Neural Implicit Monocular Visual SLAM for 3D Reconstruction in Planetary Environments

Chen Liu[1], Rong Huang[1,2], Huan Xie[1,2], Tao Tao[1], Yongjiu Feng[1,2], Xiaohua Tong[1,2]

[1] College of Surveying and Geo-Informatics, Tongji University, Shanghai, China
[2] Shanghai Key Laboratory for Planetary Mapping and Remote Sensing for Deep Space Exploration, Shanghai, China

Keywords: Neural Radiance Fields, Visual SLAM, Reconstruction, Planetary Environments.

Abstract

The application of SLAM technology in planetary environments has become a research frontier for autonomous rovers. Existing visual SLAM methods often exhibit low accuracy in pose estimation and reconstruction due to poor feature extraction and mismatched correspondences. This paper introduces a novel strategy that integrates neural implicit networks within a visual SLAM framework. By jointly optimizing camera poses and implicit scene representations using neural radiance fields, we achieve high-precision visual localization in the Mars scene without requiring loop closure. We validate our method using data from NASA's Perseverance rover and compare its performance with OV^{2}SLAM. The results demonstrate that our method significantly outperforms OV^{2}SLAM in localization accuracy, achieving an 85.16% reduction in absolute trajectory errors and maintaining translation errors within 1 m across the entire trajectory. Moreover, our framework delivers compelling novel view synthesis despite sparse inputs and a fixed forward-facing viewpoint. The 3D point cloud models, synthesized from estimated depth maps and poses, further highlight the feasibility and effectiveness of our method for reconstruction in planetary environments.

1. Introduction

3D reconstruction technology plays a pivotal role in autonomous navigation and scientific exploration tasks for rovers (Gemme et al., 2005). In planetary environments lacking GPS support, such as the surfaces of Mars and the Moon, Simultaneous Localization and Mapping (SLAM) has become a central research focus for autonomous exploration systems (Cao et al., 2012). To support rover missions, SLAM systems must not only construct real-time centimeter-level environmental models but also ensure the long-term stability of pose estimation. This is critical for subsequent tasks such as path planning (Wang et al., 2017b), scientific target identification (Barnes et al., 2009), and mission execution (Bass et al., 2005). The 3D reconstruction model generated during the rover's exploration provides strong support for both the morphological examination of Martian impact craters (Ye et al., 2025) and the global place recognition for robot navigation (Xia et al., 2021, Xia et al., 2023). However, the unique geomorphological features of planetary surfaces pose significant challenges to traditional visual SLAM methods: repetitive textures caused by weathering layers (Maimone et al., 2007), complex and variable features, and anisotropic geometries. These environmental characteristics lead to systemic risks for feature-based or direct SLAM systems, including feature mismatch and pose estimation drift.

Current mainstream visual SLAM frameworks typically utilize feature-based or direct methods as front-end processing modules. However, actual imaging data from Mars rover missions reveal that, in dust-covered regions, the spatial distribution density of SIFT features sharply decreases compared to urban environments on Earth. This feature sparsity results in an increased failure rate in co-visibility detection and severely restricts feature matching and continuous tracking between adjacent frames (Zhong et al., 2023). Although direct methods, which rely on photometric consistency, are theoretically more suitable for low-texture environments, the high dynamic range of cameras in planetary environments significantly reduces the convergence of the photometric error function, and local minima can severely degrade pose estimation accuracy. On the other hand, terrestrial SLAM systems typically incorporate loop closure for global optimization. However, given the cost constraints associated with rover missions, loop closure cannot be implemented in planetary environments (Guo et al., 2018), thereby posing additional challenges for robust SLAM estimation.

In recent years, Neural Radiance Fields (NeRF) (Mildenhall et al., 2021), have demonstrated leading performance in the field of 3D reconstruction. Unlike traditional explicit representations, this continuous scene representation method uses Multi-Layer Perceptrons (MLPs) to map spatial coordinates to volumetric density and color values, showing significant advantages in reconstructing complex geometric surfaces and synthesizing novel views. However, applying the NeRF framework typically requires Structure from Motion (SfM) preprocessing to obtain subpixel-level poses, which limits its use in autonomous navigation scenarios. Several studies have developed algorithms that jointly optimize both pose and scene representation to enhance the versatility of NeRF. One approach combines visual SLAM for pose tracking in the front end with neural implicit representation-based mapping in the back end (Zhang et al., 2023, Mao et al., 2024), while another treats pose as a learnable parameter to be optimized simultaneously with the scene representation (Bian et al., 2023). The combination of SLAM front-end and neural implicit back-end algorithms has been mainly applied to small indoor scenes (Zhu et al., 2022), where the long-term robustness of front-end tracking cannot be guaranteed in planetary environments. On the other hand, joint optimization NeRF algorithms face significant challenges when dealing with long sequences and sparse viewpoints in exploration tasks, making them ineffective for the autonomous exploration requirements of rovers.

