M-SLAM: Multimodal Visual SLAM — A Leap Towards Multimodal Perception

Submission Sumber 6730

Abstract

Visual Simultaneous Localization and Mapping (V-SLAM) serves as the cornerstone of autonomous perception, enabling real-time self-localization and dense spatial reconstruction through passive visual sensing. Traditional Visual SLAM frameworks predominantly rely on monomodal sensor inputs, wherein a single camera modality—such as monocular, stereo, or RGB-D—is utilized for scene understanding and motion estimation. However, monomodal Visual SLAM, constrained by a singular sensory input, suffers from depth estimation inaccuracies, accumulated drift, occlusion-induced feature loss, and inefficient loop closure, thereby compromising localization fidelity and scene reconstruction accuracy. To address these inherent deficiencies, in this paper, we propose a novel multimodal Visual SLAM framework that leverages dual RGB-D sensor fusion, enhancing depth perception, drift mitigation, and loop closure robustness. To the best of our discernment, and in alignment with the extant corpus of literature, this work represents the first known realization of a multimodal Visual SLAM system intrinsically equipped with a loop closure module. Grounded in the architectural principles of Co-SLAM[9], our model integrates a hybridized encoding paradigm, exploits inter-sensor redundancy for enhanced pose refinement, and ensures resilient mapping in dynamically evolving environments. correction and mitigate cumulative drift. The proposed work is substantiated with a comprehensive set of experiments and encouraging results.

CCS Concepts

• **Computing methodologies**; • **Artificial intelligence**; • **Computer vision**; • **Computer vision tasks** ; • **Vision for robotics**;

Keywords

Visual SLAM, Multi-Modal Visual SLAM, Robot Path Planning

ACM Reference Format:

Anonymous Author(s). 2025. M-SLAM: Multimodal Visual SLAM — A Leap Towards Multimodal Perception. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (ACM Multimedia '25).* ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Visual Simultaneous Localization and Mapping (V-SLAM) has long occupied a central axis in autonomous robotic perception, enabling concurrent spatial understanding and ego-motion estimation through the assimilation of visual cues. Yet, the predominant SLAM paradigm remains moored to monomodal perception frameworks [24] [8], where a singular stream—monocular, stereo, or RGB-D—is entrusted with the onerous duality of localizing and reconstructing in unconstrained, non-deterministic environments. This unimodal reliance predicates the system's robustness on fragile visual conditions and introduces systemic brittleness in the presence of perceptual aliasing, occlusion, and topological ambiguities.

Despite considerable evolution in both geometric and learning-based SLAM, the literature continues to circumscribe the sensory architecture to single-view pipelines, rendering SLAM algorithms myopic to the benefits of intra-modal redundancy. Visual-Inertial Odometry (VIO), while alleviating certain failure modes, remains limited in perceptual richness and is often orthogonal in its sensing domain. Concurrently, neural representations—e.g., implicit encoding, NeRF-based radiance fields, and hash-grid parameterizations—have markedly elevated the fidelity and semantic granularity of SLAM systems. However, even these architectures remain tethered to a solitary RGB-D stream, thereby inheriting the perceptual myopia and operational fragilities endemic to monomodal systems.

In contradistinction to the prevailing epistemological framework, this work (Fig. 1) forges an innovative path by introducing a **multi-modal, dual RGB-D visual SLAM** system wherein two spatiotemporally synchronized RGB-D sensors are fused to engender redundancy-aware, perception-rich localization and mapping. This architectural bifurcation enables mutual cross-validation, occlusion compensation, and improved loop closure saliency, thereby circumventing the epistemic limitations of single-sensor vision pipelines.

Primary Contributions

- **Multimodal Dual RGB-D Fusion:** We propose a novel SLAM architecture that simultaneously ingests and fuses dual RGB-D input streams, enabling intra-modal perceptual redundancy and enhanced geometric completeness, particularly under occlusion and adverse illumination.
- **Neural and Parametric Hybrid Encoding:** Inspired by Co-SLAM [9], our framework incorporates joint coordinate encoding and parametric hash-grid representations to achieve real-time volumetric consistency and encoding efficiency, even in dynamically evolving scenes.
- **Parallelized Multi-Threaded Architecture:** The system is decomposed into a quartet of high-efficiency threads—Tracking, Main System, Loop Closure, and Online Covisibility Graph— each optimized for minimal latency and maximal throughput in heterogeneous sensing conditions.
- **Redundancy-Aware Loop Closure:** We introduce a dual-view loop closure mechanism wherein cross-sensor correlation is exploited for candidate relocalization, significantly

ACM Multimedia '25, Dublin, Ireland

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

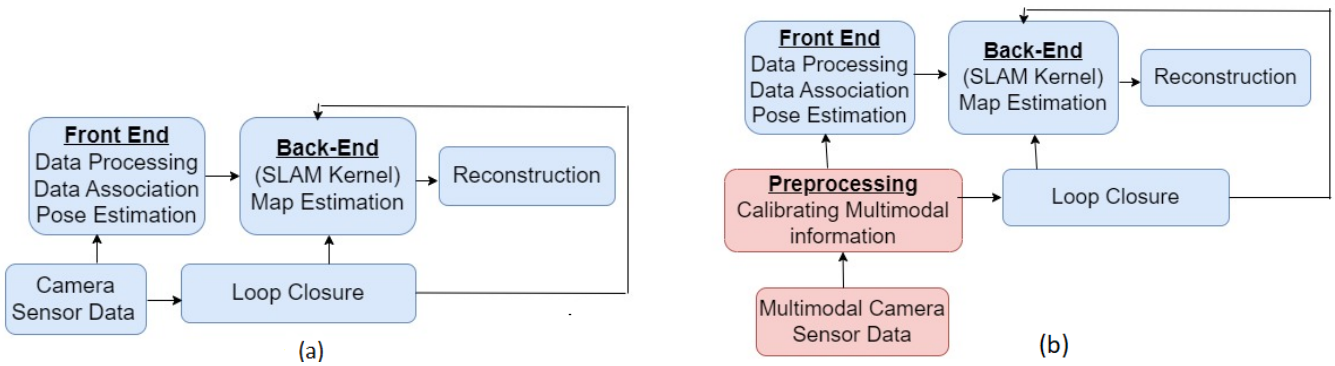

Figure 1: (a) Traditional Mono-Modal Visual SLAM (b) Proposed Multi-Modal Visual SLAM System

improving pose graph consistency and reducing cumulative drift over long sequences.

