International Journal of Applied Science; Vol. 8, No. 4; 2025
ISSN 2576-7240 E-ISSN 2576-7259
https://doi.org/10.30560/ijas.v8n4p116

A Robust SLAM System Enhanced for Degenerate Motion Scenarios via Semantic Filtering and Bayesian Motion Consistency

Jianlong Deng[1], Jing Chen[1], Sheng Xu[1] & Rongli He[1]

[1] School of Physics and Optoelectronic Engineering, Guangdong University of Technology, China

Correspondence: Rongli He, School of Physics and Optoelectronic Engineering, Guangdong University of Technology, Guangzhou, GuangDong, China.

Received: November 17, 2025; Accepted: December 11, 2025; Published: December 16, 2025

Abstract

Dynamic SLAM methods based on epipolar constraints provide a simple and efficient solution for distinguishing dynamic features. However, such constraints tend to fail under geometrically degenerate motion sce-narios, such as pure rotation, long-range linear motion, or coplanar scene structures, leading to misclassification of dynamic features, trajectory drift, and inaccurate map reconstruction. To address these challenges, a parallel SLAM framework is proposed, integrating semantic guidance, degeneracy-aware dynamic feature discrimination, and robust keyframe selection to handle degenerate motions. Without requiring inertial measurements, the proposed method relies solely on visual information to enable dynamic fea-ture removal, avoiding IMU drift. Specifically, a lightweight object detection module is introduced using the fast and compact YOLO-FASTEST model, enabling efficient semantic perception to provide prior information for dynamic point removal. Furthermore, a multi- frame Bayesian motion consistency criterion is proposed that jointly considers camera motion priors and observation residuals of feature points to enable dynamic feature discrimination in degenerate scenarios. In addition, an adaptive multi-metric keyframe insertion strategy is designed, jointly considering pose change magnitude, image entropy variation, and the ratio of constrained pixels, to enhance keyframe selection under motion-degenerate scenes. Experimental results demonstrate that the proposed method achieves superior trajectory accuracy and map completeness under various dynamic interference conditions, while maintaining real-time performance.

Keywords: Dynamic SLAM, Degenerate motion, Keyframe selection, Semantic-guided, Dynamic feature filtering

1. Introduction

ROBUST and accurate visual SLAM (Simultaneous Lo- calization and Mapping) in dynamic and structurally complex environments remains a significant challenge in the field of robotic perception. In particular, in real-world sce- narios, cameras often undergo degenerate motion (such as pure rotation, long-distance straight-line motion, or coplanar structures within the field of view) [1]. These motion patterns result in insufficient inter-frame parallax, severely weakening the geometric constraints that SLAM relies on, leading to feature matching ambiguities, amplified triangulation errors, and even causing front-end tracking failures [2]. Therefore, feature-reliability modeling and robust front-end perception under degenerate motion are essential for improving the stabil- ity of SLAM systems in high-dynamic, low-texture, or view- angle-limited environments [3-5].

To mitigate the performance degradation caused by de- generate motion in SLAM systems, various feature point elimination and enhancement strategies have been proposed. These methods can be classified into three categories: ge-ometric consistency methods, semantic perception methods, and motion residual modeling methods. Geometric consistency methods typically use epipolar constraints or disparity analysis to eliminate dynamic points. For instance, the sliding-window RANSAC method performs well in general motion scenarios, but tends to fail under extreme viewpoints (such as pure rota- tion) . Semantic perception methods introduce object detection or segmentation models, leveraging prior class information to identify potential dynamic regions. For example, YOLOv5 de- tects pedestrians and vehicles to preemptively remove dynamic features [6], which is well-suited for deployment but still prone to misclassification when the model's generalization ability is insufficient or detection boundaries are coarse. Motion residual modeling methods rely on prediction-observation error or optical flow consistency to model dynamics [7], [8]. These methods weaken reliance on geometric priors and adapt well to unknown dynamic patterns. However, under degenerate motion scenarios such

as low parallax or pure rotation, the reliabil- ity of residual estimation degrades significantly, resulting in potential misclassification of dynamic features.

To enhance the performance of the system in degenerate motion scenarios, an robust lightweight SLAM approach is proposed. The method has three key improvements: First, to address the high computational cost and limited deployability of semantic detection on embedded platforms, we introduce a lightweight semantic-guided dynamic feature elimination method. YOLO-FASTEST [9] is employed as the semantic perception module and executed with the NCNN inference framework in parallel to the main SLAM process, providing stable and efficient semantic priors for feature removal. Sec- ond, to tackle the dynamic perception problem under degener- ate motion conditions, we propose a prediction-observation modeling approach leveraging camera motion priors, and design a multi-frame Bayesian inference mechanism based on motion residuals. This enables posterior estimation of the dynamic/static state of each feature point, ensuring robust identification and suppression of potential dynamic points. Finally, an adaptive multi-metric keyframe insertion strategy is proposed to overcome the poor adaptability and redundancy issues of traditional disparity- or pose-based methods in de- generate or repetitive scenes. Two novel indicators—image entropy and constrained pixel ratio—are integrated to jointly capture semantic distinctiveness and geometric reliability, en- suring informative keyframe selection and robust trajectory tracking.

The contributions of this paper are as follows:

1) A lightweight robust visual SLAM system is enhanced for degenerate motion scenarios, enabling stable tracking and accurate mapping in challenging dynamic environ- ments.

2) A dynamic perception module leveraging predic- tion-observation residuals and multi-frame Bayesian in- ference to identify dynamic features under motion de- generacy, operating solely on visual information without inertial measurements.

3) An improved keyframe insertion strategy is proposed, leveraging multi-metric evaluation to select keyframes with adaptive thresholding, nhancing map quality under degenerate motion scenarios.

4) Extensive experiments validate the proposed method's superiority in map completeness, trajectory accuracy, and real-time performance, demonstrating its effective- ness in dynamic SLAM scenarios.

The structure of this paper is as follows: Section 2 reviews related work, Section 3 describes the proposed method, Section 4 presents experiments and result discussions, and Section 5 concludes the paper with future work prospects.

2. Related Works

2.1 Semantic Perception of Dynamic Objects in SLAM

In dynamic SLAM systems, semantic perception is widely used to identify and eliminate dynamic targets, thereby en- hancing the system's robustness and accuracy in highly dy- namic environments. Existing methods can be broadly catego- rized into three types: semantic segmentation-based methods, object detection-based methods, and hybrid discrimination- based methods.

Semantic segmentation-based methods represented by works such as DS-SLAM [10], which uses SegNet for pixel- level semantic segmentation to identify dynamic class re- gions such as "person" and filters dynamic feature points by combining optical flow detection. This method achieves high segmentation accuracy and is suitable for removing dynami- cally irregular-shaped regions, but its model is computation- ally complex and suffers from limited real-time performance. Object detection-based methods, such as Detect-SLAM [11] and YOLO-SLAM [12], which use SSD and YOLOX-S ob- ject detectors, respectively, to identify dynamic objects (e.g., people, cars) in images and eliminate ORB feature points within their bounding boxes. These methods offer high speed and flexibility for deployment, making them suitable for embedded systems, but the coarse bounding boxes can lead to misclassification and removal of static background features. Hybrid discrimination-based methods, such as DynaSLAM [13], which combines Mask R-CNN for instance segmentation with geometric motion detection to remove genuinely dynamic targets. Another example is RS-SLAM [14], which further incorporates Bayesian optimization to improve the segmen- tation accuracy of dynamic regions. These methods integrate both semantic and motion information, enabling more precise dynamic discrimination, but they typically come with a heavier computational burden.

