Leveraging 360° Camera in 3D Reconstruction: A Vision-based Approach

Hoi Chuen Cheng, Babar Hussain, Ziyang Hong, C. Patrick Yue
The Hong Kong University of Science and Technology

***Abstract*—In this paper, we present a novel vision-based approach for 3D reconstruction using a single 360° camera, aiming to offer a simplified and accessible solution for various consumer-oriented applications. Consumer-grade 360° cameras have gained significant popularity due to their affordability and ease of use. However, traditional methods for 3D reconstruction often require complex setups with multiple cameras or expensive hardware such as LiDAR. Our approach addresses the challenges associated with 360° cameras by converting the distorted equirectangular projection (ERP) into four perspective views resembling cube maps, allowing compatibility with deep learning models trained on undistorted perspective images. We leverage visual simultaneous localization and mapping (VSLAM) techniques for camera pose estimation and employ a standard 3D reconstruction pipeline for generating detailed 3D mesh representations of the indoor environment. Through experimental evaluation, we compare the performance of 360° cameras with traditional perspective cameras in 3D reconstruction, and analyze the accuracy and performance of our vision-based approach. Our findings demonstrate the potential of using 360° cameras for constructing high-quality models and facilitating efficient data collection for 3D reconstructions, opening up new possibilities for various consumer-oriented applications in multiple fields.**

***Index Terms*—360 camera, 3D reconstruction, computer vision**

I. Introduction

In recent years, using 360° cameras has become a common practice in various industries, such as the construction and automotive industry, offering a more effective means of capturing the surrounding environment in its entirety. This paper presents a novel vision-based pipeline for 3D reconstruction using a single 360° camera, aiming to harness the potential of these cameras while addressing the challenges associated with their usage. The vision-based approach is chosen over inertial measurement units (IMU) or depth sensors such as LiDAR because they require dedicated hardware setup and software, whereas relying solely on video data is more accessible for users.

3D reconstruction is a fundamental aspect in fields such as robotics [1], Augmented Reality [2], [3], Building Information Modelling (BIM) [4], [5], and autonomous navigation [7], [8]. However, traditional methods often require complex setups with multiple cameras or specialized hardware [6]. Consumer-grade 360° cameras have gained significant popularity due to their affordability and ease of use, making them ideal for a wide range of applications.

This work is in part supported by Bright Dream Robotics and the HKUST-BDR Joint Research Institute Funding Scheme under Project HBJRI-FTP-005 (OKT22EG06).

To overcome the calibration challenges of wide field-of-view cameras, we introduce a practical solution that eliminates the need for large checkerboard patterns. Additionally, we convert the distorted equirectangular projection (ERP), an image representation commonly used for 360° cameras, into four perspective views resembling cube maps. This conversion allows compatibility with deep learning models trained on undistorted perspective images, expanding the possibilities of using consumer-grade 360° cameras in 3D reconstruction.

Our vision-based approach leverages visual simultaneous localization and mapping (VSLAM) techniques, which have become prevalent in the field. VSLAM enables real-time 3D mapping by utilizing the camera's visual input, making it suitable for applications requiring accurate and up-to-date maps. Unlike traditional SLAM systems that rely on external infrastructure, our vision-based approach works in a self-contained manner, solely utilizing the camera's visual information. This simplicity and flexibility make it well-suited for more user-friendly scenarios.

By combining the camera's pose estimation obtained from VSLAM with the cube map views, we accurately determine the camera's position and orientation. This information is crucial for generating detailed 3D mesh representations of the indoor environment. Using a standard 3D reconstruction method [31], our framework facilitates the creation of realistic 3D meshes based on the extracted camera poses and corresponding images.

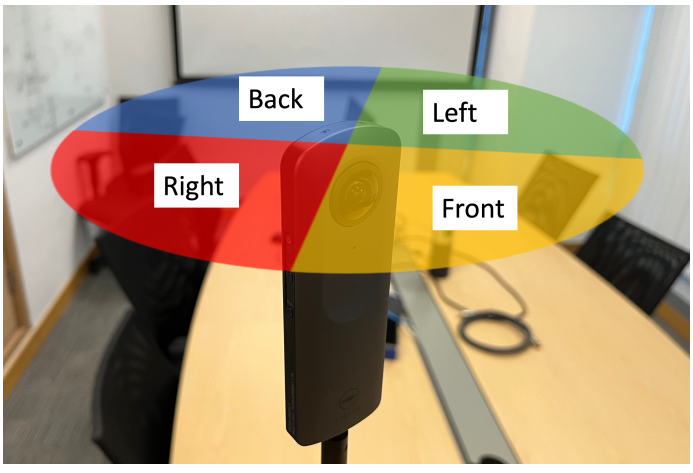

Fig. 1. Visualizing the front, back, left, right views after converting ERP into perspective images

