

## Article

# Uncertainty-Aware LiDAR–Inertial–Visual SLAM with Adaptive Fusion and Multi-Channel Geometric Loop Closure

Qixue Zhong <sup>1,2</sup>, Jing Xing <sup>1</sup>, Jian Liu <sup>1,\*</sup> and Luqing Luo <sup>1</sup>

<sup>1</sup> Institute of Microelectronics of the Chinese Academy of Sciences, Beijing 100029, China; zhongqixue@ime.ac.cn (Q.Z.); xingjing@ime.ac.cn (J.X.); luoluqing@ime.ac.cn (L.L.)

<sup>2</sup> University of Chinese Academy of Sciences, Beijing 100049, China

\* Correspondence: liujian@ime.ac.cn

## Abstract

Accurate and robust localization and mapping in complex and dynamic environments remain a fundamental challenge for autonomous systems. LiDAR–Inertial–Visual Odometry (LIVO) integrates the complementary strengths of LiDAR geometry, visual appearance, and inertial motion constraints. However, existing LIVO systems still suffer from limited adaptability to sensor degradation, weak loop-closure robustness, and insufficient cross-modal consistency modeling. This paper presents a robust multi-sensor SLAM framework that integrates an uncertainty-aware LIVO front-end, a geometry-driven loop-closure module, and a cross-modal consistency factor-graph back-end. We develop an uncertainty-aware iterated error-state Kalman filter (iESKF) to tightly fuse LiDAR, visual, and inertial measurements, with measurement covariances dynamically adjusted according to innovation statistics, feature-matching quality, and observability. To improve global consistency, we propose a multi-channel Binary Triangle Constraint (mBTC) descriptor for LiDAR-based loop detection, which enhances robustness under viewpoint changes and appearance degradation. In addition, we introduce a cross-modal consistency factor to explicitly constrain the relative motion agreement between visual and LiDAR odometries. Extensive experiments on multiple public benchmarks demonstrate improved accuracy, loop-closure reliability, and long-term consistency compared with state-of-the-art LIVO systems.

**Keywords:** LiDAR–inertial–visual SLAM; multi-sensor fusion; uncertainty-aware iESKF; loop closure; cross-modal consistency

## 1. Introduction

Accurate and robust localization and mapping are fundamental capabilities for autonomous systems operating in complex real-world environments, such as mobile robots, autonomous vehicles, and unmanned aerial platforms. These systems are required to function reliably under challenging conditions, including drastic illumination changes, dynamic objects, large-scale scenes, repetitive structures, and long-term operation. Achieving high-precision, drift-free simultaneous localization and mapping (SLAM) in such scenarios remains a long-standing research challenge.

Single-modality SLAM systems suffer from inherent limitations. Vision-based methods are highly sensitive to illumination variations, motion blur, and textureless regions, often leading to tracking failures. LiDAR-based approaches, while robust to lighting conditions, typically provide only sparse geometric information and struggle in environments with repetitive geometric patterns or structure-degenerate scenes. Inertial measurement units



Received: 9 March 2026

Revised: 19 April 2026

Accepted: 27 April 2026

Published: 29 April 2026

**Copyright:** © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

(IMUs) supply high-frequency motion constraints but inevitably accumulate drift over time due to sensor bias and noise. To overcome these individual weaknesses, multi-sensor fusion has become a promising direction, aiming to leverage the complementary strengths of LiDAR geometry, visual appearance, and inertial motion priors.

Among multi-sensor frameworks, LiDAR–Inertial–Visual Odometry (LIVO) has emerged as an effective solution that tightly integrates the metric accuracy and illumination invariance of LiDAR, the rich appearance information from cameras, and the high-rate continuity provided by IMUs. Early loosely coupled systems [1–3] are computationally efficient but fail to fully exploit cross-modal constraints. Recent tightly coupled approaches, such as LVI-SAM [4], R2LIVE [5], R3LIVE [6], FAST-LIVO [7], and FAST-LIVO2 [8], have significantly advanced real-time performance and accuracy by unifying visual and LiDAR residuals within a unified optimization or filtering framework. Nevertheless, most existing LIVO systems still rely on fixed measurement noise models that cannot adapt to time-varying sensor degradation (e.g., illumination changes, motion blur, partial occlusions, or geometric degeneracy). In addition, they often lack robust geometric loop closure under large viewpoint changes and do not explicitly enforce cross-modal consistency between visual and LiDAR odometries. These limitations frequently result in reduced robustness, accumulated drift, and suboptimal global consistency in challenging environments.

To address the above issues, this paper presents a robust multi-sensor SLAM framework that integrates an uncertainty-aware LIVO front-end, a geometry-driven loop-closure module, and a cross-modal consistency factor-graph back-end. The main contributions of this work are as follows:

The main contributions of this paper are summarized as follows:

1. **Uncertainty-aware iESKF fusion.** We propose a dynamic covariance adaptation strategy within an iterated error-state Kalman filter (iESKF) that adaptively balances LiDAR, visual, and inertial measurements under varying sensor quality and environmental conditions.
2. **Multi-channel geometric loop detection.** We introduce a novel multi-channel Binary Triangle Constraint (mBTC) descriptor that encodes richer geometric context, significantly improving loop detection robustness and discriminability in repetitive and viewpoint-variant scenes.
3. **Cross-modal consistency factor-graph optimization.** We design a cross-modal consistency factor that explicitly enforces short-baseline motion agreement between visual and LiDAR odometries, reducing long-term drift and enhancing global consistency.
4. **Comprehensive evaluation.** Extensive experiments on multiple public multi-sensor datasets demonstrate that the proposed system achieves superior accuracy, robustness, and long-term consistency compared to state-of-the-art LIVO methods.

## 2. Related Work

### 2.1. Direct Visual and LiDAR Odometry

Direct visual odometry and SLAM estimate camera motion by directly minimizing photometric inconsistencies across image frames, without explicit feature extraction. Representative sparse direct methods such as DSO [9] jointly optimize camera poses and sparse depth through photometric bundle adjustment, achieving high accuracy under stable illumination. Earlier semi-dense approaches, including LSD-SLAM [10], demonstrated the feasibility of direct mapping at larger scales, while semi-direct pipelines such as SVO [11] combine feature-based tracking with direct alignment to balance efficiency and accuracy. Despite their effectiveness, direct visual methods remain highly sensitive to illumination variations, motion blur, and weak-texture regions, often leading to tracking failures.

In parallel, feature-based visual SLAM systems such as ORB-SLAM2 [12] and its extensions are widely used due to their robust keyframe management, loop closure detection, and global bundle adjustment. Closely related visual-inertial systems, including VINS-Mono [13] and subsequent variants [14,15], improve short-term robustness by incorporating inertial constraints, yet still suffer from limited performance in visually degraded environments.

On the LiDAR side, direct geometric registration has been extensively explored. The seminal LOAM framework [16] and its variants [17,18] establish point-to-edge and point-to-plane constraints for LiDAR odometry and mapping. Recent direct LiDAR approaches avoid explicit feature extraction by operating on raw or voxelized point clouds [19,20], enabling efficient scan-to-map alignment. However, LiDAR-only systems may encounter geometric degeneracy in structure-sparse environments and lack the rich appearance cues available in vision.

## 2.2. LiDAR–Visual–Inertial SLAM

To exploit complementary sensing modalities, LiDAR–visual–inertial SLAM systems have been developed in both loosely and tightly coupled architectures. Early loosely coupled pipelines integrate visual odometry with LiDAR mapping at the state level, where visual estimates primarily serve as motion priors for LiDAR registration [1–3]. Although modular, these approaches often fail to fully exploit cross-modal constraints, resulting in suboptimal information utilization.

Tightly coupled formulations have demonstrated superior performance by jointly optimizing or filtering multi-modal measurements. Optimization-based systems such as LVI-SAM [4] incorporate LiDAR and visual factors into a unified factor graph, achieving high mapping accuracy and global consistency at the expense of increased computational cost. In contrast, filter-based pipelines have gained popularity for their real-time capability. R2LIVE [5] unifies LiDAR geometric residuals and visual photometric errors within an iterated error-state Kalman filter (iESKF), enabling high-rate propagation and fast updates. Its successor, R3LIVE [6], further improves robustness through radiance modeling and multi-threaded visual rendering.

Recent efforts have focused on enhancing efficiency and robustness of tightly coupled LIVO. FAST-LIVO [7] introduces a sparse direct fusion strategy within an iESKF framework, significantly accelerating front-end estimation. Its successor, FAST-LIVO2 [8], refines visual-to-map alignment and residual modeling, establishing a strong real-time baseline. Related tightly coupled systems, including Fast-LIO [21], Fast-LIO2 [22], Point-LIO [23], CT-ICP [24], DLIO [25], and LIO-SAM [26], further explore efficient residual formulations, continuous-time modeling, and improved map representations.

Despite these advances, most existing LIVO systems rely on static or heuristic measurement noise models that assume fixed reliability across modalities. In practice, sensor quality varies significantly due to illumination changes, partial occlusions, and dynamic objects, motivating uncertainty-aware fusion strategies that adaptively balance heterogeneous observations.

## 2.3. Loop Closure and Consistency in Multi-Modal SLAM

Loop closure detection is essential for correcting accumulated drift and ensuring long-term global consistency. In LiDAR-based place recognition, global geometric descriptors such as Scan Context [27] and projection-based methods like M2DP [28] have been widely adopted due to their computational efficiency and illumination invariance. Subsequent extensions [29–31] further improve robustness under viewpoint changes. However, global descriptors may suffer from geometric aliasing in repetitive or sparse environments.

Recent works explore geometric-invariant representations based on local structural constraints, such as binary or triangular configurations [32,33], which exhibit improved robustness under wide-baseline and appearance-degraded conditions. Nevertheless, descriptors relying on a single geometric attribute may still lack discriminative power in complex scenes.

In addition to loop closure, maintaining cross-modal consistency between LiDAR and visual estimates remains an underexplored problem. Many multi-sensor SLAM systems implicitly assume agreement between visual and LiDAR motion estimates, yet discrepancies can accumulate and degrade global optimization. Consistency-aware strategies inspired by innovation-based statistical testing [34] and adaptive covariance modeling [35] provide a principled foundation for dynamically adjusting measurement confidence and improving estimator reliability.

### 3. Materials and Methods

#### 3.1. System Overview

In the preprocessing stage, raw data streams from the IMU, LiDAR, and camera are hardware-synchronized using an STM32-based timing board (Figure 1), where the STM32 microcontroller was provided by STMicroelectronics, Geneva, Switzerland, following the synchronization scheme of FAST-LIVO2 [8]. After synchronization, IMU measurements are used for state propagation, while LiDAR and visual observations provide geometric and photometric constraints, respectively. These measurements are then tightly fused within the uncertainty-aware iESKF front-end. To improve robustness under varying sensing conditions, an adaptive dynamic covariance adjustment strategy is introduced to modulate the measurement uncertainty of each modality according to three complementary cues, namely innovation consistency (NIS), feature-matching quality, and geometric observability. In this way, the estimator is able to dynamically balance the contributions of different sensors according to their instantaneous reliability.

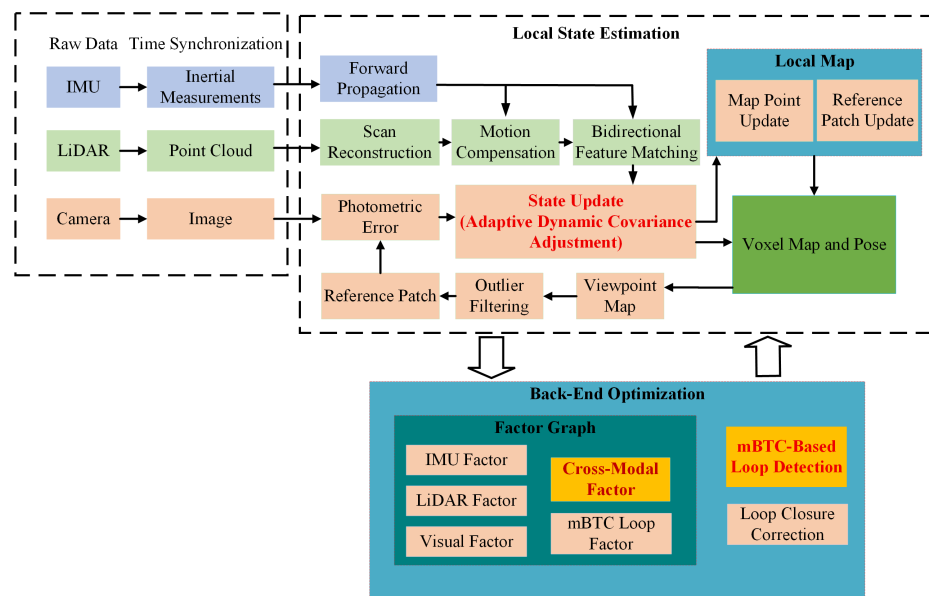


**Figure 1.** Overall framework of the proposed uncertainty-aware LiDAR-inertial-visual SLAM system. The solid arrows indicate data flow and information propagation, whereas the hollow arrows denote the interaction between the front-end local estimation and the back-end optimization. The front-end performs synchronized multi-sensor preprocessing, motion compensation, bidirectional feature matching, photometric residual construction, and uncertainty-aware state update with adaptive dynamic covariance adjustment. The back-end jointly incorporates IMU, LiDAR, visual, cross-modal consistency, and mBTC-based loop closure factors to improve global consistency.

Based on the estimated states, the front-end further maintains a local map composed of map-point updates, reference-patch updates, and voxel-map-based pose representation. This local mapping mechanism provides stable geometric and photometric references for subsequent state estimation, thereby enhancing the continuity and consistency of the front-end odometry.

To improve long-term consistency, the back-end performs global optimization within a factor-graph framework. Specifically, the factor graph jointly incorporates IMU factors, LiDAR factors, visual factors, cross-modal consistency factors, and mBTC loop factors. Meanwhile, an mBTC-based loop detection module is employed to identify revisited places, and the resulting loop closure constraints are introduced for loop closure correction and global trajectory refinement. Compared with the original single-channel BTC descriptor that encodes only geometric shape information, the proposed mBTC extends the descriptor by introducing additional multi-dimensional contextual cues, thereby improving loop robustness and distinctiveness in repetitive or partially degraded environments.