Although existing semantic perception methods for dynamic targets are effective, each has its limitations. Semantic seg- mentation methods like DS-SLAM achieve high segmentation accuracy but suffer from model complexity and poor real- time performance, making them difficult to run on embed- ded platforms. Object

detection methods like Detect-SLAM and YOLO-SLAM offer faster inference speeds but rely on relatively heavy detectors, and their rectangular bounding box elimination strategy is often too coarse, leading to the misclassification of background features. Hybrid methods like DynaSLAM improve discrimination accuracy by incorporating instance segmentation and motion detection, but they come with significant computational overhead, making them difficult to deploy. To address these issues, this paper adopts YOLO- FASTEST as the semantic perception module with very simple structure, small parameter size, and fast inference speed . Compared to existing methods, YOLO-FASTEST balances detection accuracy and operational efficiency, making it par- ticularly suitable for real-time dynamic SLAM systems [15].

2.2 Dynamic Feature Removal in SLAM

In visual SLAM systems, dynamic feature points violate the assumption of a static world, leading to data association errors and pose estimation drift. To enhance system robustness and accuracy in dynamic environments, researchers have pro- posed various dynamic feature point elimination mechanisms. These methods can be broadly classified into four categories: geometry-based methods, semantic perception-based methods, motion consistency-based methods, and multimodal fusion methods [16].

Geometry-based methods, which rely on the assumption that features in a static scene should adhere to a unified geometric model. These methods eliminate dynamic points by detecting features that violate geometric consistency. For example, DS-SLAM uses RANSAC to estimate the funda- mental matrix and applies epipolar distance to detect dynamic

points; DynaSLAM performs geometric consistency validation on dynamic regions based on depth maps; ClusterSLAM uses motion consistency-based clustering to identify and eliminate motion outliers [17]. Semantic perception-based methods, which leverage deep learning models for object detection or semantic segmentation to identify potential dynamic objects such as people or vehicles, and eliminate the feature points within them. For example, DynaSLAM integrates Mask R- CNN instance segmentation to label dynamic objects and combines motion detection for elimination ; YOLO-SLAM employs YOLOX to detect pedestrians and vehicles, filter- ing dynamic features by incorporating depth information . Motion consistency-based methods, which determine dynamic points by analyzing frame-to-frame optical flow, reprojection residuals, or motion trajectory consistency. These methods are suitable for unknown dynamic interference objects that do not rely on class labels. For example, StaticFusion uses dense optical flow residuals to determine whether independent motion exists in a local region [18]; DynaVINS combines IMU predictions with visual reprojection errors to eliminate dynamic observations [19]. Multimodal fusion methods, which combine RGB images, depth information, IMU, or LiDAR data to enhance dynamic recognition. For example, DOR- LINS uses LiDAR point cloud motion detection to filter out dynamic objects [20]; while VINS-Fusion enhances estimation consistency by fusing inertial measurements, which provide complementary motion constraints that help mitigate the im- pact of dynamic outliers [21].

These methods have their limitations: geometry-based meth- ods rely on significant disparity and tend to fail in degenerate scenarios such as pure rotation or coplanar motion [22]; semantic perception methods, although capable of detecting dynamic targets in advance, suffer from model complexity and slow inference speeds, making them difficult to deploy on embedded platforms, and they also exhibit poor generalization to unknown objects ; motion consistency-based methods are less sensitive to slow motion and can be easily affected by jitter and weak textures [23]; while multimodal fusion methods offer strong robustness, they come with high system costs and complex integration [24]. To address these challenges, this paper proposes a dynamic point elimination strategy with degenerate consistency modeling, aiming to improve the robustness of dynamic discrimination under degenerate motion conditions.

2.3 Keyframe Strategies in SLAM

In visual SLAM systems, the rational selection of keyframes plays a critical role in constructing high-quality maps, main- taining robust front-end tracking, and controlling the com- putational burden of back-end optimization. In recent years, researchers have proposed various improvements to keyframe insertion strategies, which can be broadly categorized into the following types: threshold-driven methods, feature track- ing methods, information gain methods, semantic perception methods, and multimodal fusion methods.

Threshold-driven methods, which determine keyframe in- sertion based on predefined geometric change thresholds (e.g., pose changes or inter-frame disparity) [25]. Feature tracking methods, which rely on the number of trackable feature points in the local map as a reference. A keyframe is inserted when the number of trackable features in the current frame drops below a certain threshold [26]. Information gain methods, which introduce information-theoretic metrics to evaluate the contribution of new frames. For example, a

keyframe selection strategy based on pose covariance entropy inserts keyframes only when the new frame significantly reduces system uncer- tainty [27]. Other methods evaluate the density of new dis- parity information in the image and select the frame with the highest information gain as a keyframe . Semantic perception methods, which incorporate semantic segmentation or object detection to assist in keyframe selection. For example, Sem- SLAM analyzes changes in static regions in keyframes based on a semantic map, ensuring that the keyframe represents incremental changes in the background view [28]. Multimodal fusion methods, which combine multiple observation metrics to enhance adaptability. Some works fuse various observation metrics for keyframe selection. For instance, A-VINS jointly models feature count, pose change, and IMU velocity to improve system stability in high-speed motion or occlusion scenarios [29].

The aforementioned methods still exhibit significant limita- tions under degenerate motion conditions. Specifically, during pure rotation or long-distance straight-line motion, inter-frame disparity becomes minimal, and the viewpoint changes slowly. Traditional geometric threshold methods often fail to insert keyframes due to insufficient discriminative criteria, leading to incomplete trajectory recording and sparse maps [30]. Feature tracking methods are prone to frequent triggering under weak textures or dynamic occlusions, resulting in overly dense keyframe insertion [31]. Although information gain and semantic strategies are effective, they suffer from high computational complexity or are heavily reliant on the model's generalization ability, limiting their practicality in real-world deployments [32]. To address these issues, this paper proposes a multi-metric-driven keyframe insertion strategy that consid- ers translation magnitude, rotation angle, image entropy ratio, and effective constraint pixel ratio to enable precise keyframe selection under degenerate or dynamic conditions.

3. System Description

Current visual SLAM systems face two fundamental chal- lenges in high-dynamic or geometrically degenerate environ- ments: One is the ineffective removal of dynamic feature in- terference, and the other is the failure of geometric constraints under degenerate motion conditions. Both of these issues can directly lead to front-end matching errors, back-end estimation drift, and anomalies in map construction [33].

