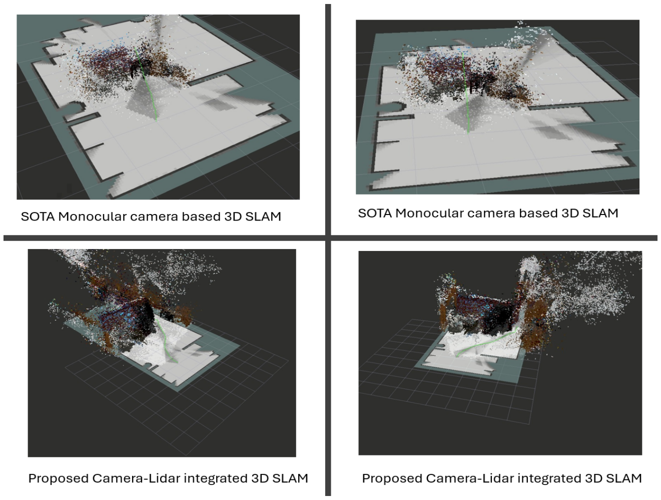
3D Reconstruction via Camera-lidar (2D) Fusion for Mobile Robots: A Gaussian Splatting Approach

Ajay Kumar Sandula[1], Shriram Damodaran[2*], Suhas Nagaraj[3*],
Debasish Ghose[4], Pradipta Biswas[5]

***Abstract*— We present a novel 3D reconstruction-based SLAM (Simultaneous Localization and Mapping) approach for robots that leverage multimodal sensory input data, including a camera and a 2D lidar. By integrating these inputs with the gaussian splatting technique, our method significantly enhances performance over traditional SLAM approaches. Traditional SLAM techniques often struggle with the limitations of monocular vision and fail to accurately map and locate objects in dynamic and cluttered environments. Purely relying on camera to localize the robot and map creation is challenging in the presence of dynamic obstacles in the scene. To address this, we proposed a multimodal sensor fusion-based 3D reconstruction. Our approach employs lidar-based localization to achieve precise positioning of both the camera and the robot, while utilizing the gaussian splatting technique for robust environmental mapping and 3D reconstruction. This approach is robust to dynamic obstacles in the scene. We have conducted extensive experiments in various real-world and simulated environments, demonstrating that our method not only outperforms traditional monocular SLAM approaches but also achieves higher accuracy in terms of localization and constructed map. Our results demonstrate substantial improvements in 3D reconstruction for mobile robots, achieving reduced computational load, higher FPS and enhanced scaling accuracy.**

I. INTRODUCTION

Rapid advancements in robotics and autonomous systems have revolutionized industries and daily life, enhancing efficiency, safety, and productivity. From autonomous vehicles and robotic manufacturing to smart surveillance and home automation, robots are increasingly performing complex tasks in dynamic environments [1]. A critical capability underpinning these applications is the ability of robots to perceive, understand, and interact with their surroundings in three dimensions. This capability is facilitated by Simultaneous Localization and Mapping (SLAM), a technology that allows robots to build a map of an unknown environment while simultaneously determining their location within it.



SOTA Monocular camera based 3D SLAM

SOTA Monocular camera based 3D SLAM

Proposed Camera-Lidar integrated 3D SLAM

Proposed Camera-Lidar integrated 3D SLAM

Fig. 1: Real-time 3D reconstruction with and without lidar-camera integration, showing unscaled 3D maps without lidar integration.

3D reconstruction and Semantic SLAM are pivotal in various applications, including robotics, augmented/virtual reality (AR/VR), and autonomous driving [2], [3]. It enables a robot to create a detailed 3D map of its environment and understand the semantic context of the objects it encounters. However, conventional SLAM techniques often rely on monocular or RGB-D cameras, which can struggle with limitations such as the inability to handle dynamic obstacles, occlusions, or environments with less features. This makes them prone to inaccuracies in both mapping and localization, particularly in cluttered and dynamic settings.

Monocular vision-based SLAM often fails in environments [4] with accurate scale (as observed in Figure 1), poor lighting, or low texture, leading to inaccurate or incomplete reconstructions. While lidar-based SLAM can provide precise depth measurements and localization, it lacks the rich color and texture information needed for detailed 3D reconstruction and semantic understanding. Additionally, pure lidar systems can struggle with small or transparent objects that do not reflect laser beams well. The integration of visual and lidar data for SLAM aims to leverage the strengths of both sensors, combining the precise depth measurements of lidar with the detailed visual information of cameras. However, existing multimodal fusion approaches still face

This work was not supported by any organization

*2 and 3 are equally contributing authors

[1]Ajay Kumar Sandula is a PhD student at the Department of Cyber-Physical Systems, Indian Institute of Science, Bangalore, Karnataka, 560012, India. ajaykumars@iisc.ac.in

[2]Shriram Damodaran is an undergraduate student at Dr B R Ambedkar National Institute of Technology, Jalandhar, India. dshriram27@gmail.com

[3]Suhas Nagaraj is a master's student at the University of Maryland, College Park, United States. suhas99@umd.edu

[4]Debasish Ghose is a professor with the Robert Bosch Centre for Cyber-Physical Systems and the Department of Aerospace Engineering, Indian Institute of Science, Bangalore, Karnataka, 560012, India. dghose@iisc.ac.in

[5]Pradipta Biswas is an associate professor with the Robert Bosch Centre for Cyber-Physical Systems and the Department of Design and Manufacturing, Indian Institute of Science, Bangalore, Karnataka, 560012, India. pradipta@iisc.ac.in

challenges in real-time processing, accurate data alignment, and robustness in dynamic environments.

