Global Structure-From-Motion Enhanced Neural Radiance Fields 3D Reconstruction

Tong Ye, He Huang, Yucheng Liu, Junxing Yang *

School of Geomatics and Urban Spatial Information, Beijing University of Civil Engineering and Architecture, Beijing, China
Correspondence: yangjunxing@bucea.edu.cn

KEY WORDS: Urban 3D reconstruction, Neural Radiation Fields, sparse point cloud, SFM algorithm.

ABSTRACT:

Urban three-dimensional modeling involves digitizing geographical elements such as urban ground and architecture, profoundly impacting urban management, planning, and development. In recent years, such models have demonstrated various advantages, but traditional challenges in acquiring comprehensive geographical information persist. The emergence of unmanned aerial vehicle (UAV) oblique photography offers a viable solution, enabling cost-effective acquisition of geographical information and facilitating the construction of three-dimensional urban models. However, UAV-based 3D reconstruction encounters certain issues.

Conventional 3D reconstruction typically begins with data collection, involving two-dimensional images or point cloud data acquisition using imagery or LiDAR technology. Subsequent steps include data preprocessing, feature extraction and matching, surface model construction, texture mapping, rendering, and generating three-dimensional models. However, traditional methods exhibit limitations and drawbacks, such as inadequate adaptation to complex scenes, increased computational demands for large-scale scenes, and difficulties with dynamic scenes.

Deep learning-based 3D reconstruction methods, like MVSNet, employ deep neural networks to infer scene depth information from multiple-view images, yielding high-quality 3D reconstruction results. However, they rely heavily on prior datasets and auxiliary information, limiting generalization.

Neural Radiance Fields (NeRF) combine neural networks with volumetric rendering techniques, excelling in reconstructing objects with high precision and detail, handling dynamic scenes, and addressing areas with sparse viewpoints. However, NeRF's input requires sparse point clouds, commonly obtained using COLMAP, which has limitations. To address these limitations, using NeRF with global Structure-from-Motion (SfM) has emerged as a promising solution. This simultaneous processing of the entire dataset estimates camera trajectories and scene structure from shared features and information among multiple images, resulting in more accurate sparse point clouds and significantly improving image quality. Experimental validation using open-source and self-generated datasets demonstrates that this algorithm markedly enhances surface texture quality compared to traditional 3D reconstruction and NeRF with incremental SfM. In summary, this algorithm enhances 3D reconstruction efficacy and exhibits superior robustness, scalability, and accuracy compared to conventional methods.

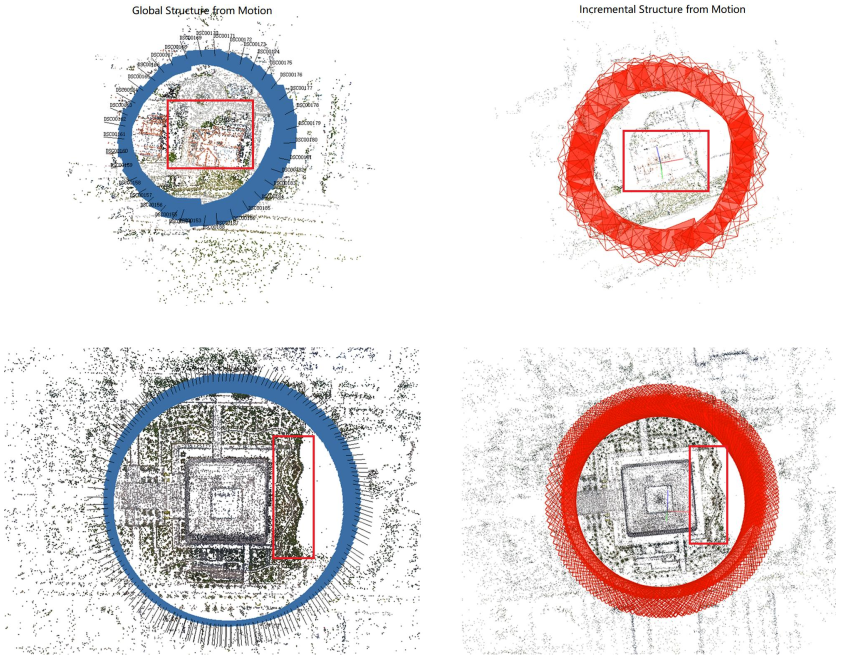
Global Structure from Motion
Incremental Structure from Motion