- **Robust Real-Time Performance:** Our method achieves real-time operation without sacrificing mapping fidelity, owing to the harmonized interplay of sparse bundle adjustment, learned priors, and redundant geometric cross-validation.

To the best of our discernment, this is the first documented realization of a neural SLAM architecture that internalizes dual RGB-D input within an end-to-end, loop-closed, parametric mapping pipeline. The proposed system redefines expectations for robustness and perceptual generality, marking a significant step forward in autonomous visual cognition.

The remainder of this manuscript is organized as follows: **Section 2** rigorously surveys the corpus of existing literature in mono modal and hybrid SLAM paradigms, identifying epistemic lacunae and methodological insufficiencies that motivate our formulation. **Section 3** delineates the architectural substrate of our proposed system, explicating the sensor fusion backbone, hybridized encoding schema, and multi-threaded parallelism across mapping, tracking, loop closure, and covisibility maintenance. **Section 4** provides an exhaustive experimental exposition. **Section 5** offers benchmarking the system's quantitative and qualitative performance under diverse operational regimes and against state-of-the-art baselines. **Section 6** engages in a principled ablation study to dissect the individual contributions of the system's constituent modules. **Section 7** concludes with a synthesis of insights and outlines avenues for future generalization into multi-agent and embodied SLAM contexts.

2 Related Work

The evolution of Visual SLAM has been inextricably linked to the progression of camera-centric perception systems, with canonical implementations leveraging monocular [14], stereo [7], or RGB-D [6] modalities. These mono-modal (Fig-1) frameworks, though operationally effective within constrained environmental priors, remain fundamentally tethered to the epistemological limitations of their singular sensor streams. Monocular SLAM systems, typified by ORB-SLAM [14], are intrinsically afflicted by scale ambiguity and brittle feature association under low-texture or non-Lambertian conditions. Similarly, stereo and depth-augmented approaches, such as ElasticFusion [20] and CoFusion [16], while delivering metrically scaled reconstructions, are prone to systemic degradation under conditions of sensor occlusion, depth quantization noise, and transient illumination perturbations.

More recent paradigms have gravitated towards heterogeneous sensor fusion, particularly visual-inertial odometry (VIO) [10], to circumvent the fragility of monomodal tracking. However, such cross-modal integrations typically exploit orthogonal sensor modalities (e.g., IMU, LiDAR) [2] and do not leverage intra-modal redundancy, which is particularly critical for temporal coherence and drift attenuation in visually ambiguous or topologically repetitive scenes. Furthermore, these hybridized systems [3] often employ loop closure as an auxiliary or deferred component—rather than an intrinsically fused module—thereby undermining global consistency during prolonged operation.

Several SLAM systems [17] have addressed the need for dynamic object handling [1], real-time surface fusion [20], and topological relocalization. Yet, conspicuously absent from the prevailing literature is any methodological framework that exploits *parallel redundant RGB-D modalities* as a co-optimized, tightly coupled visual backbone. This absence is not merely a consequence of oversight, but rather reflects the formidable algorithmic and computational challenges involved: sensor synchronization and calibration, spatio-temporal data association, fusion consistency, and redundant loop closure resolution introduce a combinatorial increase in both perceptual ambiguity and optimization dimensionality.

Moreover, contemporary SLAM literature is largely circumscribed by the implicit assumption that a single visual stream—augmented or not—is sufficient to model the geometric and topological intricacies of an unstructured 3D environment. This assumption breaks down under dynamic, non-rigid, or semi-structured environmental transitions where inter-sensor complementarity and redundant spatial observability are not luxuries but prerequisites. Notably, no prior SLAM architecture has capitalized on synchronized dual RGB-D streams to construct a joint pose graph with mutual reinforcement mechanisms or used such configuration to instantiate a loop closure strategy intrinsically informed by multimodal correlation. Recent advances in neural SLAM architectures have

