janus-like particles that were created from block copolymers. Transmission electron microscopy (TEM) gives the resolution and contrast necessary to clearly examine the nanostructure morphology and surface topology of the Janus particles(Walther & Müller, 2013) (Tang & Yang, 2017).

**Novel view Synthesis**

The term **View synthesis** refers to the process of generating new photographic viewpoints of a subject from one or more input photographs. This may be done with either a single image or many images. This allows to create unique synthetic viewpoints using only a little amount of photographic data. View synthesis is useful in a variety of contexts, including virtual reality, augmented reality, and the reconstruction of three-dimensional models(Xia & Xue, n.d.).

For view synthesis, a wide range of methods have been utilized. The multi-view stereo approach builds a three-dimensional reconstruction of a scene by piecing together a few photographs obtained with a variety of cameras (Seitz et al., 2006; Xia & Xue, n.d.). Then, this model may be displayed from any perspectives. Image-based rendering distorts and interpolates pixels depending on the original inputs to infer new viewpoints (Chen & Williams, 2023). These methods concentrate on identifying correspondences between different pictures.

The most recent deep learning algorithms develop an implicit representation of the image generation process using neural networks. The neural rendering algorithms directly produce unique views by making predictions about the values of pixels based on the attributes of the scene that they have learnt (Tewari et al., n.d.). Neural radiance fields (NeRF) are a method for efficiently encoding a scene as a continuous five-dimensional function that maps three-dimensional coordinates to volume density and view-dependent brightness(Mildenhall et al., 2020). The continuous volumetric scene representation that NeRF provides has made it possible to do photorealistic view synthesis with only a few photos.

The capacity to implicitly infer a three-dimensional structure and appearance from just two-dimensional supervision is the primary benefit offered by neural view synthesis systems. Because of this, formal three-dimensional modeling or estimate is not required. These learning-based systems continue to increase the realism and flexibility of new view creation across a wide variety of applications, including augmented reality, virtual tourism, and 3D photography (Fang et al., n.d.).

**Neural 3D shape representations:**

Major innovations in deep learning have made it possible for neural networks to automatically represent and display 3D forms. This was previously not possible. Effectively mapping 3D coordinates to shape attributes such as occupancy, signed distance, or radiance is something that neural implicit models do as opposed to explicit mesh or voxel representations (Park et al., 2019).

Early experiments were on discovering verified distance functions as a means of representing 3D surfaces for synthetic datasets.(Mescheder et al., 2018; Wu et al., 2015). In later techniques, an attempt was made to reduce the need for ground truth 3D surveillance by defining distinguishable rendering targets that could be improved with just 2D images(Sitzmann et al., 2020; Wu et al., 2015). These techniques generate a feature vector at every three-dimensional place, which is then represented as an RGB color. They have, however, been restricted to simple, smooth forms up until this point.

Neural radiance fields (NeRF), which were developed very recently, have lately shown considerable gains in modeling complicated real-world scene shape and view-dependent presentation(Mildenhall et al., 2020). Using multilayer perceptron’s, NeRF describes the radiance of the scene as well as the volume density of the scene as continuous 5D functions (3D position + 2D view direction)(Mildenhall et al., 2020). The main benefits, when compared to earlier neuronal representations of 3D space, are as follows:

1. MLPs that are based on coordinates can more accurately capture local spatial connections.
2. The production of high-quality novel perspectives is made possible by continuous scene representation.
3. The view-dependent effects such as highlights are encoded by the 5D radiance field.

In this study, we propose utilizing the capabilities of NeRF to represent and denoise 3D volumes that have been reconstructed from TEM tilt series. This will be accomplished by leveraging the strengths of NeRF. Teaching the MLP to successfully encode 3D structural priors that are crucial for biomolecular imaging might be accomplished by teaching the MLP to map noisy TEM inputs to clearer targets. It's possible that the coordinate-based volumetric modeling will be able to pick up on important local context that other 3D denoising networks overlook. The interpretability of structural features from TEM tomograms might be greatly improved because of this.

**Camera parameters**

The geometric and optical properties of a camera are referred to as its camera parameters. These parameters define how a camera constructs a picture from the 3D world(Hartley & Zisserman, 2000). Understanding the process of picture generation as well as the tasks involved in 3D computer vision relies heavily on an accurate representation of these factors.