To address the specific constraints of planetary exploration tasks, we apply a joint optimization approach for pose and scene representation in a neural implicit method as a new SLAM

framework to tackle the aforementioned challenges. This approach is notably different from existing visual SLAM and neural implicit SLAM systems in key aspects. First, instead of using the typical feature extraction and matching procedure in the front end, it employs an end-to-end differentiable rendering pipeline that jointly optimizes camera positions and scene representations using pixel-level photometric consistency and other prior geometry constraints. Second, the framework adopts a progressive optimization strategy, beginning with a small initial set of images and gradually adding images from various viewpoints to avoid reaching local minima. Finally, the entire scene is separated into bunches of local radiance fields, allowing for independent optimization while avoiding the restrictions of direct global optimization.

We applied this framework to the Perseverance rover's Mars images for practical testing and compared it with OV^{2}SLAM as a visual SLAM method. The experiments show that our approach provides more robust pose estimation results, with the full travel trajectory closely matching actual conditions and a significantly lower drift rate.

In summary, our main contributions are listed in two aspects:

- We apply the progressively optimized neural radiance field, LocalRF, to a monocular visual SLAM framework to address the challenges of low localization accuracy and insufficient robustness in traditional visual SLAM systems for planetary environments.
- We conduct autonomous rover localization and 3D reconstruction experiments using real Mars imagery from the Perseverance rover. Our approach significantly improves localization accuracy and successfully reconstructs a 3D point cloud model of the exploration scene.

2. Related Work

Visual SLAM is currently the most efficient deployment approach in planetary environments. Existing visual SLAM research has evolved into multiple branches, which can be categorized into feature-based SLAM, direct SLAM, and deep learning-based SLAM. Each offers various solutions for different scenarios. Therefore, this section will review and analyze these three types of algorithms.

Feature-based SLAM. Feature-based visual SLAM constitutes the majority of widely adopted SLAM algorithms. These methods operate on hand-crafted feature detectors (such as Harris corners and SIFT descriptors) to establish matching and tracking relationships by identifying stable feature points across consecutive frames (Schönberger et al., 2017), subsequently optimizing camera trajectories and sparse point cloud maps through multi-view geometric constraints. The early proposed MonoSLAM system (Davison et al., 2007) pioneered real-time monocular visual SLAM implementation through an Extended Kalman Filter (EKF) framework for feature tracking and map construction, laying the foundation for subsequent research. In feature extraction and matching advancements, the ORB-SLAM system (Mur-Artal et al., 2015) utilized Oriented FAST and Rotated BRIEF (ORB) features, which enhanced feature repeatability and matching robustness while maintaining real-time performance, establishing a landmark work in feature-based SLAM. Subsequent iterations including ORB-SLAM2 (Mur-Artal and Tardós, 2017) and ORB-SLAM3 (Campos et al., 2021) extended this framework through binocular vision integration and multi-sensor fusion capabilities. To address varying operational scales and frame rates, OV^{2}SLAM (Ferrera et al., 2021) demonstrates exceptional performance with its efficient front-end tracking and innovative online Bag-of-Words approach, supporting real-time localization across frequencies ranging from single-digit Hz to hundreds of Hz while achieving superior positioning accuracy across three public datasets. However, these methodologies exhibit inherent limitations when handling consecutive frames with insufficient local features. Planetary remote sensing images lack surface texture information and show considerable nonlinear radiation differences (Wan et al., 2025, Huang et al., 2024), making classic feature-matching algorithms ineffective. The tracking pipelines prove particularly vulnerable to failure in scenarios where localized regions exhibit persistent feature scarcity, revealing critical robustness deficiencies for autonomous localization tasks in planetary environments characterized by homogeneous terrain textures and sparse visual features.

Direct SLAM. Direct SLAM methods accomplish pose estimation and scene reconstruction through nonlinear optimization of pixel intensity differences between adjacent frames. The core methodology involves constructing a photometric consistency residual function. This intensity-based optimization paradigm eliminates computational overhead from feature descriptor generation, granting unique advantages for localization and mapping in low-texture environments. A seminal contribution in this domain is the DTAM system (Newcombe et al., 2011), which innovatively integrated global dense depth map optimization with direct tracking, achieving real-time 3D reconstruction under a sub-pixel level photometric error minimization framework. LSD-SLAM (Engel et al., 2014) advanced this paradigm, introducing hierarchical keyframe selection and semi-dense reconstruction to address scale drift in large-scale environments. The system implements gradient-based pixel selection to enhance tracking stability in texture-deficient scenarios. Dense Visual Odometry (DVO) (Kerl et al., 2013), proposed an intensity error optimization framework augmented with depth information. This approach leverages photometric consistency constraints between RGB and depth images for motion estimation and dense reconstruction, demonstrating improved performance in feature-sparse environments. On planetary surfaces, large areas of weak texture or repetitive geometric structures may lead to insufficient image gradient information relied upon by direct method SLAM, resulting in a deteriorated condition number of the Hessian matrix of the photometric error function and a significant decrease in pose estimation accuracy.