Through the hierarchical integration of uncertainty-aware iESKF-based front-end estimation, local map maintenance, mBTC-based loop detection, and cross-modal factor-graph optimization, the proposed framework unifies local odometry estimation and global consistency optimization within a single multi-sensor SLAM system, achieving enhanced robustness and accuracy in challenging environments.

### 3.2. Uncertainty-Aware iESKF Fusion

The proposed front-end uses an iterated error-state Kalman filter (iESKF). This filter tightly fuses LiDAR, visual, and inertial measurements in a unified probabilistic framework. It propagates the state with IMU measurements at high frequency. It then corrects the state using LiDAR and visual observations. This approach achieves real-time estimation while preserving local consistency on the SE(3) manifold.

Conventional tightly coupled filters rely on fixed and isotropic measurement covariances. In contrast, our method introduces an uncertainty-aware adaptive covariance mechanism. This mechanism dynamically adjusts the confidence of each sensing modality. It bases the adjustment on real-time statistical consistency, matching quality, and geometric observability.

This design improves robustness under changing sensing conditions. The system can maintain reliable state estimation even when illumination changes, motion blur occurs, texture is weak, or LiDAR geometry degenerates.

#### 3.2.1. State Representation and Error Definition

The system state is defined as

$$\mathbf{x} = [\mathbf{R}_{wb}, \mathbf{p}_w, \mathbf{v}_w, \mathbf{b}_g, \mathbf{b}_a], \quad (1)$$

where  $\mathbf{R}_{wb} \in \text{SO}(3)$  is the rotation matrix that maps vectors from the body frame to the world frame, and  $\text{SO}(3)$  denotes the 3D rotation group. The vectors  $\mathbf{p}_w \in \mathbb{R}^3$  and  $\mathbf{v}_w \in \mathbb{R}^3$  denote the body position and velocity in the world frame, respectively. The gyroscope bias  $\mathbf{b}_g \in \mathbb{R}^3$  and accelerometer bias  $\mathbf{b}_a \in \mathbb{R}^3$  are both three-dimensional. The corresponding 15-dimensional error state is written as

$$\delta \mathbf{x} = [\delta \boldsymbol{\theta}^\top, \delta \mathbf{p}^\top, \delta \mathbf{v}^\top, \delta \mathbf{b}_g^\top, \delta \mathbf{b}_a^\top]^\top, \quad (2)$$

where  $\delta \boldsymbol{\theta} \in \mathbb{R}^3$  is the rotation error vector (also called the attitude error) expressed in the tangent space of  $\text{SO}(3)$ . A right-invariant error formulation is adopted so that the attitude update is performed through exponential retraction on the manifold:

$$\mathbf{R}_{wb} \leftarrow \hat{\mathbf{R}}_{wb} \exp(\delta\theta), \quad (3)$$

thereby preserving local consistency on the manifold.

### 3.2.2. IMU Propagation, Measurement Models, and Iterated ESKF Update

The front-end performs tight multi-sensor fusion using an iterated error-state Kalman filter (iESKF). IMU measurements are used for high-frequency state propagation, while LiDAR point-to-plane residuals and visual photometric residuals provide complementary corrections. The continuous-time inertial propagation model and the linearized dynamics follow the standard error-state formulation commonly used in iESKF-based SLAM systems [5,21,22].

The visual and LiDAR measurement models are defined as photometric residuals and point-to-plane residuals, respectively. These heterogeneous measurements are stacked into a joint residual vector and fused within the iESKF framework through the standard innovation form. The update is iterated until convergence to improve accuracy.

The detailed mathematical derivations of IMU propagation, linearization of measurement models, Jacobian computation, covariance propagation, and the full iterated ESKF update equations are provided in Appendix A.

After each correction, the nominal state is updated on the  $SE(3)$  manifold as

$$\mathbf{R}_{wb} \leftarrow \hat{\mathbf{R}}_{wb} \exp(\delta\theta), \quad \hat{\mathbf{p}}_w \leftarrow \mathbf{p}_w + \delta\mathbf{p}, \quad \hat{\mathbf{v}}_w \leftarrow \mathbf{v}_w + \delta\mathbf{v}. \quad (4)$$

### 3.2.3. Adaptive Uncertainty Modeling and Weight Regulation Mechanism

The previous subsection formulated multi-modal fusion within the standard iESKF update as a maximum a posteriori estimation problem, where the measurement covariance matrix  $\mathbf{R}_k^m$  determines the relative contribution of each modality to the joint innovation information. However, in real-world environments, the quality of visual and LiDAR observations is highly time-varying due to illumination changes, motion blur, dynamic interference, sparse correspondences, and geometric degeneration. Under such conditions, a fixed covariance assumption may lead to statistical mismatch and improper confidence allocation across sensing modalities.

To address this issue, this subsection develops an adaptive dynamic covariance modeling mechanism. The uncertainty of each modality is decomposed into three complementary sources: statistical consistency, perception/matching quality, and geometric observability. These three factors are combined in a multiplicative form. This produces a unified adjustment strategy that is interpretable, controllable, and easily extensible. In this way, the covariance used in the iESKF update is no longer fixed. Instead, it is adaptively modulated according to the instantaneous reliability of each sensing modality.

#### Unified Adaptive Covariance Form: Three-Factor Multiplicative Decomposition

Intuitively, measurements from different modalities should not be trusted equally at all times. For example, visual observations become less reliable under blur or poor illumination, while LiDAR constraints may weaken in geometrically degenerate scenes. To reflect this behavior, we scale the nominal covariance of each modality online using three complementary indicators: innovation consistency, measurement quality, and observability. For each sensing modality  $m \in \{v, \ell\}$ , let  $\mathbf{R}_0^m$  denote its nominal covariance. The adaptive covariance is defined using a scale modulation model:

$$\mathbf{R}_k^{m*} = \eta_k^m \mathbf{R}_0^m, \quad \eta_k^m = \eta_{\text{NIS},k}^m \cdot \eta_{\text{qual},k}^m \cdot \eta_{\text{obs},k}^m. \quad (5)$$

The joint adaptive covariance is then written as

$$\mathbf{R}_k^* = \text{blkdiag}(\mathbf{R}_k^{v*}, \mathbf{R}_k^{\ell*}), \quad (6)$$

which replaces the original covariance in the innovation covariance  $\mathbf{S}$  and the Kalman gain computation, thereby enabling online redistribution of multi-modal weights within a unified filtering framework [36].

This three-factor decomposition has clear physical and statistical meaning. Specifically,  $\eta_{\text{NIS}}$  enforces statistical consistency and prevents over-confidence,  $\eta_{\text{qual}}$  reflects the reliability of the sensing and matching process, and  $\eta_{\text{obs}}$  encodes geometric observability and suppresses spurious information gain along degenerate directions. Their multiplicative combination ensures that degradation in any one dimension can trigger covariance inflation, thereby avoiding incorrect confidence assignment caused by the failure of any single indicator.

#### Statistical Consistency Driven by NIS and Chi-Square Testing

At update instant  $k$ , let  $\mathbf{r}_k^m$  denote the innovation (residual) vector of modality  $m$ . Its linearized innovation covariance is

$$\mathbf{S}_k^m = \mathbf{H}_k^m \mathbf{P}_{k|k-1} \mathbf{H}_k^{m\top} + \mathbf{R}_0^m, \quad (7)$$

where  $\mathbf{H}_k^m$  is the measurement Jacobian and  $\mathbf{P}_{k|k-1}$  is the predicted state covariance. The normalized innovation squared (NIS) is defined as [34,37]

$$\text{NIS}_k^m = \mathbf{r}_k^{m\top} \left( \mathbf{H}_k^m \mathbf{P}_{k|k-1} \mathbf{H}_k^{m\top} + \mathbf{R}_0^m \right)^{-1} \mathbf{r}_k^m. \quad (8)$$

If the noise model and linearization are statistically valid, then

$$\text{NIS}_k^m \sim \chi^2(d_m), \quad \mathbb{E}[\text{NIS}_k^m] = d_m, \quad (9)$$

where  $d_m$  is the residual dimension (degree of freedom) of modality  $m$ . When  $\text{NIS}_k^m$  persistently exceeds its expectation  $d_m$ , this indicates either underestimated measurement noise or increased model mismatch and outlier contamination. In such cases, directly using a fixed covariance  $\mathbf{R}_0^m$  would cause the filter to become over-confident.

Accordingly, the statistical consistency scaling factor is defined as

$$\eta_{\text{NIS},k}^m = \frac{\text{NIS}_k^m}{d_m}. \quad (10)$$

The NIS measures whether the current residual magnitude is consistent with the assumed noise model. If the filter is well calibrated, the NIS should stay close to its expected chi-square range. A persistently large NIS indicates that the modality is less reliable than expected, possibly due to model mismatch or outliers. In this case, we increase its covariance and reduce its update weight.

#### Information-Matrix and Observability Perspective: Suppression of Degenerate Directions

Relying only on innovation statistics may still be insufficient in geometrically degenerate cases. When the column space of  $\mathbf{H}_k^m$  becomes rank-deficient or poorly conditioned—for example, in planar structures, pure translation motion, or weak-texture visual scenes—short-term innovations may remain deceptively small, while false information gain may still be injected along weakly observable directions. As a result, the covariance may shrink unrealistically in degenerate directions.

From the information-matrix perspective, the instantaneous information contribution of modality  $m$  to the state is

$$\mathcal{I}_k^m = \mathbf{H}_k^{m\top} (\mathbf{R}_k^{m*})^{-1} \mathbf{H}_k^m, \quad (11)$$

whose spectral structure determines the constraint strength on different degrees of freedom. When the eigen-spectrum of  $\mathbf{H}_k^{m\top} \mathbf{H}_k^m$  collapses, the system becomes unobservable or weakly observable along the corresponding directions. In this case, the influence of the modality should be actively reduced to avoid over-confidence in degenerate subspaces.

Motivated by this observation, an observability-aware modulation factor is introduced [38]:

$$\eta_{\text{obs},k}^m = \frac{1}{\log \det(\mathbf{H}_k^{m\top} \mathbf{H}_k^m + \epsilon \mathbf{I})}, \quad (12)$$

where  $\epsilon > 0$  is a small positive constant for numerical stability. As  $\det(\mathbf{H}^\top \mathbf{H})$  decreases,  $\eta_{\text{obs},k}^m$  increases, which inflates the covariance and attenuates the measurement weight under geometric degeneration. In practice, to avoid numerical instability caused by extremely small or ill-conditioned log-determinant values, the observability factor is clipped to a bounded range  $[\eta_{\min}, \eta_{\max}]$  and optionally smoothed using an exponential moving average. Physically, this term acts as a gate on the amount of observable information injected into the filter, thereby preventing unrealistically high confidence along poorly constrained directions.

#### Maximum-Likelihood Interpretation and Data Quality Awareness

In addition to statistical consistency and geometric observability, multi-modal observations also exhibit substantial differences in data quality. Typical examples include visual tracking failures, abrupt illumination variation, motion blur, sparse LiDAR correspondences, and poor ICP fitting. These effects are often manifested as reduced matching reliability, heavier-tailed residual distributions, and a higher proportion of outliers.

To explicitly incorporate such information into covariance modeling, a modality quality score  $\text{score}_k^m$  is introduced, and the quality-aware modulation factor is defined as

$$\eta_{\text{qual},k}^m = \frac{1}{\text{score}_k^m + \epsilon}. \quad (13)$$

Here,  $\text{score}_k^m$  is a modality-specific quality measure. For the visual modality, it can be constructed from statistics such as feature-tracking success rate, photometric residual RMS, and image gradient variance. For the LiDAR modality, it can be characterized by ICP fitness, the number of valid correspondences, and local surface anisotropy. This term resembles the weighting mechanism in IRLS-based robust optimization [39,40]; however, unlike purely residual-driven robust weights, it is derived from interpretable sensing and matching quality indicators, thereby enabling reliability regulation *before* residual contamination fully propagates into the update step.

To make the adaptive covariance adjustment mechanism reproducible and practically implementable, the modality quality score  $\text{score}_k^m$  must be explicitly defined in mathematical form. Accordingly, this work adopts a weighted normalized fusion model:

$$\text{score}_k^m = \sum_{i=1}^{N_m} w_i^m \phi_i^m(k), \quad \sum_i w_i^m = 1, \quad (14)$$

where  $\phi_i^m(k) \in [0, 1]$  is the  $i$ -th normalized quality indicator and  $w_i^m$  is its weighting coefficient.

Visual modality quality score.

For the visual modality  $m = v$ , three quality indicators are defined as follows:

1. Feature-tracking success rate:

$$\phi_1^v(k) = \frac{N_{\text{tracked}}}{N_{\text{detected}}}. \quad (15)$$

2. Photometric residual stability:

$$\phi_2^v(k) = \exp\left(-\frac{\text{RMS}_{\text{photo}}^2}{\sigma_{\text{photo}}^2}\right), \quad (16)$$

where  $\text{RMS}_{\text{photo}}$  is the root-mean-square photometric residual of the current frame.

3. Image gradient informativeness:

$$\phi_3^v(k) = \frac{\text{Var}(\nabla I)}{\text{Var}_{\text{ref}}}. \quad (17)$$

The visual quality score is therefore defined as

$$\text{score}_k^v = w_1^v \phi_1^v + w_2^v \phi_2^v + w_3^v \phi_3^v. \quad (18)$$

LiDAR modality quality score.

For the LiDAR modality  $m = \ell$ , three quality indicators are constructed as follows:

1. ICP goodness-of-fit:

$$\phi_1^\ell(k) = \exp\left(-\frac{\text{RMSE}_{\text{ICP}}^2}{\sigma_{\text{ICP}}^2}\right). \quad (19)$$

2. Valid correspondence ratio:

$$\phi_2^\ell(k) = \frac{N_{\text{correspondence}}}{N_{\text{points}}}. \quad (20)$$

3. Surface-structure anisotropy indicator:

$$\phi_3^\ell(k) = \frac{\lambda_3}{\lambda_1 + \lambda_2 + \lambda_3}, \quad (21)$$

where  $\lambda_i$  are the eigenvalues of the local covariance matrix and are used to quantify the degree of structural non-degeneracy.