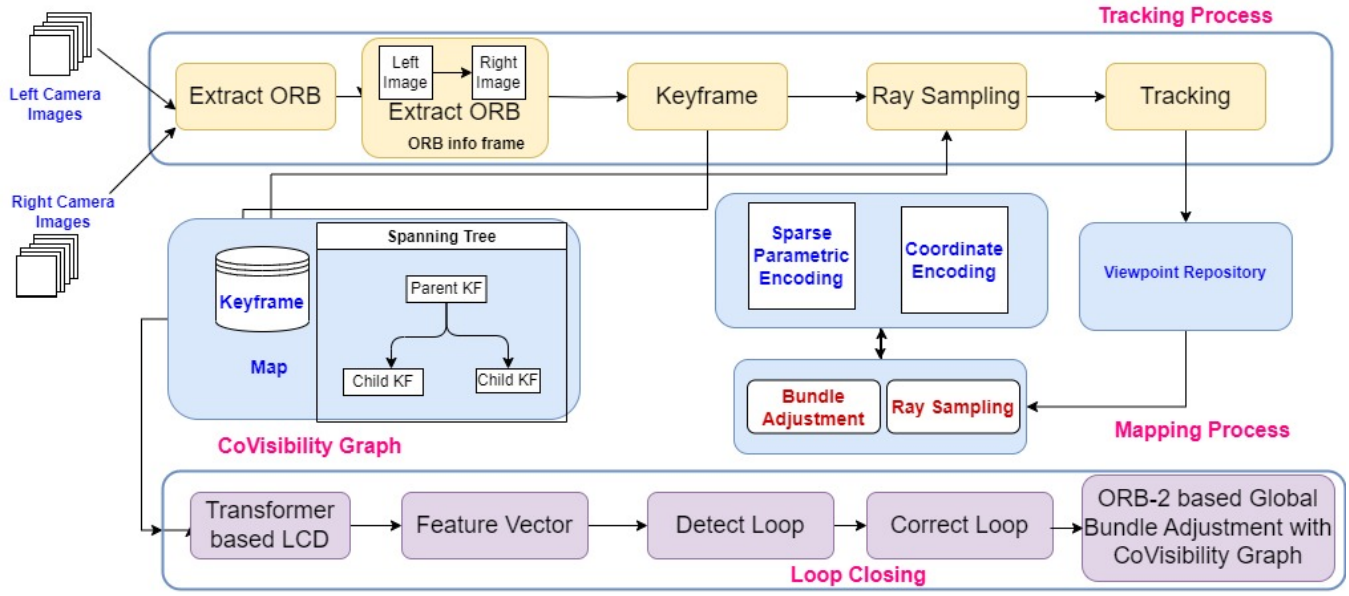

Figure 2: Overall architecture of the proposed system.

introduced a transformative lens through which SLAM can be reconceptualized—embedding implicit volumetric representations and learned scene priors directly into the mapping pipeline. Systems like Kimera [15], NICE-SLAM [26], and NeRF-SLAM [25] have demonstrated the utility of neural scene encoding, offering improvements in completeness, fidelity, and semantic integration. In particular, **Co-SLAM (2023)** [9] introduced a joint coordinate and sparse parametric encoding scheme, employing multi-resolution hash-grid representations and one-blob encodings to achieve real-time, high-fidelity RGB-D SLAM. Its neural priors enabled rapid surface convergence and resilience to visual degradation; however, the framework remained monomodal, limited to a single RGB-D stream and lacking a deeply integrated loop closure strategy.

Subsequent contributions in 2024 further enhanced neural SLAM robustness and efficiency. **EC-SLAM**[11] leveraged sparse TSDF encodings and globally constrained bundle adjustment, fusing neural priors for accurate loop closure without redundant sensing. **HERO-SLAM** [21] combined hybrid optimization and neural scene understanding to withstand adverse conditions, while **DG-SLAM** [22]integrated dynamic Gaussian Splatting for dense mapping in non-static environments. Despite these technical advancements, none of these systems employed or exploited **dual RGB-D sensor fusion**, thereby inheriting the inherent vulnerabilities of monomodal perception under occlusion, depth ambiguity, and environmental change.

This critical lacuna forms the foundational impetus of the present work. By architecting a multi-threaded, multimodal SLAM framework that fuses dual RGB-D inputs at both the front-end tracking and back-end optimization layers, we operationalize a novel class of spatial cognition. The proposed system, to the best of our discernment, represents the first-of-its-kind realization of a dual-modality visual SLAM pipeline that not only preserves real-time constraints but also integrates neural encoding priors and parametric hash-grid representations to enhance geometric fidelity, inter-frame consistency, and semantic robustness. Our method thus constitutes a transformative leap—redefining the operational envelope of Visual SLAM from monolithic sensory abstraction to redundant, cooperative, and perception-driven scene understanding.

3 Proposed Model:

The proposed Visual SLAM architecture (Fig. 2) is meticulously engineered [9] [12] [5][14] as a multi-threaded, multi-modal estimation pipeline (Fig. 3), operationalized through the concurrent execution of four interleaved yet functionally modular subsystems: the **Tracking Thread** , the **Main System Thread** , the **Loop Closure Thread**, and the **Online Covisibility Graph Thread** . The system is architected to ingest dual RGB-D sensory streams and to process them through a finely orchestrated cascade of spatial, photometric, and topological inference mechanisms, culminating in a globally consistent map and temporally coherent trajectory estimate.

At the crux of the system lies a **multi-modal dual-RGB-D sensing interface**, which simultaneously acquires redundant yet complementary visual and depth cues from stereoscopic RGB-D image pairs. These input modalities are fused at multiple stages of the computational graph to facilitate robust and redundant spatial perception, even under challenging photogeometric conditions. Each constituent thread is delineated below in terms of its functional purview and interaction within the holistic Visual SLAM ecosystem.

3.1 Tracking Thread

The **Tracking Thread** (Fig. 3) is the perceptual front-end of the system. It continuously ingests synchronized RGB-D frames from the stereo pair and fuses them into composite keyframes. Oriented FAST and Rotated BRIEF (ORB) features are extracted from both

constituent views. Depending on the system state (e.g., lost, initializing, or nominal), the thread either initiates bootstrapping, performs relocalization, or updates tracking via pose refinement using the most recent keyframe and local map. For each incoming frame, the thread estimates the camera pose via feature-based pose graph optimization. It also performs keyframe selection based on motion heuristics and map point parallax, triggering keyframe insertion into the **Online Covisibility Graph** when warranted.

3.2 Main System Thread

The central computational locus of the system, this thread is continuously attuned to the RGB-D stream encapsulated in the online map. It is event-driven, activated by keyframe insertion requests from the Tracking Thread. This module encapsulates a Neural RGB-D SLAM sub-system, leveraging sophisticated learned representations for joint depth and appearance modeling. The operational pipeline developed based on [9] is delineated below:

Ray Sampling. For every incoming RGB-D frame, pixel coordinates are unprojected into 3D camera coordinates using known intrinsics, thereby yielding rays that emanate from the camera's optical center. Each ray represents a hypothesized trajectory through the spatial extent of the scene.

Pixel Sampling. Points are discretely sampled along each ray at stratified depth intervals, serving as query loci for density and color estimation.

Parametric Encoding. The representational backbone synergistically integrates coordinate-based and parametric embeddings. Traditional MLPs (Multi-Layer Perceptrons) exhibit excellent continuity priors but suffer from sluggish convergence and catastrophic forgetting under sequential updates. Conversely, parametric encoding offers accelerated inference at the expense of reconstruction fidelity and hole-filling.

To reconcile this dichotomy, a dual encoding mechanism is employed: a coordinate-based representation augmented by sparse parametric descriptors. Specifically, One-blob encoding $\gamma(x)$ is utilized in lieu of sinusoidal positional embeddings.

