Hybrid Explicit–Implicit Dense Mapping with Quality-Guided Refinement and Residual Feedback

Xing Lin [1], Leyang Zhao [1], Weixi Wang [1], Yizhen Yan [1], Rui Yuan [1], Zirui Liu [1], Junjie Huang [1], Hao Zhang [1]

[1] Research Institute for Smart Cities, School of Architecture and Urban Planning, Shenzhen University, Shenzhen, 518060, PR China – 1833704910@qq.com

Keywords: Simultaneous Localization and Mapping; Neural Scene Representations; Hybrid Explicit–Implicit Mapping; Real-Time 3D Reconstruction

Abstract:

Real-time dense SLAM must balance geometric fidelity and computational efficiency for autonomous navigation. Explicit mapping methods provide stable global structure and fast updates, but suffer from discretization artifacts and memory overhead. Implicit neural representations capture continuous surfaces and fine details, yet require expensive optimization and are sensitive to initialization. Existing hybrid approaches combine both paradigms, but often allocate neural refinement inefficiently and remain vulnerable to pose errors. To address these limitations, we propose a selective hybrid dense mapping framework that couples a scene-wide TSDF backbone with quality-guided implicit local refinement and residual-guided sliding-window pose feedback. Neural refinement is activated only in low-quality regions identified by multi-indicator assessment, while keyframe poses are re-optimized using residuals from explicit raycasting and implicit rendering. Experiments on TUM RGB-D and Replica demonstrate improved mapping accuracy, localization robustness, and real-time efficiency.

1. Introduction

Simultaneous Localization and Mapping (SLAM) is a fundamental capability for autonomous robots. Beyond basic localization and mapping, practical robotic systems require dense and reliable scene representations to support perception, planning, and decision-making.

Compared with sparse point clouds, dense mapping can provide detailed and metrically consistent spatial representations in real time (Newcombe et al., 2011), (Whelan et al., 2015). Classical pipelines based on sparse features or photometric alignment often degrade under weak appearance cues (Kerl et al., 2013). While recent neural approaches improve expressiveness but frequently bind mapping to heavy rendering and global optimization, complicating online deployment (Sucar et al., 2021), (Zhu et al., 2022). This work targets real-time dense SLAM settings that require high-coverage mapping together with tightly coupled pose estimation.

In dense SLAM systems, explicit mapping strategies such as occupancy grid or Truncated Signed Distance Function (TSDF) remain widely adopted due to real-time performance, geometric interpretability, and compatibility with conventional robotics pipelines (Curless and Levoy, 1996), (Hornung et al., 2013), (Oleynikova et al., 2017). However, explicit mapping strategies are inherently constrained by discretization artifacts, significant memory overhead, and limited scalability when extended to high-resolution or large-scale environments. In contrast, implicit neural volumetric representations offer superior compactness and continuous surface modeling capabilities (Mildenhall et al., 2020), (Sucar et al., 2021), (Zhu et al., 2022). Nevertheless, these methods typically require substantial training demands, sensitivity to initialization conditions, and considerable computational overhead.

To overcome the limitations of purely explicit and purely implicit strategies, hybrid representations have been introduced to combine explicit volumetric structures with implicit neural fields (Huang et al., 2024), (Zhang et al., 2023), (Zhang et al., 2024). Such methods aim to preserve the efficiency and geometric reliability of voxel-based mapping while incorporating the continuity and expressiveness of learned representations. As a result, hybrid SLAM has become a promising direction for achieving accurate and efficient dense reconstruction in challenging environments.

Hybrid explicit–implicit SLAM has significantly advanced the development of real-time dense mapping; Nevertheless, current hybrid methods still face two key challenges: First, neural refinement is often allocated inefficiently, so computational resources may be wasted on already well-reconstructed regions while truly problematic areas remain under-refined (Sandström et al., 2023), (Liso et al., 2024). Second, pose errors can contaminate both explicit fusion and implicit optimization, leading to inconsistent supervisory signals and degraded reconstruction quality (Huang et al., 2024), (Zhang et al., 2023). These limitations reduce the practical effectiveness of hybrid dense SLAM.

In this work, we propose a selective hybrid dense mapping framework that integrates explicit reconstruction with implicit refinement through a quality assessment module, and further incorporates pose optimization, enhancing computational efficiency while maintaining high reconstruction accuracy.

Our contributions in this article are summarized as follows:

1) We propose an explicit assessment and implicit completion mechanism that identifies low-quality regions in the explicit map using multiple diagnostic indicators and selectively activates implicit local refinement to improve reconstruction fidelity while preserving global continuity.

2) We introduce a residual-guided sliding-window pose feedback scheme that re-optimizes keyframe poses

using residuals from both TSDF-based raycasting and implicit neural rendering, thereby reducing drift and improving pose accuracy.