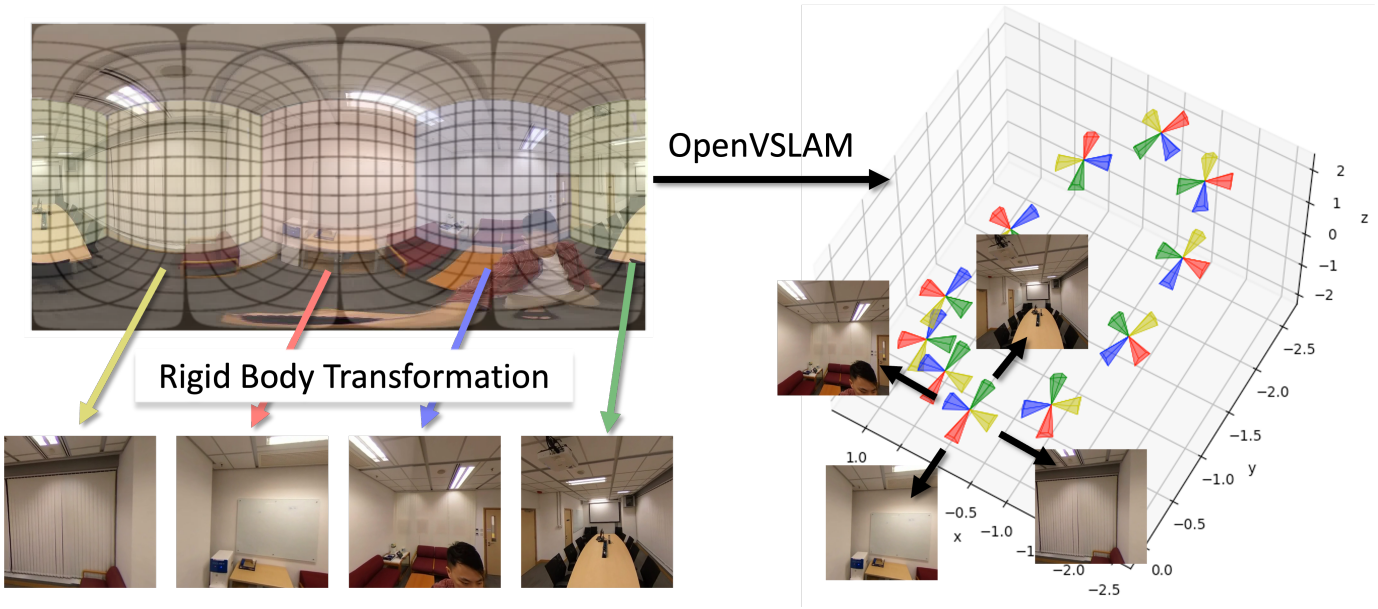

Fig. 2. Overview of the processing pipeline for converting ERP into perspective images and their corresponding poses. The top left image shows an ERP with a cube map projection overlaid on top. The cube map is aligned and registered to the ERP, providing a visualization of how the 4 cube map faces to the ERP representation. The pose visualization graph on the right depicts the changes in poses' locations and rotations for every 50 frames. The yellow, red, blue and green colour corresponds to the front, right, back and left views, which also depict the corresponding 4 perspective images on the bottom left and on the pose visualization graph.

To evaluate the effectiveness of our approach, we conducted an experimental study comparing the performance of 360° cameras with traditional perspective cameras. By capturing videos of an indoor environment while walking a complete circle, we assessed their capabilities in capturing the scene's geometry. Additionally, we compared the output 3D meshes generated by each camera type with ground truth data obtained using a LiDAR scanner. This analysis provides insights into the accuracy and performance of our vision-based approach.

Through our proposed vision-based approach using a single 360° camera, we aim to offer a simplified and accessible solution for 3D reconstruction. By leveraging the advancements in consumer technology and the advantages of 360° cameras, we unlock their potential for achieving high-quality 3D reconstructions. This opens up new possibilities for various consumer-oriented applications in multiple fields.