The spatial domain is indexed using a multi-resolution hash-based feature grid:

$$\alpha = \{V_{\alpha_l}\}_{l=1}^{L} \tag{1}$$

Each level in the hierarchy spans a resolution range between $R_{\min}$ and $R_{\max}$. Query vectors $V_\alpha(x)$ are interpolated via trilinear sampling. The geometry decoder outputs both a signed distance function (SDF) estimate s and a latent vector h:

$$f_\tau(\gamma(x), V_\alpha(x)) \rightarrow (h, s) \tag{2}$$

Subsequently, color is predicted via an MLP:

$$f_\phi(\gamma(x), h) \rightarrow c \tag{3}$$

The set of learnable parameters is defined as $\theta = \{\alpha, \phi, \tau\}$. The integration of One-blob encoding within this hierarchical structure yields rapid convergence, high memory efficiency, and robust spatial interpolation.

Depth and Color Rendering. Given camera origin o and ray direction r, depth samples $x_i = o + t_i r$, $i \in \{1, \ldots, M\}$ are generated. Colors $\{c_i\}$ and depths $\{d_i\}$ are predicted and rendered via:

$$\hat{c} = \frac{1}{\sum_{i=1}^{M} w_i} \sum_{i=1}^{M} w_i c_i, \quad \hat{d} = \frac{1}{\sum_{i=1}^{M} w_i} \sum_{i=1}^{M} w_i d_i \tag{4}$$

Weights w_i are derived from predicted SDFs s_i using a smooth approximation:

$$w_i = \sigma\left(\frac{s_i}{t_r}\right) \sigma\left(-\frac{s_i}{t_r}\right) \tag{5}$$

where $\sigma(\cdot)$ is the sigmoid function, and t_r is the truncation threshold.

Bundle Adjustment. Optimization is conducted over both the scene parameters θ and camera poses ξ_t. Losses are defined as:

$$L_{\text{rgb}} = \frac{1}{N} \sum_{n=1}^{N} (\hat{c}_n - c_n)^2, \quad L_d = \frac{1}{|R_d|} \sum_{r \in R_d} (\hat{d}_r - D[u,v])^2 \tag{6}$$

where R_d denotes rays with valid depth supervision.

To further constrain geometry:

SDF Supervision (for points within truncation region):

$$L_{\text{sdf}} = \frac{1}{|R_d|} \sum_{r \in R_d} \frac{1}{|S_r^{\text{tr}}|} \sum_{p \in S_r^{\text{tr}}} (s_p - (D[u,v] - d))^2 \tag{7}$$

Free-space Loss (for points outside truncation region):

$$L_{\text{fs}} = \frac{1}{|R_d|} \sum_{r \in R_d} \frac{1}{|S_r^{\text{fs}}|} \sum_{p \in S_r^{\text{fs}}} (s_p - t_r)^2 \tag{8}$$

Feature Smoothness (for spatial regularization):

$$L_{\text{smooth}} = \frac{1}{|G|} \sum_{x \in G} (\Delta_x^2 + \Delta_y^2 + \Delta_z^2) \tag{9}$$

where $\Delta_{x,y,z} = V_\alpha(x + \epsilon_{x,y,z}) - V_\alpha(x)$.

3.3 Loop Closure Detection via Transformer-based Encoding

The Loop Closure Detection (LCD) module (Fig. 2), adapted from TT-LCD [5], initiates by partitioning each input image into a sequence of non-overlapping patches. Each patch $x[k]$ is then embedded and propagated through a hierarchical transformer architecture to derive a compact, discriminative feature representation.

We employ a distilled Vision Transformer (DeiT) model [18], developed by Facebook Research, for patch-wise token embedding and contextual representation learning. The model follows the feature extraction paradigm established in TT-LCD [5], whereby each input image is encoded into a high-dimensional feature vector via successive transformer layers. The loop closure event is hypothesized when the cosine similarity between feature vectors corresponding to two temporally distinct frames exceeds a predefined threshold.

Formally, let V_i and V_j denote the transformer-encoded feature vectors of frames i and j, respectively. The cosine similarity score is computed as:

$$\text{Sim}(i,j) = \frac{V_i \cdot V_j}{\|V_i\| \|V_j\|} \tag{10}$$

A loop closure is triggered if $\text{Sim}(i,j) > \tau$, where τ is the similarity threshold.

Multi-Modal Visual SLAM Framework for Dual RGB-D Sensing

Algorithm 1 Tracking Thread

```
function Track(I_L, I_R)
  F ← ExtractORB(I_L, I_R);
  if NotInitialized then
    Initialize(F);
  else
    T ← TrackFrame(F);
    if Lost then
      Relocalize();
    else
      if NeedKF(F) then
        AddKF(F);
      end if
    end if
  end if
end function
```

Algorithm 2 Main System Thread

```
function ProcessKF(KF)
  rays ← SampleRays(KF);
  pts ← DepthSample(rays);
  enc ← Encode(pts);
  c, d ← Predict(enc);
  Optimize(c, d);
  UpdatePose(KF);
end function
```

Algorithm 3 Loop Closure Thread

```
function LoopDetect(KF)
  f ← Feats(KF);
  C ← Neighbors(KF);
  for all c ∈ C do
    sim ← CosineSim(f, c);
    if sim > τ then
      T ← ComputeSim3(KF, c);
      CorrectLoop(T);
      RunGBA();
    end if
  end for
end function
```

Algorithm 4 Covisibility Graph

```
function UpdateGraph(KF)
  InsertKF(KF);
  for all k_i in Graph do
    if Shared(KF, k_i) > δ then
      AddEdge(KF, k_i);
    end if
  end for
  if Redundant(KF) then
    Remove(KF);
  end if
end function
```