The LiDAR quality score is then defined as

$$\text{score}_k^\ell = w_1^\ell \phi_1^\ell + w_2^\ell \phi_2^\ell + w_3^\ell \phi_3^\ell. \quad (22)$$

The final quality-aware modulation factor becomes

$$\eta_{\text{qual},k}^m = \frac{1}{\text{score}_k^m + \epsilon}. \quad (23)$$

When the observation quality deteriorates, the score decreases and  $\eta_{\text{qual},k}^m$  increases, resulting in covariance inflation and a reduced contribution of that modality to the posterior estimate.

From a maximum-likelihood viewpoint, the covariance can be interpreted as an unknown noise-scale parameter and written as  $\mathbf{R}_k^{m*} = \eta_k^m \mathbf{R}_0^m$ . When observation quality decreases, a reasonable maximum-likelihood strategy tends to increase the noise scale so that the degraded observation remains statistically explainable. In this sense, Equation (13)

can be viewed as introducing a data-quality-driven prior adjustment on the noise scale. It is complementary to the NIS-based online scale correction:  $\eta_{\text{NIS}}$  captures statistical deviation at the *result level*, whereas  $\eta_{\text{qual}}$  captures degradation at the *process level*.

### Unified Interpretation of Fusion Weight Regulation and Implementation Considerations

In summary, the three-factor model acts as a sensor reliability scheduler:  $\eta_{\text{NIS}}$  reflects statistical agreement,  $\eta_{\text{qual}}$  captures sensing reliability, and  $\eta_{\text{obs}}$  suppresses over-confidence in weakly constrained directions. Their combined effect enables more balanced and robust multi-modal fusion.

Whenever any modality becomes degraded in one of these dimensions,  $\eta_k^m$  increases, which inflates  $\mathbf{R}_k^{m*}$  and correspondingly attenuates the information contribution  $\mathcal{I}_k^m$ . As a result, the proposed mechanism realizes adaptive redistribution of multi-modal fusion weights in an online and principled manner. In essence, it embeds a self-assessment loop of sensing reliability into the front-end filter, allowing the system to continuously rebalance heterogeneous observations under changing environmental and motion conditions.

It should be emphasized that this adaptive covariance mechanism primarily improves the local statistical reliability of front-end fusion by regulating modality confidence within the filter update. However, covariance reweighting alone does not explicitly constrain the long-term agreement among different sensing modalities at the trajectory level, nor does it directly eliminate accumulated drift over extended operation. Therefore, to further enhance global consistency, the subsequent back-end modules introduce additional geometric loop-closure constraints and cross-modal consistency modeling, so that locally reliable estimation and globally consistent optimization can be jointly achieved within a unified SLAM framework.

### 3.3. mBTC: A Stable Geometric-Skeleton-Driven Multi-Channel Structural Representation Framework

In repetitive or degenerate environments, descriptors constructed from only one type of geometric statistic tend to exhibit insufficient separability, which degrades the reliability of loop detection for global relocalization. To address this issue, this section introduces a multi-channel Binary Triangle Constraint (mBTC) framework, which enhances place representation by combining stable geometric skeletons with complementary contextual channels.

Given a raw observation space  $\mathcal{X}$ , place recognition can be abstracted as a mapping

$$f : \mathcal{X} \rightarrow \mathcal{D}, \quad (24)$$

where  $\mathcal{D}$  denotes the descriptor space used for retrieval and matching. An ideal descriptor should preserve geometric invariance while retaining as much place-discriminative information as possible. However, in conventional single-channel formulations, the descriptor is typically composed of only one class of geometric statistics,

$$\mathbf{d} = [d^{(1)}, d^{(2)}, \dots, d^{(C)}], \quad (25)$$

where the channels are often highly correlated or partially redundant. In such cases, the effective discriminative dimensionality is limited, and the place-discriminative information  $S$  cannot be significantly improved beyond that of a dominant single channel.

To overcome this limitation, the proposed mBTC framework expands the descriptor space through multiple physically interpretable and statistically complementary channels. Instead of encoding only the geometric shape of a local structure, mBTC represents each candidate loop unit as a combination of a stable triangle-based geometric skeleton and a set of contextual attributes. This design preserves rigid-body invariance while in-

creasing the effective rank of the descriptor space, thereby improving discriminability in complex scenes.

### 3.3.1. Multi-Channel Contextual Modeling

Although the triangle skeleton provides rigid-body-invariant geometric structure, shape information alone is often insufficient in repetitive or weakly distinctive environments. To improve discriminability, mBTC augments each stable triangle with a set of contextual channels extracted from its local support region.

Let the gravity-aligned LiDAR point cloud be represented as

$$\mathcal{P} = \left\{ \mathbf{p} = [x, y, z]^\top, I(\mathbf{p}) \right\}, \quad (26)$$

where  $z$  is the gravity direction and  $I(\mathbf{p})$  is the LiDAR intensity. For a triangle  $\Delta_{ijk}$  formed by vertices  $\mathbf{p}_i, \mathbf{p}_j, \mathbf{p}_k$ , a local neighborhood of radius  $r$  is defined around each vertex:

$$\mathcal{N}(u) = \{ \mathbf{p} \in \mathcal{P} \mid \|\mathbf{p} - \mathbf{p}_u\| < r \}, \quad u \in \{i, j, k\}. \quad (27)$$

The joint support region is then defined as

$$\mathcal{U}_{ijk} = \mathcal{N}(i) \cup \mathcal{N}(j) \cup \mathcal{N}(k). \quad (28)$$

This union-based support region is symmetric with respect to vertex permutation and provides more stable statistics than a single-point neighborhood.

Within  $\mathcal{U}_{ijk}$ , four complementary contextual channels are extracted.

First, the vertical span channel characterizes the local vertical structural variation:

$$\Delta h_{ijk} = \max_{\mathbf{p} \in \mathcal{U}_{ijk}} z(\mathbf{p}) - \min_{\mathbf{p} \in \mathcal{U}_{ijk}} z(\mathbf{p}), \quad (29)$$

which is normalized as

$$\tilde{h}_{ijk} = \text{clip} \left( \frac{\Delta h_{ijk}}{h_{\text{ref}}}, 0, 1 \right). \quad (30)$$

Second, the density channel measures the spatial compactness of the local structure. Considering the range-dependent sparsity of LiDAR measurements, a distance-compensated density is defined as

$$\rho_{ijk} = \frac{\sum_{\mathbf{p} \in \mathcal{U}_{ijk}} \|\mathbf{p}\|^2}{|\mathcal{U}_{ijk}| \cdot \frac{4}{3}\pi r^3}, \quad (31)$$

and normalized by

$$\tilde{\rho}_{ijk} = \text{clip} \left( \frac{\rho_{ijk}}{\rho_{\text{ref}}}, 0, 1 \right). \quad (32)$$

Third, the normal-variation channel captures local surface irregularity. Let  $\mathbf{n}(\mathbf{p})$  be the unit normal vector of point  $\mathbf{p}$ , and define the mean normal direction as

$$\bar{\mathbf{n}}_{ijk} = \frac{\sum_{\mathbf{p} \in \mathcal{U}_{ijk}} \mathbf{n}(\mathbf{p})}{\left\| \sum_{\mathbf{p} \in \mathcal{U}_{ijk}} \mathbf{n}(\mathbf{p}) \right\|}. \quad (33)$$

The corresponding normal variation is then computed as

$$\sigma_n^2(i, j, k) = \frac{1}{|\mathcal{U}_{ijk}|} \sum_{\mathbf{p} \in \mathcal{U}_{ijk}} \arccos^2(\mathbf{n}(\mathbf{p})^\top \bar{\mathbf{n}}_{ijk}), \quad (34)$$

and normalized by

$$\tilde{v}_{ijk} = \text{clip}\left(\frac{\sigma_n(i, j, k)}{v_{\text{ref}}}, 0, 1\right). \quad (35)$$

Fourth, the mean intensity channel captures local reflectance properties:

$$\bar{I}_{ijk} = \frac{1}{|\mathcal{U}_{ijk}|} \sum_{\mathbf{p} \in \mathcal{U}_{ijk}} I(\mathbf{p}), \quad (36)$$

which is standardized within each frame and clipped to  $[0, 1]$ :

$$\tilde{I}_{ijk} = \text{clip}\left(\frac{\bar{I}_{ijk} - \mu_I}{\sigma_I}, 0, 1\right). \quad (37)$$

As illustrated in Figure 2, the conventional BTC descriptor encodes only the geometric scale information of the triangle (left), whereas the proposed mBTC (right) augments the same rigid-body invariant skeleton with four additional contextual channels—vertical span  $\Delta h$ , spatial compactness  $\rho$ , surface irregularity  $v$ , and reflectance intensity  $\bar{I}$ . These multi-channel attributes provide complementary information beyond pure geometry, greatly improving place recognition robustness and discriminability in complex and repetitive environments.

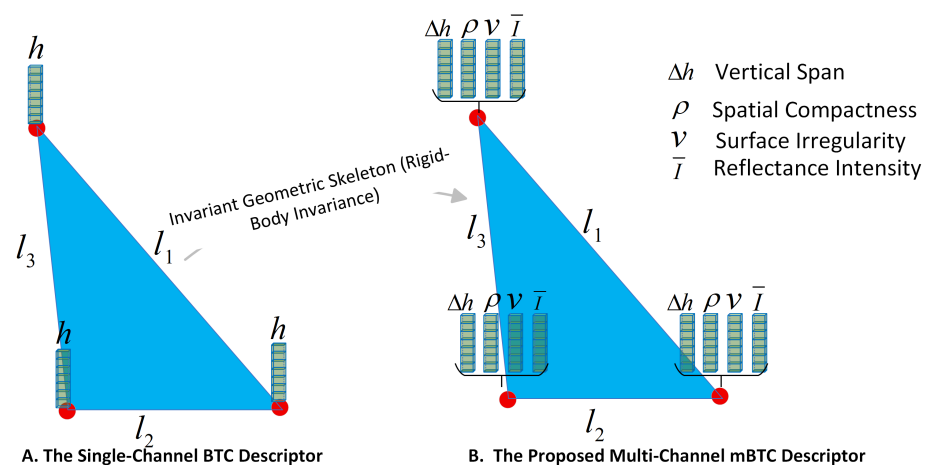


**Figure 2.** Comparison of the conventional single-channel Binary Triangle Constraint (BTC) descriptor and the proposed multi-channel Binary Triangle Constraint (mBTC) descriptor. The arrows indicate the extension from a single geometric channel to a multi-channel structural representation. **(A)** Conventional BTC encodes only the geometric scale information of the triangle, which is invariant under rigid-body transformations. **(B)** The proposed mBTC extends the same geometric skeleton with four additional contextual channels: vertical span  $\Delta h$ , spatial compactness  $\rho$ , surface irregularity  $v$ , and reflectance intensity  $\bar{I}$ . These multi-channel attributes enhance descriptor discriminability and robustness in repetitive or viewpoint-variant environments.

The complete continuous representation of a triangle can therefore be written as

$$\mathbf{f}_{ijk} = \begin{bmatrix} \mathbf{g}_{ijk} \\ \mathbf{q}_{ijk} \end{bmatrix}, \quad (38)$$

where  $\mathbf{g}_{ijk}$  denotes the geometric skeleton attributes and  $\mathbf{q}_{ijk}$  denotes the contextual channels. Since these channels describe different physical properties of the local environment, they are typically weakly correlated and jointly improve the effective rank of the descriptor space. This statistical complementarity allows mBTC to better distinguish places with similar geometric layouts but different local structural or radiometric characteristics.

Finally, the multi-channel feature vector for triangle  $(i, j, k)$  is defined as:

$$\mathbf{f}_{ijk} = [\tilde{l}_1, \tilde{l}_2, \tilde{l}_3, \tilde{h}_{ijk}, \tilde{\rho}_{ijk}, \tilde{v}_{ijk}, \tilde{I}_{ijk}]^T \in [0, 1]^7, \quad (39)$$

where the first three elements are geometric invariants from BTC, and the latter four are contextual attributes. This hybrid formulation enables mBTC to encode not only the triangle's intrinsic shape geometry but also the local environmental structure and material response.

The multi-channel continuous features are transformed into compact binary codes through subsequent quantization and encoding steps, enabling efficient large-scale retrieval while preserving the key discriminative information. In this way, mBTC achieves a balanced design between geometric invariance, descriptive richness, and matching efficiency, making it suitable for robust loop detection in long-term and large-scale SLAM scenarios.

### 3.3.2. Binary Encoding and Descriptor Formation

After constructing the multi-channel local geometric context at each triangle vertex, an efficient encoding scheme is required to transform these continuous descriptors into a compact and discriminative representation suitable for large-scale loop detection. This section describes the binary encoding and descriptor formation process, which converts the normalized multi-channel attributes into a unified binary code while preserving their relative geometric relationships. Binary encoding is adopted to reduce memory consumption and enable fast similarity computation through bitwise operations, which is critical for real-time loop detection in long-term SLAM. To preserve BTC's compactness and enable efficient loop retrieval, the continuous feature vector  $\mathbf{f}_{ijk}$  is quantized into binary codes and aggregated into a frame-level descriptor.

The mBTC binary representation and retrieval process are implemented through the following three steps:

#### 1. Adaptive Quantization

For each feature dimension  $n \in \{1, \dots, 7\}$ , adaptive thresholds  $\{\tau_{n,q}\}$  are determined by quantiles of all triangles within the current frame.

The feature value  $f_{ijk}^{(n)}$  is quantized as

$$q_n(i, j, k) = \sum_{q=1}^{Q-1} \mathbb{I}\{f_{ijk}^{(n)} > \tau_{n,q}\}, \quad \mathbf{g}_n = \text{Gray}(q_n) \in \{0, 1\}^{\lceil \log_2 Q \rceil}, \quad (40)$$

where Gray encoding ensures local continuity in Hamming space. All encoded channels are concatenated to form the mBTC binary vector:

$$\mathbf{b}_{ijk} = [\mathbf{g}_1 \parallel \mathbf{g}_2 \parallel \dots \parallel \mathbf{g}_7] \in \{0, 1\}^B, \quad B = 7 \lceil \log_2 Q \rceil. \quad (41)$$

#### 2. Histogram Pooling

Histogram pooling aggregates local binary features into a fixed-size frame-level representation, suppressing the influence of repetitive structures via term frequency-inverse document frequency (TF-IDF) [41] reweighting. This enables robust scene discrimination and similarity retrieval, which is crucial for real-time loop closure.