We propose a novel 3D reconstruction approach using multimodal sensor fusion of camera and lidar data with a gaussian splatting [5] technique. By combining visual data for texture and color with lidar's depth precision, our method creates a unified 3D representation that generates accurate, high-fidelity maps in real time. The gaussian splatting technique provides a smooth and computationally efficient representation, enhancing SLAM performance in complex environments. Our architecture overcomes the limitations of traditional SLAM by enhancing mapping accuracy, reducing computational load by 34%, and increasing FPS, resulting in more efficient and responsive 3D reconstruction.

II. Background

SLAM has evolved from basic geometric mapping with Kalman and particle filters [6], [7] to advanced Visual SLAM [8]–[14] using monocular, stereo, and RGB-D cameras. Recent innovations include event-based SLAM [15]–[19] for high-motion environments and Semantic SLAM, which integrates object recognition to enhance environmental understanding in dynamic settings. Methods like DS-SLAM [20], DynaSLAM [21], YOLO-SLAM [22], PSPNet-SLAM [23] and SGS-SLAM [24] combine semantic segmentation with traditional SLAM, improving accuracy in cluttered environments. A notable advancement is MonoGS [25], which uses monocular cameras and Gaussian splatting for real-time 3D reconstruction. Instead of traditional point clouds or voxel grids, MonoGS employs a 3D Gaussian representation, differentiable splatting, adaptive density control, and camera tracking, significantly advancing monocular visual SLAM capabilities.

At its core, gaussian representation of the environment utilizes 3D gaussians characterized by their position μ, anisotropic covariance Σ, which defines their size and orientation in 3D space, and opacity α. The 3D gaussian is mathematically defined as:

$$G(\mathbf{x}) = e^{-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)} \tag{1}$$

This representation offers several advantages over traditional methods like point clouds and voxel grids, making it ideal for visual SLAM and 3D reconstruction. It provides a smooth, continuous, and differentiable volumetric function, enabling efficient gradient computation and real-time updates in SLAM. The gaussian splatting approach uses compact memory, allowing for scalable representation of complex geometries without the high memory overhead of voxel grids. The fast, differentiable rasterization process enables efficient rendering of 3D gaussians by projecting these 3D elements onto a 2D image plane. The projection of the 3D gaussian to the 2D plane is given by the transformation of the covariance matrix Σ into image space as follows:

$$\Sigma' = JW\Sigma W^\top J^\top \tag{2}$$

where W represents the viewing transformation and J is the Jacobian of the affine approximation of the projective transformation. This projection allows for efficient optimization of scene representations in real time. The camera tracking approach in the gaussian Splatting SLAM system involves direct optimization of camera poses against 3D gaussians, enabling robust tracking through a differentiable rendering process that updates both the camera trajectory and scene geometry simultaneously. The system minimizes the photometric residual error E_{pho}, defined as: $E_{\text{pho}} = \left\|I(G, T_{CW}) - \bar{I}\right\|_1$, where $I(G, T_{CW})$ renders the gaussians G from the camera pose T_{CW}, and $\bar{I}$ is the observed image. Additionally, when depth observations are available, the system minimizes the geometric residual E_{geo}: $E_{\text{geo}} = \left\|D(G, T_{CW}) - \bar{D}\right\|_1$, where $D(G, T_{CW})$ is the depth rasterization and $\bar{D}$ is the observed depth. This system utilizes an analytic Jacobian on the Lie group of the camera pose, isotropic gaussian shape regularization, and dynamic gaussian allocation and pruning to maintain an accurate and efficient scene representation during incremental SLAM. The proposed camera tracking uses gradient descent to optimize the camera pose directly against observed 3D Gaussians. It iteratively adjusts the camera's position and orientation by minimizing the reprojection error between the Gaussians and their expected positions in the image plane, ensuring accurate alignment of the camera trajectory with the dynamic scene during SLAM.

Relying solely on RGB data for depth estimation is computationally demanding, reducing FPS and affecting real-time performance. To enhance efficiency and accuracy, we integrate monocular depth estimation methods like UniDepth [26] and Depth Anything [27]. Additionally, the integration of 2D lidar data corrects scale misalignments, ensuring the reconstructed 3D environment aligns with real-world dimensions, thereby improving SLAM accuracy and navigation reliability.