II. Related Works

A. 3D Reconstruction

The process of reconstructing a 3D model typically entails obtaining depth information from a series of images and integrating these depth maps. Traditional methods usually require specialized hardware, such as LiDAR scanners or stereo cameras, to capture the environment's geometry. In light of recent advancements in consumer-grade cameras and computer vision techniques, more accessible and cost-effective approaches have been made possible.

Structure from Motion (SfM) [9]–[11] and Multi-View Stereo (MVS) [12] are two common techniques for 3D reconstruction. SfM primarily relies on feature detection and matching methodologies to estimate camera poses and reconstruct the scene's 3D geometry. In contrast, MVS focuses directly on reconstructing 3D geometry from input images with calibrated cameras. Thanks to the rapid advances in deep learning, 3D reconstruction studies [24]–[26] have improved by enabling models to learn powerful feature representations directly from data. Recent methods such as [31]–[33] use neural networks to directly regress a truncated signed distance function (TSDF) volume for 3D model generation. Atlas [31] leverages extracted 3D features and directs them to semantic heads for scene labeling. These labeling or semantics can improve 3D reconstruction quality by incorporating understandings of objects, textures, and scenes that provide useful priors and constraints for generating more accurate models.

In the context of fisheye or 360° cameras, the main approach is to utilize the camera's wide Field-of-View (FOV) to capture the environment from different viewpoints. Previous works such as [27] devised a fisheye stereo matching algorithm. While more recently, deep learning techniques have been applied to 360° monocular depth estimation [28], [29], which provide the cornerstone for 3D reconstruction that requires depth information. In particular, existing methods have been leveraged for such application, such as applying MVSNet [24] in 360MVSNet [30].

B. Visual-based Pose Estimation

VSLAM techniques such as [13]–[15] utilize image information to create 3D environmental representations and estimate camera poses. Although these cameras can support different camera setups, such as monocular and stereo, they face challenges in dynamic environments, particularly with

monocular setups. Utilizing sensor fusion, such as integrating IMU [20] and LiDAR [21], enables the algorithms to work in environments with limited visual information. However, these sensor fusion setup requires complex calibration. Alternatively, widening the FOV will provide additional input data for potential improvements.

Wide FOV cameras, such as fisheye and 360° cameras, can capture enough visual information about the environment to reinforce a more accurate pose estimation. Approaches like [18], [19] have successfully extended existing techniques [14], [17] to work with 360° camera. Particularly, OpenVSLAM [22], an Oriented FAST and Rotated BRIEF (ORB) based algorithm, uses spatial feature matching for 360° camera pose estimation. With the increasing maturity of 360° cameras and the corresponding supporting algorithms, utilizing these cameras has emerged as a cost-efficient alternative to traditional monocular and stereo setups.

III. Proposed Framework

A. Conversion for Equirectangular Projection

Well-established deep learning models, which typically rely on undistorted perspective images, are not well-suited for handling 360° images. This is due to the inherent distortions presented in such images. Moreover, calibrating 360° cameras poses challenges and complexities, particularly when using large checkerboard patterns. To address these challenges, we propose a simple solution. Our approach involves converting ERP into four perspective views: front, back, left, and right. The conversion process maps the pixels from the surface of a 360° sphere onto a tangent plane, enabling a more manageable representation for subsequent processing. These transformed views resemble cube maps and can be treated as outputs from four virtual cameras positioned in different directions.

Typically, 360° cameras consist of 2 fisheye lenses, and the resulting images are subsequently stitched together. However, due to the high distortion and inherent limitations of the stitching process of the top and bottom views, we have chosen to exclude them from our experiment.

B. Pose Estimation via OpenVSLAM

The camera pose of ERP can be obtained by employing the VSLAM approach that exclusively utilizes imagery. In this study, we propose the utilization of an ORB feature extractor-based VSLAM algorithm, OpenVSLAM, which is specifically tailored to be compatible with 360° cameras. The algorithm encompasses three fundamental modules, namely tracking, mapping, and global optimization.

The tracking module is responsible for estimating the camera pose for each frame by extracting features using the ORB feature extractor. Moreover, this module determines whether a frame should be classified as a keyframe, which subsequently undergoes processing in the mapping module. Within the mapping module, the keyframes are utilized to triangulate 3D points, forming a comprehensive map of the environment. This step enables the reconstruction of spatial information from the captured 360° imagery. Finally, the global optimization module incorporates loop detection and global bundle adjustment techniques to refine and optimize the overall map. This stage ensures the accuracy and consistency of the reconstructed camera poses and 3D points.