Figure 1. Comparison chart

1 INTRODUCTION

Three-dimensional reconstruction is widely used in various fields and industries such as virtual reality/augmented reality (VR/AR), medical imaging, robot navigation, and smart cities, serving as the foundation for tasks in complex environments. While three-dimensional scanners based on structured light principles can achieve high-precision three-dimensional reconstruction, such devices are expensive and operationally cumbersome. Over the past decade, researchers have integrated geometry-based multi-view geometry techniques into deep learning-based approaches, directly inferring three-dimensional scene representations from two-dimensional observations for subsequent three-dimensional scene reconstruction tasks.

Recently, significant progress has been made in learning-based methods for synthesizing realistic new views. The Neural Radiance Fields (NeRF) method is particularly noteworthy. NeRF is an implicit MLP model that maps a 5D vector, consisting of 3D coordinates and 2D viewing directions, to opacity and color values by fitting the model to a set of training views. The resulting 5D function can be used to generate new views using traditional volume rendering techniques.

Currently, the methods used in multi-view reconstruction can be categorized into geometric explicit methods and neural implicit methods. Geometric explicit methods include voxel-based, point cloud-based, and surface mesh-based methods. Although these methods can effectively represent the three-dimensional features of objects, they typically sample and reconstruct only local regions, requiring a large number of image inputs and increasing the memory requirements for querying 3D geometric priors. In contrast, neural implicit methods, relying on implicit neural representations, have gained widespread attention due to their advantages such as small scene representation storage and independence from view resolution.

In recent years, one of the implicit methods that has achieved excellent reconstruction results is Neural Radiance Fields (NeRF). NeRF adopts a coarse-to-fine sampling strategy, mapping the colors and densities of objects from three-dimensional spatial positions to two-dimensional pixel points implicitly by feeding a set of sparse views into a multi-layer perceptron (MLP) model, achieving high-quality reconstruction of new views. However, the simple linear fully connected layer sampling method in NeRF may result in local information loss, leading to blurred and mixed reconstruction views.

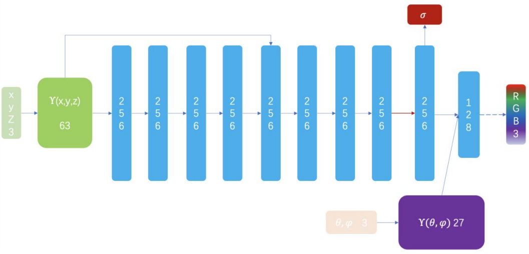

Figure 2: NeRF neural network diagram

Visual synthesis methods typically utilize an intermediate 3D scene representation to generate high-quality virtual viewpoints. The representation of this intermediate 3D scene can be analyzed in terms of "explicit representation" and "implicit representation," followed by rendering this intermediate 3D scene to generate photorealistic viewpoints.

"Explicit representation" of a 3D scene includes meshes, point clouds, voxels, etc., which can directly model the scene for display. However, due to its discrete nature, it may lead to artifacts such as overlapping and aliasing. Moreover, the large amount of data stored in explicit representations poses limitations on memory, restricting the application of high-resolution scenes.

"Implicit representation" of a 3D scene typically employs a function to describe the geometry of the scene, storing the complex 3D scene representation in the parameters of the function. As it often learns a description function for the 3D scene, the parameter volume of implicit representation is relatively smaller compared to explicit representation, and the continuous nature of implicit representation provides more expressive scene representations. NeRF achieves photorealistic synthesis using implicit representation by employing volume rendering to synthesize images from specific viewpoints. NeRF effectively converts a discrete set of images into an implicit volume representation and then utilizes this implicit volume representation along with volume rendering to generate photorealistic images from a given viewpoint.