3) We validate the proposed framework on the TUM RGB-D and Replica datasets (Sturm et al., 2012), (Straub et al., 2019), where it demonstrates superior accuracy and robustness compared with state-of-the-art methods, especially in texture-poor and geometrically complex scenes.

The remainder of this article is organized as follows. Section II reviews related work, Section III describes the proposed framework, Section IV presents the experimental results, and Sections V and VI provide the conclusion and discussion.

2. Related Work

Recent progress in dense SLAM has been driven by continuous advances in scene representation and optimization strategies. Existing approaches can be broadly categorized into three lines of research: explicit mapping methods, implicit neural representations and hybrid explicit – implicit frameworks. In this section, we briefly review these three directions and discuss their respective strengths and limitations in the context of real-time dense mapping and pose estimation.

2.1 Explicit Mapping in SLAM

Explicit dense mapping represents scene geometry and appearance with directly editable primitives such as occupancy grids, TSDF, ESDF, point sets, or Gaussian primitives. These representations support efficient updates, pruning, and integration with conventional robotics pipelines (Curless and Levoy, 1996), (Hornung et al., 2013), (Oleynikova et al., 2017). Among them, TSDF-based volumetric fusion remains a strong baseline for dense RGB-D SLAM. KinectFusion (Newcombe et al., 2011) established the paradigm of projective depth fusion and surface raycasting for real-time tracking and reconstruction. ElasticFusion (Whelan et al., 2015) further improved dense tracking and non-rigid surface consistency without relying on a global pose graph. BundleFusion (Dai et al., 2017) later demonstrated globally consistent dense reconstruction through on-the-fly surface reintegration and global optimization.

Another explicit direction is dense point mapping. MASt3R-SLAM (Murai et al., 2025) and related systems represent the environment as editable 3D points with per-point features, enabling scalable updates, rigid submap alignment, and second-order solvers. Compared with TSDF pipelines, point methods reduce discretization artifacts and preserve thin structures, though requiring outlier rejection and density regularization.

More recently, three-dimensional Gaussian splatting has been introduced as an explicit differentiable representation for high-quality scene modeling and rendering (Kerbl et al., 2023). This representation has also been adopted in dense SLAM, where Gaussian primitives are jointly optimized for geometry and appearance while maintaining fast rasterization-based rendering (Yan et al., 2024). Although Gaussian-based approaches provide strong rendering quality and local detail preservation, the number of primitives may grow rapidly without effective scheduling and pruning. In practice, voxel- and point-based explicit pipelines remain widely used because of their predictable memory usage, stable geometry updates, and ease of deployment.

2.2 Implicit Scene Representations for Dense SLAM

Recent research has revived test-time optimization with differentiable rendering, aligning frames by minimizing photometric and geometric residuals (Mildenhall et al., 2020). Some systems tightly couple tracking and mapping, while others preserve modularity. Implicit mapping represents scenes as continuous neural fields queried at arbitrary 3D coordinates. Online variants such as iMAP (Sucar et al., 2021), Co-SLAM (Wang et al., 2023) and NICE-SLAM (Zhu et al., 2022) showed that neural fields can be optimized during test time. Multi-resolution grids, tri-planes, and hash encodings reduce optimization while retaining sub-voxel fidelity (Müller et al., 2022), (Zhu et al., 2022). Implicit fields naturally model thin structures, sharp edges, and textures, while novel-view synthesis provides extra supervision (Mildenhall et al., 2020), (Yariv et al., 2021).

Despite strengths, implicit SLAM faces two challenges: sampling cost grows with resolution, requiring subsampling that limits detail; and global consistency is difficult, as drift deforms the learned field without loop closure or bundle adjustment (Sucar et al., 2021), (Zhu et al., 2022). Recent work integrates pose-graph optimization and online loop closure, but cost remains high when refinement is applied globally (Zhang et al., 2023), (Liso et al., 2024). These motivate hybrid designs where implicit models are deployed selectively.

2.3 Hybrid Explicit–Implicit Mapping

Point-SLAM (Sandström et al., 2023) uses a learnable decoder on point clouds but struggles with large-scale consistency. Loopy-SLAM (Liso et al., 2024) organizes neural point maps into submaps and performs loop closure, but depends on detection reliability. Photo-SLAM (Huang et al., 2024) integrates explicit anchors with implicit rendering, yet faces geometry–appearance trade-offs. ESLAM (Johari et al., 2023) employs a hybrid signed-distance representation for efficient dense SLAM, but does not explicitly prioritize refinement according to local map quality. GO-SLAM (Zhang et al., 2023) combines global optimization with hash-based fields but has high overhead. HI-SLAM (Zhang et al., 2024) incorporates neural fields into monocular SLAM but suffers from scale ambiguity. These systems demonstrate the complementary strengths of explicit and implicit representations, while also revealing persistent challenges in memory efficiency, global consistency, and selective refinement. Motivated by these limitations, our work develops a selective hybrid framework, which is detailed in the next section.