Deep-learning based SLAM. Deep learning-based visual SLAM models the relationship between environment geometry and camera motion implicitly through neural networks, overcoming the limitations of traditional hand-crafted features. In terms of end-to-end SLAM architecture innovations, DeepVO (Wang et al., 2017a) uses an LSTM network to model temporal pose constraints, achieving continuous motion estimation without explicit geometric modeling. TartanVO (Wang et al., 2021) introduces a cascade architecture of optical flow prediction networks and pose regression networks, making breakthroughs in cross-domain generalization. DROID-SLAM (Teed and Deng, 2021) combines the RAFT (Teed and Deng, 2020) optical flow network with an iterative updating mechanism to achieve its high accuracy, high robustness, and strong generalization. With the advancement of NeRF technology, the integration of NeRF with SLAM has opened up new pathways for

implicit scene modeling. iMAP (Sucar et al., 2021) was the first to introduce NeRF into real-time SLAM systems, achieving dense reconstruction under monocular input through keyframe sampling and radiance field online optimization. Nice-SLAM incorporates a hierarchical scene representation, integrating multi-level local information and leveraging pre-trained geometric priors to optimize scene representation, enabling dense reconstruction of large indoor scenes. NeRF-SLAM (Rosinol et al., 2023) proposes a differentiable framework that jointly optimizes camera poses and NeRF scene representations. It combines DROID-SLAM as the front-end tracking module with neural radiance fields for scene reconstruction, leading to good indoor geometric accuracy based on photometric loss and depth prior constraints.

3. Methodology

In this work, we propose a method for monocular visual SLAM that applies the joint optimization of pose and scene representation using neural radiance field to planetary images. We adopt LocalRF (Meuleman et al., 2023) as the SLAM framework, which enables large-scale and unbound outdoor scene localization and scene representation without initial pose using only monocular images. To enhance geometric constraints in complex environments, this framework employs monocular depth estimation and optical flow estimation models to generate prior information as geometric supervision, which, combined with photometric loss, constructs the overall loss function. The workflow is described in Fig. 1.

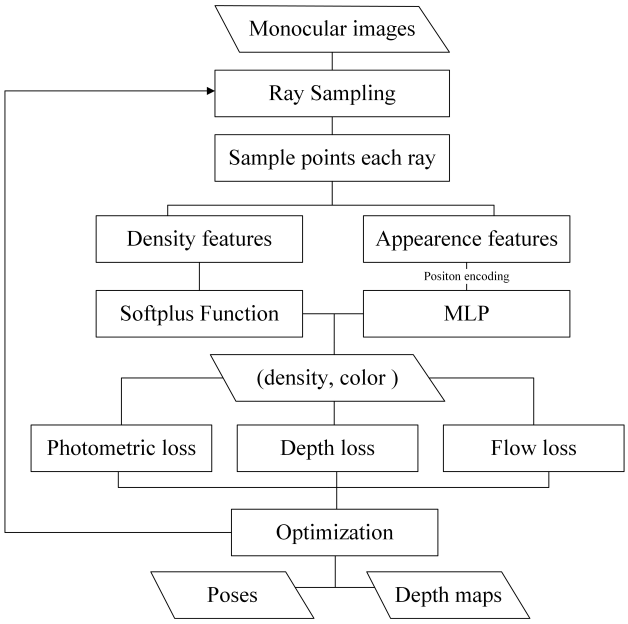

Figure 1. Workflow of our framework.

3.1 Preliminaries

NeRF is an innovative deep-learning framework that enables the synthesis of photorealistic novel views from a limited set of input images. NeRF employs an MLP-based neural network to model a continuous volumetric scene representation. This representation maps 3D spatial coordinates $\mathbf{x} = (x, y, z)$ and viewing directions $\mathbf{d} = (\theta, \phi)$ to corresponding colors $\mathbf{c} = (r, g, b)$ and densities σ, allowing the system to generate high-quality renderings of scenes from arbitrary viewpoints:

$$\varphi(\mathbf{x}), \varphi(\hat{\mathbf{d}}) \xrightarrow{\mathcal{F}} (\mathbf{c}, \sigma) \tag{1}$$

where $\hat{\mathbf{d}} = (d_x, d_y, d_z)$ is a representation of the viewing direction using the actual three-dimensional Cartesian space vector, $\varphi(\cdot)$ indicates the application of position encoding and $\mathcal{F}$ denotes the neural network function.

To generate a novel view, NeRF leverages a volume rendering technique to render input rays $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ cast from the camera to the scene. For each pixel, the color is determined by accumulating the color and density of each sample points along the ray, which is expressed as the integral:

$$C(\mathbf{r}) = \int_{t_1}^{t_2} T(t)\sigma(\mathbf{r}(t))\mathbf{c}(\mathbf{r}(t), \hat{\mathbf{d}})dt \tag{2}$$

Where $T(t) = \exp\left(-\int_{t_1}^{t} \sigma(\mathbf{r}(s))ds\right)$ indicates the overall transparency, which models the amount of light that reaches the camera after passing through the medium. The color at each point along the ray is weighted by the opacity $\sigma(\mathbf{r}(t))$, and the total color contribution is integrated over the ray's length. This integral allows the model to capture the complex interactions of light within the scene.