To train the network, a training dataset containing a large number of images with known camera parameters, along with the corresponding 3D coordinates of the camera positions and camera orientations (represented as 3D unit vectors), is required for a static scene. By inputting arbitrary camera positions and orientations into the trained neural network, volume rendering can be performed to render images as outputs.

In the traditional 3D reconstruction field, COLMAP (short for "Structure-from-Motion and Multi-View Stereo") is an extremely popular open-source tool. It offers a powerful set of tools for reconstructing 3D scenes from images captured from multiple viewpoints. COLMAP holds a high position in the traditional 3D reconstruction domain, which is manifested in several aspects:

1. Rich Functionality: COLMAP provides various functionalities, including camera localization and calibration, feature extraction and matching, 3D point cloud reconstruction, dense reconstruction, 3D model optimization, and more. This makes it a comprehensive tool for 3D reconstruction.
2. Open Source Nature: COLMAP is open-source, meaning anyone can obtain and use it for free. This open-source nature has attracted many researchers and developers to improve and extend it, leading to its widespread adoption in both academia and industry.
3. Support for Academic Research: COLMAP is widely used in academic research as it provides an effective way to perform 3D reconstruction, which is essential in computer vision.
4. Community Support: COLMAP boasts an active developer community, allowing users to seek support, make suggestions, and resolve issues. This community support contributes to the continuous development and improvement of COLMAP.

Overall, COLMAP's rich functionality, open-source nature, support for academic research, and community support make it a highly esteemed tool in the field of traditional 3D reconstruction.

The new method seems to be a promising direction, especially in improving the quality and efficiency of 3D reconstruction. By simultaneously processing the entire dataset and estimating camera trajectories and scene structures from shared features and information among multiple images, this approach can generate more accurate sparse point clouds and significantly enhance image quality. Compared to traditional 3D reconstruction methods, it is better at capturing the details and lighting effects of real scenes, thus producing more realistic images.

Through experimental validation, the algorithm shows significant improvement in surface texture quality compared to traditional 3D reconstruction and incremental Structure from Motion (SfM) combined with Neural Radiance Fields (NeRF) methods. This demonstrates the effectiveness of the algorithm in enhancing the quality of 3D reconstruction.

Overall, compared to traditional methods, the algorithm offers improvements in 3D reconstruction efficiency, robustness, scalability, and accuracy. The development of this method may have a positive impact on the field of 3D reconstruction in the future, providing higher-quality and more realistic 3D reconstruction results for various applications.

2.PRELIMINARIES

We hope to build a model that has a texture quality close to that of NeRF-generated images, and has a certain improvement in the efficiency of 3D reconstruction. It is efficient and has high robustness, scalability and accuracy

.

The effect we want to achieve in the work process is as consistent as possible with the previous method. The MVS-based 3D reconstruction technology process includes the following steps: data acquisition, sparse reconstruction, depth map estimation and dense reconstruction. Incremental SfM selection disorder The image is matched for features, and geometric correction and triangulation are performed to restore the sparse point cloud structure. The relative posture is re-estimated through the existing point cloud, and local and global BA optimization is performed. Then gradually add perspectives or images to the existing structure, perform triangulation and attitude estimation, then perform BA optimization to correct the structural data, and finally output all camera parameters and sparse 3D point clouds.

3.Related Work

We use Beijing University of Civil Engineering and Architecture Daxing Campus library and gymnasium data sets for test verification and evaluation.