Dynamic feature removal based on geometric consistency under degenerate motion conditions, such as pure rotation, forward-backward translation (resulting in vanishing dispar- ity), or large areas of coplanar structures (leading to degraded depth estimation), causes the rank of the fundamental matrix in epipolar geometry to decrease, preventing effective triangu- lation and causing motion consistency checks to fail [34]. To

address this, we do not rely on sufficient disparity information or triangulation-based structure. Instead, a statistical inference method is adopted based on prior motion information and observation residuals, relying solely on visual data without the need for inertial sensors . This ensures reliable motion dis- crimination even under geometrically degenerate conditions. First, we introduce the lightweight object detection network YOLO-FASTEST, which provides real-time semantic priors for dynamic target regions at low computational cost, offering class-level constraints for subsequent feature removal. Then, residuals are aggregated across consecutive frames using a sliding window, and a multi-frame Bayesian motion consis- tency criterion is applied. This approach enhances motion classification reliability by smoothing noise and mitigating the impact of transient occlusions or motion blur, resulting in more accurate feature filtering.

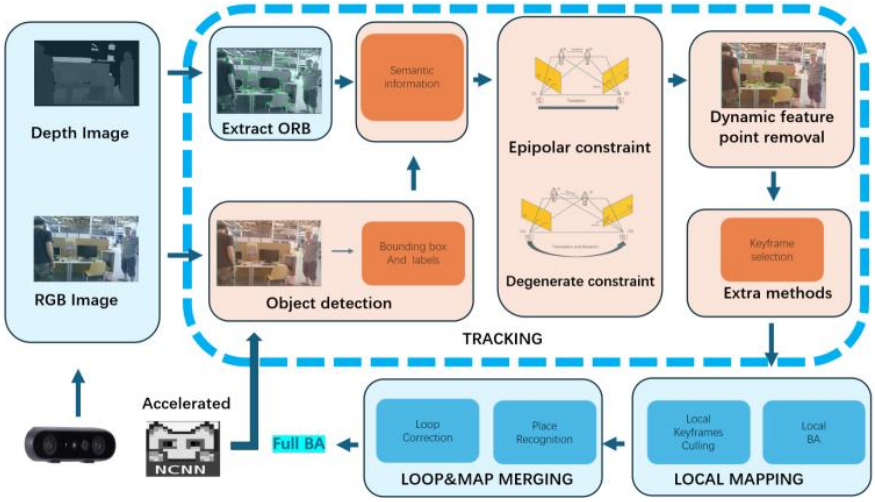

Figure 1. This is the overall framework of our SLAM system: orange blocks indicate our newly added modules, while blue blocks represent the original ORB-SLAM2 pipeline

The keyframe selection mechanism in traditional SLAM is often based on disparity thresholds or the number of features, which often fails in dynamic interference or low- texture environments due to insufficient evaluation of image quality and observability. This paper proposes a multi-metric driven keyframe insertion strategy. In addition to traditional translation and rotation thresholds, the strategy incorporates two additional observation metrics: image entropy variation and the proportion of effective constraint pixels . Image entropy variation offers non-geometric cues to capture novel visual content under geometric degradation, while the pro- portion of constrained static pixels ensures well-distributed static features, preventing planar degeneration. Furthermore, an adaptive thresholding mechanism is introduced to dynam- ically adjust keyframe insertion based on scene conditions, addressing the issue of overly sparse or dense keyframe insertion in challenging scenarios.

The system consists of three main functional modules: a lightweight semantic perception module, integrating YOLO- FASTEST, a degenerate consistency modeling module, and a multi-metric keyframe insertion module. The overall system workflow is shown in Fig 1. First, image frames are captured from the depth camera and processed by the SLAM main thread. Simultaneously, YOLO-FASTEST detects dynamic objects and provides masks to filter out dynamic feature points. Next, dynamic feature removal is performed using prediction-observation residuals and multi-frame Bayesian in- ference. Keyframes are then determined based on significant motion, image entropy increase, or static constraint features and passed to the back-end for optimization and map updating. Finally, back-end optimization and loop closure detection are performed.

3.1 Semantic Understanding of Dynamic Targets

To enhance the feature point filtering capability of the system in dynamic environments, this paper introduces the lightweight object detection network YOLO-FASTEST and constructs an independent semantic perception thread using the NCNN inference framework. This module provides efficient, real-time category priors for front-end feature removal, thereby boosting the system's ability to perceive dynamic interference regions. Compared to traditional semantic segmentation meth- ods (e.g., SegNet, Mask R-CNN) or larger model detectors (e.g., YOLOv5-L, SSD), YOLO-FASTEST offers the follow- ing advantages: it is extremely fast, with a very small param- eter size, allowing for high-frame-rate detection on embedded platforms without the need for GPU support; it maintains good recognition accuracy for common dynamic objects while preserving speed; and it can operate as an independent module running in parallel, decoupling the detection thread from the SLAM main process, thus offering strong scalability. Unlike semantic masking methods, which only provide static-only labeling of potentially dynamic regions in the image domain, the detector provides explicit category-level object bounding box information, making it easier to precisely mask feature points in potential dynamic regions and reduce mismatches caused by dynamic objects.

YOLO-FASTEST is a single-stage object detector, and its core consists of a lightweight convolutional backbone network, a feature fusion module, and a fully connected detection head. After the input image passes through the backbone network to extract multi-scale features, the detection head directly predicts the object category probabilities and bounding box parameters. The output format is:

$$\text{Output} = \left\{(x_i, y_i, w_i, h_i, c_i)\right\}_{i=1}^{N} \quad (1)$$

Here, (x_i , y_i) denotes the center coordinates of the target, w_i , $h_i$represents the width and height of the bounding box, and c_i indicates the confidence score for the target class.

To enable efficient deployment of the object detection module on GPU-less embedded platforms, YOLO-FASTEST is first converted into the ONNX format and then deployed using the NCNN (Tencent Neural Network Computing Li- brary) inference engine. This setup enables fast dynamic object detection and provides mask inputs with millisecond-level latency, ensuring stable and real-time SLAM performance on embedded or mobile platforms.

3.2 Degraded Motion Perception and Dynamic Point Removal

In degraded motion perception scenarios, geometric con- straints become unreliable, and residuals may reflect both true motion and degeneracy, often causing feature misclassification and state drift. Fig. 2 highlights the degenerate case where coplanar motion yields residuals smaller than τ, causing dynamic points to be misclassified as static. To address these challenges, we explicitly model the static/dynamic state of each feature point as a hidden variable and infer it through Bayesian filtering on residuals. This formulation enables us to fuse the prior motion information of the camera with observation residuals to dynamically estimate the motion state of feature points. Importantly, this visual-only approach elim- inates dependence on inertial

measurements, thereby avoiding drift and noise often introduced by low-cost or misaligned IMUs, thereby enhancing robustness in practical deployments.