Figure 3: Algorithm for the proposed model

The internal feature extraction process of the transformer encoder is expressed as:

$$T_0 = [x_{\text{cls}}; x_1^{\text{Emb}}; x_2^{\text{Emb}}; \ldots; x_m^{\text{Emb}}] + \text{Emb}_{\text{pos}} \tag{11}$$

$$T_k' = \text{MSA}(\text{LN}(T_{k-1})) + T_{k-1}, \quad k = 1, \ldots, K \tag{12}$$

$$T_k = \text{MLP}(\text{LN}(T_k')) + T_k', \quad k = 1, \ldots, K \tag{13}$$

$$y = \text{PCA}(\text{LN}(T_K^0)) \tag{14}$$

The Multi-Head Self-Attention (MSA) mechanism is defined as:

$$\text{MSA}(h) = \text{Linear}(\text{Concat}(h_1, h_2, \ldots, h_n)) \tag{15}$$

$$h_k = \text{softmax}\left(\frac{Q_k K_k^T}{\sqrt{d}}\right) V_k \tag{16}$$

Here, $\text{LN}(\cdot)$ denotes layer normalization, and Q_k, K_k, and V_k correspond to the query, key, and value matrices obtained through learned linear projections. The scalar d is the dimensionality of the patch embeddings.

3.4 Online Covisibility Graph

The Online Covisibility Graph functions as the global memory of the Visual SLAM system, encoding all spatial-temporal associations. Its primary constituents are:

- **KeyFrames**: Each keyframe contains the camera pose, map point associations, Bag-of-Words (BoW) vectors derived from ORB descriptors, and the constituent RGB-D images.
- **MapPoints**: These represent persistent 3D landmarks triangulated across views.

Graph connectivity is determined by the number of shared map points across keyframes, with edge weights proportional to observation overlap. Redundancy is continuously pruned via **Keyframe Culling**, ensuring that superfluous keyframes—those sharing excessive map points with neighbors—are expunged to maintain computational tractability.

Summary of Contributions and Improvements

This multi-modal SLAM framework, underpinned by dual RGB-D sensing and a modular, thread-based architecture, markedly ameliorates several critical limitations endemic to conventional monocular or single-modal SLAM systems. Specifically, the joint parametric and coordinate encoding enables robust **depth estimation** under sparsity and occlusion, while neural rendering-based bundle adjustment mitigates **accumulated drift** through globally consistent trajectory refinement. The informed **loop closure mechanism** alleviates **feature degradation due to occlusion** by exploiting inter-keyframe coherence. Finally, the integration of a dynamically updated **Online Covisibility Graph** ensures **loop closure efficiency** and real-time adaptability, rendering the system highly suitable for deployment in complex, large-scale 3D environments.

4 Experimental Setup

4.1 Challenges in Dataset

The acquisition of publicly available multi-modal datasets— particularly those incorporating more than a single RGB-D sensor—remains a substantial impediment in the advancement of multi-perspective visual SLAM frameworks. Our methodology mandates a dual RGB-D configuration satisfying several stringent criteria: (i) both cameras must operate synchronously and capture aligned RGB and depth streams at each timestamp; (ii) the sensors must be spatially separated and extrinsically calibrated; and (iii) their orientations must afford unobstructed, overlapping views of the environment, avoiding configurations constrained to skyward or ground-facing perspectives.

Datasets such as *M2DGR* [23], where in the RGBD camera's multi-modality looks only towards the sky or [13], wherein both RGB-D streams are directed toward the ground plane, fall short of these requirements. Similarly, canonical benchmarks like *Replica*, *TUM RGB-D*, and *ScanNet* are inherently mono-modal, providing a single RGB-D stream per frame, thus rendering them unsuitable for dual-perspective SLAM architectures. Conversely, *TartanAir* [19] and *New Tsukuba* [4] satisfy the aforementioned constraints, offering stereo RGB-D data with appropriate spatial coverage and temporal synchronization, and are therefore employed as the foundational datasets in our experimental evaluation.

4.2 Training Details

The proposed system was trained and evaluated on the *TartanAir* and *Tsukuba* datasets, utilizing a total of 150 images per sequence from each dataset to ensure both environmental diversity and representational robustness. For the initialization and parameterization of the main system thread, we adhered closely to the methodological prescriptions outlined in the Co-SLAM framework, thereby preserving consistency in architectural baselines and enabling a principled comparative evaluation.

4.3 Competing methods

To the best of our knowledge, there exists no prior work that addresses Visual SLAM from a genuinely **multi-modal dual RGB-D** configuration, thereby necessitating the selection of a competitive baseline that, while not directly multimodal, represents the state of the art in neural real-time SLAM. For this purpose, we adopt *Co-SLAM* [9] as our principal comparative model. Despite operating under a monocular RGB or RGB-D framework, Co-SLAM exhibits a robust integration of coordinate-based neural representations with sparse parametric encodings, enabling real-time performance with commendable scene fidelity and geometric consistency. Its architectural alignment with neural SLAM paradigms and its efficiency in online reconstruction make it a suitable, albeit unimodal, benchmark for evaluating the performance and scalability of our proposed multi-modal dual RGB-D SLAM system.

Table 1: Performance metrics for Co-SLAM using Left, Right, and the Proposed Multimodal model across different sequences.

Sequence	Co-SLAM (Left)*	Co-SLAM (Right)†	Proposed Model
Tartan Air			
Abandoned Factory Night	5.21	5.22	4.73
Ocean	3.86	4.46	3.86
Office	4.28	4.43	2.32
Amusement	3.90	3.90	3.39
Japanese Alley	3.11	3.32	3.23
New Tsukuba Dataset			
Daylight	2.39	2.61	1.89
Fluorescent	2.42	1.99	1.78
Flashlight	2.58	1.93	1.48
Lamp	1.94	2.55	1.81

* Co-SLAM using only the Left image sequence.
† Co-SLAM using only the Right image sequence.Note that Co-SLAM doesn't have loop closure

