2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

MRASfM: Multi-Camera Reconstruction and Aggregation through Structure-from-Motion in Driving Scenes

Lingfeng Xuan[†1], Chang Nie[†1], Yiqing Xu[2], Yanzi Miao[2], and Hesheng Wang[1]

***Abstract*— Structure from Motion (SfM) estimates camera poses and reconstructs point clouds, forming a foundation for various tasks. However, applying SfM to driving scenes captured by multi-camera systems presents significant difficulties, including unreliable pose estimation, excessive outliers in road surface reconstruction, and low reconstruction efficiency. To address these limitations, we propose a Multi-camera Reconstruction and Aggregation Structure-from-Motion (MRASfM) framework specifically designed for driving scenes. MRASfM enhances the reliability of camera pose estimation by leveraging the fixed spatial relationships within the multi-camera system during the registration process. To improve the quality of road surface reconstruction, our framework employs a plane model to effectively remove erroneous points from the triangulated road surface. Moreover, treating the multi-camera set as a single unit in Bundle Adjustment (BA) helps reduce optimization variables to boost efficiency. In addition, MRASfM achieves multi-scene aggregation through scene association and assembly modules in a coarse-to-fine fashion. We deployed multi-camera systems on actual vehicles to validate the generalizability of MRASfM across various scenes and its robustness in challenging conditions through real-world applications. Furthermore, large-scale validation results on public datasets show the state-of-the-art performance of MRASfM, achieving 0.124 absolute pose error on the nuScenes dataset. The code is available at `https://github.com/IRMVLab/MRASfM`.**

I. Introduction

Image-based driving scene reconstruction is commonly achieved through visual Simultaneous Localization and Mapping (vSLAM) and Structure from Motion (SfM). However, for critical downstream tasks like high-definition (HD) mapping construction and novel view synthesis [1], [2], [3], [4], the limitations of vSLAM become apparent. Specifically, its reliance on incremental, locally optimized estimations impedes global refinement, leading to accumulated drift and reduced accuracy. In contrast, SfM leverages batch processing and global Bundle Adjustment (BA), thereby yielding more accurate scene point clouds and camera trajectories. This offline nature allows SfM to overcome the accuracy trade-offs inherent in real-time vSLAM, making it more suitable for tasks requiring high precision and consistency. However, the application of SfM to driving scenes captured by multi-camera systems encounters notable obstacles, including unreliable pose estimation, excessive outliers in road surface reconstruction, and low reconstruction efficiency.

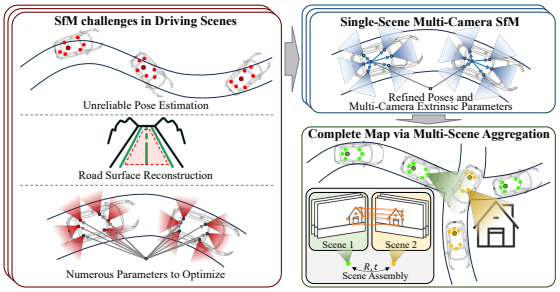


Fig. 1: **The idea of MRASfM.** To address the challenges of SfM in driving scenes, MRASfM incorporates camera set priors and semantic information into the reconstruction framework, achieving efficient and robust scene reconstruction. Through multi-scene aggregation, fragmented scenes are integrated into a complete and consistent map.

The reliability of camera pose estimation in driving scenes is inherently challenged by the characteristics of these environments. Traditional SfM typically captures images around the object, while it is difficult to implement in driving scenes. Moreover, repetitive patterns and dynamic objects also decrease the quality of correspondence search, making pose estimation unreliable in driving scenes.

Road surface reconstruction is another significant challenge for SfM in driving scenes, which is crucial for downstream tasks [5], [6]. The edges of vehicle shadows are often detected as feature points. When the vehicle shadows move, the corresponding feature points become dynamic points, introducing noise into road surface reconstruction. Moreover, the lack of texture also causes outliers.

Efficiency remains a critical bottleneck in SfM, particularly in multi-camera systems. While multi-camera setups offer a broader field of view (FOV) and the ability to capture more comprehensive environmental data, they also introduce complexities. Unlike single-camera systems, which may fail entirely when obstructed, multi-camera systems can leverage alternative perspectives to maintain functionality. However, more cameras also bring more poses to be optimized in BA, thereby diminishing reconstruction efficiency.

To overcome these challenges, we propose a Multi-camera Reconstruction and Aggregation Structure-from-Motion (MRASfM) framework for driving scenes, as shown in Fig. 1. For unreliable pose estimation, our method in-

[†]The first two authors contributed equally.

This work was supported by National Key R&D Program of China (Grant No. 2024YFB4708900). It was also supported in part by the Natural Science Foundation of China under Grant 62225309, U24A20278, 62361166632. (Corresponding Author: Hesheng Wang)

[1]Authors are with the School of Automation and Intelligent Sensing, Shanghai Jiao Tong University and Shanghai Key Laboratory of Navigation and Location Based Services, Shanghai, 200240, China. (Emails: {lingfengxuan, changnie, wanghesheng}@sjtu.edu.cn)

[2]Authors are with the Advanced Robotics Research Center, Artificial Intelligence Research Institute and School of Information and Control Engineering, China University of Mining and Technology, Xuzhou, 221116, China. (Emails: {ts22060094a31, myz}@cumt.edu.cn)

corporates a learning-based feature extractor to improve feature extraction. Moreover, prior camera poses are utilized to identify image pairs with significant visual overlap for matching, thereby enhancing the efficiency and robustness of matching. During image registration, we initially estimate the camera poses of images with rich correspondences. Leveraging the fixed spatial relationships of the multi-camera system, we can then robustly register even those images that are partially occluded. To mitigate excessive outliers in road surface reconstruction, we apply plane fitting to filter erroneous points and enhance the consistency of the reconstructed road surface. To improve the reconstruction efficiency, we enhance BA by treating the camera set as a unified unit. In this way, optimization results are more consistent, accelerating the convergence of subsequent BA. In addition, the reduction of optimization variables also boosts the efficiency of reconstruction.

In practical engineering, scenes are often captured over short durations with limited overlap [7], [8]. However, generating a comprehensive scene understanding requires the integration of these fragmented reconstructions into a unified map. Traditional SfM aggregation modules typically require shared images between different segments [9], [10], which are often unavailable in practical applications. For such disjointed reconstructed scenes, we first use reconstruction results and GNSS to associate nearby scenes. Then, images with large visual overlap are identified for matching and fine assembly. Finally, an SfM-based optimization refines the transformation matrix between associated scenes for fine assembly.