3.1 Collect photos

The collection of photo information of ruins generally requires taking multiple orthographic projection photos and multi-angle oblique photography photos. First, you need to set the control points and select a control point sign board with appropriate specifications. The specifications of control point sign boards generally include 10*10, 20*20, and 30*30. If the relative altitude of the drone is less than 30 meters, the 10*10 specification can be used; if the relative altitude is between 30-60 meters, the 20*20 specification can be used; if the relative altitude is between 60-100 meters, the 30*30 control point mark can be used . The control point marks can be obtained by yourself. Adjust the specifications of the picture below and print four or more copies of the control point mark board on coated paper. Set control points at the four corners of the building, use RTK to measure each point, and obtain elevation information. Orthographic projection images can be obtained using drones. Place the photographed building object in the middle of the camera frame, and adjust the exposure to make the picture clear. Control point marks The collected photos require clear images, and camera parameters should be set to avoid the impact of light and shadow on the ruins information as much as possible. When collecting library pictures, use aperture priority (A gear) to shoot, set the aperture value to 8-10, the sensitivity to between 200-400, and full frame. To comprehensively collect the image information of the shooting object, the shooting image should meet the overlap degree specified by photogrammetry. Generally, the heading overlap degree is 60% and the side overlap degree is 30%. In order to ensure that the brightness, contrast, and shadow conditions of each photo are consistent, the shooting must be completed in the shortest possible time. The library was photographed on a cloudy day to ensure as much light as possible. Drones take low-altitude orthographic images to supplement the bird's-eye view of the entire plane. The flying height can be determined according to the size of the building, as long as the photos are clear.

3.2 Data processing

First, import the library photos you took. Adjust the order and direction of taking photos to ensure they are all in portrait or landscape orientation. Store all photos in new folder. To start, the code has eight interface languages, including German, English, Spanish, French, Japanese, Portuguese, Russian, and Chinese. According to the language requirements, set the preferences in the tool and change the code language to Chinese. Add photos to your workflow.

The second step is to align the photos. The process of aligning photos mainly involves camera calibration and sparse reconstruction. This step generates a sparse point cloud with five accuracy parameters, which can be selected according to computer configuration and building accuracy requirements. The size and number of photos are important factors in how long it takes to align photos. The more photos and the higher the quality, the longer it will take to complete the reconstruction. . The point cloud generated in the interface after alignment can basically show the shape of the reconstructed object. At this time, check the overlap of the photos and observe whether there are structural gaps to determine whether additional photos need to be taken. Use the rectangular selection tool to delete excess point clouds around it, which can save calculation time in the subsequent process.

The third step is to establish a high-quality dense point cloud, generate a dense point cloud grid, and clearly present the structural information of the object surface. There are also five options for generating dense point cloud quality. Select quality "High". The quality will determine the time it takes to generate a dense point cloud. The higher the quality, the longer the calculation time. Considering the definition requirements of the generated model, it is more appropriate to select "medium" or "high" for photo quality. Through this step, the structural characteristics of the library can be basically displayed.

The fourth step is to generate the grid. The data source for generating the mesh is the high-quality dense point cloud created in the previous step. There are two surface type options: Arbitrary and Heightfield. Height fields are commonly used in aerial imagery and can be selected based on output quality requirements. Quantity and quality requirements are the main factors that affect the calculation time. The greater the quantity and the higher the quality required, the longer it takes. And select interpolation inference in Advanced. This step has completed the reconstruction of the shape and structure of the library.

The fifth step is to generate texture. The differences between different land objects are reflected and distinguished through surface textures. Generating textures mainly fills in and corrects the tones inside the library. Complete this operation to make the library model colorful.

Step six, correction. To create a marker, enter the control point elevation value measured with RTK in the workspace reference. If the 3D model is not ideal enough, you can take additional photos, add more photos, and repeat the above operations to generate a 3D model image that meets expectations. The library model has a complete structure and clear detailed features, so no additional shots are needed. Create markers directly, entering elevation coordinates measured during outdoor preparations.