5 Result

The proposed Multi-Modal Visual SLAM system, grounded in dual RGB-D sensing, is empirically evaluated on two representative and structurally diverse benchmarks—*Tartan Air* and *New Tsukuba Dataset*. These datasets, selected for their spatial complexity and depth variability, serve as a rigorous testbed for assessing both the fidelity and generalizability of the system under dynamic and non-trivial conditions. In the absence of publicly released evaluation metrics for dual-input configurations, we re-ran the Co-SLAM pipeline—recognized for its neural scene representation and real-time capabilities—on both datasets under carefully harmonized experimental conditions. This ensures a fair and reproducible comparison, isolating the benefits introduced by our architectural innovations. *See supplementary for architecture flow and hardware details.*

The subsequent analysis provides a comparative evaluation focused on the Root Mean Square Error (RMSE) of the Absolute Trajectory Error (ATE), serving as the principal quantitative metric for global pose consistency. Our system demonstrates substantial reductions in ATE-RMSE across both datasets relative to the Co-SLAM baseline, underscoring its enhanced capability for accurate and stable trajectory estimation under complex dual-modality conditions. It is important to underscore that both the *Tartan Air* and *New Tsukuba Dataset* datasets exhibit elevated structural and photometric complexity—characterized by frequent occlusions, abrupt viewpoint transitions, and irregular depth distributions—which render them substantially more challenging than conventional SLAM benchmarks. Consequently, the absolute performance metrics for both Co-SLAM and the proposed framework are comparatively attenuated. Nevertheless, the proposed dual-RGBD system consistently achieves lower ATE-RMSE values across both sequences, demonstrating its relative robustness and superior adaptability in the face of such adversarial sensory conditions. Quantitative results are delineated in Table 1 for empirical validation.

5.1 Evaluation on TartanAir Dataset

We evaluate the proposed dual-RGBD visual SLAM architecture against the Co-SLAM baseline on the *TartanAir* dataset—a synthetic, photo-realistic benchmark explicitly designed to challenge visual SLAM systems in complex robotic navigation scenarios. Captured in high-fidelity simulation environments, TartanAir encompasses diverse illumination conditions, weather variations, dynamic object presence, and loop closure events. It offers multimodal sensor streams—including stereo RGB and depth imagery—alongside temporally consistent, ground-truth pose annotations.

For this study, we utilized a **representative corpus of sequences**: *Abandoned Factory Night*, *Ocean*, *Office*, *Amusement*, and *Japanese Alley*, encompassing both indoor and outdoor scenes. Compared to conventional datasets such as Replica or TUM RGB-D, TartanAir exhibits markedly greater temporal and spatial variation, increased occlusion frequency, and nontrivial bounding box shifts per frame, making it an especially rigorous testbed.

Despite these elevated challenges, our method yields consistently lower ATE-RMSE values across all sequences, outperforming Co-SLAM in both internal and external scenarios. This superior

Figure 4: Interframe Consistency being maintained over Office sequence

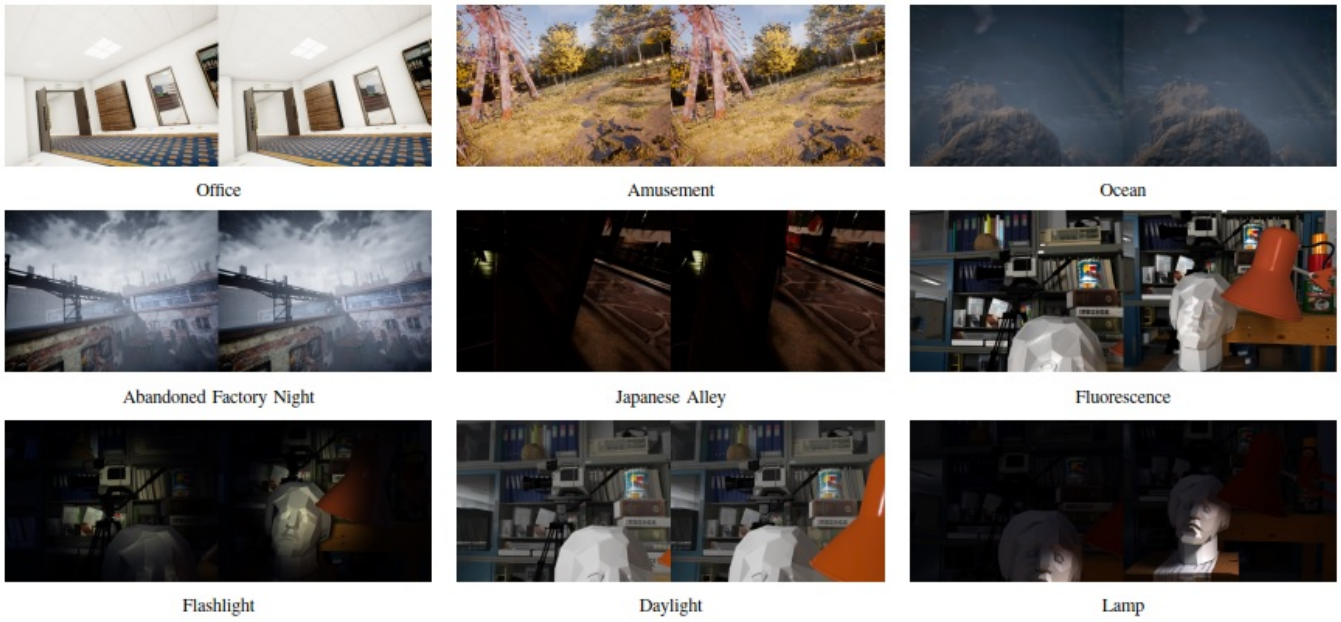

Figure 5: Visualisation of Key Frames: Portion of image not visible by one camera can be present in the other camera's image in the keyframe. This increases the view area of the environment. This helps in Dynamic Object Handling.

performance reflects the system's heightened robustness to *occlusion*, *dynamic object interference*, and large-scale appearance variation—conditions under which mono-modal baselines often degrade or fail.