III. Problem Statement

This work aims to achieve precise 3D surface reconstruction using a mobile robot equipped with a monocular camera and a 2D lidar. The camera provides 2D images with color, texture, and intrinsic parameters, while the lidar delivers accurate depth measurements of the environment. By fusing these complementary data sources, we aim to create a detailed 3D map that captures both the geometry and appearance of the environment.

Achieving real-time 3D surface reconstruction that aligns accurately with real-world dimensions is a significant challenge in mobile robotics and SLAM applications. Traditional SLAM methods, denoted as $\mathcal{S}$, rely heavily on monocular vision data I_c and intrinsic parameters $\mathbf{K} = [f_x, f_y, c_x, c_y]$ to estimate depth d_i for each pixel i within the image. However, these approaches often fail to maintain accurate scaling due to the inherent limitations of monocular depth estimation, such as ambiguities in scale and sensitivity to dynamic changes within the environment. This results in 3D reconstructions, $\mathcal{M}_{\text{est}}$, that are not correctly scaled to the

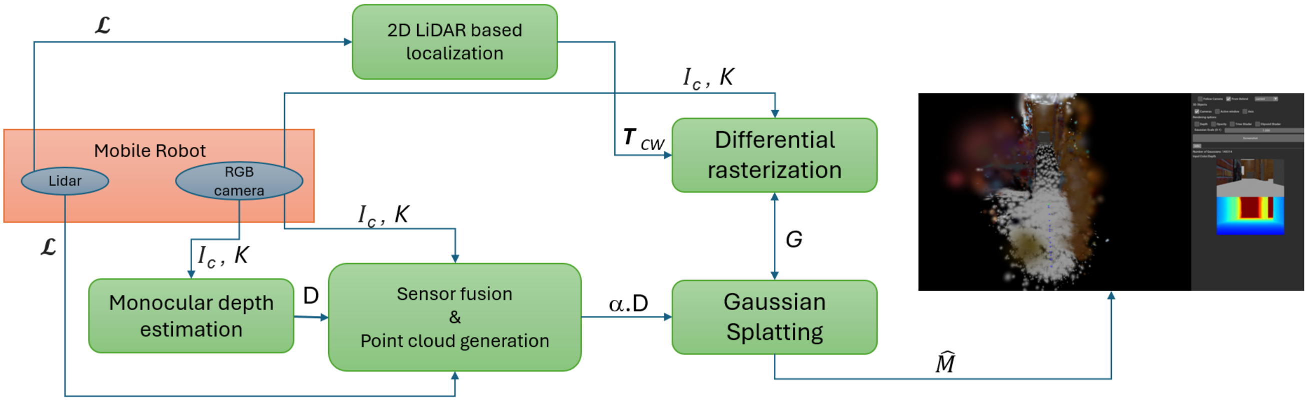


Fig. 2: Proposed architecture for real time 3D reconstruction with camera and lidar integration

real world $\mathcal{M}_{\text{real}}$, compromising their use in navigation and interaction tasks.

The input to our proposed SLAM system consists of camera images I_c, intrinsic camera parameters $\mathbf{K}$, and lidar scan data $\mathcal{L} = \{l_j\}$, where l_j represents a range measurement at a specific angle θ_j. Our goal is to compute an accurate 3D reconstruction $\hat{\mathcal{M}}$ that closely matches the true environment $\mathcal{M}_{\text{real}}$ while simultaneously determining the robot's pose $\mathbf{T}_{CW}$. The challenge with existing methods that rely solely on RGB data is that they use computationally intensive processes such as gradient descent for pose optimization $\mathbf{T}_{CW}$, which can severely impact the system's real-time performance due to the high computational load and reduced frames per second (FPS).

To address these limitations, we propose a multimodal sensor fusion approach that integrates advanced monocular depth estimation techniques like UniDepth [26] and Depth Anything [27] to enhance depth accuracy while maintaining computational efficiency. Additionally, by incorporating lidar data $\mathcal{L}$, we adjust the scale of the reconstruction using a scaling factor $\alpha_i = \frac{d_i}{l_j}$, where d_i is the depth estimated from the camera and l_j is the corresponding lidar measurement. This scaling factor refines the depth map, ensuring that the reconstructed 3D points $P_i = [x_i, y_i, z_i]$ align correctly with real-world dimensions. By employing a clustering-based approach to remove outliers in the scaling factors, we improve the robustness of the depth adjustment, resulting in a more accurate and reliable 3D surface reconstruction $\hat{\mathcal{M}}$. This integrated sensor fusion strategy significantly enhances SLAM performance, providing precise and reliable navigation capabilities in both static and dynamic environments, and making the system highly adaptable for real-world robotic applications.

IV. Methodology

The proposed architecture leverages multimodal sensory inputs from a 2D lidar and a monocular camera mounted on a mobile robot to achieve precise, real-time 3D reconstruction of the environment. The data flow across different stages involves key modules such as 2D LiDAR-based Localization, Monocular Depth Estimation, Sensor Fusion & Point Cloud Generation, Differential Rasterization, and Gaussian Splatting. This integration of camera and LiDAR data enhances SLAM performance, as illustrated in Figure 2, which provides a comprehensive overview of how each module interacts within the system.