C. Pose Extraction of Cube Map Views

To determine the pose of the four perspective views derived from ERP, we leveraged the pose calculated by OpenVSLAM and applied rigid body rotation. The rotation incorporates four distinct sets of rotation matrices corresponding to four perspective views. Fig. 2 shows the corresponding mapping (yellow, red, blue and green) of each perspective image on a pose visualization graph.

The rigid body rotation is done by applying the 3× rotation matrix

$$R = \begin{bmatrix} cos\theta & -sin\theta & 0 \\ sin\theta & cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

to the rotation component $\bar{R}$ of a pose matrix P. The right, left and back views are created through rotation θ of 90°, 180° and 270° along z-axis.

The pose matrix P, a combination of a 3×3 rotation matrix component $\bar{R}$ and a translation vector component t is defined as

$$P = \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \bar{R} & t \\ 0 & 1 \end{bmatrix}$$

.

Finally, the pose is updated as follows:

$$\bar{R}' = \begin{bmatrix} cos\theta & -sin\theta & 0 \\ sin\theta & cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix}$$

$$P' = \begin{bmatrix} R' & t \\ 0 & 1 \end{bmatrix}$$

D. 3D Mesh Generation

With the perspective images available from ERP conversion, we can combine the extracted poses with the corresponding images. The combined posed images are then passed to our 3D reconstruction system. To support applications such as VR and BIM solely by visual perception, we require a 3D reconstruction system that does not utilize any depth inputs. For such a requirement, we adopted Atlas [31], an end-to-end 3D reconstruction model that directly regresses truncated signed distance function (TSDF) from posed images (or input RGB image sequences). Atlas uses a 2D convolutional neural network (CNN) backbone to extract features from each image. Utilizing the camera intrinsics and extrinsics, these 2D extracted features are then back-projected and aggregated into a voxel volume. Finally, the aggregated voxel volume undergoes a 3D CNN-based refinement to predict TSDF values and the final 3D model.

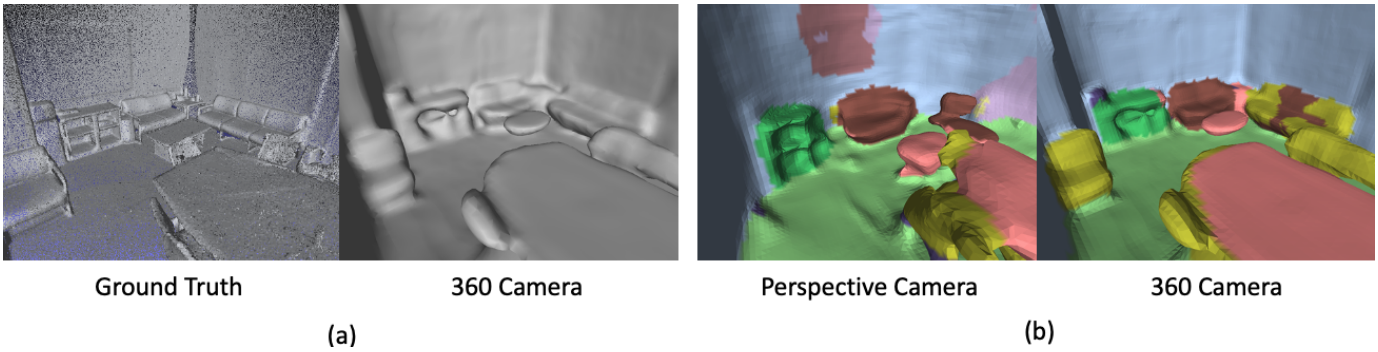

Fig. 3. Qualitative 3D reconstruction results. (a) Ground truth LiDAR point cloud vs 3D model (without semantics) generated by 360° camera's data (b) 3D model (with semantics) generated by perspective camera's data vs 3D model (with semantics) generated by 360° camera's data

IV. Experiment and Evaluation

In order to evaluate the effectiveness of using 360° cameras, we conducted several experimental studies to compare a 360° camera and a perspective camera and the 3D mesh quality from a 360° camera under different conditions. The experiment involved capturing videos of an indoor environment using both camera types. The test environment, a conference room in our laboratory, is where the cameraman walks in one complete circle while the camera captures the scene. We have chosen a popular 360° camera, Ricoh THETA V, for our experiment, and iPhone 13 for our choice of perspective camera (with a FOV of approximately 72 degrees). Each camera took five videos, and the ones that yielded the best 3D meshes were selected for the final comparison. The selected video from the 360° camera contains 484 ERP image frames, and that of the perspective camera contains 986 image frames.