5.2 Evaluation on New Tsukuba Dataset

We further assess the efficacy of the proposed framework on the *New Tsukuba* dataset, a synthetic stereo benchmark comprising photorealistically rendered video sequences under diverse illumination regimes. This evaluation emphasizes the model's robustness under dynamic lighting conditions, wherein we benchmark performance against the mono-modal Co-SLAM baseline. Empirical observations reveal that our dual-RGBD system consistently outperforms the baseline across all lighting scenarios— *Daylight*, *Lamps*, *Fluorescent*, and *Flash Light*—thereby affirming its resilience to photometric perturbations and its capacity to maintain spatial-temporal consistency across heterogeneous visual domains. This substantiates the model's effectiveness under *varying illumination conditions*.

5.3 Interframe Consistency

To ensure global structural coherence across temporal frames, our proposed system maintains a dynamically evolving **spanning tree of keyframes**, wherein the edge connectivity is governed by the inferred photometric and geometric affinities among candidate frames. In this context, we present the selected **RGB keyframe exemplars**, automatically extracted during SLAM execution, which

Table 2: Run Time of Proposed Model

Sequence	Proposed Model (Sec)
Tartan Air	
Abandoned Factory Night	0.62
Ocean	0.65
Amusement	0.58
Japanese Alley	0.62
Office	0.64
New Tsukuba Dataset	
Daylight	0.64
Fluorescent	0.64
Flashlight	0.62
Lamp	0.65

Table 3: Ablation Study: ATE RMSE Comparison – Running Proposed Model in Multimodal vs Monomodal Mode

Sequence	Monomodal *	Multimodal
Tartan Air		
Abandoned Factory Night	3.84	4.73
Ocean	3.44	3.86
Office	3.88	2.32
Amusement	3.76	3.39
Japanese Alley	3.71	3.23
New Tsukuba Dataset		
Daylight	2.28	1.89
Fluorescent	1.91	1.78
Flashlight	1.88	1.93
Lamp	1.81	1.81

* Proposed model executed in monomodal mode (using either left or right camera input only)

visually underscore the interframe relational coherence. The qualitative correspondence between adjacent keyframes substantiates the system's ability to preserve interframe consistency, thereby corroborating the fidelity of the underlying similarity metric that governs keyframe association. These visual results are illustrated in (Fig. 4) wherein the perceptual alignment across temporally distinct keyframes further validates the robustness of our interframe similarity mechanism.

5.4 Occlusion and Dynamic Lighting Conditions

To address the challenges posed by occlusion and fluctuating illumination, our proposed model employs a dual-camera keyframe selection strategy. This design integrates input imagery from both RGB-D sensors, thereby facilitating a more comprehensive and robust scene representation. As shown in Figure 5, our system is capable of capturing regions in the environment that may be occluded or entirely invisible to one camera, but are revealed through the alternate viewpoint provided by the second camera. This redundancy enhances both spatial completeness and inter-frame consistency.

Experimental validation using the Tsukuba dataset—specifically constructed for evaluating performance under varying lighting conditions—demonstrates the superior adaptability of our multimodal SLAM system in comparison to its mono-modal counterpart. The dataset includes identical scene sequences rendered under a range of illumination types, including fluorescence, flashlight, daylight, and tungsten-lamp lighting. For example, in the scene labeled *Office*, a segment of a cabinet that is occluded in the left camera's view is clearly visible in the right camera's perspective, thus expanding the effective observational coverage of the system. Likewise, in the *Fluorescence* sequence, portions of the environment hidden from the second camera are distinctly resolved in the first camera's view. These findings underscore the model's ability to handle occlusions through complementary viewpoints.

Furthermore, the performance of our model under dynamically shifting and low-light conditions, as evidenced by the results from Tsukuba's diverse lighting scenarios, affirms its robustness and photometric resilience. This versatility highlights the model's capacity to operate effectively in real-world, unconstrained environments where lighting conditions are neither static nor predictable.

5.5 Real-Time Performance

The proposed multi-modal SLAM with loop closure framework demonstrates stable and efficient real-time performance, with a median per-frame tracking latency of approximately 600 milliseconds (Table 2) across diverse sequences and environments. This consistency in runtime, irrespective of scene complexity or dynamic variation, underscores the robustness of our system's architectural design and computational pipeline. The sustained temporal regularity in tracking not only affirms the operational viability of the model in real-time deployments but also validates its scalability for long-duration SLAM tasks under variable conditions. These empirical findings substantiate our claim of robust real-time performance and reinforce the practicality of deploying the system in real-world.

6 Ablation: Monomodal vs. Multimodal

To empirically evaluate the contribution of sensor multiplicity, we conducted an ablation study Table 3 contrasting the proposed system operating in monomodal mode—i.e., utilizing a single RGB-D stream as input—with its full multimodal configuration. This analysis was performed across nine representative sequences. The results reveal that in five out of the nine sequences, the multimodal variant demonstrably outperformed the monomodal baseline by exhibiting a marked reduction in ATE, indicating superior drift correction. One sequence showed performance parity between the two modes, while the remaining three sequences registered a marginal increase in ATE under the multimodal setting—potentially attributable to minor inter-sensor redundancy or scene-specific ambiguities. Overall, these findings substantiate the efficacy of multimodal integration in improving spatial consistency and trajectory accuracy, especially in environments where occlusion, viewpoint diversity, or photometric variation play a critical role.

7 Conclusion

In this work, we proposed a novel Visual SLAM frameworks by integrating both multimodal input streams and a loop closure mechanism. Our approach introduces a dual-camera configuration, leveraging complementary views to enhance spatial completeness, and demonstrates the feasibility of multimodal integration within a real-time, neural SLAM architecture. While the system exhibits promising results—particularly in terms of inter-frame consistency and drift correction—there remain several avenues for future optimization. The tight coupling between the tracking and main system threads, although beneficial for calibration consistency, introduces latency that may impede real-time responsiveness in more computationally constrained environments. Employing more efficient data structures, particularly for rapid retrieval of keyframes from the covisibility graph, could significantly mitigate this overhead. Furthermore, augmenting the main system thread with dynamic object detection capabilities may enhance robustness in non-static environments, thereby expanding the operational scope of the system. Overall, this work serves as a foundational step toward more adaptive, resilient, and semantically aware Visual SLAM systems leveraging multimodal visual input.