During training, the color loss quantifies the photometric error between predicted images and ground truth. This loss term is calculated through pixel-wise comparison of RGB values in the rendered image $C(r)$ and the corresponding ground truth image $C^{\text{gt}}(r)$. Minimization of this photometric discrepancy during optimization enhances the model's capacity to reconstruct photorealistic scene representations. The color loss $\mathcal{L}_{\text{color}}$ is formally defined as:

$$\mathcal{L}_{\text{color}} = \sum_{r \in \mathcal{R}} \left\| C(r) - C^{\text{gt}}(r) \right\|_2^2 \tag{3}$$

where $\mathcal{R}$ denotes the set of sampled rays, N represents the total number of rays in each batch, and $\| \cdot \|_2$ indicates the L2-norm operation. This formulation ensures dense supervision across all observed pixels while maintaining computational efficiency.

3.2 LocalRF

LocalRF is an incremental NeRF algorithm that jointly optimizes camera poses and scene representations. It partitions the input images into multiple localized radiance fields, with only the current localized radiance field being activated for training and optimization at each stage. LocalRF employs three types of loss functions for optimization: photometric loss, depth loss, and optical flow loss. Each local radiance field will independently calculate its own photometric loss, while the last part of the optimized consecutive frames from the previous local radiance field will be used for supervision.

To handle unbounded scenes, LocalRF employs a unique dynamic radiance field setup to avoid reliance on predefined global scaling parameters used in Mip-NeRF360 (Barron et al., 2022). The use of the L_∞ norm uniformly extends the spatial range of the scene to the maximum side length of a square bounding box, with all 3D points in each local scene being scaled to the [-2, 2] space, facilitating subsequent MLP infer-

ence for the radiance field:

$$\text{contract}(\mathbf{p}) = \begin{cases} \mathbf{p} & \text{if } \|\mathbf{p}\|_\infty \leq 1, \\ \left(2 - \frac{1}{\|\mathbf{p}\|_\infty}\right)\left(\frac{\mathbf{p}}{\|\mathbf{p}\|_\infty}\right) & \text{otherwise.} \end{cases} \quad (4)$$

To enhance joint optimization performance in long-sequence large-scale scene reconstruction, LocalRF employs a progressive optimization strategy. The framework initializes with a predefined frame subset and incrementally incorporates new images from the subsequent sequence. Camera poses of newly added frames are initialized as identical to their preceding counterparts.

$$\mathbf{T}_p \Rightarrow \mathbf{T}_{p+1}^{(0)} \quad (5)$$

The initialized pose $\mathbf{T}_{p+1}^{(0)}$ is subsequently treated as learnable parameters during joint optimization—this scheme ensures robust optimization convergence while preventing local minima trapping. This progressive optimization strategy is particularly beneficial for sequence images captured from a fixed viewpoint on planetary rovers.

Another core process lies in the dynamic instantiation of discrete local radiance fields, decomposing extensive scenes into spatially adjacent neural representations with partial frame overlap. This architecture ensures meticulous local optimization while preserving global scene consistency across viewpoints. Each new local radiance field is established where the last frame is in the currently optimized sequence, thereby fully utilizing the optimized frames as effective supervision for the new radiance field.

Monocular depth and dense optical flow provide strong constraints when reconstructing Neural Radiance Fields from consecutive frames captured from sparse viewpoints, effectively maintaining geometric consistency during progressive optimization. In this study, DPT (Ranftl et al., 2021) is used for monocular depth estimation, and RAFT is used for optical flow estimation. Depth loss leverages the fine geometry information from pre-trained models to improve the neural radiance field, while optical flow loss utilizes pixel-level correspondence constraints to optimize depth and poses. Dense depth maps $\hat{\mathbf{D}}(\mathbf{r})$ are estimated by the volumetric density, both the estimated depth and the prior depth are then normalized to compute loss $\mathcal{L}_\text{d}$:

$$\hat{\mathbf{D}}(\mathbf{r}) = \sum_{i=1}^{N} T_i \left(1 - \exp(-\sigma_i \delta_i)\right) d_i, \quad (6)$$

$$\mathcal{L}_\text{d} = \left| \hat{\mathbf{D}}^* - \mathbf{D}^* \right|. \quad (7)$$

By incorporating intrinsic and extrinsic camera parameters with depth, the forward optical flow is derived as $\hat{\mathcal{F}}_{k\to k+1}$, while the backward optical flow $\mathcal{F}_{k\to k-1}$ is computed in the same way. The 3D points in the scene are reprojected using the relative poses between adjacent frames to calculate the actual optical flow. The loss is then computed as $\mathcal{L}_\text{f}$ as follows:

$$\hat{\mathcal{F}}_{k\to k+1} = (u, v) - \Pi\left([R|t]_{k\to k+1} \Pi^{-1}(u, v, \hat{D})\right), \quad (8)$$

$$\mathcal{L}_\text{f} = \left\| \hat{\mathcal{F}}_{k\to k+1} - \mathcal{F}_{k\to k+1} \right\|_1. \quad (9)$$

We utilize the depth maps and poses estimated by LocalRF to perform depth fusion with the consistency check for better reconstruction (Cheng et al., 2020), thus exporting the final 3D point clouds.