Each binary code  $\mathbf{b}_{ijk}$  is mapped to a discrete index using a locality-sensitive hash function  $\phi(\cdot)$ , and the frame-level histogram descriptor is accumulated:

$$\mathbf{h}_t = \sum_{(i,j,k) \in \mathcal{T}_t} \mathbf{e}_{\phi(\mathbf{b}_{ijk})}, \quad \hat{\mathbf{h}}_t = \frac{\mathbf{h}_t}{\|\mathbf{h}_t\|_1}. \quad (42)$$

The weighting scheme follows the principle of TF-IDF in information retrieval, where frequently occurring but less informative patterns are downweighted:

$$\tilde{\mathbf{h}}_t = \frac{\mathbf{w} \odot \hat{\mathbf{h}}_t}{\|\mathbf{w} \odot \hat{\mathbf{h}}_t\|_1}, \quad w_k = \log \frac{N}{1 + \text{df}_k}. \quad (43)$$

### 3. Similarity Computation and Fast Matching

After TF-IDF reweighting, the frame descriptor can be used for global similarity computation. In this work, cosine similarity is adopted as the similarity measure:

$$\text{Sim}(c, r) = \frac{\mathbf{h}_c^\top \mathbf{h}_r}{\|\mathbf{h}_c\|_2 \|\mathbf{h}_r\|_2}, \quad (44)$$

where  $c$  and  $r$  denote the current frame and the candidate frame, respectively.

#### 3.3.3. Loop Retrieval and Verification

At runtime, the current frame descriptor  $\tilde{\mathbf{h}}_c$  is compared against historical descriptors  $\{\tilde{\mathbf{h}}_r\}$  stored in the loop database. Top-ranked loop candidates are then validated through geometric registration using point-to-plane ICP, producing the loop constraint.

$$E_{\text{loop}} = \mathbf{e}_{\text{loop}}^\top \mathbf{\Omega}_{\text{loop}} \mathbf{e}_{\text{loop}}, \quad \mathbf{e}_{\text{loop}} = \text{Log} \left( \left( \mathbf{T}_{ij}^{\text{loop}} \right)^{-1} \mathbf{T}_{iw}^{-1} \mathbf{T}_{jw} \right), \quad (45)$$

where  $\mathbf{T}_{ij}^{\text{loop}}$  denotes the ICP-estimated relative pose between frame  $i$  and  $j$ , and  $\mathbf{\Omega}_{\text{loop}}$  is derived from the registration covariance. Confirmed loop closures are inserted into the global factor graph to correct accumulated drift and maintain long-term global consistency.

This reliability-aware loop closure leads to superior performance over appearance-only retrieval methods such as Scan Context, especially in geometrically degenerate or perceptually challenging environments.

### 3.4. Cross-Modal Consistency Factor Graph Optimization

To achieve globally consistent trajectory estimation, the proposed framework incorporates a joint factor-graph optimization module that consolidates inertial, visual, LiDAR, and loop-closure constraints. Nevertheless, independently integrating modality-specific odometry factors may lead to local inconsistencies, particularly in the presence of transient sensor degradation. To alleviate modality drift divergence, we introduce a cross-modal consistency factor to explicitly enforce agreement between the visual and LiDAR relative motions.

#### 3.4.1. Factor Graph Formulation

Let  $\mathbf{T}_{wb,k} \in SE(3)$  denote the body pose at keyframe  $k$ , and  $\mathbf{T}_{bc}, \mathbf{T}_{bl} \in SE(3)$  represent the known extrinsic calibrations for the camera and LiDAR, respectively. The back-end optimization seeks the maximum a posteriori (MAP) estimate by minimizing the sum of all residuals defined over the factor graph:

$$\mathcal{L}^* = \arg \min_{\{\mathbf{T}_{wb,k}\}} \left( \sum_{(i,j) \in \mathcal{E}_{imu}} \|\mathbf{e}_{imu}^{ij}\|_{\Omega_{imu}}^2 + \sum_{(i,j) \in \mathcal{E}_v} \|\mathbf{e}_v^{ij}\|_{\Omega_v}^2 + \sum_{(i,j) \in \mathcal{E}_\ell} \|\mathbf{e}_\ell^{ij}\|_{\Omega_\ell}^2 + \sum_{(i,j) \in \mathcal{E}_{loop}} \|\mathbf{e}_{loop}^{ij}\|_{\Omega_{loop}}^2 + \sum_{(i,j) \in \mathcal{E}_{cm}} \|\mathbf{e}_{cm}^{ij}\|_{\Omega_{cm}}^2 \right) \quad (46)$$

Here,  $\mathcal{E}_{imu}$ ,  $\mathcal{E}_v$ ,  $\mathcal{E}_\ell$ ,  $\mathcal{E}_{loop}$ , and  $\mathcal{E}_{cm}$  denote the sets of IMU pre-integration, visual reprojection, LiDAR point-to-plane, loop closure, and cross-modality constraints, respectively;  $\Omega$  represents the associated information matrix of each factor. As illustrated in Figure 3, the back-end optimization jointly integrates IMU, LiDAR, visual, loop-closure, and cross-modal consistency factors within a unified factor-graph framework.

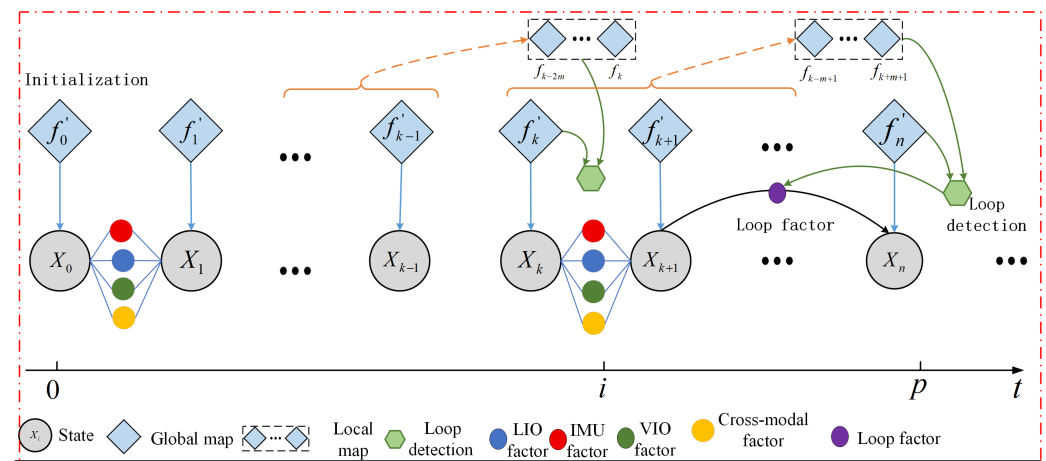


**Figure 3.** Overview of the factor-graph optimization back-end with the proposed cross-modal consistency factor. The graph jointly incorporates: IMU factors (red), LiDAR-Inertial Odometry (LIO) factors (blue), Visual-Inertial Odometry (VIO) factors (green), the proposed cross-modal consistency factors (yellow), and loop closure factors (purple). Green hexagons indicate loop detection results, and purple loop factors are added when a revisited place is detected. The cross-modal consistency factors (yellow) explicitly enforce short-baseline motion agreement between visual and LiDAR odometries, bridging front-end estimation and global optimization.

### 3.4.2. Relative Motion Constraint Modeling

For a keyframe pair  $(i, j)$ , the front-end provides relative pose measurements from both camera and LiDAR:

$$\hat{\mathbf{T}}_{ij}^c, \hat{\mathbf{T}}_{ij}^\ell \in \text{SE}(3),$$

with associated covariances  $\Sigma_{ij}^c$  and  $\Sigma_{ij}^\ell$ .

The predicted relative motions derived from current trajectory estimates are:

$$\mathbf{T}_{ij}^c = \mathbf{T}_{bc}^{-1} \mathbf{T}_{wb,i}^{-1} \mathbf{T}_{wb,j} \mathbf{T}_{bc}, \quad \mathbf{T}_{ij}^\ell = \mathbf{T}_{bl}^{-1} \mathbf{T}_{wb,i}^{-1} \mathbf{T}_{wb,j} \mathbf{T}_{bl}. \quad (47)$$

Then, the visual and LiDAR odometry residuals are formulated as:

$$\mathbf{e}_v^{ij} = \log((\hat{\mathbf{T}}_{ij}^c)^{-1} \mathbf{T}_{ij}^c), \quad \mathbf{e}_\ell^{ij} = \log((\hat{\mathbf{T}}_{ij}^\ell)^{-1} \mathbf{T}_{ij}^\ell). \quad (48)$$

### 3.4.3. Cross-Modal Consistency Factor

Although visual and LiDAR odometry provide independent relative motion estimates, they describe the same underlying physical motion of the robot over the same keyframe interval. Instead of introducing an additional absolute-motion measurement, we define

a soft agreement factor that penalizes the discrepancy between the two modality-specific relative-motion residuals in the  $\mathfrak{se}(3)$  local error space.

$$\mathbf{e}_{\text{cm}}^{ij} = \mathbf{e}_v^{ij} - \mathbf{e}_\ell^{ij}, \quad (49)$$

where  $\mathbf{e}_v^{ij}$  and  $\mathbf{e}_\ell^{ij}$  denote the visual and LiDAR odometry residuals derived from the predicted relative motion in Section 3.4.2.

The associated Mahalanobis cost is defined as

$$E_{ij}^{cm} = (\mathbf{e}_{ij}^{cm})^\top \Omega_{ij}^{cm} \mathbf{e}_{ij}^{cm}, \quad (50)$$

where, under the common approximation that the visual and LiDAR relative-motion uncertainties are locally independent in the linearized error space, the covariance of their residual difference can be approximated as the sum of the two covariance terms. Accordingly, the information matrix is defined as

$$\Omega_{ij}^{cm} = (\Sigma_{ij}^v + \Sigma_{ij}^\ell)^{-1}. \quad (51)$$

The above formulation naturally reflects the confidence levels of both modalities: when visual tracking degrades (e.g., motion blur or low texture), its uncertainty increases, reducing its influence in the consistency factor. Conversely, in geometrically degenerate LiDAR environments (e.g., planar corridors), the visual modality provides stronger constraints. As a result, the consistency factor becomes less influential when one modality is highly uncertain, while still encouraging agreement when both modalities remain informative.

#### 3.4.4. Linearization and Jacobian Structure

We adopt a right-multiplicative perturbation model for pose updates:

$$\mathbf{T}_{wb,k} \leftarrow \mathbf{T}_{wb,k} \exp(\delta \boldsymbol{\xi}_k), \quad \delta \boldsymbol{\xi}_k \in \mathfrak{se}(3). \quad (52)$$

Linearizing the visual and LiDAR residuals around the current estimate yields:

$$\mathbf{e}_v^{ij} \approx \mathbf{e}_v^{ij,0} + \mathbf{J}_{v,i}^{ij} \delta \boldsymbol{\xi}_i + \mathbf{J}_{v,j}^{ij} \delta \boldsymbol{\xi}_j, \quad (53)$$

$$\mathbf{e}_\ell^{ij} \approx \mathbf{e}_\ell^{ij,0} + \mathbf{J}_{\ell,i}^{ij} \delta \boldsymbol{\xi}_i + \mathbf{J}_{\ell,j}^{ij} \delta \boldsymbol{\xi}_j. \quad (54)$$

Since the cross-modal error is defined as the difference between two odometry residuals, its first-order linearization becomes:

$$\mathbf{e}_{\text{cm}}^{ij} \approx \mathbf{e}_{\text{cm}}^{ij,0} + \mathbf{J}_{\text{cm},i}^{ij} \delta \boldsymbol{\xi}_i + \mathbf{J}_{\text{cm},j}^{ij} \delta \boldsymbol{\xi}_j, \quad \mathbf{e}_{\text{cm}}^{ij,0} = \mathbf{e}_v^{ij,0} - \mathbf{e}_\ell^{ij,0}, \quad (55)$$

$$\mathbf{J}_{\text{cm},i}^{ij} = \mathbf{J}_{v,i}^{ij} - \mathbf{J}_{\ell,i}^{ij}, \quad \mathbf{J}_{\text{cm},j}^{ij} = \mathbf{J}_{v,j}^{ij} - \mathbf{J}_{\ell,j}^{ij}. \quad (56)$$

The Jacobians of the cross-modal factor are obtained by a direct subtraction of existing Jacobians from visual and LiDAR odometry factors, without requiring any additional symbolic derivation or re-computation.

The proposed cross-modal consistency factor effectively binds the visual and LiDAR odometry in a unified error-space, suppressing modality-dependent drift while respecting their respective uncertainties. This joint constraint yields smoother trajectories, improved resilience under single-modality degradation, and stronger global consistency, particularly over long-term navigation.

## 4. Results

All methods were implemented in C++ and executed on Ubuntu 20.04 with ROS Noetic on a platform equipped with an Intel Core i7-10750H CPU and 24 GB RAM. For all baseline methods, we used their publicly available implementations and kept the recommended default settings whenever possible.

For the proposed method, the main implementation-related parameters were fixed across all datasets unless otherwise stated. In the uncertainty-aware fusion module, the covariance scaling factor was bounded within  $[\eta_{\min}, \eta_{\max}] = [0.5, 5.0]$ , and the observability-related scaling term was smoothed using an exponential moving average with coefficient  $\alpha = 0.8$ . In the quality-aware modulation, the visual quality weights were set to  $(w_1^v, w_2^v, w_3^v) = (0.4, 0.3, 0.3)$ , and the LiDAR quality weights were set to  $(w_1^l, w_2^l, w_3^l) = (0.4, 0.3, 0.3)$ . The numerical stability constant was fixed to  $\epsilon = 10^{-6}$ .

For loop detection, the top- $K$  retrieval candidates were selected with  $K = 10$ , and a loop pair was regarded as valid only when the temporal separation was larger than 50 frames to avoid trivial neighboring matches. In the Precision–Recall evaluation, two frames were considered a true loop pair if the Euclidean distance between their ground-truth poses was below 5 m. No dataset-specific parameter tuning was performed for the proposed method unless explicitly stated.