In summary, the main contributions of our work are:

- We propose MRASfM, a novel framework for multi-camera reconstruction, which overcomes the limitations of traditional SfM in pose estimation robustness and computational efficiency in driving scenes.
- MRASfM introduces a camera set registration module, enhancing pose estimation robustness. Moreover, the camera set BA module greatly improves efficiency and achieves online calibration. For fragmented scenes, the multi-scene aggregation module seamlessly binds them together in a coarse-to-fine fashion.
- Experimental results on real-world applications demonstrate the generalizability of MRASfM across various environments and its robustness under challenging conditions. Moreover, experiments on public datasets highlight the state-of-the-art performance of MRASfM.

II. Related Work

A. Multi-camera-Based Reconstruction

Multi-camera systems offer enhanced perception capabilities for driving scene reconstruction. Traditional vSLAM-based multi-camera reconstruction methods typically require precise internal relative poses and intrinsic parameters of the cameras as input. Systems like BAMF-SLAM [11] and MAVIS [12] exemplify this reliance, requiring pre-calibration of their multi-camera systems for efficient localization and mapping. Moreover, these vSLAM methods struggle to be compatible with different multi-camera configurations, restricting their flexibility in practical deployments.

In the field of SfM, COLMAP [13] ignores camera set priors during incremental reconstruction. Instead, it applies these priors during Rigid Bundle Adjustment (RIG BA) after the reconstruction is complete. However, RIG BA is unable to rectify errors as they accumulate during the incremental reconstruction. MMA [14], a global SfM method, seeks to improve accuracy and robustness by explicitly leveraging fixed relative camera poses. Despite its global nature, MMA is sensitive to feature match outliers and often degenerates when the estimated relative translations are collinear. MGSfM [15] achieves effective utilization of multi-camera constraints through decoupled rotation averaging and hybrid translation averaging. Chen *et al.*[16] presented an efficient odometry-guided SfM approach that directly uses odometry poses and camera extrinsics for image registration. This method, however, demands highly precise extrinsic calibration and accurate odometry measurements, requirements that can be challenging to consistently meet in real-world driving scenes. MCSfM [17] is a multi-camera-based incremental SfM system that does not require prior internal relative poses and intrinsic parameters. It allows for simultaneous camera set calibration during the reconstruction process. Inspired by these works, we treat the camera set as the fundamental unit for registration and refinement, thus effectively integrating multi-camera systems into SfM and achieving efficient and robust reconstruction.

B. Incremental SfM

Incremental SfM [13], [18], [19] begins by establishing connections between images through correspondence searches. Next, an initial model is built using select seed images. Then, SfM iteratively performs image registration, triangulation, refinement, and outlier removal until all images are integrated into the system.

In correspondence search, traditional SIFT feature descriptors [20] generate limited feature points in low-texture images. In contrast, learning-based methods [21], [22], [23], [24] can detect abundant feature points in challenging environments. Methods[16], [7], [25] all utilized learning-based correspondence search methods.

In incremental reconstruction, efforts to improve the refinement step have been key to enhancing effectiveness. In the classic COLMAP system [13], bundle adjustment (BA) optimizes camera poses, intrinsic parameters, and the positions of scene points by minimizing reprojection error. Weber *et al.*[26] connected the bundle adjustment problem to power series theory and applied inverse expansion methods to achieve efficient large-scale BA. Zheng *et al.*[27] developed a distributed bundle adjustment approach using the exact Levenberg-Marquardt algorithm for extremely large datasets. Lindenberger *et al.*[25] incorporated learned image features as optimization variables, eliminating errors from feature extraction which is difficult to address in traditional BA. We add rigid multi-camera set constraints to BA, ensuring that the optimized camera poses maintain a consistent relative

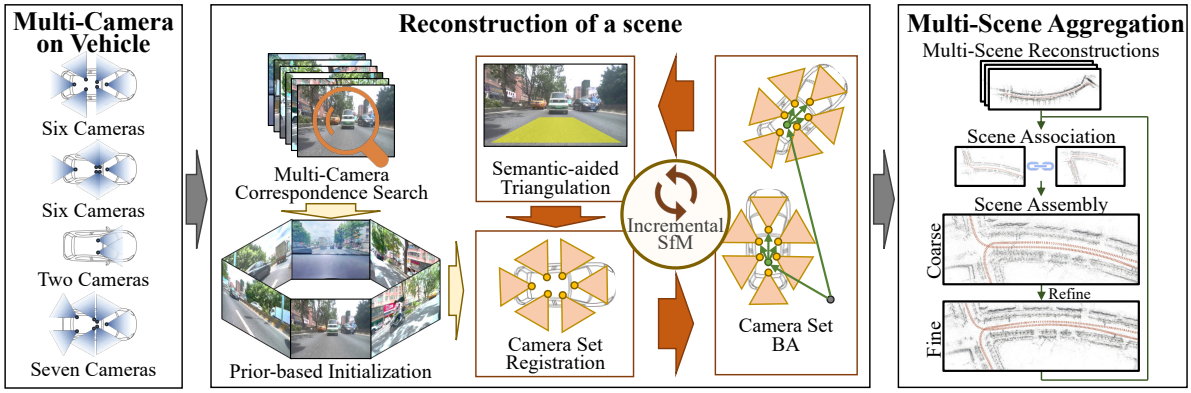


Fig. 2: **The pipeline of MRASfM.** MRASfM takes multi-camera images, semantic information, rough trajectories, and calibrations as input. When reconstructing a scene, MRASfM first utilizes prior information to perform correspondence search and initialization (see section III-C). In camera set registration, MRASfM robustly registers new images with rigid unit constraints (see section III-D). During triangulation, MRASfM improves road surface reconstruction quality using semantic information for fewer outliers (see section III-E). In camera set BA, MRASfM refines the reconstruction using rigid units as optimization variables, enhancing efficiency and robustness (see section III-F). When aggregating multiple scenes, MRASfM assembles nearby scenes with SfM in a coarse-to-fine fashion (see section III-G).

relationship. This can accelerate the convergence of BA and enhance its robustness.