4. Experiments

4.1 Dataset

We use a sequence of images captured by the navigation camera of NASA’s Perseverance rover in sol 200 during its AutoNav operation as the dataset to demonstrate the practical performance of our algorithm. The dataset consists of grayscale images captured by the stereo navigation cameras, each with the size of 1280×960. We utilize rectified images provided by NASA and perform cropping and filtering to ensure their suitability. After preprocessing, the dataset comprises 180 frames, with each image resized to 886×665. For the entire framework, we fully rely on the left navigation camera images as input.

4.2 Baseline and Metrics

Considering that the Perseverance rover image dataset lacks ground-truth trajectories similar to those in simulated datasets on earth, we use the most popular baseline algorithm COLMAP (Schönberger and Frahm, 2016, Schönberger et al., 2016) to recover image poses as the reference ground truth. Among open-source visual SLAM algorithms, OV^{2}SLAM has demonstrated superior localization accuracy compared to ORB-SLAM2 on the KITTI visual odometry benchmark, owing to its strong adaptability to different scales and frame rates. OV^{2}SLAM is used as the baseline classical visual SLAM algorithm for comparison with our method to evaluate the performance of both approaches on real planetary datasets. Monocular visual SLAM systems, including COLMAP, always suffer from scale ambiguity. Therefore, we use the binocular trajectory from OV^{2}SLAM as a reference and apply the Umeyama (Umeyama, 1991) algorithm to recover the scale of the monocular poses from OV^{2}SLAM, LocalRF, and COLMAP. The localization accuracy evaluation is then conducted using the scale-recovered poses. For scene representation, we use PSNR, SSIM, and LPIPS$_\text{VGG}$ to evaluate the quality of novel view synthesis generated by the model.

4.3 Implementation

Our experiments follow the default settings of LocalRF, with the first 5 frames of the selected sequence used for initialization. Poses of these frames are initialized as identity and incorporated into the first TensoRF model, which is then gradually optimized while adding new frames. The 5-frame poses initialized as identities are optimized and updated according to the scene representation, and once a new frame added its initial pose is set to be the same as the optimized pose of the previous frame to ensure fast and robust training convergence. The learning rates for pose optimization are set to $5 \cdot 10^{-3}$ for rotation and $5 \cdot 10^{-4}$ for translation. The TensoRF resolution is initially set to 64^3 and is later upsampled to 640^3. The weights for the photometric loss, optical flow loss, and depth loss are set to 0.25, 1.0, and 0.1 respectively. We recover the photogrammetric camera model intrinsics from NASA’s linearly corrected images using CAHVOR camera parameters (Di and Li, 2004). Based on this, we compute the key training parameter FOV for LocalRF. The MLP used for regressing rendered color values adopts Fea_late_view model. Decomposed features are first passed through two fully connected layers with 128 hidden

units and ReLU activation. The view direction is then concatenated with the output and processed by a final linear transform layer. The ray sampling batch size is set to 4096, and the entire training is conducted on an NVIDIA RTX 4090.

4.4 Results of Perseverance dataset

Fig. 2 presents the camera pose estimation results of the two algorithms in the XZ plane view. The dataset used in this experiment follows a forward trajectory without sharp turns. After performing the same Sim(3) Umeyama alignment, the trajectory of LocalRF nearly overlaps with that of COLMAP. Although the OV^{2}SLAM trajectory also maintains a consistent forward trend, there are some significant initial localization ambiguities, and it shows a larger drift error in subsequent frames. We further present the per-frame translation error of both algorithms compared to COLMAP to quantitatively visualize the drift in local regions, as shown in Fig. 3.

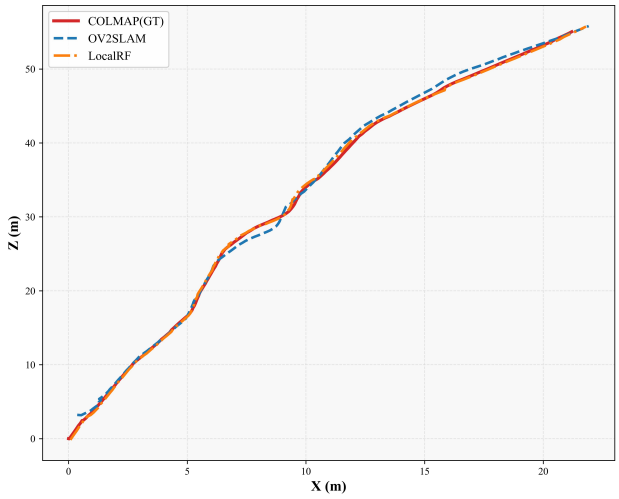

Figure 2. Trajectory of OV^{2}SLAM and LocalRF in XZ plane.

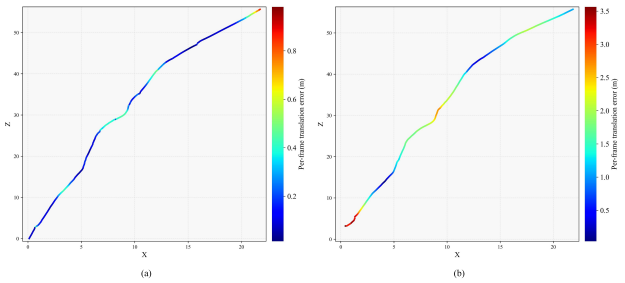

Figure 3. Translation error of each method. (a) LocalRF, (b) OV^{2}SLAM.

