CSIAM Trans. Appl. Math.
doi: 10.4208/csiam-am.SO-2025-0070

Vol. 7, No. 3, pp. 507-528
May 2026

Structure-Aware Indoor RGB-D SLAM via Manhattan-Constrained 2D Gaussian Splatting

Wenwu Guo[1], Xia Yuan[1,2,*], Yanli Liu[2], Xiangyu Wu[1], Wenyi Ge[1], Guanyu Xing[2], Jing Hu[1] and Xi Wu[1]

[1] *School of Computer Science, Chengdu University of Information Technology, Chengdu 610225, China.*
[2] *School of Computer Science, Sichuan University, Chengdu 610065, China.*

Received 1 July 2025; Accepted 8 August 2025

Abstract. Accurate and layout-consistent reconstruction remains a key challenge in indoor simultaneous localization and mapping (SLAM) due to the prevalence of planar and axis-aligned structures. Traditional visual and RGB-D SLAM methods often suffer from incomplete geometry and weak structural reasoning, while NeRF-based SLAM improves fidelity but is computationally expensive and unsuitable for real-time use. 3D Gaussian splatting offers improved efficiency but lacks structural priors, often resulting in distortions in structured scenes. To address these issues, we propose a structure-aware SLAM framework based on 2D Gaussian splatting, which provides efficient, view-consistent mapping. We introduce a lightweight regularization scheme under the Manhattan-world assumption to align Gaussian orientations and positions with dominant axes, improving layout consistency and geometric fidelity. Extensive experiments on Replica and TUM-RGBD datasets demonstrate that our method consistently outperforms existing SLAM baselines in terms of geometric accuracy and edge preservation across multiple indoor scenes.

AMS subject classifications: 68T45, 93C85

Key words: SLAM, 2DGS, Manhattan-regularized.

1 Introduction

Indoor simultaneous localization and mapping is a core technology for applications such as autonomous navigation, augmented reality (AR), and indoor robotics. It supports real-time 3D perception and spatial understanding, forming the foundation for tasks like object interaction and human-environment collaboration. Classic visual SLAM systems, such as ORB-SLAM3 [3], rely on sparse keypoint-based maps and perform well in

[*] Corresponding author. *Email address:* `xyuan623@163.com` (X. Yuan)

feature-rich outdoor scenarios. However, in indoor environments characterized by repetitive layouts and textureless surfaces, these methods often suffer from tracking failure, poor reconstruction completeness, and limited structural reasoning.

Depth-based SLAM methods attempt to mitigate the limitations of sparse visual systems by directly leveraging geometric data from RGB-D sensors. Representative approaches such as KinectFusion [20] and ElasticFusion [31] support real-time dense reconstruction through volumetric or surfed fusion and improve robustness in texture-sparse indoor environments. However, these methods often suffer from surface blending artifacts caused by viewpoint inconsistency, and their underlying representations struggle to preserve structural boundaries and geometric detail, particularly in large planar scenes. Moreover, the lack of global structural priors makes it difficult to model spatial regularity in structured or repetitive indoor environments.

Neural radiance fields (NeRF) have been introduced into SLAM systems to jointly optimize camera poses and continuous volumetric scene representations, enabling photorealistic and dense reconstruction from sparse inputs [16, 21]. While offering high-quality rendering, NeRF-based SLAM methods suffer from expensive volumetric sampling and MLP-based inference, making them computationally intensive and unsuitable for real-time applications. In addition, their implicit scene representation prevents efficient map updates and hinders integration with standard pose graph optimization. More fundamentally, these methods lack structural priors and struggle to capture the spatial regularities inherent to indoor environments, such as orthogonal walls, planar boundaries, and repetitive room layouts – leading to degraded geometric consistency and layout coherence.

Recently, 3D Gaussian splatting (3DGS) has emerged as an efficient and expressive alternative to NeRF-based scene representations for real-time SLAM. By modeling volumetric geometry with view-dependent anisotropic Gaussian primitives, 3DGS significantly reduces computational overhead while maintaining high rendering fidelity, and has been adopted in dense SLAM systems such as SplaTAM [13] and MonoGS [18]. Despite their efficiency, existing 3DGS-based methods remain limited in structured indoor environments. Unconstrained primitive placement often leads to geometric distortion, especially in scenes with dominant planar and orthogonal layouts. Moreover, the absence of structure-aware regularization and high-level spatial priors results in multi-view inconsistencies and degraded layout fidelity.

2D Gaussian splatting (2DGS) [10] replaces volumetric Gaussian spheres with 2D Gaussian disks, yielding improved rendering efficiency and view-consistent modeling. Its compact, differentiable formulation makes it particularly effective for representing thin and planar structures in indoor environments, while supporting real-time optimization – an essential property for SLAM. These characteristics make 2DGS a promising scene representation for efficient indoor mapping. Despite its advantages, combining 2DGS with Manhattan-world priors for SLAM remains unexplored. This integration presents several challenges. First, 2DGS lacks explicit structural primitives, making it difficult to directly impose axis-aligned constraints. Second, its flexibility in modeling