A. Comparison between 360° Camera and Perspective Camera

In the first experiment, we compared the 3D reconstruction performance between a 360° camera and a conventional perspective camera to evaluate which imaging modality achieves higher quality results. The captured videos from the 360° camera are first converted into perspective images, and the camera poses from the camera are extracted using OpenVSLAM and rigid body transformation. Similarly, to generate a 3D mesh using the perspective camera, we converted the perspective video into image sequences and extracted the perspective camera poses using OpenVSLAM in monocular mode. Subsequently, we utilized Atlas to generate the 3D mesh representations of the indoor environment based on the extracted camera poses and the corresponding images for the two types of cameras.

In order to effectively assess the accuracy of the 3D models, we obtained ground truth data of our test environment with a LiDAR scanner, which is presented as the point clouds overlaid on the 3D meshes with semantic information in Fig. 3a. To evaluate the performances and compare the accuracy of the 3D meshes generated by each camera type, we calculate the F-score of the output 3D meshes with against the ground truth data.

Fig. 4 shows that the 3D mesh from the 360° camera outperforms that of perspective cameras in terms of peak F-score (0.297), and this is mainly due to the additional information ERP offers. A qualitative comparison is presented in Fig. 3b. In our 360° camera processing pipeline, a single ERP is converted into four perspective images, i.e., given the same number of raw images available, the processing pipeline can provide four times the data than the perspective camera. This data collection efficiency of 360° cameras can enable practical, real-world applications, as an operator (human or robot) can reduce the video capture time required to model an environment in 3D.

B. Quantifying Data Requirements for 3D Reconstruction using 360° Camera

Other than evaluating the optimal performance of 360° camera in 3D reconstruction), we investigated how well the model performs with different amounts of data. We then varied the amount of data used for 3D reconstruction with random selection. Then, we can further examine what the minimum amount of data needed for a near-optimal performance (optimal F-score $\pm 5\%$) for 3D reconstruction per area is. The results are presented in Fig. 4, where it takes 400 perspective images or 100 raw ERP from 360° camera to generate a near-optimal 3D mesh. Specifically, with a 360° camera, capturing 3.34 ERP frames$/m^2$ provides sufficient image data for near-optimal 3D reconstruction quality, given our test environment size of $4.4m \times 6.8m = 29.92m^2$.

C. Impact of Camera Man Removal on 3D Mesh Quality

This part of the experiment aims to investigate the impact of removing the camera man from 360° camera images on the quality of 3D models. To achieve this, we implemented a simple approach where the camera man was fixed at the i-th image for every four images (a certain view at the horizontal cube map faces), and every i-th image was removed during the filtering process. Our results indicate that removing the camera man can indeed improve the quality of 3D models, but only when there are fewer frames available. Specifically, we observed that the filtered version performs better when the amount of available data is limited, as the interference of

the camera man is removed. However, as more data on the environment becomes available, the 3D mesh without filtering performs better, presumably due to the loss of information caused by the removal of frames. The final results are presented in Fig. 4.

Overall, our experiment provides valuable insights into the performance of 360° cameras in 3D reconstruction and highlights the importance of considering factors such as data amount and camera artifacts in achieving accurate and reliable results.

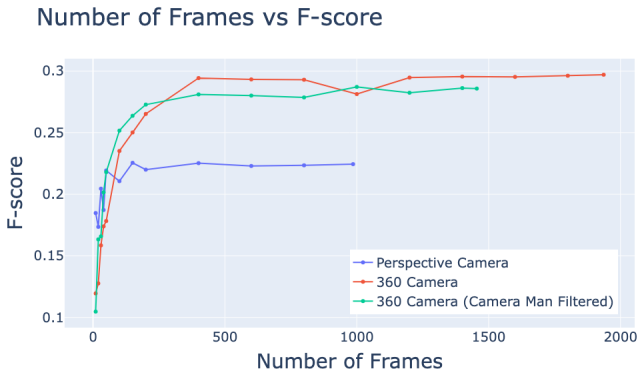

Fig. 4. Comparing F-score between perspective camera, 360° camera and 360° camera with camera man filtered, when different number of frames are applied.