As illustrated in Fig. 3, LocalRF maintains translation errors below 1 m for all frames, with a maximum deviation of 0.98 m in the end, while other frames exhibit satisfactory precision. The significant initial deviation observed in OV^{2}SLAM results from trajectory alignment to ensure global scale and positional consistency with COLMAP ground truth. This dataset from the Perseverance rover presents unique challenges compared to conventional benchmarks like KITTI: 1) Substantial inter-frame translation intervals due to mission-imposed acquisition constraints; 2) Variable frame rates significantly lower than the 10Hz standard in terrestrial datasets; 3) Non-constant speed motion patterns. Furthermore, the characteristic low-texture Martian terrain poses inherent difficulties for OV^{2}SLAM's front-end optical flow tracking module. Global trajectory analysis reveals that OV^{2}SLAM exhibits considerable mid-sequence drift, with translational errors exceeding 1 m throughout most frames.

Method	$RPE_r \downarrow$	RPE_t (m)$\downarrow$	ATE (m)$\downarrow$
OV^{2}SLAM	0.040	0.278	1.779
LocalRF	**0.014**	**0.188**	**0.264**

Table 1. Pose evaluation of Perseverance dataset. All metrics are presented in RMSE.

We used the EVO toolkit for quantitative evaluation, and the results are given in Table 1, with the best performances highlighted in bold. LocalRF achieves a superior ATE of 0.264m, demonstrating significant improvement over OV^{2}SLAM. RPE is measured with 1-meter interval to assess the drift of both algorithms within each segment which further confirms LocalRF's advantages in both rotational and translational components. This enhanced performance stems from LocalRF's dynamic local radiance field creation mechanism, which allows fine adjustments to the poses within each local scene. Depth supervision from DPT and optical flow constraints from RAFT collaboratively guide geometrically consistent representation learning. In addition, the joint optimization strategy based on NeRF replaces explicit mesh structures with implicit volumetric representations, allowing for better multi-view consistency in complex scenes through neural networks. Although camera poses lack dedicated loss terms, the continuous refinement through backpropagation during scene representation learning ensures progressive accuracy enhancement. This symbiotic optimization between pose parameters and neural scene representation proves particularly effective in handling the Perseverance dataset's challenges of sparse viewpoints and weak texture.

To evaluate the synthesized novel view results of LocalRF, we adopted the same approach as the original LocalRF, extracting one frame every 10 frames as the test frame. Qualitative results of synthesized novel views are presented in Fig. 4, demonstrating that the generated views effectively preserve the actual geometric information of test frames without structural distortions or reconstruction failures. The rendering of variably sized Martian surface rocks indicates LocalRF's capability for reasonable depth estimation along rover trajectories, providing a valuable reference for subsequent path planning in autonomous exploration missions. Nevertheless, reconstruction quality in planetary environments remains inferior to terrestrial scenarios, with observable blur artifacts in multiple near-field regions. This indicates some targeted optimizations need to be made for the environment migration.

Method	PSNR $\uparrow$	SSIM $\uparrow$	$LPIPS_{VGG} \downarrow$
LocalRF	25.750	0.684	0.460

Table 2. Quantitive novel view synthesis results.

The quantitative metrics of novel view synthesis on Perseverance dataset are shown in Table 2. While the PSNR consistently maintains a high level, both the SSIM and LPIPS exhibit observable challenges. This limitation stems from two primary factors. There is a notable reduction of high-frequency variations in planetary scenes, where most high-frequency signals originate from rock boundaries. The repetitive geological structures fail to provide sufficient distinguished features for neural

Figure 4. Novel view synthesis results of our proposed method. The top row shows the rendered results and the bottom row provides ground truth.

network learning. This feature scarcity may cause the neural radiance field to misinterpret distinct terrain points as identical locations, resulting in blur artifacts and erroneous depth estimation. Another factor is the insufficient inter-frame overlap (Yu et al., 2021) in the Perseverance dataset, where optical flow priors struggle to establish reliable pixel-wise correspondences, particularly in near-field regions with extensive occlusions that degrade constraint effectiveness. Despite these challenges, LocalRF demonstrates promising practical potential for planetary environment applications.

We reconstructed the Martian scene based on the results from LocalRF to demonstrate the feasibility of the current framework for 3D reconstruction. Following the post-processing approach from MVSNet, we performed depth fusion on the estimated depth maps and poses to generate a 3D point cloud. The reconstruction results are shown in Fig. 5.