To evaluate the proposed framework, experiments are conducted on four public multi-sensor datasets: KITTI [42], M2DGR [43], M3DGR [44] and NCLT [45]. These datasets cover indoor and outdoor environments, dynamic scenes, repeated structures, and long-term trajectories, providing synchronized LiDAR, visual, and inertial measurements for benchmarking LiDAR–Inertial–Visual SLAM systems.

### 4.1. Loop Detection Performance

To evaluate the retrieval performance of the proposed multi-channel geometric structural descriptor (mBTC) in loop detection, comparative experiments are first conducted on several public datasets. The classical LiDAR loop detection method STD and the triangle-geometry-based BTC descriptor are selected as baseline methods. Their retrieval performance is comprehensively evaluated using Precision–Recall (P–R) curves, Recall@N, and Average Precision (AP).

The experiments are conducted on three datasets, namely KITTI Odometry, NCLT, and M2DGR. For each LiDAR scan, the system retrieves the most similar candidate frames from the historical database, and the retrieved results are then judged as true or false loop closures according to the ground-truth trajectories. Specifically, two frames are regarded as a true loop pair if their spatial distance is smaller than a predefined threshold.

Figure 4 shows the Precision–Recall curves of different methods on all test sequences. It can be observed that the proposed mBTC consistently achieves improved performance across all datasets. Compared with STD, mBTC maintains higher precision even in the high-recall region, indicating that the proposed descriptor has stronger discriminative capability and can effectively reduce false matches caused by structurally similar scenes. In comparison with the single-channel geometric descriptor BTC, the curves of mBTC are generally shifted toward the upper-right region, demonstrating that the introduction of multi-channel structural information improves the stability and reliability of loop detection.

To further quantify the retrieval performance of different methods, Table 1 reports the Recall@N values and Average Precision (AP) for different values of  $N$ . Recall@N represents the probability that at least one true loop closure is included within the top- $N$  retrieved candidates, and directly reflects the retrieval capability of the descriptor. As shown in the table, mBTC achieves the highest Recall@1 and Recall@10 on all evaluated datasets. For example, on the KITTI dataset, the Recall@1 of mBTC is improved by more than 10%

compared with STD, while also showing a clear advantage over BTC. This result indicates that the introduction of multi-channel geometric contextual information can enhance the descriptor's ability to characterize structural differences in the environment. The AP results further confirm the superiority of mBTC in overall retrieval performance. Since AP jointly considers precision and recall over different thresholds, it provides a more comprehensive evaluation of the descriptor. On all three datasets, mBTC achieves the highest AP values, demonstrating stable and reliable retrieval performance under different scene structures and environmental conditions.

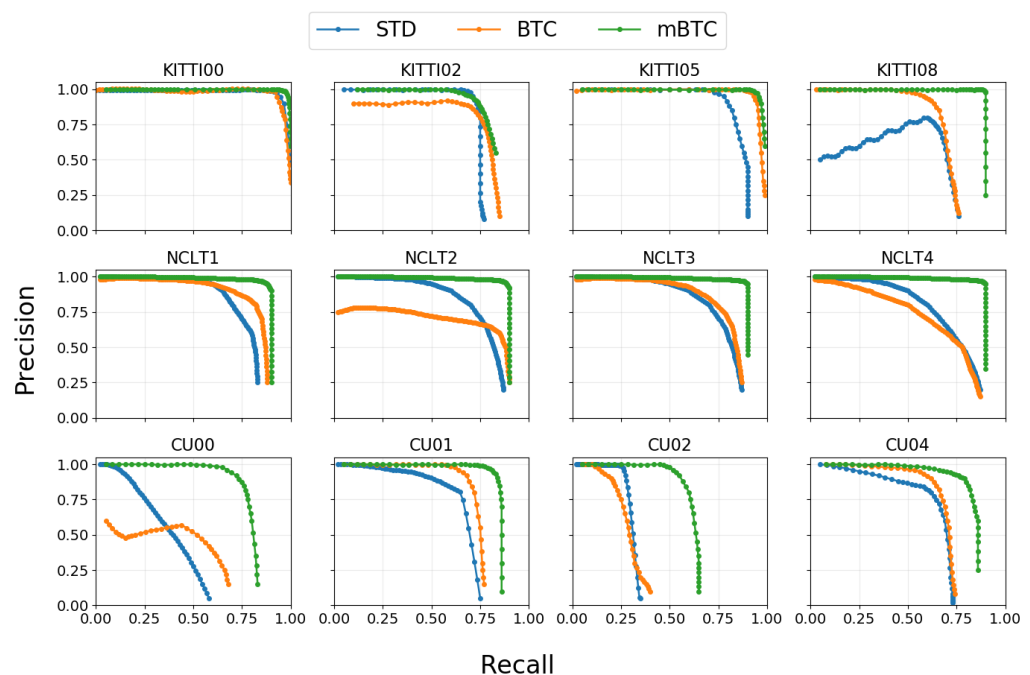


**Figure 4.** PR curves on KITTI, NCLT and M2DGR datasets.

**Table 1.** Recall@N and AP summaries on KITTI00 and NCLT1 datasets.

| Dataset | Methods | Recall@1 | Recall@5 | Recall@10 | Recall@20 | AP   |
|---------|---------|----------|----------|-----------|-----------|------|
| KITTI00 | STD     | 0.75     | 0.85     | 0.89      | 0.92      | 0.80 |
|         | BTC     | 0.82     | 0.90     | 0.93      | 0.95      | 0.86 |
|         | mBTC    | 0.88     | 0.94     | 0.96      | 0.97      | 0.91 |
| NCLT1   | STD     | 0.78     | 0.86     | 0.90      | 0.93      | 0.82 |
|         | BTC     | 0.83     | 0.91     | 0.94      | 0.96      | 0.88 |
|         | mBTC    | 0.87     | 0.93     | 0.95      | 0.97      | 0.90 |

The performance gain of mBTC mainly comes from two aspects. First, the geometric skeleton constructed from stable triangle structures ensures invariance to viewpoint changes and rigid-body transformations, thereby improving matching stability. Second, the introduced multi-channel structural attributes, including vertical structural distribution, point cloud density, surface irregularity, and reflectance intensity, provide richer contextual information for each geometric structure, enabling the descriptor to better distinguish structurally repetitive scenes.

#### 4.2. Ablation Study on mBTC Channel Design

To further validate the effectiveness of the proposed multi-channel geometric structural representation for loop detection, this subsection presents an ablation study on different encoding strategies to systematically analyze the influence of different combinations of

geometric information channels on descriptor performance. By progressively introducing different geometric attributes and comparing the loop detection performance under different encoding strategies, the contribution of each geometric channel to structural representation can be quantitatively evaluated. To analyze the contribution of different channels to the overall performance, the experiments adopt a progressive accumulation design with the following encoding strategies:

- A0 (Only Triangle): only the basic triangle structural information is used for encoding;
- A1 (+Height): height information is added on top of the triangle structure;
- A2 (+Density): point cloud density information is further incorporated;
- A3 (+Normal): surface normal information is added;
- A4 (+Intensity): LiDAR reflectance intensity is further introduced;
- A5 (All Channels): all geometric information channels are jointly fused.

The experiments are conducted on the KITTI Odometry dataset and the NCLT1 sequence. To comprehensively evaluate loop detection performance, two metrics are adopted, namely Recall@N ( $N = 1, 5, 10, 20$ ) and Average Precision (AP). Recall@N measures the proportion of queries for which at least one true loop closure is successfully retrieved within the top- $N$  candidates, while AP reflects the overall trade-off between precision and recall.

From Table 2, it can be observed that the retrieval performance improves consistently as more geometric channels are incorporated. On KITTI, Recall@10 increases from 0.91 for A0 to 0.96 for A5, while AP improves from 0.85 to 0.91. On NCLT, Recall@10 increases from 0.89 to 0.94, and AP improves from 0.80 to 0.88. These results indicate that each additional channel provides useful complementary information for structural representation.

**Table 2.** Comparison of Recall and AP performance for different channel combinations on the KITTI and NCLT datasets.

| Channel Combination | Recall@1 | Recall@5 | Recall@10 | Recall@20 | AP   |
|---------------------|----------|----------|-----------|-----------|------|
| <b>KITTI 00</b>     |          |          |           |           |      |
| Only Triangle (A0)  | 0.82     | 0.89     | 0.91      | 0.94      | 0.85 |
| +Height (A1)        | 0.83     | 0.90     | 0.92      | 0.95      | 0.87 |
| +Density (A2)       | 0.84     | 0.91     | 0.93      | 0.96      | 0.88 |
| +Normal (A3)        | 0.85     | 0.92     | 0.94      | 0.97      | 0.89 |
| +Intensity (A4)     | 0.86     | 0.93     | 0.95      | 0.97      | 0.90 |
| All Channels (A5)   | 0.88     | 0.94     | 0.96      | 0.97      | 0.91 |
| <b>NCLT</b>         |          |          |           |           |      |
| Only Triangle (A0)  | 0.79     | 0.87     | 0.89      | 0.92      | 0.80 |
| +Height (A1)        | 0.81     | 0.88     | 0.91      | 0.93      | 0.83 |
| +Density (A2)       | 0.82     | 0.89     | 0.92      | 0.94      | 0.84 |
| +Normal (A3)        | 0.83     | 0.90     | 0.93      | 0.95      | 0.86 |
| +Intensity (A4)     | 0.84     | 0.91     | 0.94      | 0.96      | 0.87 |
| All Channels (A5)   | 0.85     | 0.92     | 0.94      | 0.96      | 0.88 |

In particular, height, density, and normal channels contribute stable performance gains by enriching the descriptor with vertical, spatial, and local geometric cues, while intensity further improves discriminability by capturing reflectance differences between surfaces. The best results are achieved by A5, where all channels are jointly fused, demonstrating that the proposed multi-channel representation can significantly improve descriptor discriminability and robustness compared with single-channel encoding.

The ablation results confirm that the proposed mBTC framework effectively benefits from multi-channel geometric information and achieves better stability and generalization across different datasets.

While adding more channels improves descriptor discriminability and retrieval accuracy, it also introduces additional computation in channel extraction, binary encoding, and histogram construction. In our implementation, the runtime increases progressively from A0 to A5, but the overhead remains moderate because loop detection is executed at a lower frequency than the front-end state estimation. In particular, the full multi-channel configuration (A5) incurs an additional runtime of approximately 15–20% compared with the triangle-only baseline, while providing clear gains in both Recall@N and AP on KITTI and NCLT. The additional cost mainly comes from the extraction of contextual attributes (height, density, normal, and intensity) and the increased descriptor dimensionality. This trade-off is acceptable for practical loop-closure applications, where robustness and retrieval reliability are more critical than minimal per-frame descriptor cost.

#### 4.3. Ablation Studies on Fusion and Consistency Mechanisms

##### 4.3.1. Analysis of Adaptive Uncertainty Modeling

To evaluate the effectiveness of the proposed adaptive covariance regulation mechanism, two representative degradation scenarios are considered, and the temporal variations of the three scaling factors,  $\eta_{\text{NIS}}$ ,  $\eta_{\text{qual}}$ , and  $\eta_{\text{obs}}$ , are analyzed. Here,  $\eta_{\text{NIS}}$  measures statistical consistency of the innovation,  $\eta_{\text{qual}}$  reflects observation quality, and  $\eta_{\text{obs}}$  describes geometric observability.

The first scenario is an outdoor–corridor–outdoor sequence, where the system moves from an open area into a corridor and then returns outdoors. Since corridor environments usually exhibit strong structural repetition and limited lateral constraints, this scenario is suitable for evaluating the response of the proposed method under geometric degeneration. The second scenario is a nighttime scene with headlight interference, in which local illumination changes caused by vehicle headlights significantly affect visual matching quality and photometric residual statistics. This scenario is used to analyze the dynamic response of the method under visual degradation.

As shown in Figure 5, all three factors remain low and stable in the open outdoor region, indicating reliable LiDAR and visual constraints. After the system enters the corridor,  $\eta_{\text{obs}}$  increases significantly, showing that geometric observability is substantially reduced. Meanwhile,  $\eta_{\text{NIS}}$  also rises, indicating that the innovation statistics deviate from the theoretical expectation. In contrast,  $\eta_{\text{qual}}$  changes only slightly, suggesting that the main degradation source in this scenario is reduced geometric observability rather than poor matching quality. Once the system returns to the outdoor environment, all three factors gradually decrease to stable levels, demonstrating that the proposed mechanism can automatically reduce the covariance when the environment becomes informative again.

Figure 6 shows the factor variations in the nighttime headlight scenario. Under normal low-light conditions, the three factors stay at relatively high levels because weak illumination limits image contrast and feature stability. When headlights illuminate the scene, local texture and structural details become clearer, leading to a noticeable decrease in  $\eta_{\text{qual}}$ , which indicates improved matching quality. At the same time,  $\eta_{\text{NIS}}$  decreases, showing better statistical consistency, and  $\eta_{\text{obs}}$  also drops slightly, indicating improved observability. After the illumination returns to the original nighttime condition, all factors rise again. These results demonstrate that the proposed adaptive uncertainty modeling mechanism can both suppress unreliable information under degradation and automatically increase observation confidence when sensing conditions improve.

#### 4.3.2. Drift Suppression by the Cross-Modal Consistency Factor

To verify the effectiveness of the proposed cross-modal consistency factor, comparative experiments are conducted to analyze drift separation between visual and LiDAR odometry under challenging conditions.