V. Discussion

Our findings demonstrate the potential of using 360° cameras for achieving high-quality 3D reconstructions, opening up new possibilities for various consumer-oriented applications in multiple fields. Despite the promising results, there are still some limitations to our approach. First, our method relies on the conversion of ERP into cube maps, which may introduce artifacts and inaccuracies in the final 3D model. Future research could explore alternative representations or develop models that can directly handle equirectangular images. Second, our approach is currently limited to indoor environments, and its performance in outdoor or dynamic scenes remains to be investigated. Lastly, our method does not explicitly handle occlusions or reflections, which could affect the quality of the 3D reconstruction.

More broadly, this work helps highlight the potential of consumer-level 360° cameras for 3D modeling applications. While the proposed pipeline still relies on model architectures that are designed for perspective images, it still demonstrates the ability to densely sample scenes with a single moving 360 camera. To fully realize this potential, continued research on reconstruction algorithms designed specifically for 360° imagery and public datasets capturing diverse environments will be invaluable.

VI. Conclusion

In this paper, we presented a novel approach for 3D reconstruction using a single 360° camera. Our method addresses the challenges associated with ERP by converting them into four perspective views resembling the horizontal views of a cube map, allowing compatibility with deep learning models trained on undistorted perspective images. We also employed the state-of-the-art VSLAM technique for camera pose estimation and a proven 3D reconstruction method for generating detailed 3D mesh representations of the indoor environment. Through experimental evaluation, we compared the performance of 360° cameras with traditional perspective cameras in 3D reconstruction and analyzed the accuracy and performance of our vision-based approach. In conclusion, this paper offers a simplified and accessible solution for various consumer-oriented applications and contributes to the growing body of research on 3D reconstruction using 360° cameras.

VII. Acknowledgment

We extend our gratitude to Prof. Ling Shi (The Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology) for his expertise and assistance throughout all aspects of our study.

References

[1] Y. Tao, M. Popović, Y. Wang, S. T. Digumarti, N. Chebrolu, and M. Fallon, "3D Lidar Reconstruction with Probabilistic Depth Completion for Robotic Navigation," in 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 5339-5346, 2022.

[2] F. Gherardini, M. Santachiara, and F. Leali, "3D virtual reconstruction and augmented reality visualization of damaged stone sculptures," in IOP Conference Series: Materials Science and Engineering, vol. 364, p. 012018, 2018.

[3] S. González Izard, R. Sánchez Torres, O. Alonso Plaza, J. A. Juanes Mendez, and F. J. García-Peñalvo, "Nextmed: automatic imaging segmentation, 3D reconstruction, and 3D model visualization platform using augmented and virtual reality," Sensors, vol. 20, no. 10, pp. 2962, 2020.

[4] B. Wang, Q. Wang, J. C. Cheng, C. Song, and C. Yin, "Vision-assisted BIM reconstruction from 3D LiDAR point clouds for MEP scenes," Automation in Construction, vol. 133, pp. 103997, 2022.

[5] J. Mahmud, T. Price, A. Bapat, and J.-M. Frahm, "Boundary-aware 3D building reconstruction from a single overhead image," in CVPR, 2020, pp. 441-451.

[6] R. Ren, H. Fu, H. Xue, Z. Sun, K. Ding, and P. Wang, "Towards a fully automated 3D reconstruction system based on LiDAR and GNSS in challenging scenarios," Remote Sensing, vol. 13, no. 10, p. 1981, 2021.

[7] Y. Huang, W. Zheng, Y. Zhang, J. Zhou, and J. Lu, "Tri-perspective view for vision-based 3D semantic occupancy prediction," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 9223-9232.

[8] X. Yan, J. Gao, J. Li, R. Zhang, Z. Li, R. Huang, and S. Cui, "Sparse single sweep LiDAR point cloud segmentation via learning contextual shape priors from scene completion," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 4, pp. 3101-3109, 2021.

[9] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. Seitz, and R. Szeliski, "Building Rome in a day," in ICCV, 2009.

[10] J. M. Frahm, P. Fite-Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y. H. Jen, E. Dunn, B. Clipp, S. Lazebnik, and M. Pollefeys, "Building Rome on a Cloudless Day," in ECCV, 2010.