3. Methods

This section presents the proposed selective hybrid dense SLAM framework. As illustrated in Figure 1, the method consists of four main components: a scene-wide explicit TSDF mapping backbone, a quality-guided region selection module, an implicit local refinement module, and a residual-guided pose feedback optimization module. Given sequential RGB-D frames, the system first synchronizes color and depth streams using camera calibration, preprocesses the observations with depth denoising and photometric normalization, and estimates per-pixel confidence for subsequent matching, mapping, and optimization.

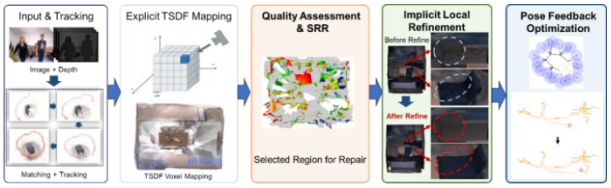

Figure 1. Proposed method integrates explicit TSDF mapping with implicit refinement and residual feedback, achieving robust real-time tracking, accurate mapping, and efficient optimization.

3.1 Scene-Wide Explicit TSDF Mapping Backbone

The explicit backbone maintains a globally consistent scene representation and provides the geometric prior for subsequent region assessment and refinement (Curless and Levoy, 1996), (Newcombe et al., 2011). Pose refinement before volumetric integration. Upon the selection of a new keyframe according to temporal or motion criteria, a point-map matching procedure is executed to refine its pose prior to volumetric integration (Chen and Medioni, 1992). The depth image of the keyframe is first back-projected using the current pose estimate to generate an observation point map. In parallel, the zero-crossings of the TSDF volume are raycast to obtain a model point map $\mathcal{M} = \{q_k\}$ together with corresponding surface normals $\mathcal{N} = \{n_k\}$. Correspondences $\mathcal{C} = \{(p, q, n)\}$ are then established through projective association, with visibility determined by the TSDF z-buffer. Candidate pairs that violate the distance threshold τ_d or a normal-angle gate τ_n are discarded. Pose refinement is subsequently achieved by minimizing the point-to-plane objective:

$$E_{pm}(T) = \sum_{(p,q,n)\in\mathcal{C}} w(p)\rho(n^{\top}(Tp - q)) \quad (1)$$

where w(p) combines the observation confidence and normal consistency, $\rho(\cdot)$ is a robust loss. Optimization proceeds with Gauss-Newton over a 3-5 level pyramid, typically converging in a few iterations. This step performs pose-only micro-adjustment to improve keyframe-to-map geometric consistency. After pose refinement, the aligned depth observation is integrated into the global TSDF volume using confidence-aware projective fusion.

3.2 Quality-Guided Selection for Explicit–Implicit Mapping

To identify regions where explicit reconstruction is insufficient, we define four complementary indicators that measure different aspects of reconstruction quality. These indicators are used to prioritize regions for implicit refinement.

The first indicator, Coverage and Geometric Novelty (CGN), measures the incremental spatial coverage provided by new depth observations and evaluates the degree of geometric novelty relative to the existing explicit reconstruction. It highlights regions where surfaces have not yet been sufficiently observed or where new viewpoints reveal structural variations not represented in the current map. The score is defined as:

$$\mathrm{CGN}(x) = \alpha_c \cdot C_{\mathrm{new}}(x) + \alpha_g \cdot (1 - \cos\theta_{\mathrm{normals}}(x)) \quad (2)$$

Where α_c and α_g balance coverage and novelty. High CGN indicates that refinement in such regions can enrich map completeness and geometric diversity, which are often underrepresented in explicit TSDF-based mapping.

The second indicator, Uncertainty Fusion (UF), quantifies the reliability of fused geometry by analyzing variance in TSDF updates or inconsistencies in depth integration. High uncertainty reflects unstable geometry due to noisy depth measurements or insufficient observation redundancy, signaling regions where explicit fusion alone is inadequate. Explicit uncertainty is measured by the variance of the TSDF field σ_{TSDF}, while implicit uncertainty comes from the variance of depth predictions in the neural voxel field σ_{Neural}. The fused metric is:

$$\mathrm{UF}(x) = w_e \cdot \sigma_{\mathrm{TSDF}}(x) + w_i \cdot \sigma_{\mathrm{Neural}}(x) \quad (3)$$

where w_e and w_i control the relative importance of the two sources. Large UF values imply that explicit fusion suffers from noise or incomplete evidence, and that implicit refinement can leverage learned priors to stabilize reconstruction.

The third indicator, Occlusion-Boundary Gain (OBG), targets depth discontinuities and occlusion boundaries where explicit reconstruction tends to blur or lose detail. It combines depth variance across multiple viewpoints $D_{\mathrm{var}}(x)$ with a boundary indicator $B_{\mathrm{occ}}(x)$ that flags voxels at occlusion edges:

$$\mathrm{OBG}(x) = \beta_d \cdot D_{\mathrm{var}}(x) + \beta_o \cdot B_{\mathrm{occ}}(x) \quad (4)$$

By refining these edge-adjacent regions, the implicit representation can preserve sharp transitions and recover partially missing geometry caused by limited triangulation in explicit TSDF fusion.