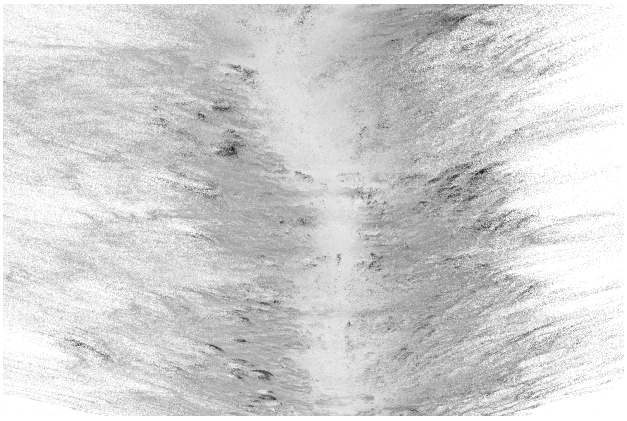

Figure 5. Reconstruction result of Perseverance dataset.

Our reconstruction results successfully capture fundamental geometric structures surrounding the rover's trajectory in Martian scenarios while reconstructing most rocks within the field of view. The lack of detail in central point cloud regions may stem from overexposure artifacts in Perseverance's imagery. Furthermore, the depth maps estimated through LocalRF exhibit notable noise artifacts when integrated with pose estimation results, particularly manifesting as degraded reconstruction quality for distant geological features. The framework also demonstrates limitations in depth consistency filtering for sky regions, resulting in disordered point cloud distributions in upper scene areas. These observations reveal inherent constraints when directly applying existing neural radiance field frameworks to extraterrestrial environments, necessitating further optimizations for improved depth estimation.

Nevertheless, the achieved 3D reconstruction results validate the methodological feasibility and reveal promising application prospects. The framework demonstrates essential capabilities for reconstructing navigation-critical features such as rock distributions and terrain undulations. Future enhancements could incorporate adaptive exposure compensation modules and hybrid depth estimation strategies combining neural rendering with geometric priors to address current limitations. These improvements would better align the system with the stringent requirements of planetary exploration missions.

5. Conclusion

In this work, we applied the LocalRF to navigation images captured by the Perseverance rover, enabling a neural implicit monocular visual SLAM system for 3D reconstruction tasks in planetary environments. Our experiments indicate that traditional visual SLAM approaches struggle to handle real planetary images effectively, whereas our incremental neural implicit SLAM framework, powered by LocalRF, achieves significantly more robust and more accurate pose estimation. Furthermore, we generated a fused 3D point cloud of the actual scene by leveraging the poses and dense depth maps estimated by LocalRF. Although the overall performance is constrained by the characteristics of planetary images and the limited availability of data, we believe that neural implicit 3D reconstruction holds substantial promise for applications in planetary exploration.

References

Barnes, D., Pugh, S., Tyler, L., 2009. Autonomous science target identification and acquisition (astia) for planetary exploration. *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 3329–3335.

Barron, J. T., Mildenhall, B., Verbin, D., Srinivasan, P. P., Hedman, P., 2022. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5460–5469.

Bass, D., Wales, R., Shalin, V., 2005. Choosing mars time: analysis of the mars exploration rover experience. *2005 IEEE Aerospace Conference*, 4174–4185.

Bian, W., Wang, Z., Li, K., Bian, J.-W., 2023. Nope-nerf: Optimising neural radiance field with no pose prior. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4160–4169.

Campos, C., Elvira, R., Rodríguez, J. J. G., Montiel, J. M., Tardós, J. D., 2021. ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual–Inertial, and Multimap SLAM. *IEEE Transactions on Robotics*, 37(6), 1874–1890.

Cao, F., Wang, R., Zhang, L., 2012. Three-Dimensional Positioning Algorithm for Lunar Rover Based on Binocular Stereoscopic Vision. *Traffic Information and Safety*, 30(04), 28–33.

Cheng, S., Xu, Z., Zhu, S., Li, Z., Li, L. E., Ramamoorthi, R., Su, H., 2020. Deep stereo using adaptive thin volume representation with uncertainty awareness. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2521–2531.

Davison, A. J., Reid, I. D., Molton, N. D., Stasse, O., 2007. MonoSLAM: Real-Time Single Camera SLAM. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6), 1052–1067.

Di, K., Li, R., 2004. CAHVOR Camera Model and Its Photogrammetric Conversion for Planetary Applications. *Journal of Geophysical Research: Planets*, 109(E4).

Engel, J., Schöps, T., Cremers, D., 2014. Lsd-slam: Large-scale direct monocular slam. *Computer Vision – ECCV 2014*, 834–849.

Ferrera, M., Eudes, A., Moras, J., Sanfourche, M., Le Besnerais, G., 2021. OV^{2}SLAM: A Fully Online and Versatile Visual SLAM for Real-Time Applications. *IEEE Robotics and Automation Letters*, 6(2), 1399–1406.

Gemme, S., Bakambu, J., Rekleitis, I., 2005. 3D reconstruction of environments for planetary exploration. *The 2nd Canadian Conference on Computer and Robot Vision (CRV'05)*, 594–601.

Guo, Y., Feng, Z., Ma, G., Guo, Y., Zhang, M., 2018. Advances and Trends in Visual Navigation and Autonomous Control of a Planetary Rover. *Journal of Astronautics*, 39(11), 1185–1196.

Huang, R., Wan, G., Zhou, Y., Ye, Z., Xie, H., Xu, Y., Tong, X., 2024. Fast Double-Channel Aggregated Feature Transform for Matching Planetary Remote Sensing Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17, 9282–9293.