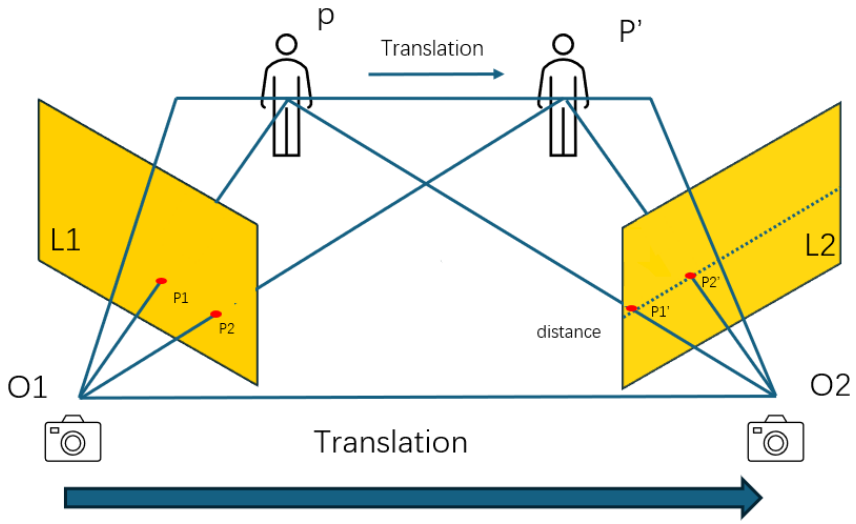

Figure 2. Degeneracy constraint: coplanar motion yields epipolar distance < τ for dynamic features, causing false-static labels

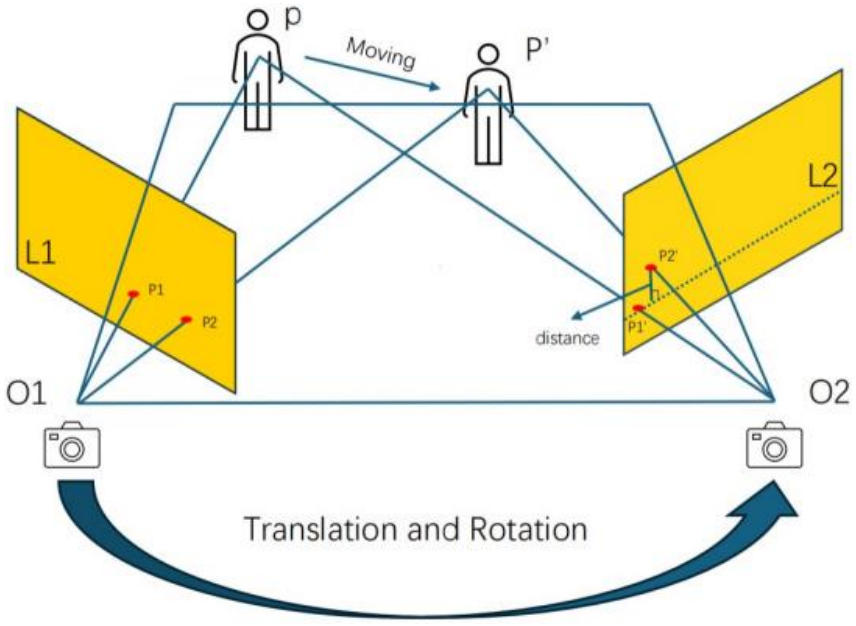

Figure 3. The epipolar constraint serves as the geometric criterion for SLAM dynamic-outlier removal: points with epipolar distance > τ are labeled dynamic.

Let the motion state of each feature point be a discrete variable $zi \in \{0, 1\}$, where $zi = 0$ represents a static point and $zi = 1$ represents a dynamic point. This state variable is not directly observable but can be inferred indirectly through image pixels. Let the feature position transformation from time t to $t + \Delta t$ be

$$\mathbf{T}_{t\to t+\Delta t}^{\text{prior}} \in SE(3). \qquad (2)$$

The current observation coordinates of the feature point are denoted $\mathbf{x}(i\ t)$, and its inverse depth is ρi. Fig. 3 illustrates how the epipolar constraint yields the predicted position and the residual. Based on this prior transformation, the predicted position in the next frame is:

$$\hat{\mathbf{x}}_i^{(t+\Delta t)} = \pi\left(\mathbf{T}_{t\to t+\Delta t}^{\text{prior}} \cdot \pi^{-1}\left(\mathbf{x}_i^{(t)}, \rho_i\right)\right), \tag{3}$$

where $\pi(\cdot)$ and $\pi^{-1}(\cdot)$ are the projection and inverse projection functions, respectively. The residual between the predicted position and the actual observed position is defined as:

$$\mathbf{r}_i = \mathbf{x}_i^{\text{obs}(t+\Delta t)} - \hat{\mathbf{x}}_i^{(t+\Delta t)}. \tag{4}$$

Under different motion state assumptions, the residual fol- lows distinct Gaussian distributions:

$$P(\mathbf{r}_i \mid z_i^{(t)} = 0) = \mathcal{N}(\mathbf{0}, \Sigma_0), \tag{5}$$

$$P(\mathbf{r}_i \mid z_i^{(t)} = 1) = \mathcal{N}(\mathbf{0}, \Sigma_1), \quad \Sigma_1 \gg \Sigma_0, \tag{6}$$

where Σ0 corresponds to the expected variance of staticpoint residuals, while Σ1 is set as an enlarged covariance to account for dynamic uncertainty. In practice, we let $\Sigma 0 = \sigma 2\, I$ and $\Sigma 1 = \kappa \Sigma 0$ with $\kappa \gg 1$ (set to $\kappa = 5$ in our work), ensuring sufficient separation between static and dynamic likelihoods. We estimate $\sigma 2$ as the empirical variance of residuals from points inferred to be static over the past W frames. Initially, we set $\sigma = 5$ pixels.

To enhance the robustness of motion state inference against transient outliers and measurement noise, we adopt a multiframe Bayesian fusion scheme in place of the conventional single-frame update. Instead of relying solely on the residual from the current frame $\mathbf{r}(i\,t)$, our method aggregates evidential cues over a temporal window spanning the past W frames. Let n $\mathbf{r}(i\,t{-}w)$ o W $w{=}0$ -1 denote the residuals of feature point i over the most recent W frames. Assuming conditional independence of residuals given the state z, the joint likelihood of observing this residual sequence under state z is:

$$P\left(\left\{\mathbf{r}_i^{(t-w)}\right\}_{w=0}^{W-1} \mid z_i^{(t)} = z\right) = \prod_{w=0}^{W-1} P\left(\mathbf{r}_i^{(t-w)} \mid z_i^{(t)} = z\right) \tag{7}$$

Exponential Memory Weighting emphasize recent observations while retaining long-term context, we introduce an exponential decay factor $\alpha \in (0,\ 1]$, leading to a weighted loglikelihood. This formulation effectively gives more weight to newer frames and can be tuned via α (set $\alpha = 0.9$ in our work).

$$\log P\left(\{\mathbf{r}_i\} \mid z\right) = \sum_{w=0}^{W-1} \alpha^w \log P\left(\mathbf{r}_i^{(t-w)} \mid z\right) \tag{8}$$