The seventh step is to take a plan view. Based on the generated three-dimensional library model, plane orthophotos, cross sections, side sections and other multi-angle cross-section views can be exported. The software view provides multiple predefined views from top, bottom, left, right, and front angles to choose from. Click the View drop-down menu, set the predefined view to the top view, rotate the object and adjust it to the appropriate position to get the floor plan of the library. If you need a left view, use the rectangular selection tool to intercept the left part, set the preview to the left view, click to rotate the object, and then use the navigation to adjust the computer science to the appropriate position, which is the left section view. The operation methods for other views are the same as above, so that you can obtain multi-angle orthophotos and intercept cross-sections from different angles.

Step 8: Output the file. Save the model and export the orthophoto model image according to the required file format. The new algorithm can export files in JPEG, TIFF, PNG and other formats as well as digital surface models, which can be selected according to needs.

4.Experiment

4.1 Evaluation indicators

It can be seen that the image quality and resolution produced by the new algorithm on the same data set are significantly higher. Compared with colmap, the new algorithm can generate quite realistic three-dimensional reconstructed images, while accepting the same external condition factors, including the quality of the input image, the complexity of the scene, etc. Overall, the image effects generated by the new algorithm reveal the detail and structure of the scene, often with a high degree of fidelity. The reconstructed image may reveal textures, lighting, and shapes in the scene, as well as possible occlusions and geometric distortions.

Even in non-orthorectographic situations, if the image quality of the data set is not high, the images produced by colmap may be slightly distorted.

Specifically, the poor image quality may be reflected in :
1.Poor image quality: image blur, underexposure or overexposure, noise and other issues.
2.Insufficient perspective coverage: The lack of images from multiple angles in the data set will lead to occlusion and missing parts in the reconstruction.
3.Insufficient number of images: Insufficient number of images may result in a lack of comprehensive coverage of the scene, thus affecting the completeness and accuracy of the reconstruction results.
4.Insufficient overlapping areas: Images lacking overlapping areas may lead to difficulties in feature matching, thereby affecting the generation and reconstruction of 3D point clouds.
5.Changes in lighting conditions: Images taken under different lighting conditions may cause inconsistencies in brightness and color, thus affecting the consistency and authenticity of the reconstruction results.
6. Occlusion objects: Occlusion objects will affect feature extraction and matching in the image, thereby affecting the accuracy and completeness of reconstruction.
7. Motion blur: The movement of the camera or scene can cause motion blur in the image, making feature extraction and matching difficult, thus affecting the reconstruction results.
8.Specular reflection and transparent objects ： Specular reflections and transparent objects may cause occlusion and distortion in the image, affecting feature extraction and matching, thereby affecting the accuracy of the reconstruction results.