[11] S. B. Kang and R. Szeliski, "3-D Scene Data Recovery Using Omnidirectional Multibaseline Stereo," Int. J. Comput. Vis., vol. 25, no. 2, pp. 167-183, 1997.

[12] J. L. Schonberger, E. Zheng, M. Pollefeys, and J.-M. Frahm, "Pixelwise view selection for unstructured multi-view stereo," in ECCV, 2016.

[13] C. Forster, M. Pizzoli, and D. Scaramuzza, "SVO: Fast semi-direct monocular visual odometry," in 2014 IEEE International Conference on Robotics and Automation (ICRA), 2014, pp. 15-22.

[14] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," in ECCV, 2014, pp. 834-849.
[15] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 3, pp. 611-625, 2017.
[16] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: a versatile and accurate monocular SLAM system," IEEE Transactions on Robotics, vol. 31, no. 5, pp. 1147-1163, 2015.
[17] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras," IEEE Transactions on Robotics, vol. 33, no. 5, pp. 1255-1262, 2017.
[18] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, "ORB-SLAM3: An accurate open-source library for visual, visual–inertial, and multimap SLAM," IEEE Transactions on Robotics, vol. 37, no. 6, pp. 1874-1890, 2021.
[19] D. Caruso, J. Engel, and D. Cremers, "Large-scale direct SLAM for omnidirectional cameras," in 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 141-148, 2015.
[20] T. Qin, S. Cao, J. Pan, and S. Shen, "A general optimization-based framework for global pose estimation with multiple sensors," arXiv preprint arXiv:1901.03642, 2019.
[21] T. Shan, B. Englot, D. Meyers, W. Wang, C. Ratti, and D. Rus, "Lio-sam: Tightly-coupled lidar inertial odometry via smoothing and mapping," in 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 5135-5142, 2020.
[22] S. Sumikura, M. Shibuya, and K. Sakurada, "OpenVSLAM: A versatile visual SLAM framework," in Proceedings of the 27th ACM International Conference on Multimedia, pp. 2292-2295, 2019.
[23] B. Curless and M. Levoy, "A volumetric method for building complex models from range images," in Proceedings of the 23rd annual conference on Computer graphics and interactive techniques, pp. 303-312, 1996.
[24] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "Mvsnet: Depth inference for unstructured multi-view stereo," in Proceedings of the European conference on computer vision (ECCV), pp. 767-783, 2018.
[25] R. Chen, S. Han, J. Xu, and H. Su, "Point-based multi-view stereo network," in Proceedings of the IEEE International Conference on Computer Vision, pp. 1538-1547, 2019.
[26] P. H. Huang, K. Matzen, J. Kopf, N. Ahuja, and J. B. Huang, "Deepmvs: Learning multi-view stereopsis," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2821-2830, 2018.
[27] C. Hane, L. Heng, G. H. Lee, A. Sizov, and M. Pollefeys, "Real-time direct dense matching on fisheye images using plane-sweeping stereo," in 2014 2nd International Conference on 3D Vision, vol. 1, pp. 57-64, 2014.
[28] N. H. Wang, B. Solarte, Y. H. Tsai, W. C. Chiu, and M. Sun, "360SD-Net: 360° stereo depth estimation with learnable cost volume," in Proc. International Conference on Robotics and Automation (ICRA), 2020.
[29] F. E. Wang, Y. H. Yeh, M. Sun, W. C. Chiu, and Y. H. Tsai, "BiFuse: Monocular 360 depth estimation via bi-projection fusion," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.
[30] C. Y. Chiu, Y. T. Wu, I. Shen, and Y. Y. Chuang, "360MVSNet: Deep Multi-View Stereo Network With 360deg Images for Indoor Scene Reconstruction," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3057-3066, 2023.
[31] Z. Murez, T. van As, J. Bartolozzi, A. Sinha, V. Badrinarayanan, and A. Rabinovich, "Atlas: End-to-end 3D scene reconstruction from posed images," in ECCV, pp. 414-431, 2020.
[32] J. Sun, Y. Xie, L. Chen, X. Zhou, and H. Bao, "Neuralrecon: Real-time coherent 3D reconstruction from monocular video," in CVPR, pp. 15598-15607, 2021.
[33] Z. Hong and C. P. Yue, "Cross-Dimensional Refined Learning for Real-Time 3D Visual Perception from Monocular Video," in ICCVW, 2023.