A. System Workflow

Starting with the mobile robot, the 2D lidar provides precise distance measurements $\mathcal{L}$ that are used by the 2D lidar based localization module. We used the cartographer [28] technique, well-known for its robust localization and mapping capabilities. The cartographer uses lidar scans to generate 2D maps and estimates the pose of camera $\mathbf{T}_{\text{CW}}$, effectively anchoring the overall mapping process by providing accurate localization information. Simultaneously, the monocular camera captures RGB images I_c, which, along with the intrinsic parameters $\mathbf{K}$, serve as inputs for monocular depth estimation models like UniDepth or Depth Anything. These models output an initial depth map $\mathbf{D}$, which represents the depth information derived solely from visual data.

However, monocular depth estimation often suffers from scale ambiguity and inaccuracies in real-world dimension representation. To mitigate these issues, the Sensor Fusion and Point Cloud Generation module combines the depth map $\mathbf{D}$ with the lidar measurements $\mathcal{L}$. This fusion process adjusts the scale of the depth map by computing a scaling factor α that aligns the camera-derived depth values with lidar data, producing a scaled depth map $\alpha \cdot \mathbf{D}$. This scaled depth map is then used to generate a point cloud, which accurately represents the spatial layout of the environment.

The generated point cloud is subsequently passed to the gaussian Splatting module, where it undergoes further refinement and 3D mapping. gaussian splatting, a technique that represents the scene with continuous gaussian functions rather than discrete points, enables the creation of high-fidelity 3D reconstructions $\hat{\mathcal{M}}$. The output from gaussian Splatting feeds into the Differential Rasterization module, which dynamically optimizes the scene by adjusting both

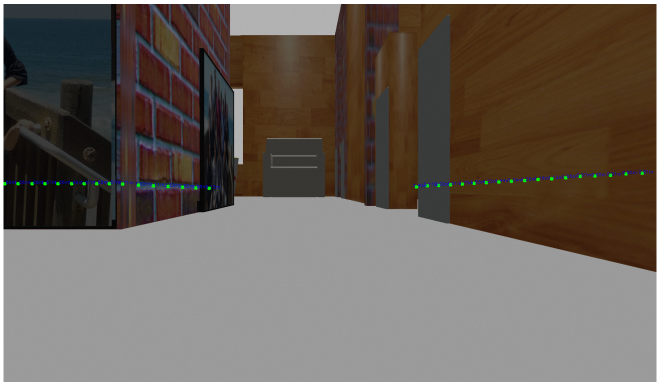
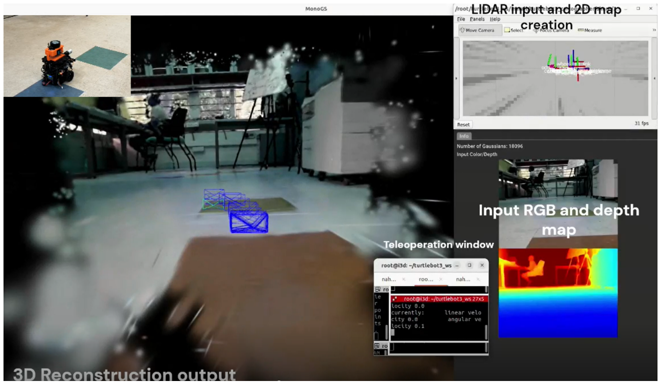


Fig. 3: Lidar points projected onto the camera image, demonstrating depth alignment for 3D mapping.



Number of Gaussians: 18096
Input Color/Depth

```
root@i3d: ~/turtlebot3_ws
nah...   roo...   nah...
root@i3d: ~/turtlebot3_ws 27x5
locity 0.0
currently:      linear velo
city 0.0         angular ve
locity 0.1
```

Fig. 4: Real-time 3D reconstruction in indoor environments using a real robot

camera poses and scene geometry. This feedback loop ensures continuous refinement of the 3D map $\hat{\mathcal{M}}$, integrating new sensory data and maintaining accurate and up-to-date representations of the environment.

Overall, this architecture combines the complementary strengths of lidar and monocular vision, addressing the limitations of each to produce a robust, scalable, and accurate 3D reconstruction. By effectively fusing sensory data and leveraging advanced techniques like gaussian splatting, the system enhances SLAM performance, providing a reliable framework for navigation and interaction in dynamic and complex environments.

B. Camera-lidar Integration

The integration of camera and lidar data enhances the accuracy and scale alignment of 3D surface reconstructions in SLAM systems. This process involves transforming lidar scan data into the camera frame, projecting the transformed points onto the image plane, and calculating orthogonal distances relative to the camera as shown in the Figure 3 By combining these data sources, our system produces a 3D map $\hat{\mathcal{M}}$ that accurately reflects real-world dimensions, overcoming common scaling limitations of monocular SLAM.