The fourth indicator, Reconstruction Mismatch (RM), evaluates the discrepancy between reprojected or rendered observations and the current explicit map, reflecting local misalignments or photometric inconsistencies. Elevated mismatch values indicate areas where explicit updates fail to reconcile observations with the model, necessitating refinement.

For each voxel projection x in the image domain, photometric error is computed as $\|I_{\mathrm{input}}(x) - \hat{I}_{\mathrm{render}}(x)\|_1$, and geometric error as $|z_{\mathrm{input}}(x) - z_{\mathrm{render}}(x)|$. The combined RM score is

$$\mathrm{RM}(x) = \gamma_p \cdot \|I_{\mathrm{input}}(x) - \hat{I}_{\mathrm{render}}(x)\|_1 + \gamma_d \cdot |z_{\mathrm{input}}(x) - z_{\mathrm{render}}(x)| \quad (5)$$

High RM indicates that explicit volumetric fusion fails to reproduce either the correct appearance or depth, signaling a strong potential benefit from implicit neural refinement.

To integrate these complementary indicators, each score is first normalized to [0,1] via min-max scaling,

$$\bar{m}_k(x) = \frac{m_k(x) - \min(m_k)}{\max(m_k) - \min(m_k)} \quad (6)$$

and the total refinement priority is computed as

$$S(x) = \lambda_{\mathrm{CGN}} \cdot \overline{\mathrm{CGN}}(x) + \lambda_{\mathrm{UF}} \cdot \overline{\mathrm{UF}}(x) + \lambda_{\mathrm{OBG}} \cdot \overline{\mathrm{OBG}}(x) + \lambda_{\mathrm{RM}} \cdot \overline{\mathrm{RM}}(x) \quad (7)$$

Regions assigned the highest scores are prioritized for implicit refinement, whereas the remaining areas are preserved within the explicit representation. This scoring strategy enables implicit

refinement to be allocated only to the most informative regions, thereby improving reconstruction quality without unnecessary computational overhead.

3.3 Implicit Local Refinement for Irreparable Regions

For regions identified as low-quality by the quality assessment module, we perform implicit local refinement to recover geometric and appearance details that cannot be reliably restored through explicit fusion alone. The refinement module operates only on the selected ROIs by initializing a neural voxel field with feature priors from the explicit map, optimizing it under photometric, depth, and normal consistency constraints, and reintegrating the refined outputs into the TSDF backbone. Through this targeted mechanism, the hybrid framework achieves enhanced reconstruction fidelity without compromising real-time performance.

For each ROI $\mathscr{R}_j$, we initialize an implicit neural voxel field $\mathscr{F}_j$ by fusing geometric and photometric priors from the explicit map. TSDF values and normals are encoded into a multi-scale geometric feature volume F_{geo}, while per-voxel photometric descriptors are encoded into a texture feature volume F_{tex}. A learnable fusion function Φ_{fuse} projects and combines these volumes:

$$z(x) = \Phi_{fuse}\left(F_{geo}(x), F_{tex}(x)\right) \tag{8}$$

The initialization stage endows the implicit model with informative priors, thereby substantially decreasing the number of iterations required for convergence.

The implicit neural voxel field $\mathscr{F}_j$ is parameterized by a Multi-Layer Perceptron that maps each three-dimensional coordinate x to a signed distance value s and color c:

$$\mathcal{F}_j : \mathrm{x} \mapsto (\mathrm{s}, \mathrm{c}) \tag{9}$$

This continuous formulation provides sub-voxel accuracy and effectively represents high-frequency surface details, thereby overcoming the resolution constraints inherent to discrete voxel grids.

Training $\mathscr{F}_j$ involves jointly optimizing photometric, depth, normal, and regularization terms:

$$\mathcal{L} = \lambda_p \mathcal{L}_{photo} + \lambda_d \mathcal{L}_{depth} + \lambda_n \mathcal{L}_{normal} + \lambda_e \mathcal{L}_{eik} \tag{10}$$

Here, $\mathcal{L}_{\text{photo}}$ enforces multi-view color consistency, $\mathcal{L}_{\text{depth}}$ aligns to observed depths, $\mathcal{L}_{\text{normal}}$ enforces surface normal consistency with explicit geometry, and $\mathcal{L}_{\text{eik}}$ regularizes the SDF gradient norm to one (Gropp et al., 2020). Differentiable volume rendering is performed only within ROI bounds to ensure computational efficiency.

During optimization, rays are sampled from all keyframes observing $\mathscr{R}_j$ and intersected with the implicit field via sphere tracing (Hart, 1996). SDF and color predictions are accumulated along each ray using alpha compositing. The process is strictly confined to the ROI bounding box, preventing unnecessary updates to well-reconstructed regions and ensuring that neural resources are focused on problematic areas.