C. Driving Scene Reconstruction and Aggregation

In practical engineering, driving data is often collected in a fragmented manner. For driving scenes, MCSfM [17] realizes scale-free driving scene reconstruction based on multi-camera systems. Aziza *et al.*[28] evaluate and compare the applicability and limitations of open-source SfM algorithms for mapping urban scenes. For aggregating fragmented data, Chen *et al.*[29] propose a graph-based scene merging algorithm, which constructs a minimum spanning tree to find accurate similarity transformations and a minimum height tree to avoid error accumulation. Merge-SfM [30] solves the problem of finding overlapping regions and the 7-DOF transformation between multiple reconstructions. We first associate nearby scenes. Then, the scene assembly binds them together in a coarse-to-fine fashion.

III. Methodology

A. Problem Setting

Given a set of RGB images $I_1, \ldots, I_N$, where N is the number of images and $I_i \in \mathbb{R}^{3\times H\times D}$, SfM estimates their corresponding intrinsic parameters $\mathbf{K}_i$, camera poses$\{\mathbf{R}_i, \mathbf{t}_i\}$ and the 3D scene represented by a point cloud$\{\mathbf{X}_1, \ldots, \mathbf{X}_M\}$, where M is the number of scene points. 3D scene points $\mathbf{X}_j$ can be projected as 2D points $\mathbf{x}'_{ij}$ onto the pixel plane of image I_i:

$$\mathbf{x}'_{ij} = \pi(\mathbf{K}_i, \mathbf{R}_i, \mathbf{t}_i, \mathbf{X}_j), \tag{1}$$

where the notation $\pi(\cdot)$ is the projection function.

In MRASfM, multiple images captured at the same time construct a rigid unit. The pose of a rigid unit is defined as the pose of the vehicle at that moment. Let U_{I_i} be the rigid unit that image I_i belongs to, *i.e.*, $I_i \in U_{I_i}$, and $\{\mathbf{R}_{U_{I_i}}, \mathbf{t}_{U_{I_i}}\}$ be the pose of rigid unit U_{I_i}. Let $\{\mathbf{R}_i^{rel}, \mathbf{t}_i^{rel}\}$ be the internal relative pose from image I_i to its rigid unit U_{I_i}. The relationship between the camera pose of image I_i and the pose of its corresponding rigid unit U_{I_i} is as follows:

$$\mathbf{R}_i^{rel} = \mathbf{R}_i \mathbf{R}_{U_{I_i}}{}^{\mathrm{T}}, \tag{2}$$

$$\mathbf{t}_i^{rel} = \mathbf{R}_{U_{I_i}}(\mathbf{t}_i - \mathbf{t}_{U_{I_i}}). \tag{3}$$

B. System Framework

The pipeline of our SfM system is illustrated in Fig. 2. MRASfM takes multi-camera images, semantic information, rough trajectories and calibrations as input.

In single-scene reconstruction, MRASfM first performs multi-camera correspondence search, which involves feature extraction, feature matching, and geometric validation. Then, incremental reconstruction begins with prior-based initialization, reconstructing images from selected rigid units with the aid of prior information. Subsequently, MRASfM iteratively integrates each rigid unit into the reconstruction. Each iteration step includes **(a)** Camera Set Registration: MRASfM robustly registers new images using rigid unit constraints; **(b)** Semantic-aided Triangulation: MRASfM triangulates scene points and removes outliers in the road surface area; **(c)** Camera Set BA: MRASfM refines the reconstruction with rigid units as optimization variables.

In multi-scene aggregation, MRASfM first associates scenes together using GNSS locations and selects two suitable scenes for aggregation. They are coarsely assembled using reconstruction results. Then, an SfM-based optimization refines the transformation matrix between candidate scenes iteratively. The fine assembly result serves as the new reference scene for the next iteration. The process will repeat until all scenes are integrated.

C. Correspondence Search and Initialization

In correspondence search, feature points are detected by the Superpoint [21] model. With the known rough internal relative poses and vehicle poses, MRASfM calculates the approximate field of view for each image. Then, images with more visual overlap are selected for matching. By filtering image pairs, both the efficiency and accuracy of the correspondence search are improved. The Superglue model [24] is used for feature matching, and matched point pairs

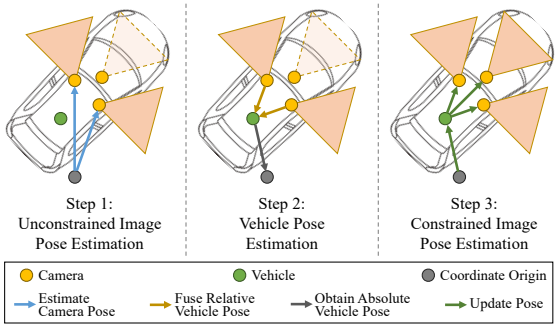


Fig. 3: **The pipeline of Camera Set Registration:** camera pose estimation via PnP solving, multi-view fused vehicle pose derivation, and camera pose updating using intra-unit constraints and vehicle poses.

of different semantic categories are removed. Outliers of matching are then filtered via geometric verification [31].

After the multi-camera correspondence search, initial rigid units used for reconstruction are selected based on the total number of correspondences within the unit. MRASfM directly registers images within initial rigid units using prior poses. Then, the initial model is constructed through triangulation and BA.

D. Camera Set Registration

Despite the refinement in correspondence search, occlusions in driving scenes can still lead to errors in pose estimation. Camera set registration combines rigid constraints from camera sets and reliable estimation from correspondence-rich images to address this issue, as shown in Fig. 3.

In unconstrained image pose estimation, MRASfM first organizes images using the next best view selection module in COLMAP [13]. The camera poses of the highest-ranked images are determined by solving the Perspective-n-Point (PnP) problem [32] using feature correspondences. Let $\{\mathbf{R}_k, \mathbf{t}_k\}$ $(k = 1, \ldots, K)$ be the newly estimated camera poses of image I_k, where K is the number of estimations. According to Eq. 2 and Eq. 3, these estimations can be used to calculate the poses of the rigid units where the aforementioned images belong:

$$\mathbf{R}_{U_{I_k}} = {\mathbf{R}_k^{rel}}^{\mathrm{T}} \mathbf{R}_k, \tag{4}$$

$$\mathbf{t}_{U_{I_k}} = \mathbf{t}_k - {\mathbf{R}_k^{rel}}^{\mathrm{T}} \mathbf{t}_k^{rel}. \tag{5}$$