The lidar provides radial distance measurements $\mathcal{L} = \{l_j\}$ at angles θ_j, initially expressed in polar coordinates. These measurements are first transformed into cartesian coordinates through the mapping function $f_{\text{polar}\to\text{cart}}$, defined as: $f_{\text{polar}\to\text{cart}} : (l_j, \theta_j) \mapsto \mathbf{P}_l = [x_l \;\; y_l \;\; 0 \;\; 1]^\top$, where (x_l, y_l) are the Cartesian coordinates of the lidar point. Next, these points are mapped into the camera frame using the transformation matrix $\mathbf{T}_{\text{lidar}\to\text{camera}}$ via the function $f_{\text{transform}}$: $f_{\text{transform}} : \mathbf{P}_l \mapsto \mathbf{P}_c = \mathbf{T}_{\text{lidar}\to\text{camera}} \cdot \mathbf{P}_l = [x_c \; y_c \; z_c \; 1]^\top$,, where $\mathbf{P}_c$ represents the transformed coordinates in the camera frame. The z-component, z_c, provides the orthogonal distance $d_\perp = z_c$ from the camera to each point. The transformed 3D points are then projected onto the 2D image plane using the camera's intrinsic matrix $\mathbf{K}$ through the mapping function $f_{\text{proj}} : \mathbf{P}_c \mapsto \mathbf{P}_{\text{image}} = \mathbf{K}[x_c \; y_c \; z_c]^\top = [f_x x_c + c_x z_c \; f_y y_c + c_y z_c \; z_c]^\top$, which outputs the 2D projection of the lidar points in the camera's image space. Finally, the pixel coordinates (u_i, v_i) are obtained by normalizing the projected coordinates via the function f_{norm}:$\mathbf{P}_{\text{image}} \mapsto (u_i, v_i) = \left(\frac{f_x x_c + c_x z_c}{z_c}, \frac{f_y y_c + c_y z_c}{z_c}\right)$.

This sequence of transformations ensures that the lidar points are accurately aligned with the camera's perspective, enhancing the overall accuracy of the 3D reconstruction $\hat{\mathcal{M}}$. To accurately scale the camera-derived depth map $\mathbf{D} : (u_j, v_j) \to d_j$ which is obtained from the UNIDepth or Depth AnyThing AI models, lidar measurements $\mathcal{L} = (u_j, v_j, z_j)$ are used, where (u_j, v_j) are pixel coordinates and z_j are the corresponding lidar depths in camera coordinate frame. For each (u_j, v_j), the corresponding camera depth d_j is extracted, forming the mapping function $\mathbf{D} = d_j$. Valid pairs are filtered to avoid errors, and a scale factor $\alpha = \frac{\sum_j l_j}{\sum_j d_j}$ is computed, aligning the camera depth values with the lidar measurements. The scaled depth map is then $\mathbf{D}_{\text{scaled}} = \alpha \cdot \mathbf{D}$, correcting scale discrepancies to ensure that the 3D reconstruction $\hat{\mathcal{M}}$ matches real-world dimensions. This approach mitigates the inherent scale ambiguities of monocular depth estimation, resulting in a robust and scalable SLAM framework capable of precise navigation in dynamic environments as shown in the Figure 4.

V. Results

In this study, we developed four distinct approaches by varying two key parameters: the depth estimation model and the localization method. The first parameter, the depth estimation model, involved two advanced monocular depth estimation techniques: UniDepth and Depth Anything. The second parameter revolves around the localization method, which we varied between camera-based (explained in Section II) and lidar-based localization (explained in Section IV). We benchmarked our proposed approaches against MonoGS, a state-of-the-art monocular SLAM method for 3D reconstruction. Additionally, we integrated MonoGS with lidar-based localization, excluding any depth estimation models, to provide a better comparative baseline with our proposed architectures. To evaluate the proposed approaches, we set up a ROS2-Gazebo simulation environment featuring a mobile robot equipped with a monocular camera and a 2D lidar sensor. The robot was deployed in a realistic house environment

TABLE I: RMSE Values for Different Configurations with lidar and Non-lidar for position and orientation

Configuration	X (m)	Y (m)	Z (m)	Roll (rad)	Pitch (rad)	Yaw (rad)	Position	Orientation
UniDepth with lidar	0.0466	0.0491	0.1178	5.78e-05	5.178e-03	0.4195	0.0784	0.2422
UniDepth without lidar	3.0497	1.5722	0.9188	5.48e-05	5.178e-03	0.4142	2.0507	0.2392
Depth Anything with lidar	0.2643	0.1075	0.1040	5.68e-05	5.173e-03	0.7058	0.1753	0.4075
Depth Anything without lidar	5.2179	2.0217	0.3157	5.51e-05	5.179e-03	0.7337	3.2359	0.4236
MonoGS with lidar	0.0306	0.0452	0.1085	5.60e-05	5.179e-03	0.4514	0.0701	0.2606
MonoGS without lidar	3.4925	2.1109	0.8798	5.65e-05	5.178e-03	0.4231	2.4102	0.2443