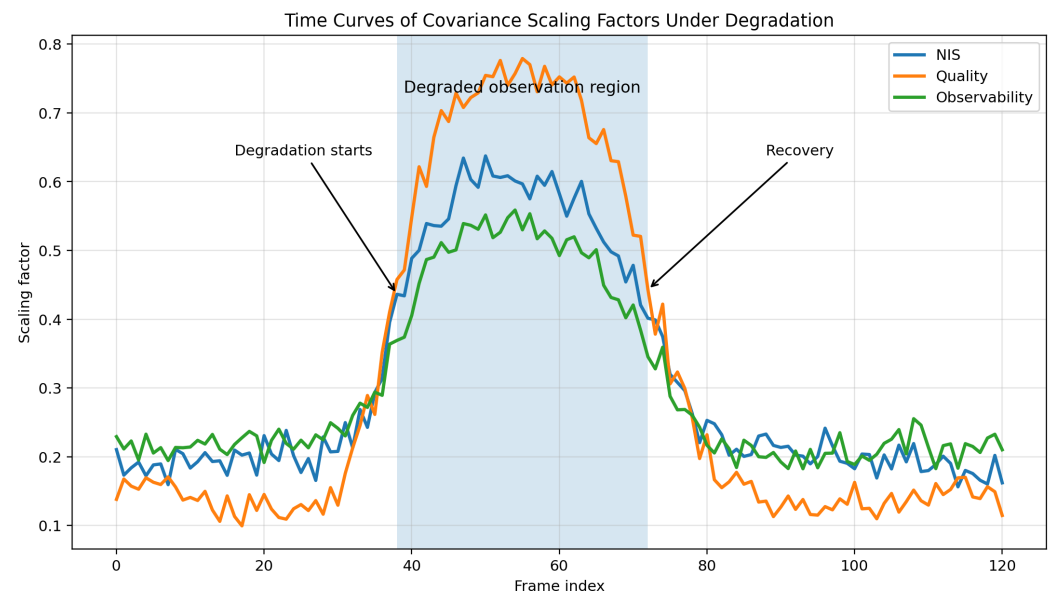


**Figure 5.** Time curves of covariance scaling factors under degraded observation conditions. The blue, orange, and green curves represent the NIS-based factor, quality-based factor, and observability-based factor, respectively. The light-blue shaded region indicates the degraded observation interval considered in this example.

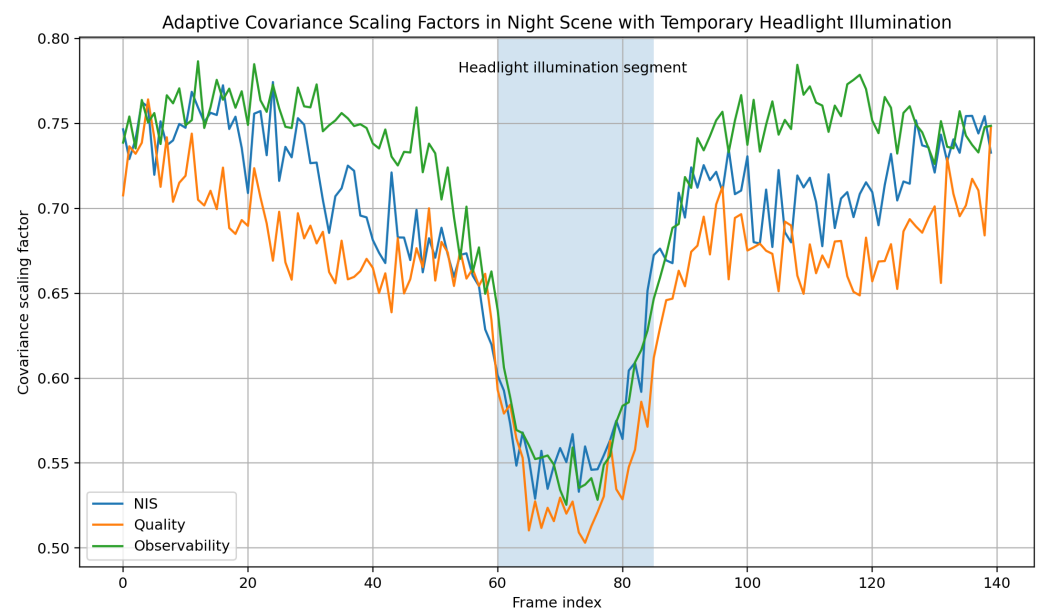


**Figure 6.** Time curves of covariance scaling factors in the nighttime scene with headlight interference. The blue, orange, and green curves represent the NIS-based factor, quality-based factor, and observability-based factor, respectively. The light-blue shaded region indicates the headlight illumination segment considered in this example.

In multi-modal SLAM, differences in sensing models and noise characteristics may gradually lead to inconsistent estimates between visual and LiDAR modalities, especially in dynamic scenes, weak-texture environments, and turning motions. To evaluate this effect, three configurations are compared:

- Baseline-0: LiDAR, visual, and IMU factors only;
- Baseline-2: Baseline-0 with loop closure constraints;
- Ours-Full: Baseline-2 further augmented with the proposed cross-modal consistency factor.

The cross-modal relative motion error is defined as

$$e_{cm}(k) = \left\| \log \left( (\mathbf{T}_k^L)^{-1} \mathbf{T}_k^V \right) \right\|, \quad (57)$$

where  $\mathbf{T}_k^L$  and  $\mathbf{T}_k^V$  denote the LiDAR and visual relative pose estimates at time  $k$ , respectively. A larger value of  $e_{cm}(k)$  indicates stronger cross-modal drift.

Figure 7 shows the temporal evolution of  $e_{cm}(k)$ , and Table 3 reports the corresponding statistics.

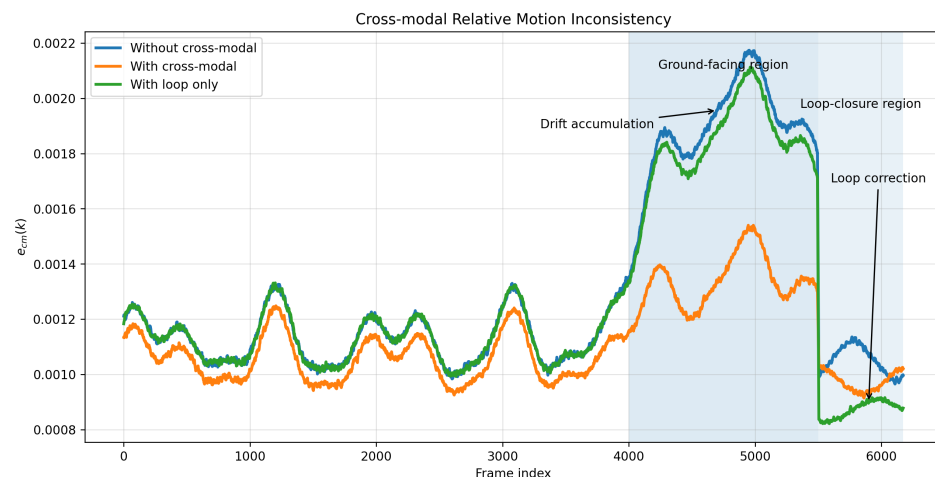


**Figure 7.** Time curve of cross-modal motion inconsistency  $e_{cm}(k)$ . The light-blue shaded region highlights the challenging interval used for detailed analysis, including the ground-facing and loop-closure regions.

**Table 3.** Cross-modal consistency error statistics.

| Method                  | Mean $e_{cm}$ | 95% $e_{cm}$ |
|-------------------------|---------------|--------------|
| Without cross-modal     | 0.001359      | 0.002082     |
| With loop only          | 0.001294      | 0.002028     |
| Ours (with cross-modal) | 0.001140      | 0.001432     |

As shown in Figure 7, the cross-modal error remains small in the early stage of the sequence, indicating good consistency between visual and LiDAR odometry. When the robot enters a visually degraded region, the visual estimate gradually accumulates drift, causing  $e_{cm}(k)$  to increase. Without the proposed cross-modal factor, this error remains high for a long period. In contrast, with the cross-modal consistency constraint, the visual state is continuously regularized by LiDAR observations, and the error increase is effectively suppressed. After loop closure and global optimization, the accumulated inconsistency is further reduced.

Table 3 shows that the proposed method achieves the lowest mean and 95% percentile errors, with reductions of approximately 16.1% and 31.2%, respectively. These results confirm that the proposed cross-modal consistency factor effectively improves the consistency and stability of multi-modal state estimation.

#### 4.4. Overall System Performance and Mapping Results

##### 4.4.1. Quantitative Evaluation on the M2DGR Dataset

Table 4 reports the quantitative localization results on the M2DGR dataset without loop closure. The evaluation metric is the Absolute Trajectory Error (ATE), measured in meters. The compared methods include VINS-Mono, A-LOAM, LIO-SAM, LVI-SAM, and the proposed method.

**Table 4.** Localization accuracy (ATE, meters) on the M2DGR dataset without loop closure.

| Method    | Walk_01     | Hall_02     | Door_02     | Room_02     | Gate_03     | Street_05   |
|-----------|-------------|-------------|-------------|-------------|-------------|-------------|
| VINS-Mono | 10.76       | fail        | 1.53        | 0.42        | 5.38        | 4.32        |
| A-LOAM    | 4.33        | 0.40        | 0.18        | 0.21        | 1.26        | 0.57        |
| LIO-SAM   | 1.81        | 0.29        | 0.15        | 0.18        | 0.91        | 0.43        |
| LVI-SAM   | 1.45        | 0.27        | 0.16        | 0.15        | 0.93        | 0.44        |
| Ours      | <b>1.02</b> | <b>0.21</b> | <b>0.10</b> | <b>0.10</b> | <b>0.34</b> | <b>0.37</b> |

Note: Bold values indicate the best results.

Without loop closure, all methods rely solely on front-end odometry and local optimization, making them more susceptible to drift accumulation. As shown in the upper part of Table 4, VINS-Mono fails on several sequences (e.g., Hall\_02 and Street\_05), indicating limited robustness in visually challenging or large-scale environments. This behavior is expected because VINS-Mono depends strongly on stable visual feature tracking and geometric initialization; when illumination changes, weak texture, motion blur, or long-term drift degrade the visual constraints, the estimator can no longer maintain reliable pose estimation.

LiDAR-based and LiDAR–visual methods exhibit improved stability, yet still suffer from noticeable drift in longer or more complex sequences such as Walk\_01 and Street\_05. In contrast, the proposed method consistently achieves the lowest ATE across most sequences, including Hall\_02, Room\_02, Door\_02, Gate\_03, Walk\_01, and Street\_05. These results demonstrate that the uncertainty-aware iESKF fusion effectively balances LiDAR and visual measurements, significantly reducing estimation error even in the absence of global loop constraints.

##### 4.4.2. Quantitative Evaluation on the M3DGR Dataset

Table 5 reports the quantitative results on the M3DGR dataset under diverse challenging scenarios, including visual degradation, LiDAR degeneration, and wheel slippage. In particular, the “Wheel Slip” column evaluates the robustness of different methods against wheel slippage on uneven terrain. Overall, the proposed method indicates robustness across all evaluated conditions and achieves the best or near-best performance in most scenarios.

**Table 5.** Quantitative results on the M3DGR dataset under six challenging scenarios. The columns correspond to Dynamic, Dark, Occlusion, Outdoor, LiDAR Degeneration, and Wheel Slip conditions, respectively. “Wheel Slip” refers to scenarios with significant wheel slippage on uneven or slippery terrain, which introduces additional motion disturbances commonly encountered by wheeled robots. Lower values indicate better performance.

| Method     | Visual Challenge |             |             | Standard    | LiDAR Deg.  | Wheel Slip.   |
|------------|------------------|-------------|-------------|-------------|-------------|---------------|
|            | Dynamic01        | Dark01      | Occlusion01 | Outdoor01   | Corridor01  | Wheel-Float01 |
| LVI-SAM    | 1.43             | 4.53        | 25.45       | 41.12       | 9.06        | 0.74          |
| R2LIVE     | 0.61             | 1.73        | 7.60        | 17.36       | 5.72        | 0.31          |
| R3LIVE     | 2.76             | 1.12        | 9.00        | 1.75        | 6.07        | 1.27          |
| Fast-LIVO  | 0.64             | 0.62        | 9.49        | 0.94        | 7.96        | 0.88          |
| SR-LIVO    | 1.23             | <b>0.10</b> | 1.71        | 2.45        | –           | 0.94          |
| Fast-LIVO2 | 0.51             | 0.18        | 0.29        | 0.87        | 2.15        | 0.61          |
| Ours       | <b>0.19</b>      | 0.11        | <b>0.15</b> | <b>0.76</b> | <b>1.07</b> | <b>0.10</b>   |

Note: Bold values indicate the best results. “–” indicates failure.

Under visual challenges, including Dynamic01, Dark01, and Occlusion01, our method consistently yields the lowest errors among all compared approaches. In particular, it achieves an error of 0.19 on Dynamic01 and 0.15 on Occlusion01, significantly outperforming existing LiDAR–visual–inertial systems. These results indicate that the proposed uncertainty-aware fusion and cross-modal consistency optimization effectively mitigate the impact of dynamic objects, illumination degradation, and partial occlusions. On the standard Outdoor01 sequence, our method attains the lowest error (0.76), slightly improving upon Fast-LIVO2 (0.87) and Fast-LIVO (0.94), demonstrating enhanced long-term global consistency. For the Corridor01 sequence, which suffers from severe LiDAR degeneration, our method achieves the best performance (1.07), while SR-LIVO fails to produce valid results. This highlights the advantage of integrating loop closure and cross-modal constraints in structurally repetitive environments. In the Wheel-float01 sequence, which simulates wheel slippage, our method significantly outperforms all baselines with an error of 0.10, showing superior robustness against motion model degradation. Overall, these results validate that the proposed system maintains accurate and stable localization across a wide range of adverse sensing conditions.

Figure 8 shows the trajectory comparison on the Outdoor01 sequence of the M3DGR dataset, with different regions highlighting distinct evaluation aspects. In region A, which corresponds to the endpoint of the trajectory, the comparison focuses on the final positional deviation among different methods. It can be observed that several baseline approaches exhibit noticeable endpoint drift relative to the ground truth, while the proposed method terminates closer to the true endpoint, indicating reduced long-term error accumulation. In regions B, C, and D, the evaluation emphasizes the spatial deviation between the estimated trajectories and the ground truth along different segments of the path. In these regions, the trajectory produced by the proposed method consistently stays closer to the ground truth compared to other approaches. Specifically, in the curved segment (region B), our method maintains a smaller deviation during rapid directional changes. In the long straight section (region C), accumulated drift is more pronounced for baseline methods, whereas the proposed approach exhibits reduced deviation over extended travel distances. Similarly, in region D, the proposed trajectory aligns more closely with the ground truth, reflecting improved consistency after global optimization. The regional comparisons in B, C, and D demonstrate that the proposed method achieves smaller trajectory-to-ground-truth distances across diverse motion patterns, resulting in more accurate and consistent localization performance throughout the sequence.