In vehicle pose estimation, for each rigid unit that has newly calibrated images, MRASfM uses the internal relative poses to infer its pose. For rigid unit U_j, the rotation of U_j is computed by the local rotation averaging:

$$\mathbf{R}_{U_j} = \min_{\mathbf{R}_{U_j}} \sum \rho_r(\|\mathbf{R}_{U_j} \mathbf{R}_{U_{I_k}}\|), I_k \in U_j, \tag{6}$$

where the notation $\rho(\cdot)$ is a robust loss function. The translation of U_j is computed by the local translation averaging:

$$\mathbf{t}_{U_j} = \min_{\mathbf{t}_{U_j}} \sum \rho_t(\|\mathbf{t}_{U_j} - \mathbf{t}_{U_{I_k}}\|), I_k \in U_j, \tag{7}$$

In constrained image pose estimation, the camera poses of all images belonging to U_j can be obtained through the following formula:

$$\begin{aligned} \mathbf{R}_i &= \mathbf{R}_i^{rel} \mathbf{R}_{U_j}, I_i \in U_j, \\ \mathbf{t}_i &= {\mathbf{R}_{U_j}}^{\mathrm{T}} \mathbf{t}_i^{rel} + \mathbf{t}_{U_j}, I_i \in U_j. \end{aligned} \tag{8}$$

where image I_i belongs to the rigid unit U_j.

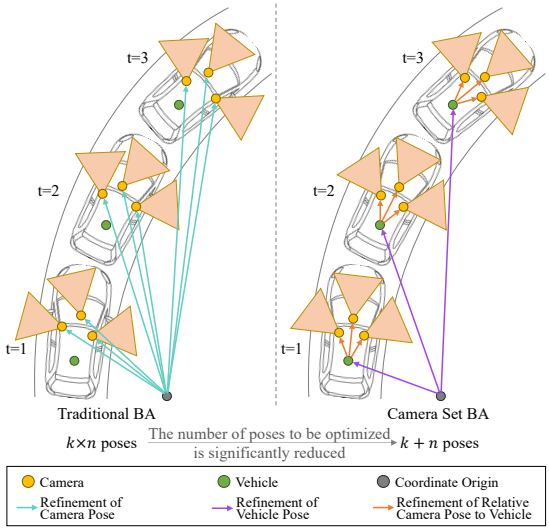


Fig. 4: **Comparison between traditional BA and Camera Set BA.** Traditional BA optimizes the absolute camera pose of each image individually. Camera Set BA, by contrast, optimizes vehicle poses and internal relative poses, significantly reducing optimization variables and enhancing robustness.

In this way, although some images are difficult to register in driving scenes, MRASfM can still robustly estimate their camera poses in camera set registration.

E. Semantic-aided Triangulation

After registering a new rigid unit, MRASfM triangulates 3D points but faces challenges with dynamic feature points from vehicle shadows and insufficient road texture, leading to significant triangulation noise for the road surface. To address this, the Locally Optimized Random Sample Consensus (LO-RANSAC) method is applied to filter outliers by fitting a plane model to the 3D road points, thereby improving road surface reconstruction quality.

F. Camera Set BA

In incremental SfM, BA optimizes registration and triangulation results. Traditional BA optimizes each image individually, which can lead to inconsistent internal relative poses across frames. To address this, we introduce the Camera Set Bundle Adjustment (CSBA) module to enhance consistency. CSBA treats the camera set as a unit, optimizing vehicle poses and internal relative poses. Assuming that the multi-camera system comprises k cameras and images from n different timestamps are integrated into BA, traditional BA optimizes $k \times n$ poses, while CSBA only requires $k + n$. As the number of frames increases, CSBA significantly reduces the number of optimizing poses compared to traditional BA, conserving computational resources. The comparison between traditional BA and CSBA is shown in Fig. 4.

To enhance system efficiency, the CSBA module is divided into Local CSBA and Global CSBA. The system activates Global CSBA when there are a large number of unoptimized rigid units; otherwise, it executes Local CSBA. The key difference is their optimization scope: Local CSBA optimizes only newly registered and their connected rigid units, while Global CSBA optimizes all registered units.

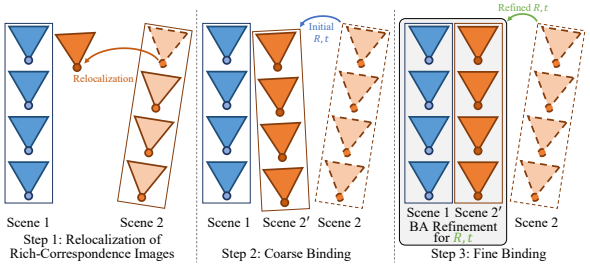


Fig. 5: **Pipeline of Scene Assembly.** MRASfM begins by selecting images with rich correspondences for relocalization. By comparing the original and updated camera poses, the transformation matrix is initialized, which is then optimized by BA. The refined matrix is used to register remaining rigid units into reconstruction. Registration and refinement alternate until all rigid units are successfully integrated.

According to Section III-D, the camera pose of an image can be derived from its corresponding rigid unit pose and internal relative pose. Therefore, the parameters to be optimized include intrinsic camera parameters, vehicle poses, internal relative poses and scene points. In local CSBA, intrinsic camera parameters and internal relative poses are fixed . Let $\{\mathbf{R}_l, \mathbf{t}_l\}$ $(l = 1, \ldots, L)$ be the camera poses of images integrated in local CSBA, which can be defined as:

$$\min_{\mathbf{R}_{U_{I_l}}, \mathbf{t}_{U_{I_l}}, \mathbf{X}_j} \sum \rho(\|\mathbf{x}_{lj} - \pi(\mathbf{K}_l, \mathbf{R}_l, \mathbf{t}_l, \mathbf{X}_j)\|^2) \quad \text{s.t.} \quad \mathbf{R}_l = \mathbf{R}_l^{rel}\mathbf{R}_{U_{I_l}}; \quad \mathbf{t}_l = \mathbf{R}_{U_{I_l}}{}^{\mathrm{T}}\mathbf{t}_l^{rel} + \mathbf{t}_{U_{I_l}}. \tag{9}$$

where the notation $\pi(.)$ is the projection function in Eq. 1, $\rho(\cdot)$ is the Cauchy loss function, $\mathbf{x}_{lj}$ are the 2D points in image I_l corresponding to the 3D scene point X_j and K_l is the intrinsic parameter matrix of image I_l.