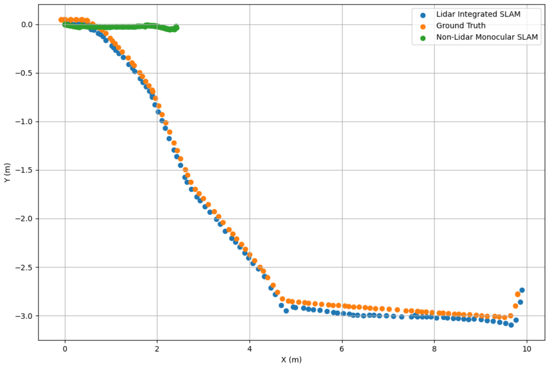


Fig. 5: Location of robot as perceived by lidar integrated and unintegrated monocular 3D SLAM

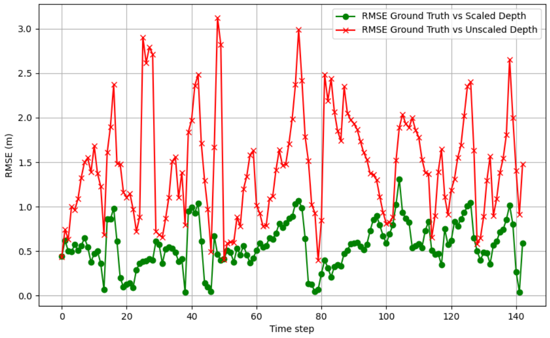


Fig. 6: RMSE Comparison of Scaled vs. Unscaled Point Clouds with Ground Truth lidar for UniDepth with lidar case

where it navigated to collect data streams from both sensors. These recorded data streams were then used as consistent inputs across all the pipelines mentioned above, ensuring a fair and comprehensive benchmarking of our approaches against each other and against the baseline methods.

TABLE II: Computational Load Comparison of Different Configurations (i9-12th gen processor, RTX 4090 GPU)

Configuration	Computational Load across all cores (%)
Depth Anything (DAT) with lidar	1305
Depth Anything (DAT) without lidar	1445
UniDepth with lidar	1024
UniDepth without lidar	1243
MonoGS with lidar	1343
MonoGS without lidar	1549

Key performance indicators (KPIs) such as computational load, point cloud scaling accuracy, Frames Per Second (FPS), and Root Mean Square Error (RMSE) values are crucial in 3D reconstruction of environments as they directly impact the system's efficiency, accuracy, and real-time capabilities. We collected continuous sensory data streams and processed them through each configuration of interest for 3D reconstruction. Table I and the Figure 5 shows the RMSE errors relative to the ground truth within the gazebo simulator. Computational load determines the feasibility of deploying the system in resource-constrained settings, while point cloud scaling accuracy ensures that the reconstructed map aligns with real-world dimensions, essential for precise navigation. Table II shows the computational load details. In the Table III, FPS measures the system's responsiveness, affecting how quickly it can update the environment model, and RMSE values of the Table IV highlight the accuracy improvements achieved by integrating lidar data, ensuring that the reconstructed 3D environment is reliable and suitable for real-world robotic applications.

TABLE III: FPS Comparison among benchmarks

Configuration	FPS
UniDepth with lidar	2.176
UniDepth without lidar	1.201
Depth Anything (DAT) with lidar	2.566
Depth Anything (DAT) without lidar	1.564
MonoGS with lidar	1.593
MonoGS without lidar	1.418

To evaluate the point cloud scaling accuracy, which reflects the performance of the sensor fusion model, we compared the averaged Root Mean Square Error (RMSE) between the scaled and unscaled point clouds against the ground truth lidar measurements as shown in the Figure 6. The scaled point cloud results from integrating lidar measurements with monocular depth estimation, providing an adjusted representation that aligns with real-world dimensions. Table IV has the RMSE results for each configuration.

TABLE IV: RMSE Comparison of Scaled and Unscaled Point Clouds with Ground Truth lidar Measurements

Configuration	Avg RMSE (Scaled Z)	Avg RMSE (Unscaled Z)
Depth Anything with lidar	0.5404	1.4588
UniDepth with lidar	0.3980	1.4600

VI. Discussion

The results presented in this study demonstrate the effectiveness of integrating lidar with monocular depth estimation models for enhancing the 3D reconstruction capabilities of SLAM systems. By varying the depth estimation techniques and localization methods, we evaluated four distinct configurations and compared them against the established MonoGS SLAM benchmark. The primary KPIs, including computational load, FPS, and RMSE values, provide critical insights into the performance and applicability of each configuration in real-world scenarios.