Posterior Inference via Bayes Rule The posterior belief over the motion state $z_i^{(t)} \in \{0,\ 1\}$ is then updated using the aggregated likelihood:

$$P\left(z_i^{(t)} = z \mid \left\{\mathbf{r}_i^{(t-w)}\right\}\right) = \frac{P\left(\left\{\mathbf{r}_i^{(t-w)}\right\} \mid z_i^{(t)} = z\right) \cdot P\left(z_i^{(t)} = z\right)}{\sum_{z' \in \{0,1\}} P\left(\left\{\mathbf{r}_i^{(t-w)}\right\} \mid z_i^{(t)} = z'\right) \cdot P\left(z_i^{(t)} = z'\right)} \tag{9}$$

Final Motion Classification The final decision is made via maximum a posteriori (MAP) estimation:

$$\hat{z}_i^{(t)} = \arg\max_{z \in \{0,1\}} P\left(z_i^{(t)} = z \mid \left\{\mathbf{r}_i^{(t-w)}\right\}_{w=0}^{W-1}\right) \tag{10}$$

Posterior Inference via Bayes Rule The posterior belief over the motion state $zi\ (t) \in \{0,\ 1\}$ is then updated using the aggregated likelihood:

3.3 Keyframe Selection Strategies

To enhance system efficiency and mapping quality under degenerate motion or dynamic disturbances, we present a multi-criteria keyframe selection strategy that refines the ORB- SLAM2 policy for more complex dynamic and degraded environments. Centered on the concept of information gain, the proposed scheme jointly accounts for pose variation, image content change, and static region quality, and introduces three novel indicators to improve keyframe selection under degeneracy:

1) **Absolute Image Entropy Difference**: reflects changes in information density and provides a non-geometric cue for assessing inter-frame information increment, allowing the system to detect newly appeared objects or structures even under geometric degradation. It is defined as

$$\Delta H = |H(I_t) - H(I_k)| \tag{11}$$

where $H(I) = -\sum_i p(i) \log p(i)$ is the image entropy computed from the grayscale intensity histogram, and I_k denotes the most recent keyframe.

2) **Proportion of Effectively Constrained Static Pixels** *(ρ_S):* evaluates the spatial distribution quality of static features outside dynamically removed regions, avoiding feature points being overly clustered or confined to a single plane. It is computed as the proportion of reliable static pixels that remain after dynamic region filtering relative to the total number of pixels.

3) **View-Perspective Change Intensity (VPCI)**: a unified metric that fuses both translational and rotational motion cues into a single scalar, defined as

$$\mathrm{VPCI} = \sqrt{\alpha\,\Delta T^2 + \beta\,\Delta R^2} \tag{12}$$

where ΔT is the Euclidean translation distance to the last keyframe (in meters), ΔR is the angular rotation difference (in radians), and α and β are learned weighting coefficients that balance scale and contribution between translation and rotation.

Keyframe candidacy is established through a tri-criterial trigger mechanism. A frame is promoted to keyframe status if any of the following conditions is satisfied:

$$\delta(I_t) = 1\left(\Delta H_t > \theta_H(t) \vee \rho_{s,t} < \theta_\rho(t) \vee \mathrm{VPCI}_{I_t} > \theta_V(t)\right) \tag{13}$$

where $\delta(I_t) = 1$ indicates that I_t is selected as a keyframe; $1(\cdot)$ denotes the indicator function, which outputs 1 if the condition inside is satisfied and 0 otherwise. The thresholds θH, $\theta\rho$, and θV are determined by an adaptive thresholding scheme based on the Exponentially Weighted Moving Average (EWMA), which is designed to enhance robustness and maintain controllability of the target false alarm rate. For each statistical indicator, ΔH, ρs, and VPCI, the recursive estimation of the mean and variance is given by:

$$\begin{aligned}
\mu_H(t) &= (1-\beta)\mu_H(t-1) + \beta\Delta H_t \\
\sigma_H^2(t) &= (1-\beta)\sigma_H^2(t-1) + \beta(\Delta H_t - \mu_H(t))^2 \\
\mu_\rho(t) &= (1-\beta)\mu_\rho(t-1) + \beta\rho_{s,t} \\
\sigma_\rho^2(t) &= (1-\beta)\sigma_\rho^2(t-1) + \beta(\rho_{s,t} - \mu_\rho(t))^2 \\
\mu_V(t) &= (1-\beta)\mu_V(t-1) + \beta\mathrm{VPCI}_t \\
\sigma_V^2(t) &= (1-\beta)\sigma_V^2(t-1) + \beta(\mathrm{VPCI}_t - \mu_V(t))^2
\end{aligned} \tag{14}$$

where β is the smoothing coefficient balancing historical statistics and current observations (set to 0.2 in this work), while $\mu(\cdot)$ and $\sigma(\cdot)$ denote the EWMA-based mean and variance, respectively.

Accordingly, the adaptive thresholds are defined as:

$$\begin{aligned}
\theta_H(t) &= \mu_H(t) + k\sigma_H(t) \\
\theta_\rho(t) &= \mu_\rho(t) - k\sigma_\rho(t) \\
\theta_V(t) &= \mu_V(t) + k\sigma_V(t)
\end{aligned} \tag{15}$$

Given a target significance level α (set to 0.05 in this work), the control limit k is obtained from the standard normal distribution: $k = z1-\alpha = z1-0.05 = 1.65$. For ΔH and VPCI, thresholds are set as $\mu+k\sigma$ since larger values indicate stronger abnormality, while for ρs they are set as $\mu - k\sigma$ as smaller values imply weaker static constraints. The proposed keyframe-selection workflow (executed within the ORB-SLAM2 framework) is summarised below:

1) Initialization: The first frame is selected as the reference keyframe Iref, and for subsequent frames, the reference is updated to the most recently inserted keyframe.

2) Indicator calculation: For each new frame I_t, the indicators ΔT, ΔR, ΔH, and ρs are computed.

3) Insertion Decision: A keyframe is inserted if Eq.(12) holds, where the thresholds θH, $\theta\rho$, and θV are are computed according to Eq.(14).

4. Experiental Results

This section presents a comprehensive evaluation of the proposed method from five aspects. First, trajectory accuracy is evaluated on the TUM RGB-D benchmark dataset [35]. Second, robustness and generalization are assessed on the more challenging BONN dataset [35], which features complex dynamic scenes with significant motion degeneracy. Third, ablation studies are carried out to quantify the contribution of each module in the proposed framework. Fourth, runtime analysis is performed to evaluate computational efficiency and real-time capability. Fifth, a case study is conducted on rep- resentative sequences (crowd1 and crowd2) to analyze system performance under severe motion degeneracy and co-moving dynamic objects— conditions under which conventional meth- ods typically suffer from substantial trajectory drift. Finally, the results are thoroughly interpreted, and their implications for robust visual SLAM in real-world dynamic environments are discussed.

The experiments were conducted on a laptop equipped with an Intel i5-9600 CPU, NVIDIA GeForce RTX 1650 GPU, and 8 GB RAM, running Ubuntu 20.04.