Kerl, C., Sturm, J., Cremers, D., 2013. Dense visual slam for rgb-d cameras. *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2100–2106.

Maimone, M., Cheng, Y., Matthies, L., 2007. Two Years of Visual Odometry on the Mars Exploration Rovers. *Journal of Field Robotics*, 24(3), 169–186.

Mao, Y., Yu, X., Zhang, Z., Wang, K., Wang, Y., Xiong, R., Liao, Y., 2024. Ngel-slam: Neural implicit representation-based global consistent low-latency slam system. *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 6952–6958.

Meuleman, A., Liu, Y.-L., Gao, C., Huang, J.-B., Kim, C., Kim, M. H., Kopf, J., 2023. Progressively optimized local radiance fields for robust view synthesis. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16539–16548.

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.

Mur-Artal, R., Montiel, J. M. M., Tardos, J. D., 2015. ORB-SLAM: A Versatile and Accurate Monocular SLAM System. *IEEE Transactions on Robotics*, 31(5), 1147–1163.

Mur-Artal, R., Tardós, J. D., 2017. ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras. *IEEE Transactions on Robotics*, 33(5), 1255–1262.

Newcombe, R. A., Lovegrove, S. J., Davison, A. J., 2011. Dtam: Dense tracking and mapping in real-time. *2011 International Conference on Computer Vision*, 2320–2327.

Ranftl, R., Bochkovskiy, A., Koltun, V., 2021. Vision transformers for dense prediction. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 12159–12168.

Rosinol, A., Leonard, J. J., Carlone, L., 2023. Nerf-slam: Real-time dense monocular slam with neural radiance fields. *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3437–3444.

Schönberger, J. L., Zheng, E., Frahm, J.-M., Pollefeys, M., 2016. Pixelwise view selection for unstructured multi-view stereo. *Computer Vision – ECCV 2016*, 501–518.

Schönberger, J. L., Frahm, J.-M., 2016. Structure-from-motion revisited. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4104–4113.

Schönberger, J. L., Hardmeier, H., Sattler, T., Pollefeys, M., 2017. Comparative evaluation of hand-crafted and learned local features. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6959–6968.

Sucar, E., Liu, S., Ortiz, J., Davison, A. J., 2021. imap: Implicit mapping and positioning in real-time. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 6209–6218.

Teed, Z., Deng, J., 2020. Raft: Recurrent all-pairs field transforms for optical flow. *Computer Vision – ECCV 2020: 16th European Conference*, 402–419.

Teed, Z., Deng, J., 2021. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34, 16558–16569.

Umeyama, S., 1991. Least-Squares Estimation of Transformation Parameters Between Two Point Patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4), 376–380.

Wan, G., Huang, R., Xu, Y., Ye, Z., You, Q., Yan, X., Tong, X., 2025. Efficient Phase Congruency-Based Feature Transform for Rapid Matching of Planetary Remote Sensing Images. *IEEE Geoscience and Remote Sensing Letters*, 22, 1–5.

Wang, S., Clark, R., Wen, H., Trigoni, N., 2017a. Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2043–2050.

Wang, W., Hu, Y., Scherer, S., 2021. Tartanvo: A generalizable learning-based vo. *Proceedings of the Conference on Robot Learning*, 1761–1772.

Wang, Y., Zhang, W., An, P., 2017b. A survey of simultaneous localization and mapping on unstructured lunar complex environment. *AIP Conference Proceedings*, 1890(1), 030010.

Xia, Y., Gladkova, M., Wang, R., Li, Q., Stilla, U., Henriques, J. F., Cremers, D., 2023. CASSPR: Cross Attention Single Scan Place Recognition . *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 8427–8438.

Xia, Y., Xu, Y., Li, S., Wang, R., Du, J., Cremers, D., Stilla, U., 2021. Soe-net: A self-attention and orientation encoding network for point cloud based place recognition. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11343–11352.

Ye, P., Huang, R., Xu, Y., Li, W., Ye, Z., Tong, X., 2025. 3D Morphometry of Martian Craters from HRSC DEMs Using a Multi-Scale Semantic Segmentation Network and Morphological Analysis. *Icarus*, 426, 116358.

Yu, A., Ye, V., Tancik, M., Kanazawa, A., 2021. pixelnerf: Neural radiance fields from one or few images. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4576–4585.

Zhang, Y., Tosi, F., Mattoccia, S., Poggi, M., 2023. Go-slam: Global optimization for consistent 3d instant reconstruction. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 3704–3714.

Zhong, J., Yan, J., Li, M., Barriot, J.-P., 2023. A Deep Learning-Based Local Feature Extraction Method for Improved Image Matching and Surface Reconstruction from Yutu-2 PCAM Images on the Moon. *ISPRS Journal of Photogrammetry and Remote Sensing*, 206, 16–29.

Zhu, Z., Peng, S., Larsson, V., Xu, W., Bao, H., Cui, Z., Oswald, M. R., Pollefeys, M., 2022. Nice-slam: Neural implicit scalable encoding for slam. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12776–12786.