Performance Analysis of Depth Estimation Models: The RMSE values presented in Table I highlight the substantial impact of incorporating lidar-based localization on point cloud scaling accuracy. The configurations using lidar consistently demonstrated lower RMSE values compared to their non-lidar counterparts, particularly in position accuracy. For instance, UniDepth with lidar achieved a significantly lower RMSE for position (0.0784) compared to UniDepth without lidar (2.0507), indicating that lidar integration effectively mitigates scaling errors inherent in monocular depth estimation methods. The consistency in orientation errors across all configurations suggests that while lidar primarily enhances spatial accuracy, it does not significantly alter the orientation estimation.

Computational Load Considerations: Table II reveals that configurations using lidar tend to have lower computational loads compared to those relying solely on camera-based localization. For example, UniDepth with lidar exhibited a computational load of 1024%, while the non-lidar variant consumed 1243%. This reduction is crucial for real-time performance, as lower computational demands enable higher responsiveness and faster data processing, which are essential for applications requiring real-time 3D reconstruction. The integration of lidar appears to streamline the localization process, reducing the need for computationally intensive pose optimization techniques typically required in purely vision-based systems.

FPS Analysis and Real-Time Capabilities: As shown in Table III, the FPS performance of lidar-integrated configurations consistently outperforms those without lidar. Depth Anything with lidar achieved the highest FPS (2.566), demonstrating superior efficiency in processing sensory data compared to its camera-only counterpart (1.564 FPS). This finding underscores the advantage of lidar in enhancing system responsiveness, making it more suitable for dynamic environments where rapid updates to the 3D map are necessary.

Impact on Scaling Accuracy: The RMSE comparisons between scaled and unscaled point clouds, as detailed in Table IV, further validate the benefits of lidar integration. The scaled configurations (e.g., Depth Anything with lidar and UniDepth with lidar) showed substantial reductions in RMSE when compared to unscaled point clouds, emphasizing the critical role of lidar in correcting depth inaccuracies. This adjustment not only improves the overall accuracy of the 3D map but also ensures that the spatial data aligns with real-world measurements, which is essential for precise navigation and interaction in complex environments.

Benchmarking Against MonoGS: The results demonstrate that while MonoGS serves as a robust benchmark, the integration of lidar enhances both computational efficiency and spatial accuracy, particularly in environments with complex geometries and dynamic obstacles. The lower RMSE values and computational load observed in the proposed configurations suggest that integrating lidar with advanced depth estimation models provides a more balanced and scalable approach for real-time 3D reconstruction.

Practical Implications: The observed improvements in scaling accuracy, computational efficiency, and real-time performance have significant implications for SLAM applications in autonomous navigation, robotic manipulation, and augmented reality. The reduction in computational load and enhanced FPS make the proposed configurations suitable for deployment in resource-constrained settings, such as mobile robots operating in cluttered and dynamic environments. By aligning the reconstructed 3D environment with real-world dimensions, the integrated approach ensures reliable spatial understanding, crucial for task execution and decision-making in autonomous systems.

In summary, integrating lidar with monocular depth estimation enhances camera-only SLAM systems, offering a robust, scalable solution for precise 3D reconstruction. Combining advanced depth models with lidar-based localization improves SLAM performance, paving the way for more accurate and efficient autonomous systems.

VII. Future work & Conclusion

The integration of lidar with monocular depth estimation models, such as UniDepth and Depth Anything, significantly enhances the accuracy and scalability of 3D reconstruction in SLAM systems. Our results demonstrate that lidar-based localization reduces scaling errors and improves spatial accuracy compared to camera-only methods, as shown by the lower RMSE values. The reduced computational load and increased FPS in lidar-integrated configurations highlight the importance of multimodal sensor fusion, particularly in dynamic environments where rapid updates and efficient processing are crucial. Benchmark comparisons with MonoGS further validate that our approach offers superior performance, especially in complex and cluttered settings.

Future work would aim to explore incorporating semantic perception models to enhance the environmental understanding of SLAM systems. Integrating advanced object detection and segmentation could provide valuable contextual information, aiding in better decision-making and obstacle avoidance. This direction aligns with the goal of developing intelligent SLAM systems that not only reconstruct but also interpret complex environments, enhancing their adaptability and effectiveness in autonomous navigation and robotic applications.

References

[1] Farinelli, A., Iocchi, L. and Nardi, D., 2004. Multirobot systems: a classification focused on coordination. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 34(5), pp.2015-2028.

[2] Mukhopadhyay, A. and Biswas, P., 2024. Advancements in Deep-Learning-Based Object Detection in Challenging Environments. Wireless World Research and Trends Magazine, pp.1-6.

[3] Wei, Y., Zhao, L., Zheng, W., Zhu, Z., Zhou, J. and Lu, J., 2023. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 21729-21740).

[4] Macario Barros, A., Michel, M., Moline, Y., Corre, G. and Carrel, F., 2022. A comprehensive survey of visual slam algorithms. Robotics, 11(1), p.24.

[5] Fei, B., Xu, J., Zhang, R., Zhou, Q., Yang, W. and He, Y., 2024. 3d gaussian splatting as new era: A survey. IEEE Transactions on Visualization and Computer Graphics.