**Intrinsic** parameters are those that are unique to a camera and are not affected by the scene:

* **Focal length** - The distance from the optical center to the image plane when the image is sharp. A primary component that determines both the field of view and the magnification (Heikkila & Silven, 1997). When dealing with non-square pixels, the x and y axes may have unique values.
* **Principal point** - The coordinates of the image's center on the plane of the sensor. enables the use of lenses that are not centered (Heikkila & Silven, 1997). It is dependent on how the lens is aligned.
* **Skew coefficient** - A rotation of the axis between the pixel grid and the sensor that considers non-rectangular pixel shapes (Heikkila & Silven, 1997). Produces a shearing transformation when applied.
* **Distortion coefficients** - This model simulates optical distortions such as radial, tangential, and narrow prism effects. Radial is the most noticeable and gives an impression like a barrel or pincushion (Heikkila & Silven, 1997).

The **extrinsic** parameters are determined by the position of the camera in relation to the world:

* **Rotation matrix** - Orientation of the camera's coordinate frame in 3 dimensions with respect to a fixed world frame (Zhang, 2000). A representation of a sequence of rotations based on the Euler angle.
* **Translation vector** - The 3D origin point of the camera center in the space for world coordinates (Zhang, 2000).

Intrinsic and extrinsic parameters collaborate to completely define the camera projection matrix (Lepetit et al., 2009), which is responsible for mapping 3D world points into 2D picture coordinates. To perform computer vision tasks such as 3D reconstruction, posture estimation, and new view synthesis, it is necessary to have an accurate assessment of these parameters.

Applications such as augmented reality (Zhang et al., 1995), autonomous navigation (Zhang et al., 1995), and computational photography (Zhang et al., 1995) rely heavily on accurate camera calibration to perform optimally. Adapting camera models to new modalities such as light field imaging is still a research challenge that is being actively worked on.

**Colmap**

COLMAP is an open-source pipeline that uses structure-from-motion (SfM) and multi-view stereo (MVS) to generate 3D models from 2D images (Schönberger & Frahm, n.d.). Through solid correspondence construction, global optimization, and volumetric fusion, it features state-of-the-art reconstructions.

* **Feature Extraction and Matching**

First, appearance - based image features that can be paired between views are found and described. Based on local gradients, SIFT is frequently used to locate scale- and rotation-invariant key points (Lowe, 1999). Each key point has a high-dimensional descriptor vector that is insensitive to noise, perspective, and illumination (Lowe, 1999).

Based on similarity measures like Euclidean or cosine distance, an effective closest neighbor search matched characteristics between image pairings. Uncertain matches can be eliminated with the ratio test (Lowe, 1999). Outlier matches that are inconsistent with a single 3D point are eliminated by geometric verification using RANSAC (Lowe, 2004).

* **Incremental Structure from Motion (SfM)**

In an incremental SfM method, the registered 2D-2D matches create initial sparse 3D point clouds (Fischler & Bolles, 1981). An initial point cloud is plotted using an initial image pair. Which views to update next are efficiently chosen by robust visibility constraints (Fischler & Bolles, 1981). With points recursively mapped from fresh views (Snavely et al., n.d.), camera poses are predicted using a Straight Linear Transform within a RANSAC cycle.

* **Global SfM Optimization**

Utilizing bundle adjustment, the progressive reconstruction is globally improved to simultaneously improve camera poses and 3D point coordinates. Scale drift is reduced with regularization. Bundle adjustment reduces the top view error between the positions of anticipated and actual 2D features in all perspectives (Heinly et al., n.d.). This enhances accuracy and comprehensiveness.

* **Multi-View Stereo (MVS) Depth Map Estimation**

The optimal cameras and points start the estimate of the multi-view stereo depth map. Using photo consistency metrics such normalized cross correlation between distorted picture patches, dense correspondence is created each view (Triggs et al., n.d.). Accuracy is improved by regularization using filtering such Gaussian smoothing (Galliani et al., n.d.). The per-view depth maps that are constructed include geometric detail.

* **Surface Reconstruction**

When creating a final 3D surface mesh, volumetric fusion methods such as screening Poisson reconstruction are utilized to merge the depth data to produce the mesh (Facciolo et al., 2015). It accomplishes this by interpolating an indicator function to provide a continuous and smooth surface. Additional post-processing steps, such as graph cuts-based optimization (Kazhdan et al., n.d.), may be utilized to improve details even further. Realism and color are added when texturing with the use of input images.

The fact that the source code for COLMAP is freely available has made it possible for several different extensions to be developed, such as semantic 3D modeling, augmented reality, and fusion with other modalities (Zhou & Koltun, n.d.). The pipeline is a shining example of some of the most effective methods for visual reconstruction, including robust matching, global optimization, and volumetric fusion.

**Related Work**

**Method**

**Datasets:**

**Experiments**

**Results**

**Limitation and Future work**

**Discussion and Conclusion**

**faasf**