Upon convergence, the refined implicit surface is discretized back into the explicit TSDF volume:

$$d_{\text{new}}(\mathbf{v}_i) = \text{Truncate}\left(\text{SDF}_{\mathcal{F}_j}(\mathbf{v}_i), \mu\right) \tag{11}$$

Where $\boldsymbol{\mu}$ is the truncation distance. Photometric information is also rasterized into the texture representation of the explicit map, ensuring smooth transitions at the boundaries between refined and unrefined regions. By writing the refined local geometry back to the TSDF volume, the system preserves the global consistency of the explicit map while improving local surface fidelity.

3.4 Residual-Guided Pose Feedback Optimization

We refine camera poses in a fixed-size sliding window by directly minimizing map-to-frame residuals derived from two complementary sources: (i) explicit TSDF raycasting, which provides geometry-anchored depth and synthesized intensity, and (ii) implicit neural rendering, which offers continuous, high-fidelity predictions for color and depth within recently refined regions. Unlike traditional frame-to-frame formulations, this design closes the loop between mapping and localization: whenever mapping (explicit or implicit) changes the predicted appearance/geometry, the updated residuals immediately feedback to pose estimation.

Let $\mathcal{W} = \{\mathrm{K}_{t-N+1}, \ldots, \mathrm{K}_t\}$ denote a sliding window of N keyframes with poses $\{\mathrm{T}_i \in \mathrm{SE}(3)\}$. For any frame $\mathrm{K}_j \in \mathcal{W}$, denote by I_j and D_j its observed intensity and depth. Given the fixed explicit TSDF map $\mathscr{M}_{\text{exp}}$ and the fixed implicit field $\mathscr{F}$ (as produced by the mapping modules before the current refinement), we can synthesize from pose T_j an explicit rendering $(\hat{\mathrm{I}}_j^{\text{exp}}, \hat{\mathrm{D}}_j^{\text{exp}})$ via TSDF raycasting and an implicit rendering $(\hat{\mathrm{I}}_j^{\text{imp}}, \hat{\mathrm{D}}_j^{\text{imp}})$ via differentiable neural rendering. Poses are the only optimization variables in this stage, while both maps remain fixed to stabilize estimation.

For each active frame, we define four residual channels, including explicit photometric residuals, explicit depth residuals, implicit photometric residuals, and implicit depth residuals. These residuals are evaluated at a selected pixel set $\Omega_j \subset$ the image lattice of frame j :

$$\begin{aligned}
\mathrm{r}_{\text{exp}}^{\text{photo}}(\mathrm{p}, \mathrm{j}) &= \mathrm{I}_j(\mathrm{p}) - \hat{\mathrm{I}}_j^{\text{exp}}(\mathrm{p}; \mathrm{T}_j), \\
\mathrm{r}_{\text{exp}}^{\text{depth}}(\mathrm{p}, \mathrm{j}) &= \mathrm{D}_j(\mathrm{p}) - \hat{\mathrm{D}}_j^{\text{exp}}(\mathrm{p}; \mathrm{T}_j), \\
\mathrm{r}_{\text{imp}}^{\text{photo}}(\mathrm{p}, \mathrm{j}) &= \mathrm{I}_j(\mathrm{p}) - \hat{\mathrm{I}}_j^{\text{imp}}(\mathrm{p}; \mathrm{T}_j), \\
\mathrm{r}_{\text{imp}}^{\text{depth}}(\mathrm{p}, \mathrm{j}) &= \mathrm{D}_j(\mathrm{p}) - \hat{\mathrm{D}}_j^{\text{imp}}(\mathrm{p}; \mathrm{T}_j).
\end{aligned} \tag{12}$$

The explicit terms constrain camera poses using the fused geometry and coarse appearance, while the implicit terms enhance the reconstruction by introducing sub-voxel detail in regions that have recently undergone refinement. To maintain computational efficiency, Ω_j is constructed through residual-guided sampling: a uniformly distributed base set is supplemented with the pixels exhibiting the largest residual magnitudes from the previous iteration across all available channels. This strategy directs optimization toward the most informative discrepancies without relying on manually designed

structural heuristics.

The sliding-window objective is then formulated as the weighted sum of the four residual channels over all frames:

$$\min_{\{T_j\}} \sum_{j\in} \Big(\lambda_p^{exp} \sum_{p\in\Omega_j} \tilde{r}_{exp}^{photo}(p,j) + \lambda_d^{exp} \sum_{p\in\Omega_j} \tilde{r}_{exp}^{depth}(p,j) + \lambda_p^{imp} \sum_{p\in\Omega_j} \tilde{r}_{imp}^{photo}(p,j) + \lambda_d^{imp} \sum_{p\in\Omega_j} \tilde{r}_{imp}^{depth}(p,j)\Big) \quad (13)$$

Pose refinement is formulated as a nonlinear least-squares problem and solved using Levenberg–Marquardt optimization. The Jacobians with respect to camera pose follow the standard photometric derivative,$\partial I/\partial T = \nabla I \cdot \partial \pi/\partial T$ and are computed consistently for both explicit and implicit renderings. For depth constraints, derivatives are derived either from the projective geometry of ray–surface intersections in the explicit map or from differentiable rendering gradients in the implicit field, both evaluated at the current pose estimate T_j.