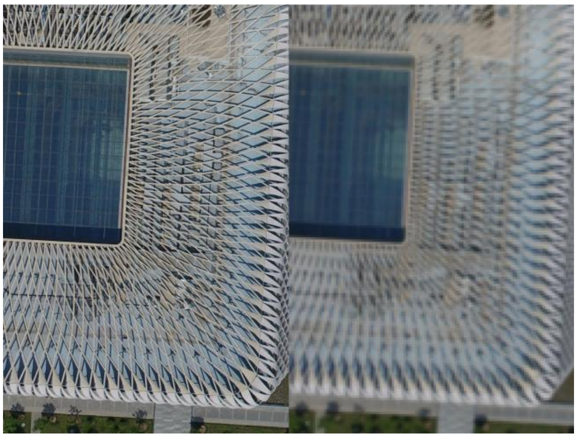

Figure3 . New algorithm vs Colmap

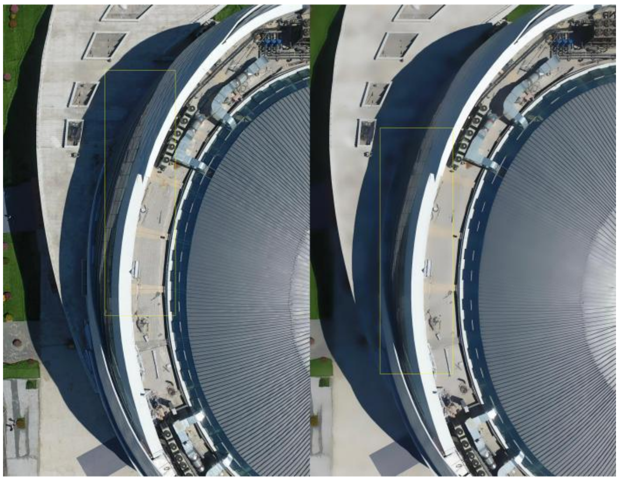

Figure4 . New algorithm vs Colmap

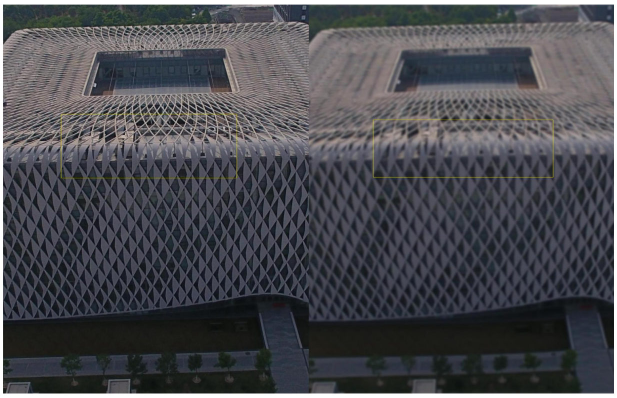

Figure5 . New algorithm vs Colmap

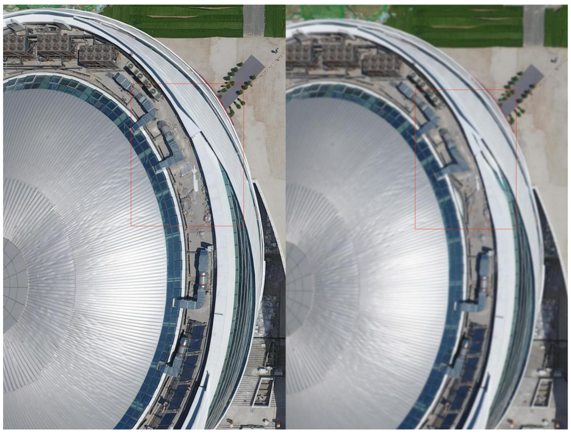

Figure 6. New algorithm vs Colmap

Even in Figures 5 and 6, we can find that the reconstruction of some of the framed new algorithms is not only higher in resolution but also more realistic in detail and texture.

5.Conclusion

In recent years, 3D modeling technology has made significant progress, especially new algorithms that use ordinary digital cameras as image acquisition tools, demonstrating its outstanding advantages and broad application prospects in many aspects. The new algorithm captures multiple high-quality digital images from different angles around the object being photographed, and combines computer vision principles and 3D modeling software to generate accurate 3D point clouds and load image information, thereby achieving the reconstruction of the object's 3D model. Compared with the traditional manual measurement and drawing method, this new algorithm can provide higher accuracy and reliability. The front section model it generates has extremely small errors, effectively improving the accuracy and detail restoration of the model. In addition, the new algorithm can not only fully display the three-dimensional image of the building, but also present orthographic views of the cultural relics from multiple angles and cross-sectional views of any part, providing more comprehensive and detailed visual information. This multi-angle and multi-dimensional display greatly enhances researchers and engineers' understanding of the structure and characteristics of objects. Simplicity and ease of use are another important advantages of this algorithm. Users only need to use ordinary digital cameras to collect images, without the need for professional equipment or complex operations, which greatly lowers the technical threshold and expands the user base. In terms of cost-effectiveness, compared with expensive traditional 3D scanning equipment, the new algorithm significantly reduces the cost of using ordinary digital cameras and 3D modeling software. It is highly cost-effective and is especially suitable for small and medium-sized projects. Mature technology and wide application are also notable features of this algorithm. By combining low-cost and widely used equipment such as electronic total stations and PTK, the three-dimensional modeling process is more efficient and reliable, and is suitable for construction, cultural relics, engineering surveying, and film and television production. , game development and other fields, showing its wide application potential. The use of ordinary digital cameras gives this method a high degree of flexibility and adaptability. It can be photographed in various environments, whether indoors or outdoors, and can generate high-quality three-dimensional models to meet the needs of different projects. In addition, the combination of computer vision and 3D modeling software makes the image matching and point cloud generation process more automated and efficient, significantly shortening modeling time and improving work efficiency. In summary, ordinary digital cameras are new algorithms for image acquisition tools. Through their multiple advantages such as high precision, comprehensive display, easy operation, low cost, mature technology, flexible adaptation, and efficient modeling, they have demonstrated the advantages of 3D modeling technology. With huge potential and broad application prospects in the field, it is expected to be promoted and applied in more fields in the future, promoting the development and popularization of 3D modeling technology.