In Global CSBA, all aforementioned parameters are optimized. Let $\{\mathbf{R}_g, \mathbf{t}_g\}$ $(g = 1, \ldots, G)$ be poses of images to be optimized by global CSBA, which can be defined as:

$$\min_{\substack{\mathbf{K}_g, \mathbf{R}_{U_{I_g}}, \mathbf{t}_{U_{I_g}}, \\ \mathbf{R}_g^{rel}, \mathbf{t}_g^{rel}, \mathbf{X}_j}} \sum \rho(\|\mathbf{x}_{gj} - \pi(\mathbf{K}_g, \mathbf{R}_g, \mathbf{t}_g, \mathbf{X}_j)\|^2) \quad \text{s.t.} \quad \mathbf{R}_g = \mathbf{R}_g^{rel}\mathbf{R}_{U_{I_g}}; \quad \mathbf{t}_g = \mathbf{R}_{U_{I_g}}{}^{\mathrm{T}}\mathbf{t}_g^{rel} + \mathbf{t}_{U_{I_g}}. \tag{10}$$

By combining local CSBA and global CSBA, MRASfM can efficiently and robustly optimize the reconstruction process.

G. Multi-scene Aggregation

In practical engineering, scenes are frequently collected in a fragmented manner [7], [8], [5]. Each scene typically contains only a short segment of driving information, and there are no shared images between different scenes. Additionally, associated scenes may be collected at long intervals. To create a comprehensive and complete map, it is essential to seamlessly integrate multiple nearby scenes.

In the scene association module, MRASfM first employs GNSS to position reconstructed scenes within a global coordinate system. The locations of these scenes are represented by the midpoints of their trajectories and the scene located at the geometric center is considered the reference scene S_r. The scenes closest to the reference scene S_r (typically three in experiments) are identified as candidate scenes for aggregation. Using homography-guided spatial pairs (HSP) from CAMAv2 [16], MRASfM assesses the visual overlap between two scenes and selects the candidate with the highest overlap as the merge scene S_m.

In the scene assembly module, shown in Fig. 5, associated scenes S_r and S_m are first coarsely assembled within the global coordinate system. The camera pose of image I_i in the coarsely assembled scene S_{fine} is defined as $\mathbf{P}_i^{coarse}$. However, this assembly may differ from the actual situation by a transformation matrix $\mathbf{T}_{trans}$ due to GNSS errors. Therefore, an iterative optimization for the transformation matrix $\mathbf{T}_{trans}$ is conducted for fine assembly. To reduce computation time and memory usage, rigid units with large visual overlap are chosen for constructing SfM. During optimization, MRASfM first initializes SfM by preserving the camera poses of the images from scene S_r and triangulating them. Next, MRASfM selects the most suitable image I_i from scene S_m and performs unconstrained pose estimation (see Fig. 3) to obtain its pose $\mathbf{P}_i^{fine}$ in the refined assembled scene S_{fine}. The initial transformation matrix $\mathbf{T}_{trans}$ can be obtained through the following formula:

$$\mathbf{T}_{trans} = \mathbf{P}_i^{fine}\mathbf{P}_i^{coarse-1} \tag{11}$$

where $\mathbf{P}_i^{coarse}$ is the camera pose of image I_i from the coarsely assembled scene S_{coarse}.

Using transformation matrix $\mathbf{T}_{trans}$, we can register rigid unit U_{I_i}, which image I_i belongs to:

$$\mathbf{P}_{U_{I_i}}^{fine} = \mathbf{T}_{trans}\mathbf{P}_{U_{I_i}}^{coarse} \tag{12}$$

where $\mathbf{P}_{U_{I_i}}^{coarse}$ is the pose of rigid unit U_{I_i} in scene S_{coarse}, and $\mathbf{P}_{U_{I_i}}^{fine}$ is the pose of rigid unit U_{I_i} in scene S_{fine}. After triangulation outlined in Section III-E, $\mathbf{T}_{trans}$ and the scene points will be optimized by transformation-based CSBA. Camera poses of images from the reference scene S_r are fixed in BA.

Let $\{\mathbf{R}_m^{coarse}, \mathbf{t}_m^{coarse}\}$ $(m = 1, \ldots, M)$ be the camera poses of registered images from scene S_m. The transformation-based BA can be defined as:

$$\min_{\mathbf{T}_{trans}, \mathbf{X}_j} \sum \rho(\|\mathbf{x}_{mj} - \pi(\mathbf{K}_m, \mathbf{R}_m, \mathbf{t}_m, \mathbf{X}_j)\|^2) \quad \text{s.t.} \quad \mathbf{R}_m = \mathbf{R}_{\text{trans}}\mathbf{R}_m^{\text{coarse}}; \mathbf{t}_m = \mathbf{R}_{\text{trans}}\mathbf{t}_m^{\text{coarse}} + \mathbf{t}_{\text{trans}}. \tag{13}$$

where $\mathbf{R}_{trans}$ and $\mathbf{t}_{trans}$ are the rotation matrix and the translation vector decomposed from $\mathbf{T}_{trans}$. Camera set registration, triangulation, and transformation-based BA will iteratively take place until all rigid units in scene S_m are integrated into scene S_{fine}. Thereby scene S_r and scene S_m are seamlessly aggregated as scene S_{fine}, which serves as the new reference scene for the remaining ones. In a circular manner, all scenes are integrated into a complete map through scene association and scene assembly.

IV. Real-World Applications

We evaluate MRASfM using datasets from two multi-camera systems: a six-camera and a seven-camera surround-view setup (see Fig. 6 (a)). These experimental platforms operate at speeds of 10 to 60 km/h, capturing high-resolution (1920 × 1080) surround-view imagery at 30 Hz, with simultaneous GNSS data recorded. All the experiments were conducted on a computer equipped with a 3.4 GHz CPU.