Within the sliding-window framework, the problem is expressed as a factor graph, where nodes correspond to keyframe pose states and factors encode residuals from explicit TSDF raycasting and implicit neural photometric, depth, and geometric consistency terms (Dellaert and Kaess, 2017). Optimization is carried out jointly over the active window, after which the oldest keyframe is marginalized and the window advances. This formulation establishes a closed optimization loop in which updates to the map induce residual variations, the factor graph optimization refines the poses accordingly, and subsequent integrations and renderings benefit from improved alignment. Because the map is held fixed during this stage, the system remains well-conditioned and computationally efficient.

4. Results

The experiments conducted on TUM RGB-D and Replica demonstrate that the proposed hybrid system, combining an explicit TSDF backbone with implicit-on-demand refinement and residual-guided pose feedback, achieves consistent improvements in both mapping and localization. In mapping (Table 1), the method obtains the highest F1-scores across datasets, with particularly pronounced gains on sequences containing thin structures and high-frequency textures. This reflects the system’s design philosophy: the explicit map ensures global coverage and rapid updates, while implicit refinement enhances regions where explicit fusion is insufficient. Visual comparisons (Figure 2 and Figure 3) further highlight sharper edges, such as chair legs and table boundaries, and fewer artifacts near occlusion regions, confirming that local implicit refinement contributes sub-voxel detail without incurring the computational burden of fully implicit pipelines.

Runtime and memory analysis (Table 2) further supports the efficiency of the approach. Since implicit parameters are instantiated only for selected regions and the explicit TSDF maintains predictable memory growth, throughput remains within the 20–25 FPS range with moderate VRAM consumption, even for extended sequences. This property is particularly important for long-duration deployments, such as UAVs and mobile robots, where sustained frame rate and bounded memory are essential for safe autonomy. Furthermore, the explicit backbone retains compatibility with conventional robotics pipelines, including ESDF-based planning, collision checking, and map editing, while the refined appearance supports downstream perception tasks such as texture-aware re-localization and semantic processing.

On TUM Dataset	fr1-desk			fr2-xyz			fr3_office		
	Acc. (%)	Comp. (%)	F1 (%)	Acc. (%)	Comp. (%)	F1 (%)	Acc. (%)	Comp. (%)	F1 (%)
Ours	**85.8**	**83.6**	**84.7**	86.9	**84.4**	**85.6**	**88.1**	**85.9**	**87.0**
Loopy-SLAM	81.4	77.9	79.6	**92.0**	79.1	85.5	83.5	80.3	81.9
Point-SLAM	54.3	30.2	42.2	78.7	76.5	75.1	56.2	33.7	42.2
GO-SLAM	38.6	35.7	37.1	60.2	57.9	59.0	41.5	38.0	39.7
On Replica Dataset	**office0**			**office1**			**room0**		
Ours	84.7	**92.1**	88.4	**86.2**	**93.0**	**89.5**	**85.4**	**91.5**	**88.3**
Loopy-SLAM	80.2	76.5	78.3	81.0	77.4	79.1	79.8	75.9	77.8
Point-SLAM	**92.5**	88.4	**90.4**	93.6	89.1	**91.3**	92.1	87.9	**90.0**
GO-SLAM	67.3	63.6	65.4	68.1	64.2	66.1	66.5	62.8	64.6

Table 1. Quantitative mapping evaluation on Replica and TUM

Dataset	Method	Operation Time (s)	Mapping FPS	Explicit Mapping FPS	Rendering FPS	GPU Memory (GB)
Replica office0	Ours	0.191	12.5	16.2	5.8	**16.1**
	Loopy-SLAM	6.212	1.4	–	–	29.8
	Point-SLAM	10.505	1.0	–	–	30.2
	GO-SLAM	**0.055**	**7.1**	–	–	18.6
Replica office1	Ours	0.130	13.1	16.7	6.1	**15.9**
	Loopy-SLAM	4.378	1.5	–	–	30.0
	Point-SLAM	8.902	0.9	–	–	30.3
	GO-SLAM	**0.042**	**9.4**	–	–	18.4
Replica room0	Ours	0.219	11.9	15.8	5.9	**16.3**
	Loopy-SLAM	7.725	1.3	–	–	29.7
	Point-SLAM	11.208	0.8	–	–	30.5
	GO-SLAM	**0.063**	**6.9**	–	–	19.1