The algorithm proposed in this paper is built upon ORB- SLAM2. To ensure a fair and consistent experimental baseline, we selected SG-SLAM and YOLO-SLAM—both of which are also based on ORB-SLAM2—as comparative baselines. Furthermore, to comprehensively evaluate the improvements of our approach, we conducted extensive comparisons with various state-of-the-art SLAM algorithms.

In the experiments, we rigorously evaluate the SLAM algorithm using Absolute Trajectory Error (ATE) and Relative Pose Error (RPE). For each metric we report the Root Mean Square Error (RMSE), Standard Deviation (S.D.), Mean, and Median to provide a comprehensive assessment of accuracy and consistency.

All quantitative results in the tables follow a unified format- ting convention: **bold** denotes the best value and underline the second-best value across all compared methods.

4.1 Evaluation on the TUM RGB-D Dataset

The TUM RGB-D Dataset is a widely recognized bench- mark in the field of SLAM. The dataset contains a large number of image sequences captured by an RGB-D camera in dynamic environments, along with corresponding ground truth trajectories. We selected four sequences from highly dy- namic scenarios and one from relatively static environments to evaluate system performance. These sequences are labeled as: ”W/xyz”, ”W/static”, ”W/rpy”, ”W/half”, and ”S/static”. In these labels, ”W” denotes sequences involving walking per- sons, while ”S” refers to sequences with stationary individuals. The second part of each label indicates the camera motion pattern: ”xyz” represents translational motion along the three axes, ”static” indicates no movement, ”rpy” denotes rotational motion around different axes, and ”half” corresponds to hemi- spherical motion. This combination of sequences enables a comprehensive and accurate evaluation of the system across various dynamic scenarios.

As shown in Table I, compared to ORB-SLAM2, [36] our SLAM system achieves more than a 94 % improvement on most metrics in highly dynamic sequences. As shown in Fig. 4(a)-(d) and Fig. 5(a)-(d), the trajectory estimation of our system in high-dynamic scenarios is significantly more accurate than that of ORB- SLAM2.

As shown in Fig. 6(c), the trajectory estimated by our method on the walking_rpy sequence closely follows the ground truth, whereas competing algorithms exhibit substantial tracking losses. In this scenario, the camera performs contin- uous, large-amplitude rotations while two subjects alternately occupy the foreground and background, creating a canonical “pure-rotation + heavy pedestrian dynamics” configuration that severely challenges SLAM systems. Leveraging a degeneracy- aware constraint that accurately suppresses dynamic features and an information-driven keyframe selection strategy that pre- serves representative frames, we eliminate most dynamic out- liers without sacrificing static landmarks, yielding an accurate and continuous trajectory under such demanding conditions. In complex dynamic environments, the system effectively eliminates feature points associated with moving objects while simultaneously selecting keyframes that capture significant motion changes and rich map information. As reported in Table 2 and Table 3, this strategy yields substantial reductions in both relative rotation and translation errors compared to ORB-SLAM2.

To further validate the proposed method, we compare it with YOLO-SLAM, DS-SLAM, SG-SLAM [37], RDS-SLAM [38], and CA-SLAM [39]; quantitative results are summarized in Table 4. Among all baselines, only SG-SLAM holds a slight edge on low-dynamic sequences, whereas our system consistently maintains the highest accuracy in high-dynamic scenarios.

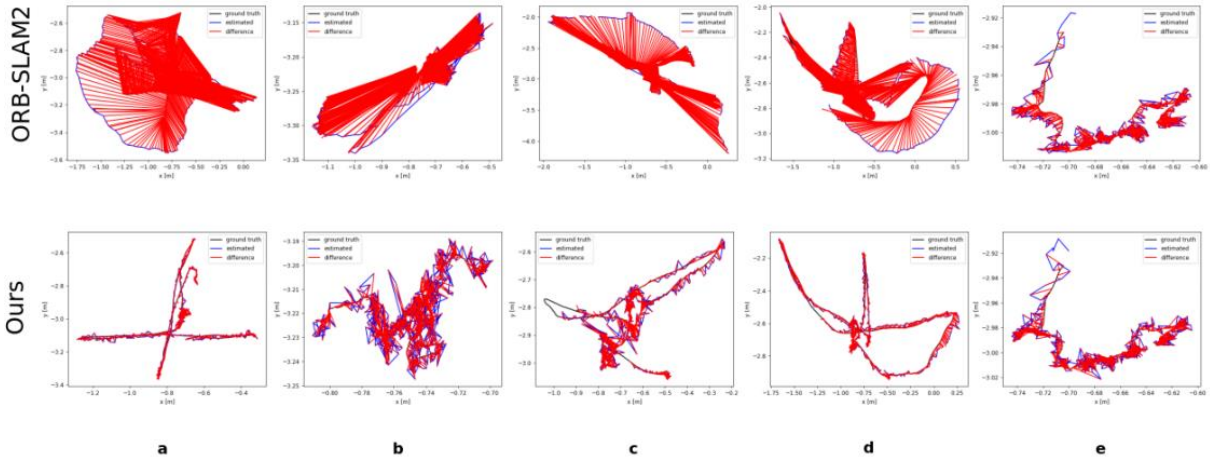

Figure 4. ATE comparison on the TUM dataset between ORB-SLAM2 and our SLAM system

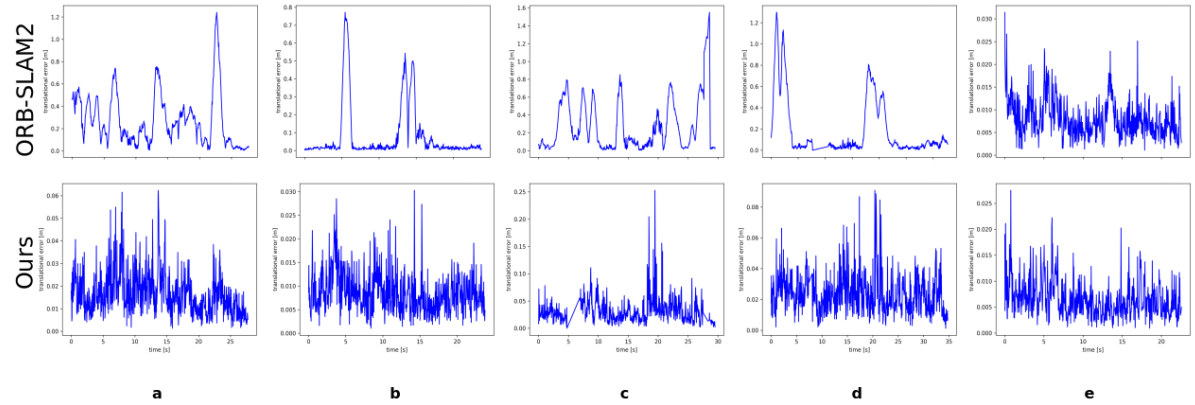

Figure 5. RPE comparison on the TUM dataset between ORB-SLAM2 and our SLAM system

4.2 Evaluation on the Bonn Dataset

The Bonn RGB-D Dynamic Dataset [40] contains a total of 24 dynamic sequences, in which people perform various ac- tivities such as crowd walking, box moving, and synchronized motion. For each sequence, ground truth camera trajectories are provided, recorded using an Optitrack Prime 13 motion capture system.