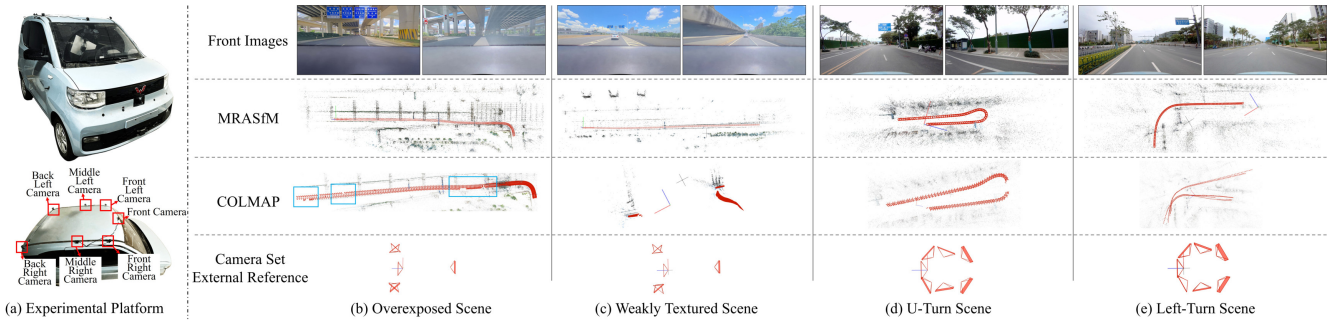
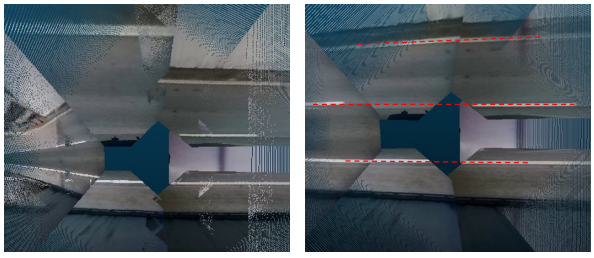


Fig. 6: **Reconstruction results of self-collected datasets.** The calibrated internal relative poses of our system are shown in the last row. The camera poses are shown in red.



(a) Original calibration (b) Refined calibration

Fig. 7: **BEV perspectives generated with calibrations.** The refined calibrations generate more consistent line markings in BEV perspectives.

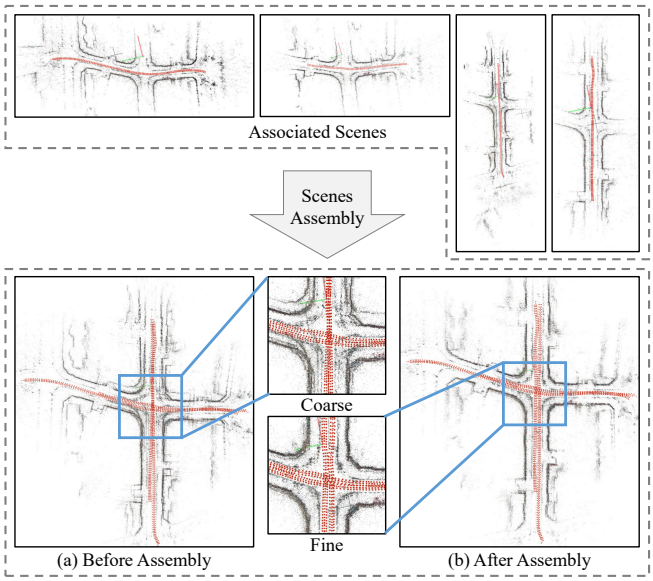


Fig. 8: **Results of Multi-Scene Aggregation.** The associated scenes are integrated into a complete and consistent scene through the multi-scene aggregation module.

Sample reconstructed scenes and camera poses are shown in Fig. 6. Scene (b) and (c) were captured with a six-camera system in an urban driving environment characterized by weak textures and slope changes, and were successfully reconstructed by MRASfM. Scene (d) and (e), captured with a seven-camera system, further demonstrate MRASfM's robustness in complex U-turn and left-turn scenarios. Compared to COLMAP [13], MRASfM shows superior generalizability and robustness under challenging conditions. This improved performance is primarily attributed to accurate correspondence search and effective integration of rigid multi-camera constraints.

A key feature of MRASfM is its ability to mitigate calibration inaccuracies. Due to the difficulty of vehicle calibration, initial calibration inherently contains errors, which can worsen when driving on uneven terrain. The bird's-eye views (BEV) in Fig. 7, generated using calibration data, demonstrate improved lane marking alignment, visually confirming MRASfM's recalibration capability. By treating rigid camera sets as fundamental units within BA, MRASfM effectively corrects initial calibration errors, yielding a more accurate and consistent multi-camera system calibration.

For scenes collected in a fragmented manner, the Multi-Scene Aggregation module effectively integrates them into a cohesive whole. Fig. 8 illustrates the difference between coarse binding and fine binding. Due to initial GNSS errors, fragmented scenes cannot form a coherent map directly. However, after the iterative optimization of transformation matrices, a complete and consistent map is finally achieved.

V. Large-Scale Validation Results

A. *KITTI odometry*

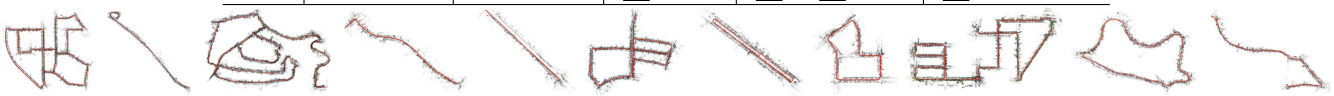
The KITTI odometry benchmark dataset is collected using a car equipped with a stereo camera, encompassing 11 sequences of urban driving scenes. The quantitative experiment results are presented in TABLE I. The median absolute pose error (APE) is calculated for evaluation. Our method achieves state-of-the-art performance in terms of pose estimation accuracy. This is mainly due to iterative refinement and effective utilization of prior information. Meanwhile, MRASfM significantly outperforms the previous state-of-the-art incremental SfM method, MCSfM[17], in terms of reconstruction efficiency. This improvement is largely attributed to our selection of high-quality matching image pairs during the correspondence search. Moreover, the accurate registration and triangulation accelerate the convergence of refinement. The qualitative experiment results shown in Fig. 9 demonstrate that MRASfM is able to achieve consistent reconstruction across various scenes

B. *NuScenes*

The nuScenes dataset is a public large-scale dataset for autonomous driving, which is collected using a vehicle

TABLE I: **Vehicle pose accuracy of KITTI odometry benchmark.** e_r is the median absolute rotation error (degrees); e_t is the median absolute translation error (meters); T is running time of pose estimation (minutes). The best results are shown in **bold**; the second best are underlined.