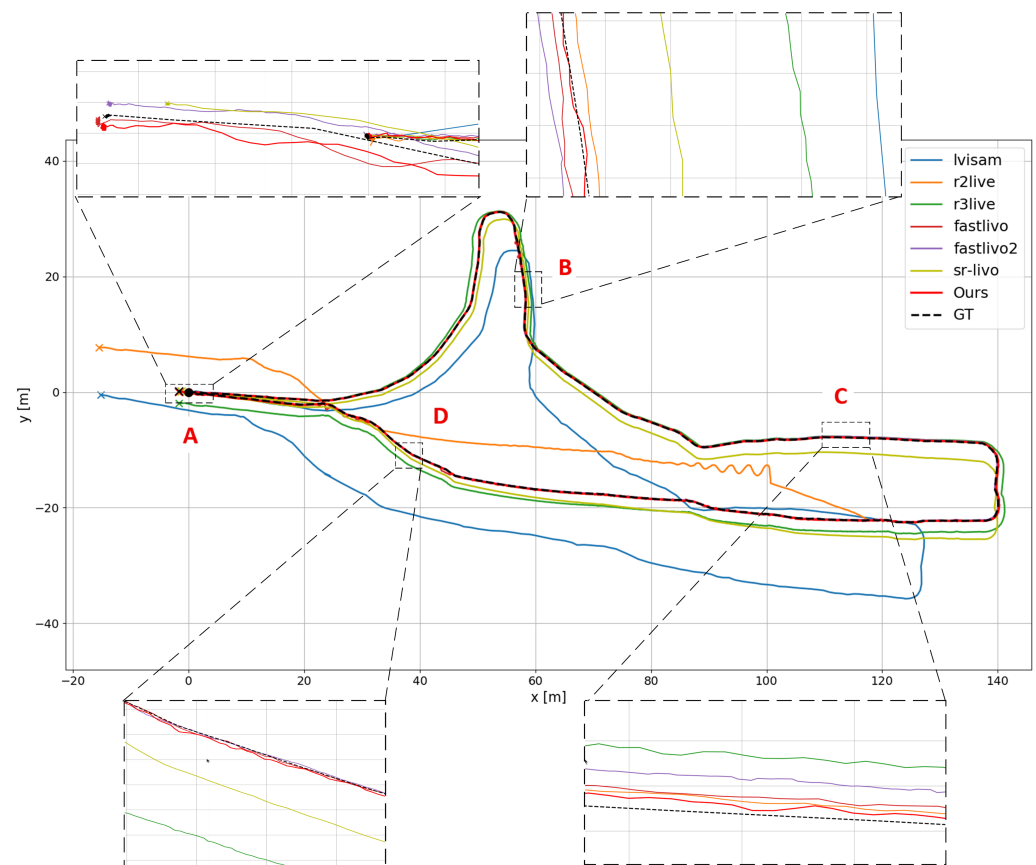
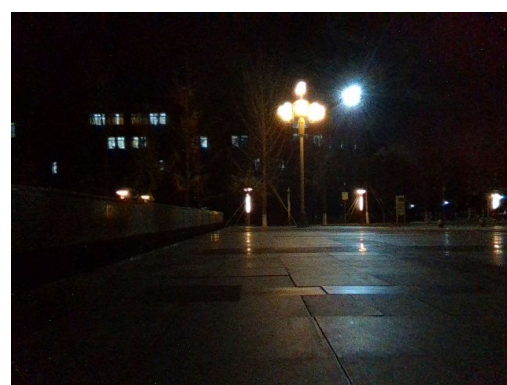
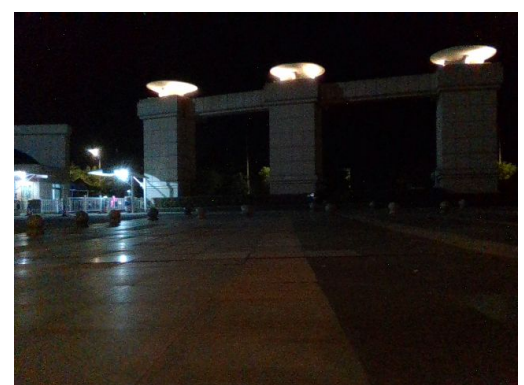


**Figure 8.** Qualitative trajectory comparison on the Outdoor1 sequence. The trajectories are projected onto the horizontal  $x$ - $y$  plane, with the axes measured in meters. The dashed rectangles show zoomed-in regions, and the dashed arrows indicate the corresponding locations of these enlarged views in the full trajectory. The letters A–D mark representative local regions used for visual comparison. GT denotes the ground-truth trajectory.

Figure 9 shows the low-light outdoor environment used for the real-world experiment. The corresponding mapping results of the proposed method and FAST-LIVO2 are compared in Figure 10. As highlighted in regions A and B, our approach reconstructs significantly denser and more complete point clouds with clearer structural boundaries and fewer missing regions. These qualitative improvements validate that the uncertainty-aware iESKF fusion and adaptive measurement weighting enable more reliable utilization of LiDAR measurements under severe visual degradation, leading to superior map completeness and structural fidelity.



(a) Real World on Dark01 (View 1)



(b) Real World on Dark01 (View 2)

**Figure 9.** Camera-view observations of the Dark01 sequence.

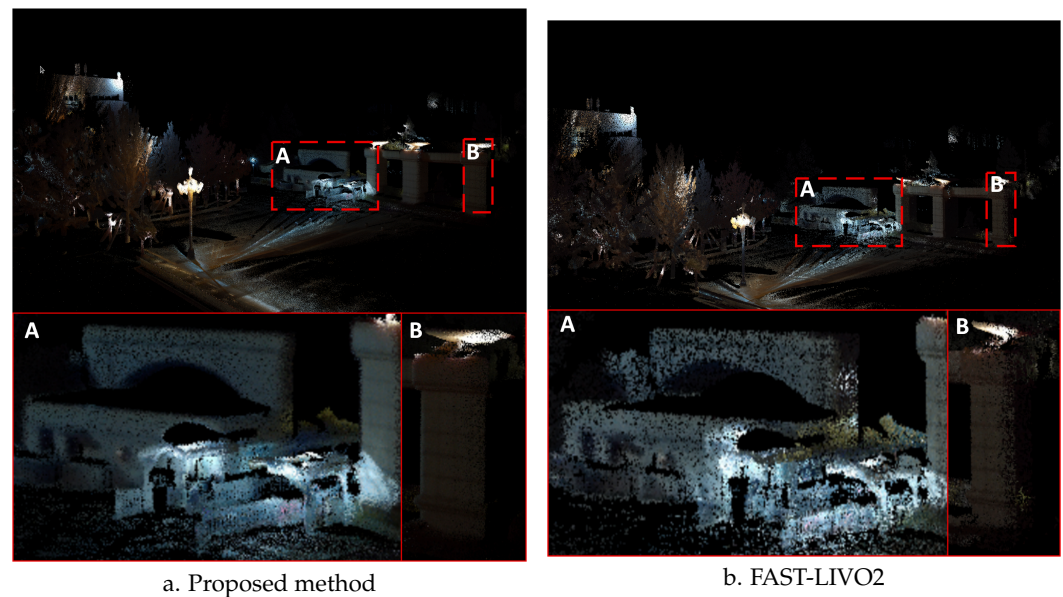


**Figure 10.** Qualitative comparison of point cloud mapping results in a challenging low-light outdoor environment; see Figure 9 for the corresponding scene. (a) The proposed method generates denser and more complete maps. (b) FAST-LIVO2 produces relatively sparse and incomplete mapping results. The differences are particularly visible in regions A, corresponding to the building facade, and B, corresponding to the vertical pillar. This comparison illustrates the effectiveness of uncertainty-aware fusion and adaptive covariance weighting under severe illumination degradation.

## 5. Conclusions

This paper presented a robust multi-sensor SLAM framework that tightly integrates LiDAR, visual, and inertial measurements to achieve accurate and consistent localization and mapping in complex and dynamic environments. By revisiting the limitations of existing LiDAR–Inertial–Visual Odometry (LIVO) systems, we identified three key challenges related to uncertainty modeling, loop-closure robustness, and cross-modal consistency. To address these issues, an uncertainty-aware iterated error-state Kalman filter (iESKF) was developed to dynamically balance heterogeneous measurements through adaptive covariance adjustment guided by innovation statistics, feature quality, and observability. In addition, a geometry-driven loop-closure module based on the proposed multi-channel Binary Triangle Constraint (mBTC) descriptor was introduced to improve loop detection robustness under viewpoint changes, illumination variation, and appearance degradation. Furthermore, a cross-modal consistency factor was incorporated into the global factor-graph optimization to explicitly constrain the relative motion agreement between visual and LiDAR odometries, thereby mitigating long-term drift and improving global consistency.

Extensive experiments on multiple public benchmarks demonstrated that the proposed system consistently outperforms state-of-the-art LIVO approaches in terms of localization accuracy, loop-closure reliability, and long-term consistency, particularly in dynamic and visually challenging scenarios. The results validate the effectiveness of combining uncertainty-aware front-end estimation with geometry-based loop detection and consistency-aware back-end optimization within a unified probabilistic framework.

Future work will focus on extending the proposed framework to large-scale long-term autonomy, including online adaptation to extreme sensor degradation, learning-assisted geometric descriptor enhancement, and tighter integration with semantic perception. Additionally, incorporating continuous-time modeling and exploring more expressive cross-modal constraints may further improve robustness in highly dynamic and complex environments.

**Author Contributions:** Conceptualization, Q.Z.; methodology, Q.Z., and J.L.; software, Q.Z.; validation, Q.Z.; formal analysis, Q.Z., J.X., and L.L.; investigation, Q.Z., and J.X.; resources, Q.Z.; data curation, Q.Z.; writing—original draft preparation, Q.Z.; writing—review and editing, J.X., J.L., and L.L.; visualization, Q.Z.; supervision, J.L.; project administration, J.X., and J.L.; funding acquisition, J.L., and L.L. All authors read the manuscript, contributed to the writing, and approved it. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Natural Science Foundation of China (NSFC), grant number 52278448.

**Data Availability Statement:** The public KITTI dataset analyzed during this study is available at <https://www.cvlibs.net/datasets/kitti/> (accessed on 26 April 2026). The public M2DGR dataset analyzed during this study is available at <https://github.com/SJTU-ViSYS/M2DGR> (accessed on 26 April 2026). The public M3DGR dataset analyzed during this study is available at <https://github.com/sjtuyinjie/M3DGR> (accessed on 26 April 2026). The public NCLT dataset analyzed during this study is available at <https://robots.engin.umich.edu/nclt/> (accessed on 26 April 2026). The custom datasets generated and analyzed during the real-world experiments are not publicly available due to privacy considerations, but are available from the corresponding author upon reasonable request.

**Conflicts of Interest:** The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

## Nomenclature

| Abbreviation/Symbol          | Description                                                   |
|------------------------------|---------------------------------------------------------------|
| LIVO                         | LiDAR–inertial–visual odometry                                |
| iESKF                        | Iterated error-state Kalman filter                            |
| mBTC                         | Multi-channel Binary Triangle Constraint                      |
| BTC                          | Binary Triangle Constraint                                    |
| NIS                          | Normalized innovation squared                                 |
| IMU                          | Inertial measurement unit                                     |
| LiDAR                        | Light detection and ranging                                   |
| VIO                          | Visual–inertial odometry                                      |
| LIO                          | LiDAR–inertial odometry                                       |
| ROS                          | Robot Operating System                                        |
| SE(3)                        | Special Euclidean group in three dimensions                   |
| SO(3)                        | Special orthogonal group in three dimensions                  |
| $\mathbf{R}_{wb}$            | Body-to-world rotation matrix ( $3 \times 3$ )                |
| $\mathbf{p}_w$               | Position of the body in the world frame ( $\mathbb{R}^3$ )    |
| $\mathbf{v}_w$               | Velocity of the body in the world frame ( $\mathbb{R}^3$ )    |
| $\mathbf{b}_g, \mathbf{b}_a$ | Gyroscope and accelerometer biases ( $\mathbb{R}^3$ )         |
| $\delta\theta$               | Rotation error vector, also referred to as the attitude error |
| $\eta_m^{\text{NIS}}$        | Statistical consistency scaling factor                        |
| $\eta_m^{\text{qual}}$       | Perception/matching quality scaling factor                    |
| $\eta_m^{\text{obs}}$        | Geometric observability scaling factor                        |
| $\mathbf{R}_m^*$             | Adaptive covariance matrix of modality $m$                    |

## Appendix A. Detailed Mathematical Derivations

This appendix provides the complete mathematical derivations that were summarized in Section 3.2 for brevity.

### Appendix A.1. IMU Propagation and Linearized Dynamics

Given the measured angular velocity  $\omega_m$  and acceleration  $\mathbf{a}_m$ , the continuous-time inertial propagation model is expressed as

$$\dot{\mathbf{R}}_{wb} = \mathbf{R}_{wb}[\boldsymbol{\omega}_m - \mathbf{b}_g - \mathbf{n}_g]_{\times}, \quad (\text{A1})$$

$$\dot{\mathbf{v}}_w = \mathbf{R}_{wb}(\mathbf{a}_m - \mathbf{b}_a - \mathbf{n}_a) + \mathbf{g}, \quad (\text{A2})$$

$$\dot{\mathbf{p}}_w = \mathbf{v}_w, \quad (\text{A3})$$

where  $[\cdot]_{\times}$  denotes the skew-symmetric operator and  $\mathbf{g}$  is the gravity vector. Linearizing around the nominal trajectory yields

$$\delta\dot{\mathbf{x}} = \mathbf{F}\delta\mathbf{x} + \mathbf{G}\mathbf{n}, \quad (\text{A4})$$

where  $\mathbf{F}$  is the state transition Jacobian and  $\mathbf{G}$  maps the process noise

$$\mathbf{n} = [\mathbf{n}_g^{\top}, \mathbf{n}_a^{\top}, \mathbf{n}_{wg}^{\top}, \mathbf{n}_{wa}^{\top}]^{\top}. \quad (\text{A5})$$

After discretization, the covariance propagation becomes

$$\mathbf{P}_{k|k-1} = \mathbf{F}_d \mathbf{P}_{k-1|k-1} \mathbf{F}_d^{\top} + \mathbf{Q}_d, \quad (\text{A6})$$

where  $\mathbf{Q}_d$  is obtained from the time-integrated process noise.