Table 2. Runtime and Memory Analysis on Replica

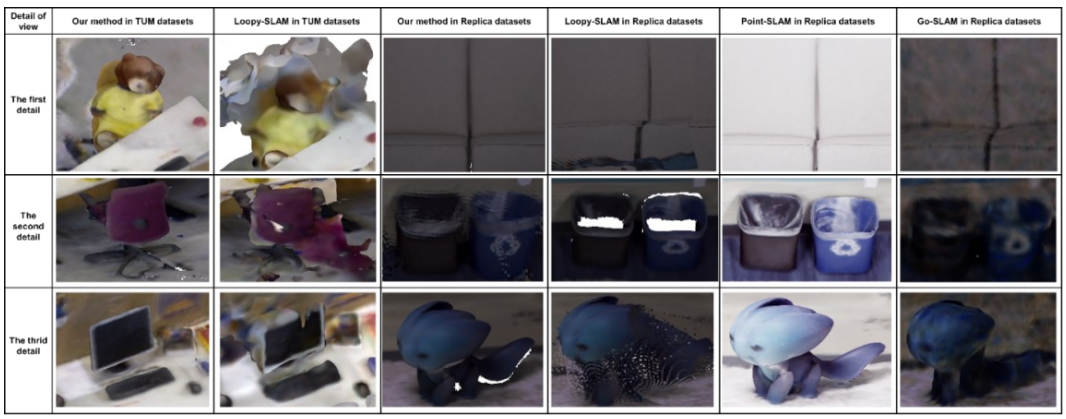

Figure 2. Local object refinement comparison on TUM and Replica datasets, where independent objects such as teddy bear, chair, and desktop computer in TUM, as well as sofa, trash bin, and doll in Replica, are reconstructed, demonstrating that our method and Loopy-SLAM preserve local details better than Point-SLAM and GO-SLAM

Figure 3. Comparative visualization of mapping results on Replica and TUM datasets, where our method, Loopy-SLAM, Point-SLAM, and GO-SLAM are evaluated, and the reconstructions are displayed from top, left, and right viewpoints.

5. Conclusion

This work has introduced a selective hybrid framework for real-time dense SLAM that integrates explicit volumetric mapping with targeted implicit refinement and residual-guided pose feedback. Central to the design is an explicit assessment–implicit completion mechanism, where multi-indicator diagnostics—including coverage gain, reprojection residuals, and photometric uncertainty—are used to evaluate map quality and identify regions where explicit fusion is insufficient. In these regions, implicit local refinement is activated: a neural voxel field is initialized via feature-level fusion and optimized under photometric, depth, and normal consistency constraints, with the refined outputs integrated back into the explicit backbone to preserve global continuity. Complementing this module is a residual-guided sliding-window pose feedback scheme, which re-optimizes keyframe poses by jointly minimizing residuals from TSDF raycasts and implicit renderings. This feedback loop effectively reduces drift and maintains pose accuracy, thereby ensuring the stability required for high-quality reconstruction.

Extensive experiments on TUM RGB-D and Replica confirm that this integrated framework achieves consistent improvements in both mapping fidelity and localization accuracy. The explicit backbone ensures scalable updates and compatibility with loop closure, the implicit refinement injects sub-voxel detail only where it is most needed, and the pose feedback mechanism tightly couples mapping with trajectory optimization. Collectively, these contributions provide a principled approach to balancing geometric completeness, local detail, and computational efficiency, enhancing the robustness of dense SLAM in complex observational conditions and improving its suitability for autonomous navigation and downstream semantic perception.

Nevertheless, challenges remain. The diagnostic weights and thresholds used in the assessment module are currently dataset-specific; adaptive scheduling strategies, such as learned weighting or automatic percentile tuning, could enhance generalization to unseen environments. In addition, the tracking front-end relies exclusively on photometric alignment, which is sensitive to illumination variation and non-Lambertian effects. Incorporating exposure-invariant cues or lightweight geometric information, such as confidence-filtered depth, represents a promising direction to increase robustness without sacrificing efficiency. Beyond these aspects, integrating online semantic signals to guide refinement prioritization and leveraging event or inertial measurements to handle fast motion are natural extensions for future research.

References

Chen, Y., Medioni, G., 1992. Object modelling by registration of multiple range images. Image Vis. Comput., 10(3), 145-155. doi:10.1016/0262-8856(92)90066-C.

Curless, B., Levoy, M., 1996. A volumetric method for building complex models from range images. Proceedings of SIGGRAPH 1996, 303-312.

Dai, A., Nießner, M., Zollhöfer, M., Izadi, S., Theobalt, C., 2017. BundleFusion: Real-Time Globally Consistent 3D Reconstruction Using On-the-Fly Surface Reintegration. ACM Trans. Graph., 36(3), Article 24. doi:10.1145/3054739.

Dellaert, F., Kaess, M., 2017. Factor Graphs for Robot Perception. Found. Trends Robot., 6(1-2), 1-139. doi:10.1561/2300000043.