Data	COLMAP[13]		GLOMAP[33]		MGSfM[15]		MCSfM[17]			MRASfM		
Name	$e_r \downarrow$	$e_t \downarrow$	$e_r \downarrow$	$e_t \downarrow$	$e_r \downarrow$	$e_t \downarrow$	$e_r \downarrow$	$e_t \downarrow$	$T \downarrow$	$e_r \downarrow$	$e_t \downarrow$	$T \downarrow$
data00	0.4	0.9	0.4	0.8	0.4	0.5	**0.3**	0.5	286	0.5	**0.3**	192
data01	**0.2**	1.5	0.5	4.5	**0.2**	**0.6**	0.4	1.0	47	**0.2**	**0.6**	34
data02	0.5	4.0	0.4	5.4	0.4	0.9	**0.3**	1.0	355	0.4	**0.7**	276
data03	**0.1**	0.2	0.2	0.4	0.2	0.2	0.2	0.2	77	0.2	**0.1**	48
data04	**0.1**	**0.1**	**0.1**	0.2	**0.1**	**0.1**	**0.1**	**0.1**	4	**0.1**	**0.1**	3
data05	0.6	0.8	**0.2**	0.3	**0.2**	0.2	0.3	0.2	116	0.3	**0.1**	81
data06	0.2	0.2	**0.1**	0.3	**0.1**	**0.1**	**0.1**	0.2	34	**0.1**	**0.1**	28
data07	1.1	0.5	0.3	0.3	0.2	**0.2**	0.4	0.4	62	**0.1**	**0.2**	43
data08	0.7	6.1	0.7	3.1	0.4	0.9	0.4	1.2	276	**0.3**	**0.5**	188
data09	0.6	1.1	**0.3**	1.7	**0.3**	0.5	**0.3**	0.5	74	**0.3**	**0.2**	54
data10	0.8	2.5	**0.3**	0.7	0.4	0.6	0.4	0.4	53	0.4	**0.2**	38

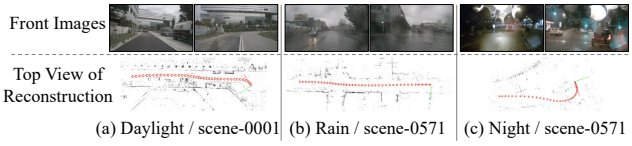


(a) data00 (b) data01 (c) data02 (d) data03 (e) data04 (f) data05 (g) data06 (h) data07 (i) data08 (j) data09 (k) data10

Fig. 9: **Reconstruction results of KITTI odometry benchmark produced by MRASfM.**

TABLE II: **Vehicle pose accuracy of nuScenes dataset.** e_t is the RMSE absolute translation error in meters. The best results are shown in bold.

Method	Input	$e_t \downarrow$
ORBSLAM3 [34]	Front camera	0.199
DROID-SLAM [35]	Front camera	0.282
OCC-VO [36]	Surround cameras	0.140
MRASfM	Surround cameras	**0.124**



(a) Daylight / scene-0001 (b) Rain / scene-0571 (c) Night / scene-0571

Fig. 10: **Reconstruction results on the nuScenes dataset.**

equipped with six surrounding cameras. The dataset covers over 1000 scenes, each lasting about 20 seconds, which spread across different countries, lighting settings, weather variations, and environments. The quantitative experiment results are presented in TABLE II. Our method also achieves state-of-the-art performance in terms of pose estimation accuracy. Furthermore, qualitative results in Fig. 10 showcase the capacity of MRASfM to consistently reconstruct scenes across the diverse and challenging environments within the nuScenes dataset. MRASfM demonstrates robust performance across both quantitative metrics and qualitative visual assessments. These accurate and robust results are primarily attributed to the improved correspondence research and the effective utilization of rigid unit constraints during the reconstruction process.

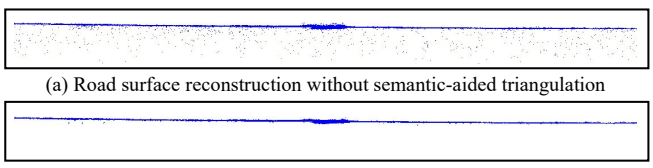
C. Ablation Study

To validate each module of MRASfM, reconstructions of KITTI odometry benchmark data00 and data01 under different conditions are presented in TABLE III.

The effects of camera set BA are demonstrated in TABLE III (a). The absence of CSBA leads to a noticeable decline in both efficiency and accuracy, primarily due to the increased number of optimization variables. Additionally, the lack of camera set constraints increases optimization errors and prolongs convergence times.

TABLE III: **Ablation study results of MRASfM on vehicle pose estimation.**

Exp.	Method	data00			data01		
		$e_r \downarrow$	$e_t \downarrow$	$T \downarrow$	$e_r \downarrow$	$e_t \downarrow$	$T \downarrow$
(a)	Ours (w/o camera set BA)	1.8	2.7	8720	0.7	1.0	534
	Ours (full, w/ camera set BA)	**0.5**	**0.3**	**192**	**0.2**	**0.6**	**34**
(b)	Ours (w/o camera set registration)	0.6	0.4	203	0.3	0.7	40
	Ours (full, w/ camera set registration)	**0.5**	**0.3**	**192**	**0.2**	**0.6**	**34**
(c)	Ours (w/o semantic-aided triangulation)	0.6	**0.3**	197	**0.2**	**0.6**	37
	Ours (full, w/ semantic-aided triangulation)	**0.5**	**0.3**	**192**	**0.2**	**0.6**	**34**



(a) Road surface reconstruction without semantic-aided triangulation

(b) Road surface reconstruction with semantic-aided triangulation

Fig. 11: **Comparison between triangulated road points.** The semantic-aided triangulation module effectively filters outliers of the road surface.

The impact of camera set registration is evaluated in TABLE III (b). While it has minimal effect on pose estimation accuracy and reconstruction efficiency, it is particularly useful for handling occluded viewpoints. In scenes with rich correspondences, its impact is less noticeable, though it still improves reconstruction by enhancing image registration.

The influences of semantic-aided triangulation are examined in TABLE III (c). Semantic-aided triangulation improves reconstruction accuracy and accelerates optimization convergence by removing outliers. Although its impact on pose estimation and efficiency is minimal, its effectiveness is evident in the enhanced road surface reconstruction shown in Fig. 11.

VI. Conclusion

In this work, we propose MRASfM, a novel multi-camera reconstruction framework for driving scene reconstruction. Our framework uses camera sets as atomic units for registration and refinement, enabling robust pose estimation in complex environments and higher efficiency than conventional

per-camera methods. During triangulation, the quality of road surface reconstruction is greatly enhanced using semantic information. For fragmented scenes, the proposed multi-scene aggregation module seamlessly binds nearby scenes into a complete map using a coarse-to-fine approach. In real-world applications, MRASfM demonstrates its generalizability and robustness through its superior performance compared to COLMAP. Public datasets evaluations validate the state-of-the-art performance of MRASfM. MRASfM enables intelligent vehicles to better perceive environments and estimate ego poses, which is crucial for downstream tasks.