[6] Dissanayake, M.G., Newman, P., Clark, S., Durrant-Whyte, H.F. and Csorba, M., 2001. A solution to the simultaneous localization and map building (SLAM) problem. IEEE Transactions on robotics and automation, 17(3), pp.229-241.

[7] Luo, J. and Qin, S., 2018. A fast algorithm of simultaneous localization and mapping for mobile robot based on ball particle filter. IEEE Access, 6, pp.20412-20429.

[8] Se, S., Lowe, D. and Little, J., 2002. Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks. The international Journal of robotics Research, 21(8), pp.735-758.

[9] Klein, G. and Murray, D., 2007, November. Parallel tracking and mapping for small AR workspaces. In 2007 6th IEEE and ACM international symposium on mixed and augmented reality (pp. 225-234). IEEE.

[10] Howard, A., 2008, September. Real-time stereo visual odometry for autonomous ground vehicles. In 2008 IEEE/RSJ international conference on intelligent robots and systems (pp. 3946-3952). IEEE.

[11] Henry, P., Krainin, M., Herbst, E., Ren, X., Fox, D. (2014). RGB-D Mapping: Using Depth Cameras for Dense 3D Modeling of Indoor Environments. In: Khatib, O., Kumar, V., Sukhatme, G. (eds) Experimental Robotics. Springer Tracts in Advanced Robotics, vol 79. Springer, Berlin, Heidelberg.

[12] Xu, J., Cao, H., Li, D., Huang, K., Qian, C., Shangguan, L. and Yang, Z., 2020, July. Edge assisted mobile semantic visual SLAM. In IEEE INFOCOM 2020-IEEE Conference on computer communications (pp. 1828-1837). IEEE.

[13] Peng, Q., Xiang, Z., Fan, Y., Zhao, T. and Zhao, X., 2022. RWT-SLAM: Robust visual SLAM for highly weak-textured environments. arXiv preprint arXiv:2207.03539.

[14] Naveed, K., Anjum, M.L., Hussain, W. and Lee, D., 2022. Deep introspective SLAM: Deep reinforcement learning based approach to avoid tracking failure in visual SLAM. Autonomous Robots, 46(6), pp.705-724.

[15] Rebecq, H., Horstschaefer, T. and Scaramuzza, D., 2017. Real-time visual-inertial odometry for event cameras using keyframe-based nonlinear optimization.

[16] Chamorro, W., Sola, J. and Andrade-Cetto, J., 2022. Event-based line SLAM in real-time. IEEE Robotics and Automation Letters, 7(3), pp.8146-8153.

[17] Huang, K., Zhang, S., Zhang, J. and Tao, D., 2023. Event-based simultaneous localization and mapping: A comprehensive survey. arXiv preprint arXiv:2304.09793.

[18] Liu, P., Zuo, X., Larsson, V. and Pollefeys, M., 2021. MBA-VO: Motion blur aware visual odometry. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 5550-5559).

[19] Hidalgo-Carrió, J., Gallego, G. and Scaramuzza, D., 2022. Event-aided direct sparse odometry. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5781-5790).

[20] Yu, C., Liu, Z., Liu, X.J., Xie, F., Yang, Y., Wei, Q. and Fei, Q., 2018, October. DS-SLAM: A semantic visual SLAM towards dynamic environments. In 2018 IEEE/RSJ international conference on intelligent robots and systems (IROS) (pp. 1168-1174). IEEE.

[21] Bescos, B., Campos, C., Tardós, J.D. and Neira, J., 2021. DynaSLAM II: Tightly-coupled multi-object tracking and SLAM. IEEE robotics and automation letters, 6(3), pp.5191-5198.

[22] Wu, W., Guo, L., Gao, H., You, Z., Liu, Y. and Chen, Z., 2022. YOLO-SLAM: A semantic SLAM system towards dynamic environment with geometric constraint. Neural Computing and Applications, pp.1-16.

[23] Long, X., Zhang, W. and Zhao, B., 2020. PSPNet-SLAM: A semantic SLAM detect dynamic object by pyramid scene parsing network. IEEE access, 8, pp.214685-214695.

[24] Li, M., Liu, S. and Zhou, H., 2024. Sgs-slam: Semantic gaussian splatting for neural dense slam. arXiv preprint arXiv:2402.03246.

[25] Matsuki, H., Murai, R., Kelly, P.H. and Davison, A.J., 2024. Gaussian splatting slam. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 18039-18048).

[26] Piccinelli, L., Yang, Y.H., Sakaridis, C., Segu, M., Li, S., Van Gool, L. and Yu, F., 2024. UniDepth: Universal Monocular Metric Depth Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10106-10116).

[27] Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J. and Zhao, H., 2024. Depth anything: Unleashing the power of large-scale unlabeled data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10371-10381).

[28] Hess, W., Kohler, D., Rapp, H. and Andor, D., 2016, May. Real-time loop closure in 2D LIDAR SLAM. In 2016 IEEE international conference on robotics and automation (ICRA) (pp. 1271-1278). IEEE.