Nine representative sequences were selected from the Bonn dataset for performance evaluation. The ”crowd” sequence captures a scene where three people walk randomly indoors. The ”moving_no_box” sequence depicts a person moving an unobstructed box from the floor to a table. In the ”per-son_tracking” sequence, the camera follows a person walking slowly. The ”synchronous” sequence shows two individuals moving in the same direction and at the same speed.

Table 1. Tum Dataset Ate Comparison

Sequence	ORB-SLAM2				Our Method				Improvements(%)			
	RMSE	Mean	Median	S.D.	RMSE	Mean	Median	S.D.	RMSE	Mean	Median	S.D.
walking xyz	0.6826	0.6086	0.6661	0.3091	0.0143	0.0124	0.0115	0.0072	97.97	98.00	98.33	97.70
walking static	0.4032	0.3690	0.3164	0.1627	0.0074	0.0062	0.0058	0.0032	98.27	98.34	98.23	98.04
walking rpy	0.5396	0.5012	0.4974	0.1999	0.0299	0.0223	0.0189	0.0192	94.48	95.41	96.23	90.40
walking half	0.4462	0.4096	0.3800	0.1770	0.0258	0.0222	0.0219	0.0135	94.24	94.65	94.99	92.49
sitting static	0.0087	0.0078	0.0072	0.0039	0.0063	0.0055	0.0050	0.0030	27.59	29.49	30.58	23.08

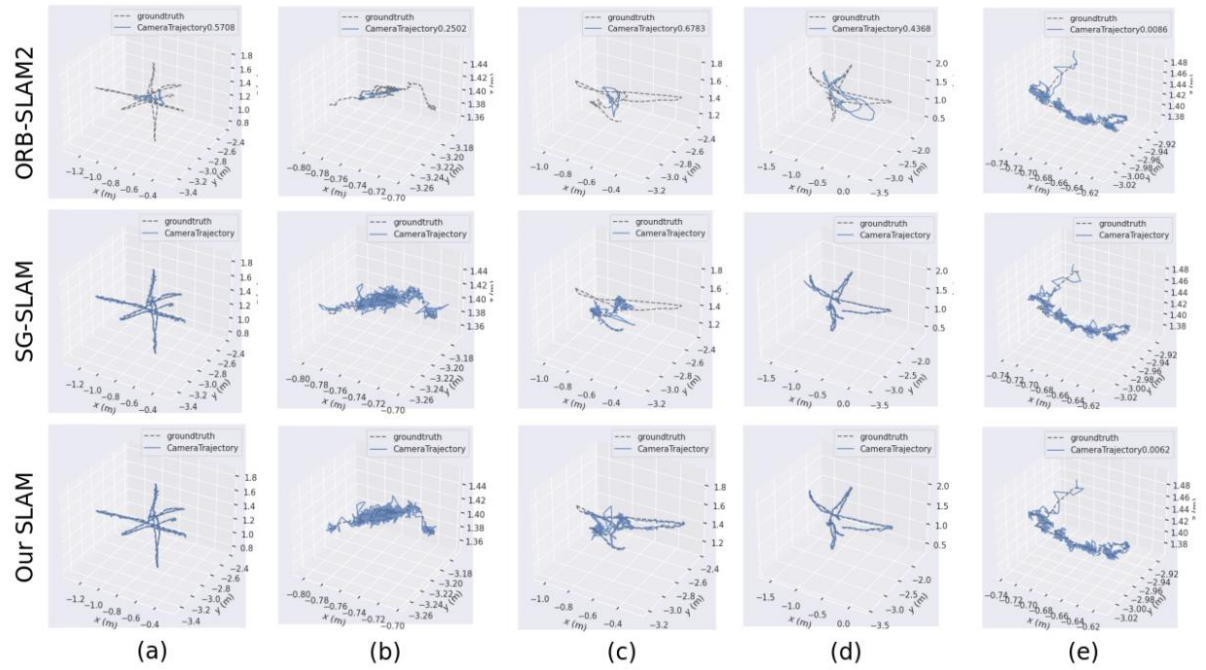

Figure 6. Trajectory comparison of ORB-SLAM2, SG-SLAM, and our SLAM system on the TUM dataset.

Table 2. Tum Dataset Rre Comparison

Sequence	ORB-SLAM2				Our Method				Improvements(%)			
	RMSE	Mean	Median	S.D.	RMSE	Mean	Median	S.D.	RMSE	Mean	Median	S.D.
walking xyz	7.1415	5.6403	4.6159	4.3804	0.0096	0.0068	0.0057	0.0067	99.87	99.88	99.83	99.85
walking static	3.8068	1.6993	0.3888	3.4065	0.0046	0.0040	0.0037	0.0022	99.87	99.77	99.04	99.93
walking rpy	6.4220	4.5134	2.2990	4.5685	0.0144	0.0103	0.0079	0.0101	99.78	99.77	99.66	99.78
walking half	7.9219	4.4695	1.2568	6.5406	0.0101	0.0086	0.0076	0.0052	99.88	99.81	99.39	99.92
sitting static	0.2899	0.2606	0.2484	0.1271	0.0040	0.0033	0.0028	0.0021	98.58	98.73	98.92	98.35

These nine sequences represent highly dynamic environ- ments that severely challenge traditional SLAM systems. In Fig. 7 we compare our method with ORB-SLAM2 and SG- SLAM: the latter incorrectly retains dynamic points due to degeneracy constraints, whereas our system removes them and, as Table 5 confirms, achieves the best metrics on most of the high-dynamic Bonn sequences.

Quantitative results in Table 5 demonstrate that our method achieves the lowest absolute trajectory error (ATE) across both the crowd1 and crowd2 sequences among all compared methods, clearly highlighting its superior resilience to both dynamic disturbances and motion degeneracy.

Table 3. Tum Dataset Rte Comparison

Sequence	ORB-SLAM2				Our Method				Improvements(%)			
	RMSE	Mean	Median	S.D.	RMSE	Mean	Median	S.D.	RMSE	Mean	Median	S.D.
walking_xyz	0.3752	0.2944	0.2394	0.2326	0.0124	0.0101	0.0088	0.0064	96.80	96.58	96.34	97.24
walking static	0.2182	0.0950	0.0169	0.1965	0.0073	0.0063	0.0056	0.0036	96.67	93.37	67.46	98.19
walking rpy	0.3374	0.2344	0.1137	0.2426	0.0245	0.0176	0.0129	0.0173	92.73	92.58	87.70	92.87
walking half	0.3685	0.2072	0.0491	0.3047	0.0145	0.0121	0.0101	0.0088	96.09	94.17	79.43	97.36
sitting static	0.0093	0.0082	0.0074	0.0044	0.0055	0.0048	0.0043	0.0026	40.86	41.46	41.89	40.91