#### Appendix A.2. Measurement Models and Linearization

For the visual modality, a 3D landmark  $\mathbf{X}_i$  is projected onto the image plane as

$$\mathbf{u}_i = \pi(\mathbf{K}\mathbf{R}_{cw}\mathbf{X}_i + \mathbf{t}_{cw}), \quad (\text{A7})$$

and the photometric residual is

$$r_i^v = I(\mathbf{u}_i) - I_{\text{ref}}(\mathbf{u}_{i,\text{ref}}). \quad (\text{A8})$$

Linearization around the current estimate gives

$$r_i^v \approx \nabla I^{\top} \mathbf{J}_{\pi} \mathbf{J}_T \delta\boldsymbol{\xi} + n_i^v. \quad (\text{A9})$$

For the LiDAR modality, each point  $\mathbf{p}_{\ell}^j$  is constrained to a local plane  $(\mathbf{n}_j, d_j)$  through the point-to-plane residual

$$r_j^{\ell} = \mathbf{n}_j^{\top} (\mathbf{R}_{wb} \mathbf{p}_{\ell}^j + \mathbf{t}_{wb}) + d_j. \quad (\text{A10})$$

Its Jacobian is

$$\mathbf{H}_j^{\ell} = [-\mathbf{n}_j^{\top} \mathbf{R}_{wb} [\mathbf{p}_{\ell}^j]_{\times} \quad \mathbf{n}_j^{\top}]. \quad (\text{A11})$$

#### Appendix A.3. Iterated ESKF Update

The iESKF update follows the standard innovation form:

$$\mathbf{S} = \mathbf{H} \mathbf{P}_{k|k-1} \mathbf{H}^{\top} + \mathbf{R}, \quad (\text{A12})$$

$$\mathbf{K} = \mathbf{P}_{k|k-1} \mathbf{H}^{\top} \mathbf{S}^{-1}, \quad (\text{A13})$$

$$\delta\hat{\mathbf{x}} = \mathbf{K} \mathbf{r}, \quad (\text{A14})$$

$$\mathbf{P}_{k|k} = (\mathbf{I} - \mathbf{K} \mathbf{H}) \mathbf{P}_{k|k-1} (\mathbf{I} - \mathbf{K} \mathbf{H})^{\top} + \mathbf{K} \mathbf{R} \mathbf{K}^{\top}. \quad (\text{A15})$$

After each correction, the nominal state is updated on the manifold:

$$\mathbf{R}_{wb} \leftarrow \hat{\mathbf{R}}_{wb} \exp(\delta\theta), \quad \mathbf{p}_w \leftarrow \mathbf{p}_w + \delta\mathbf{p}, \quad \mathbf{v}_w \leftarrow \mathbf{v}_w + \delta\mathbf{v}. \quad (\text{A16})$$

The update is iterated until convergence.

#### Appendix A.4. Adaptive Uncertainty Modeling (Three-Factor Multiplicative Decomposition)

The adaptive covariance for each modality  $m \in \{v, \ell\}$  is

$$\mathbf{R}_m^* = \eta_m \mathbf{R}_m^0, \quad \eta_m = \eta_m^{\text{NIS}} \cdot \eta_m^{\text{qual}} \cdot \eta_m^{\text{obs}}. \quad (\text{A17})$$

The detailed definitions of  $\eta_m^{\text{NIS}}$ ,  $\eta_m^{\text{qual}}$ , and  $\eta_m^{\text{obs}}$  (including NIS, observability factor, and quality score) follow the expressions presented in Section 3.2.3..

## References

1. Graeter, J.; Wilczynski, A.; Lauer, M. LIMO: Lidar-monocular visual odometry. In *Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*; IEEE: New York, NY, USA, 2018; pp. 7872–7879.
2. Zhu, Y.; Zheng, C.; Yuan, C.; Huang, X.; Hong, Y. CamVox: A low-cost and accurate LiDAR-assisted visual SLAM system. In *Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA)*; IEEE: New York, NY, USA, 2021; pp. 5049–5055.
3. Shao, W.; Vijayarangan, S.; Li, C.; Kantor, G. Stereo visual inertial LiDAR simultaneous localization and mapping. In *Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*; IEEE: New York, NY, USA, 2019; pp. 370–377.
4. Shan, T.; Englot, B.; Ratti, C.; Rus, D. LVI-SAM: Tightly-coupled lidar-visual-inertial odometry via smoothing and mapping. In *Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA)*; IEEE: New York, NY, USA, 2021; pp. 5692–5698.
5. Lin, J.; Zheng, C.; Xu, W.; Zhang, F. R2LIVE: A robust, real-time, LiDAR-inertial-visual tightly-coupled state estimator and mapping. *IEEE Robot. Autom. Lett.* **2021**, *6*, 7469–7476. [\[CrossRef\]](#)
6. Lin, J.; Zhang, F. R3LIVE: A robust, real-time, RGB-colored, LiDAR-inertial-visual tightly-coupled state estimation and mapping package. In *Proceedings of the 2022 IEEE International Conference on Robotics and Automation (ICRA)*; IEEE: New York, NY, USA, 2022; pp. 10672–10678.
7. Zheng, C.; Zhu, Q.; Xu, W.; Liu, X.; Guo, Q. FAST-LIVO: Fast and tightly-coupled sparse-direct LiDAR-inertial-visual odometry. In *Proceedings of the 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*; IEEE: New York, NY, USA, 2022; pp. 4003–4009.
8. Zheng, C.; Xu, W.; Zou, Z.; Hua, T.; Yuan, C.; He, D.; Zhou, B.; Liu, Z.; Lin, J.; Zhu, F.; et al. FAST-LIVO2: Fast, direct LiDAR-inertial-visual odometry. *arXiv* **2024**, arXiv:2408.14035. [\[CrossRef\]](#)
9. Engel, J.; Koltun, V.C.D. Direct sparse odometry. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 611–625. [\[CrossRef\]](#)
10. Engel, J.; Schöps, T.; Cremers, D. LSD-SLAM: Large-scale direct monocular SLAM. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 834–849.
11. Forster, C.; Pizzoli, M.; Scaramuzza, D. SVO: Fast semi-direct monocular visual odometry. In *Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA)*; IEEE: New York, NY, USA, 2014; pp. 15–22.
12. Mur-Artal, R.; Tardós, J.D. ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras. *IEEE Trans. Robot.* **2017**, *33*, 1255–1262. [\[CrossRef\]](#)
13. Qin, T.; Li, P.S.S. VINS-Mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Trans. Robot.* **2018**, *34*, 1004–1020. [\[CrossRef\]](#)
14. Qin, T.; Pan, J.; Cao, S.; Shen, S. A general optimization-based framework for local odometry estimation with multiple sensors. *arXiv* **2019**, arXiv:1901.03638. [\[CrossRef\]](#)
15. Cao, S.; Lu, X.; Shen, S. GVINS: Tightly coupled GNSS–visual–inertial fusion for smooth and consistent state estimation. *IEEE Trans. Robot.* **2022**, *38*, 2004–2021. [\[CrossRef\]](#)
16. Zhang, J.; Singh, S. LOAM: Lidar odometry and mapping in real-time. In *Proceedings of the Robotics: Science and Systems*, Berkeley, CA, USA, 12–16 July 2014; Volume 2, pp. 1–9.

17. Shan, T.; Englot, B. LEGO-LOAM: Lightweight and ground-optimized lidar odometry and mapping on variable terrain. In *Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*; IEEE: New York, NY, USA, 2018; pp. 4758–4765.
18. Lin, J.; Zhang, F. LOAM LIVOX: A fast, robust, high-precision LiDAR odometry and mapping package for LiDARs of small FoV. In *Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA)*; IEEE: New York, NY, USA, 2020; pp. 3126–3131.
19. Yuan, C.; Xu, W.; Liu, X.; Hong, X.; Zhang, F. Efficient and probabilistic adaptive voxel mapping for accurate online lidar odometry. *IEEE Robot. Autom. Lett.* **2022**, *7*, 8518–8525. [[CrossRef](#)]
20. Liu, Z.; Li, H.; Yuan, C.; Liu, X.; Lin, J.; Li, R.; Zheng, C.; Zhou, B.; Liu, W.; Zhang, F. Voxel-SLAM: A complete, accurate, and versatile lidar-inertial SLAM system. *arXiv* **2024**, arXiv:2410.08935.
21. Xu, W.; Zhang, F. FAST-LIO: A fast, robust LiDAR-inertial odometry package by tightly-coupled iterated Kalman filter. *IEEE Robot. Autom. Lett.* **2021**, *6*, 3317–3324. [[CrossRef](#)]
22. Xu, W.; Cai, Y.; He, D.; Lin, J.; Zhang, F. FAST-LIO2: Fast direct LiDAR-inertial odometry. *IEEE Trans. Robot.* **2022**, *38*, 2053–2073. [[CrossRef](#)]
23. He, D.; Xu, W.; Chen, N.; Kong, F.; Yuan, C.; Zhang, F. Point-LIO: Robust high-bandwidth light detection and ranging inertial odometry. *Adv. Intell. Syst.* **2023**, *5*, 2200459. [[CrossRef](#)]
24. Dellenbach, P.; Deschaud, J.E.; Jacquet, B.; Goulette, F. CT-ICP: Real-time elastic LiDAR odometry with loop closure. In *Proceedings of the 2022 IEEE International Conference on Robotics and Automation (ICRA)*; IEEE: New York, NY, USA, 2022; pp. 5580–5586.
25. Devarajan, H.; Zheng, H.; Kougkas, A.; Sun, X.H.; Vishwanath, V. DLIO: A data-centric benchmark for scientific deep learning applications. In *Proceedings of the 2021 IEEE/ACM 21st International Symposium on Cluster, Cloud and Internet Computing (CCGrid)*; IEEE: New York, NY, USA, 2021; pp. 81–91.
26. Shan, T.; Englot, B.; Meyers, D.; Wang, W.; Ratti, C.; Rus, D. LIO-SAM: Tightly-coupled LiDAR inertial odometry via smoothing and mapping. In *Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*; IEEE: New York, NY, USA, 2020; pp. 5135–5142.
27. Kim, G.; Kim, A. Scan context: Egocentric spatial descriptor for place recognition within 3D point cloud map. In *Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*; IEEE: New York, NY, USA, 2018; pp. 4802–4809.
28. He, L.; Wang, X.; Zhang, H. M2DP: A novel 3D point cloud descriptor and its application in loop closure detection. In *Proceedings of the 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*; IEEE: New York, NY, USA, 2016; pp. 231–237.
29. Kim, G.; Choi, S.; Kim, A. Scan context++: Structural place recognition robust to rotation and lateral variations in urban environments. *IEEE Trans. Robot.* **2021**, *38*, 1856–1874. [[CrossRef](#)]
30. Wang, H.; Wang, C.; Xie, L. Intensity scan context: Coding intensity and geometry relations for loop closure detection. In *Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA)*; IEEE: New York, NY, USA, 2020; pp. 2095–2101.
31. Wang, Y.; Sun, Z.; Xu, C.; Sarma, S.; Yang, J.; Kong, H. LiDAR iris for loop-closure detection. In *Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*; IEEE: New York, NY, USA, 2020; pp. 5769–5775.
32. Yuan, C.; Lin, J.; Zou, Z.; Hong, X.; Zhang, F. STD: Stable triangle descriptor for 3D place recognition. *arXiv* **2022**, arXiv:2209.12435.
33. Yuan, C.; Lin, J.; Liu, Z.; Wei, H.; Hong, X.; Zhang, F. BTC: A binary and triangle combined descriptor for 3-D place recognition. *IEEE Trans. Robot.* **2024**, *40*, 1580–1599. [[CrossRef](#)]
34. Bar-Shalom, Y.; Li, X.R.K.T. *Estimation with Applications to Tracking and Navigation: Theory, Algorithms and Software*; John Wiley & Sons: Hoboken, NJ, USA, 2001.
35. Maybeck, P.S. *Stochastic Models, Estimation, and Control*; Academic Press: Cambridge, MA, USA, 1982; Volume 3.
36. Goodman, J.; Hilal, W.; Gadsden, S.; Eggleton, C.D. Adaptive SVSF-KF estimation strategies based on the normalized innovation square metric and IMM strategy. *Results Eng.* **2022**, *16*, 100785. [[CrossRef](#)]
37. Groves, P.D. *Principles of GNSS, Inertial, and Multisensor Integrated Navigation Systems*, 2nd ed.; Artech House: Boston, MA, USA, 2013.
38. Lee, E. M.; Marsim, K.; Myung, H. LODESTAR: Degeneracy-aware LiDAR-inertial odometry with adaptive Schmidt-Kalman filter and data exploitation. *IEEE Robot. Autom. Lett.* **2025**, *11*, 922–929. [[CrossRef](#)]
39. Huber, P.J. Robust estimation of a location parameter. *Ann. Math. Stat.* **1964**, *35*, 73–101. [[CrossRef](#)]
40. Holland, P.W.; Welsch, R.E. Robust regression using iteratively reweighted least-squares. *Commun. Stat. Theory Methods* **1977**, *6*, 813–827. [[CrossRef](#)]
41. Salton, G.; Buckley, C. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag.* **1988**, *24*, 513–523. [[CrossRef](#)]
42. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets robotics: The KITTI dataset. *Int. J. Robot. Res.* **2013**, *32*, 1231–1237. [[CrossRef](#)]
43. Yin, J.; Li, A.; Li, T.; Yu, W.; Zou, D. M2DGR: A multi-sensor and multi-scenario SLAM dataset for ground robots. *IEEE Robot. Autom. Lett.* **2021**, *7*, 2266–2273. [[CrossRef](#)]

44. Zhang, D.; Zhang, J.; Sun, Y.; Li, T.; Yin, H.; Xie, H.; Yin, J. Towards robust sensor-fusion ground SLAM: A comprehensive benchmark and a resilient framework. *arXiv* **2025**, arXiv:2507.08364. [[CrossRef](#)]
45. Carlevaris-Bianco, N.; Ushani, A.K.E.R.M. University of Michigan North Campus long-term vision and LiDAR dataset. *Int. J. Robot. Res.* **2016**, *35*, 1023–1035. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.