References

Cui, Z., Tan, P., 2015. Global structure-from-motion by similarity averaging. *Proceedings of the IEEE International Conference on Computer Vision*, 864–872.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., Ng, R., 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1), 99–106.

Schonberger, J. L., Frahm, J.-M., 2016. Structure-from-motion revisited. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4104–4113.

Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L., 2018. Mvsnet: Depth inference for unstructured multi-view stereo. *Proceedings of the European conference on computer vision (ECCV)*, 767–783

Jiang, C., Sud, A., Makadia, A., Huang, J., Niener, M., Funkhouser, T.: Local implicit grid representations for 3d scenes. In: CVPR (2020)

Genova, K., Cole, F., Sud, A., Sarna, A., Funkhouser, T.: Local deep implicit functions for 3d shape. In: CVPR (2020)

Liu, S., Li, T., Chen, W., Li, H.: Soft rasterizer: A differentiable renderer for image based 3D reasoning. In: ICCV (2019)

Zhong, E.D., Bepler, T., Davis, J.H., Berger, B.: Reconstructing continuous distri butions of 3D protein structure from cryo-EM images. In: ICLR (2020)

Penner, E., Zhang, L.: Soft 3D reconstruction for view synthesis. ACM Transactions on Graphics (SIGGRAPH Asia) (2017)

https://doi.org/10.5194/isprs-archives-XLVIII-4-W10-2024-199-2024

Xu Chen, Jie Song, and Otmar Hilliges. Monocular neu ral image based rendering with continuous view control. In ICCV, pages 4090–4100, 2019.

J. Flynn, I. Neulander, J. Philbin, and N. Snavely. Deep stereo: Learning to predict new views from the world's im agery. In CVPR, pages 5515–5524, 2016.

Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In ECCV, 2018.

Gernot Riegler and Vladlen Koltun. Free view synthesis. In ECCV, pages 623–640, 2020.

Jonathan Shade, Steven Gortler, Li-wei He, and Richard Szeliski. Layered depth images. In SIGGRAPH, pages 231 242, 1998.

Shenchang EricChenand LanceWilliams.View Interpolation for Image Synthesis. In SIGGRAPH, 1993.

Richard Tucker and Noah Snavely. Single-view view syn thesis with multiplane images. In CVPR, 2020.

Shubham Tulsiani, Tinghui Zhou, Alexei A. Efros, and Ji tendra Malik. Multi-view supervision for single-view recon struction via differentiable ray consistency. In CVPR, 2017.

Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Ma lik, and Alexei A Efros. View synthesis by appearance flow. In ECCV, pages 286–301, 2016.

ZhouWang, A.C.Bovik, H.R.Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. IEEE TIP, 13(4):600–612, 2004.

Shichen Liu, Shunsuke Saito, Weikai Chen, and Hao Li. Learning to infer implicit surfaces without 3d supervision. In NeurIPS, 2019.

Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learn ing implicit 3d representations without 3d supervision. In CVPR, 2020.