Gropp, A., Yariv, L., Haim, N., Atzmon, M., Lipman, Y., 2020. Implicit geometric regularization for learning shapes. Proceedings of the 37th International Conference on Machine Learning (ICML), 3789-3799.

Hart, J.C., 1996. Sphere tracing: A geometric method for the antialiased ray tracing of implicit surfaces. Vis. Comput., 12(10), 527-545. doi:10.1007/s003710050084.

Hornung, A., Wurm, K.M., Bennewitz, M., Stachniss, C., Burgard, W., 2013. OctoMap: An Efficient Probabilistic 3D Mapping Framework Based on Octrees. Auton. Robots, 34(3), 189-206. doi:10.1007/s10514-012-9321-0.

Huang, H., Li, L., Cheng, H., Yeung, S.-K., 2024. Photo-SLAM: Real-time Simultaneous Localization and Photorealistic Mapping for Monocular Stereo and RGB-D Cameras. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 21584-21593.

Johari, M.M., Carta, C., Fleuret, F., 2023. ESLAM: Efficient Dense SLAM System Based on Hybrid Representation of Signed Distance Fields. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 17408-17419.

Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. ACM Trans. Graph., 42(4). doi:10.1145/3592433.

Kerl, C., Sturm, J., Cremers, D., 2013. Dense visual SLAM for RGB-D cameras. Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2100-2106.

Liso, L., Sandström, E., Yugay, V., Van Gool, L., Oswald, M.R., 2024. Loopy-SLAM: Dense Neural SLAM with Loop Closures. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20363-20373.

Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R., 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. Proceedings of the European Conference on Computer Vision (ECCV), 405-421.

Müller, T., Evans, A., Schied, C., Keller, A., 2022. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. ACM Trans. Graph., 41(4). doi:10.1145/3528223.3530127.

Murai, R., Dexheimer, E., Davison, A.J., 2025. MASt3R-SLAM: Real-Time Dense SLAM with 3D Reconstruction Priors. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 16695-16705.

Newcombe, R.A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A.J., Kohli, P., Shotton, J., Hodges, S., Fitzgibbon, A., 2011. KinectFusion: Real-Time Dense Surface Mapping and Tracking. Proceedings of the 10th IEEE International Symposium on Mixed and Augmented Reality (ISMAR), 127-136. doi:10.1109/ISMAR.2011.6092378.

Oleynikova, H., Taylor, Z., Fehr, M., Siegwart, R., Nieto, J., 2017. Voxblox: Incremental 3D Euclidean Signed Distance Fields for On-Board MAV Planning. Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 1366-1373. doi:10.1109/IROS.2017.8202315.

Sandström, E., Li, Y., Van Gool, L., Oswald, M.R., 2023. Point-SLAM: Dense Neural Point Cloud-based SLAM. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 18433-18444. doi:10.1109/ICCV51070.2023.01690.

Straub, J., Whelan, T., Ma, L., Chen, X., Wijmans, E., Green, S., et al., 2019. The Replica Dataset: A Digital Replica of Indoor Spaces. arXiv preprint, arXiv:1906.05797.

Sturm, J., Engelhard, N., Endres, F., Burgard, W., Cremers, D., 2012. A Benchmark for the Evaluation of RGB-D SLAM Systems. Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 573-580. doi:10.1109/IROS.2012.6385773.

Sucar, E., Liu, S., Ortiz, J., Davison, A.J., 2021. iMAP: Implicit Mapping and Positioning in Real-Time. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 6229-6238. doi:10.1109/ICCV48922.2021.00617.

Wang, H., Wang, J., Agapito, L., 2023. Co-SLAM: Joint Coordinate and Sparse Parametric Encodings for Neural Real-Time SLAM. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 13293-13302.

Whelan, T., Leutenegger, S., Salas-Moreno, R., Glocker, B., Davison, A.J., 2015. ElasticFusion: dense SLAM without a pose graph. Robotics: Science and Systems (RSS). doi:10.15607/RSS.2015.XI.001.

Yan, C., Qu, D., Xu, D., Zhao, B., Wang, Z., Wang, D., Li, X., 2024. GS-SLAM: Dense Visual SLAM with 3D Gaussian Splatting. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 19595-19604. doi:10.1109/CVPR52733.2024.01853.

Yariv, L., Gu, J., Kasten, Y., Lipman, Y., 2021. Volume Rendering of Neural Implicit Surfaces. Adv. Neural Inf. Process. Syst. (NeurIPS), 34, 4805-4815.

Zhang, W., Sun, T., Wang, S., Cheng, Q., Haala, N., 2024. HI-SLAM: Monocular Real-Time Dense Mapping with Hybrid Implicit Fields. IEEE Robot. Autom. Lett., 9(2), 1548-1555.

Zhang, Y., Tosi, F., Mattoccia, S., Poggi, M., 2023. GO-SLAM: Global Optimization for Consistent 3D Instant Reconstruction.

Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 3727-3737. doi:10.1109/ICCV51070.2023.00345.

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. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 12786-12796