REFERENCES

[1] Z. Xie, Z. Pang, and Y.-X. Wang, "Mv-map: Offboard hd-map generation with multi-view consistency," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8658–8668.

[2] L. Li, S. Peng, Z. Yu, S. Liu, R. Pautrat, X. Yin, and M. Pollefeys, "3d neural edge reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 219–21 229.

[3] X. Zhao, B. Chen, M. Sun, D. Yang, Y. Wang, X. Zhang, M. Li, D. Kou, X. Wei, and L. Zhang, "Hybridocc: Nerf enhanced transformer-based multi-camera 3d occupancy prediction," *IEEE Robotics and Automation Letters*, 2024.

[4] T. Deng, X. Chen, Y. Chen, Q. Chen, Y. Xu, L. Yang, L. Xu, Y. Zhang, B. Zhang, W. Huang, and H. Wang, "Gaussiandwm: 3d gaussian driving world model for unified scene understanding and multi-modal generation," *arXiv preprint arXiv:2512.23180*, 2025.

[5] W. Wu, Q. Wang, G. Wang, J. Wang, T. Zhao, Y. Liu, D. Gao, Z. Liu, and H. Wang, "Emie-map: Large-scale road surface reconstruction based on explicit mesh and implicit encoding," *arXiv preprint arXiv:2403.11789*, 2024.

[6] P. Dai, J. Xu, W. Xie, X. Liu, H. Wang, and W. Xu, "High-quality surface reconstruction using gaussian surfels," in *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–11.

[7] R. Mei, W. Sui, J. Zhang, X. Qin, G. Wang, T. Peng, T. Chen, and C. Yang, "Rome: Towards large scale road surface reconstruction via mesh representation," *IEEE Transactions on Intelligent Vehicles*, 2024.

[8] J. Zhang, S. Chen, H. Yin, R. Mei, X. Liu, C. Yang, Q. Zhang, and W. Sui, "A vision-centric approach for static map element annotation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 15 861–15 867.

[9] A. Cohen, T. Sattler, and M. Pollefeys, "Merging the unmatchable: Stitching visually disconnected sfm models," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2129–2137.

[10] Y. Chen, Z. Yu, S. Song, T. Yu, J. Li, and G. H. Lee, "Adasfm: From coarse global to fine incremental adaptive structure from motion," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 2054–2061.

[11] W. Zhang, S. Wang, X. Dong, R. Guo, and N. Haala, "Bamf-slam: Bundle adjusted multi-fisheye visual-inertial slam using recurrent field transforms," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 6232–6238.

[12] Y. Wang, Y. Ng, I. Sa, A. Parra, C. Rodriguez-Opazo, T. Lin, and H. Li, "Mavis: Multi-camera augmented visual-inertial slam using se2(3) based exact imu pre-integration," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 1694–1700.

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

[14] H. Cui and S. Shen, "Mma: Multi-camera based global motion averaging," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 1, 2022, pp. 490–498.

[15] P. Tao, H. Cui, D. Tu, and S. Shen, "MGSfM: Multi-Camera Driven Global Structure-from-Motion," in *IEEE International Conference on Computer Vision (ICCV)*, 2025.

[16] S. Chen, J. Zhang, R. Mei, Y. Cai, H. Yin, T. Chen, W. Sui, and C. Yang, "Camav2: A vision-centric approach for static map element annotation," *arXiv preprint arXiv:2407.21331*, 2024.

[17] H. Cui, X. Gao, and S. Shen, "Mcsfm: multi-camera-based incremental structure-from-motion," *IEEE Transactions on Image Processing*, vol. 32, pp. 6441–6456, 2023.

[18] O. Özyeşil, V. Voroninski, R. Basri, and A. Singer, "A survey of structure from motion*." *Acta Numerica*, vol. 26, pp. 305–364, 2017.

[19] C. Wu, "Towards linear-time incremental structure from motion," in *2013 International Conference on 3D Vision-3DV 2013*. IEEE, 2013, pp. 127–134.

[20] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[21] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 224–236.

[22] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, "Lightglue: Local feature matching at light speed," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17 627–17 638.

[23] H. Chen, Z. Luo, L. Zhou, Y. Tian, M. Zhen, T. Fang, D. Mckinnon, Y. Tsin, and L. Quan, "Aspanformer: Detector-free image matching with adaptive span transformer," in *European Conference on Computer Vision*. Springer, 2022, pp. 20–36.

[24] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4938–4947.

[25] P.-E. Sarlin, P. Lindenberger, V. Larsson, and M. Pollefeys, "Pixel-perfect structure-from-motion with featuremetric refinement," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[26] S. Weber, N. Demmel, T. C. Chan, and D. Cremers, "Power bundle adjustment for large-scale 3d reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 281–289.

[27] M. Zheng, N. Chen, J. Zhu, X. Zeng, H. Qiu, Y. Jiang, X. Lu, and H. Qu, "Distributed bundle adjustment with block-based sparse matrix compression for super large scale datasets," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 18 152–18 162.

[28] A. Zhanabatyrova, C. S. Leite, and Y. Xiao, "Structure from motion-based mapping for autonomous driving: Practice and experience," *ACM Transactions on Internet of Things*, vol. 5, no. 1, pp. 1–25, 2024.

[29] Y. Chen, S. Shen, Y. Chen, and G. Wang, "Graph-based parallel large scale structure from motion," *Pattern Recognition*, vol. 107, p. 107537, 2020.

[30] M. Fang, T. Pollok, and C. Qu, "Merge-sfm: Merging partial reconstructions," in *British Machine Vision Conference*, 2019.

[31] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[32] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[33] L. Pan, D. Baráth, M. Pollefeys, and J. L. Schönberger, "Global structure-from-motion revisited," in *European Conference on Computer Vision*. Springer, 2024, pp. 58–77.

[34] 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.

[35] Z. Teed and J. Deng, "DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras," *Advances in neural information processing systems*, 2021.

[36] H. Li, Y. Duan, X. Zhang, H. Liu, J. Ji, and Y. Zhang, "Occ-vo: Dense mapping via 3d occupancy-based visual odometry for autonomous driving," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 17 961–17 967.