References

[1] Berta Bescos, Jose M Facil, Javier Civera, and Jose Neira. 2018. DynaSLAM: Tracking, mapping, and inpainting in dynamic scenes. *IEEE Robotics and Automation Letters* 3, 4 (2018), 4076–4083.

[2] Peng Chen, Xinyu Zhao, Lina Zeng, Luxinyu Liu, Shengjie Liu, Li Sun, Zaijin Li, Hao Chen, Guojun Liu, Zhongliang Qiao, et al. 2025. A Review of Research on SLAM Technology Based on the Fusion of LiDAR and Vision. *Sensors* 25, 5 (2025), 1447.

[3] Mohammed Chghaf, Sergio Rodriguez, and Abdelhafid El Ouardi. 2022. Camera, LiDAR and multi-modal SLAM systems for autonomous ground vehicles: A survey. *Journal of Intelligent & Robotic Systems* 105, 1 (2022), 2.

[4] CVLAB, University of Tsukuba. 2014. New Tsukuba Stereo Dataset. https://home.cvlab.cs.tsukuba.ac.jp/dataset. Accessed: 2025-04-12.

[5] Chenchen Ding, Hongwei Ren, Zhiru Guo, Minjie Bi, Changhai Man, Tingting Wang, Shuwei Li, Shaobo Luo, Rumin Zhang, and Hao Yu. 2023. TT-LCD: Tensorized-Transformer based Loop Closure Detection for Robotic Visual SLAM on Edge. In *2023 International Conference on Advanced Robotics and Mechatronics (ICARM)*. IEEE, 166–172.

[6] Felix Endres, Jürgen Hess, Jürgen Sturm, Daniel Cremers, and Wolfram Burgard. 2014. An evaluation of the RGB-D SLAM system. *IEEE International Conference on Robotics and Automation (ICRA)* (2014).

[7] Andreas Geiger, Michael Roser, and Raquel Urtasun. 2011. StereoScan: Dense 3D reconstruction in real-time. In *IV*.

[8] Tin Lai. 2022. A review on visual-slam: Advancements from geometric modelling to learning-based semantic scene understanding using multi-modal sensor fusion. *Sensors* 22, 19 (2022), 7265.

[9] Sungbae Lee, Jaesik Kim, and Jongwoo Park. 2023. Co-SLAM: Joint coordinate and sparse parametric encodings for neural real-time SLAM. In *CVPR*.

[10] Stefan Leutenegger, Simon Lynen, Michael Bosse, Roland Siegwart, and Paul Furgale. 2015. Keyframe-based visual–inertial odometry using nonlinear optimization. In *IJRR*.

[11] Guanghao Li, Qi Chen, YuXiang Yan, and Jian Pu. 2024. EC-SLAM: Effectively Constrained Neural RGB-D SLAM with Sparse TSDF Encoding and Global Bundle Adjustment. *arXiv preprint arXiv:2404.13346* (2024).

[12] Shenghao Li, Luchao Pang, and Xianglong Hu. 2024. Multicam-slam: Non-overlapping multi-camera slam for indirect visual localization and navigation. *arXiv preprint arXiv:2406.06374* (2024).

[13] Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. 2017. 1 year, 1000 km: The oxford robotcar dataset. *The International Journal of Robotics Research* 36, 1 (2017), 3–15.

[14] Raul Mur-Artal, JMM Montiel, and JD Tardos. 2015. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics* 31, 5 (2015), 1147–1163.

[15] Antoni Rosinol, Marcus Abate, Yun Chang, and Luca Carlone. 2021. Kimera: an open-source library for real-time metric-semantic localization and mapping. In *IEEE International Conference on Robotics and Automation (ICRA)*.

[16] Martin Runz, Mickael Buffier, and Lourdes Agapito. 2017. CoFusion: Real-time segmentation, tracking and fusion of multiple objects. In *IEEE International Conference on Robotics and Automation (ICRA)*.

[17] Ali Rida Sahili, Saifeldin Hassan, Saber Muawiyah Sakhrieh, Jinane Mounsef, Noel Maalouf, Bilal Arain, and Tarek Taha. 2023. A survey of visual SLAM methods. *IEEE Access* 11 (2023), 139643–139677.

[18] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. 2021. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, Vol. 139. 10347–10357.

[19] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. 2020. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 4909–4916.

[20] Thomas Whelan, Rafael F Salas-Moreno, Ben Glocker, Andrew J Davison, and Stefan Leutenegger. 2015. ElasticFusion: Dense SLAM without a pose graph. In *Robotics: Science and Systems*.

[21] Zhe Xin, Yufeng Yue, Liangjun Zhang, and Chenming Wu. 2024. Hero-slam: Hybrid enhanced robust optimization of neural slam. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 8610–8616.

[22] Yueming Xu, Haochen Jiang, Zhongyang Xiao, Jianfeng Feng, and Li Zhang. 2024. DG-SLAM: Robust Dynamic Gaussian Splatting SLAM with Hybrid Pose Optimization. *arXiv preprint arXiv:2411.08373* (2024).

[23] Jie Yin, Ang Li, Tao Li, Wenxian Yu, and Danping Zou. 2021. M2dgr: A multi-sensor and multi-scenario slam dataset for ground robots. *IEEE Robotics and Automation Letters* 7, 2 (2021), 2266–2273.

[24] Zengrui Zheng, Kainan Su, Shifeng Lin, Zhiquan Fu, and Chenguang Yang. 2024. Development of vision–based SLAM: from traditional methods to multimodal fusion. *Robotic Intelligence and Automation* 44, 4 (2024), 529–548.

[25] Zihan Zhu and Matthias Nießner. 2023. NeRF-SLAM: Real-time dense monocular SLAM with neural radiance fields. In *CVPR*.

[26] Zihan Zhu, Songyou Peng, Martin Larsson, Yiyi Bao, Angela Dai, and Matthias Nießner. 2022. NICE-SLAM: Neural implicit scalable encoding for SLAM. In *CVPR*.