Table 4. ATE COMPARISON ACROSS SLAM SYSTEMS ON THE TUM DATASET

Sequence	YOLO-SLAM		DS-SLAM		ORB-SLAM2		SG-SLAM		CA-SLAM		Our Method	
	RMSE	S.D.	RMSE	S.D.	RMSE	S.D.	RMSE	S.D.	RMSE	S.D.	RMSE	S.D.
walking_xyz	0.0146	0.0070	0.0247	0.0161	0.0571	0.0229	0.0152	0.0075	0.0154	0.0073	0.0143	0.0072
walking_static	0.0073	0.0035	0.0081	0.0036	0.0206	0.0120	0.0073	0.0034	/	/	0.0070	0.0032
walking_rpy	0.2164	0.1001	0.4442	0.2350	0.1604	0.0873	0.0324	0.0187	0.0408	0.0209	0.0299	0.0192
walking_half	0.0283	0.0138	0.0303	0.0159	0.0807	0.0454	0.0268	0.0134	0.0105	0.0149	0.0258	0.0135
sitting_static	0.0066	0.0033	0.0065	0.0033	0.0084	0.0043	0.0060	0.0029	0.0085	0.0047	0.0063	0.0030

Table 5. COMPARISON ON BONN DYNAMIC SEQUENCES (ATE, UNIT: M).

Sequence	ORB-SLAM2				SG-SLAM				RDS-SLAM				Our Method			
	RMSE	Mean	Median	S.D.	RMSE	Mean	Median	S.D.	RMSE	Mean	Median	S.D.	RMSE	Mean	Median	S.D.
crowd1	0.8632	0.6284	0.3592	0.5918	0.0234	0.0185	0.0161	0.0143	0.018	/	/	0.0099	0.0178	0.0170	0.0144	0.0099
crowd2	1.3573	1.2071	1.1163	0.6207	0.0584	0.0420	0.0301	0.0406	0.0363	/	/	0.013	0.0325	0.0251	0.0198	0.0207
crowd3	1.0772	1.0070	0.9733	0.3823	0.0319	0.0231	0.0187	0.0219	0.036	/	/	0.020	0.0318	0.0273	0.0231	0.0164
moving box1	no	0.1174	0.0935	0.0785	0.0710	0.0192	0.0174	0.0156	0.0081	0.064	/	/	0.070	0.018	0.0148	0.0129	0.0102
moving box2	no	0.1142	0.0973	0.0955	0.0598	0.0299	0.0275	0.0261	0.0119	0.035	/	/	0.011	0.0278	0.0265	0.0266	0.0084
person tracking1		0.7959	0.7090	0.7410	0.3617	0.040	0.0375	0.038	0.0139	0.044	/	/	0.015	0.0369	0.0327	0.030	0.017
person tracking2		1.0679	0.9590	0.8732	0.4699	0.0376	0.0343	0.0312	0.0154	0.048	/	/	0.016	0.0373	0.0342	0.0351	0.015
synchronous1		1.1411	0.9884	0.9179	0.5703	0.3229	0.2665	0.1722	0.1824	0.043	/	/	0.0079	0.0134	0.0107	0.0092	0.0079
synchronous2		1.4069	1.3201	1.3259	0.4864	0.0105	0.0073	0.0073	0.0126	0.009	/	/	0.005	0.0072	0.0062	0.0054	0.0037

4.3 Ablation Study

In this section, we conduct a systematic ablation study on the proposed SLAM system to gain deeper insight into the contribution of each module to the overall performance. By progressively introducing the semantic module (S), epipolar constraint (G), and degeneracy constraint (H), we evaluate the system's behavior under complex scenarios across different configurations. Experimental results show that the localization accuracy improves significantly as each component is added sequentially.

Specifically, on the *walking_xyz* sequence, the RMSE de- creases from 0.0172 with only the semantic module (S), to 0.0157 when combining semantics and epipolar constraints (S+G), and further drops to 0.0143 after adding the degeneracy constraint (S+G+H). Similarly, on the *walking_rpy* sequence, the RMSE is initially 0.0666, which is reduced to 0.0411 with the addition of epipolar constraints, and then significantly decreases to 0.0299 after incorporating the degeneracy con- straint— corresponding to a relative error reduction of up to 29.7%.

The complete ablation results for all tested sequences are compiled in Table 6.

This substantial improvement not only demonstrates the effectiveness of the epipolar constraint in handling dynamic feature points, but also highlights the critical role of the degen- eracy constraint in accurately distinguishing between dynamic and static features. By predicting the theoretical displacement of dynamic points based on the camera's rotation matrix **R** and translation vector **t**, the degeneracy constraint effectively identifies and removes potential misclassified points, thereby significantly enhancing the system's robustness and accuracy in highly dynamic environments.

In addition, consistent improvements are observed in other sequences such as *sitting_static* and *walking_half*, where the system demonstrates notable enhancements in terms of mean, median, and standard deviation. This further validates the gen- erality and stability of the proposed modules across different motion patterns. In summary, the experimental results fully demonstrate the effectiveness of our dynamic feature rejection strategy in improving both the accuracy and robustness of the SLAM system.

Table 6. ABLATION STUDY ON DYNAMIC SEQUENCES.

Sequence	Our Method (S)				Our Method (S+G)				Our Method (S+G+H)			
	RMSE	Mean	Median	S.D.	RMSE	Mean	Median	S.D.	RMSE	Mean	Median	S.D.
walking_xyz	0.0172	0.0147	0.0128	0.0088	0.0157	0.0132	0.0110	0.0086	0.0143	0.0124	0.0113	0.0072
walking_static	0.0078	0.0069	0.0064	0.0036	0.0081	0.0068	0.0062	0.0042	0.0074	0.0062	0.0056	0.0032
walking_rpy	0.0666	0.0420	0.0278	0.0516	0.0411	0.0292	0.0214	0.0288	0.0299	0.0230	0.0189	0.0192
walking_half	0.0303	0.0259	0.0219	0.0156	0.0299	0.0252	0.0206	0.0160	0.0258	0.0224	0.0190	0.0135
sitting_static	0.0068	0.0060	0.0053	0.0032	0.0076	0.0066	0.0059	0.0037	0.0063	0.0055	0.0050	0.0030

Table 7. AVERAGE PROCESSING TIME PER FRAME (MS) AND HARDWARE PLATFORM

System	Avg. Time (ms)	Hardware Platform
SG-SLAM	57.81	i5-9400F + NVIDIA GTX 1650
ORB-SLAM2	**22.21**	i5-9400F + NVIDIA GTX 1650
Our SLAM	30.52	i5-9400F + NVIDIA GTX 1650
YOLO-SLAM	696.09	Intel i5-4228U CPU
DS-SLAM	59.40	Intel i7 + P4000 GPU
RDS-SLAM	57.50	RTX 2080 Ti GPU

4.4 Comparision Of Runtime

We benchmark the per-frame latency of tracking and map- ping on the TUM RGB-D dataset and compare our system against representative dynamic SLAM methods as well as the original ORB-SLAM2. As reported in Table 7, our approach retains real-time throughput while sustaining high localization accuracy. Thanks to a carefully engineered multi- threaded architecture, an lightweight object-detection module, and an improved keyframe-selection policy, the extra runtime introduced by our extensions is only 6-12 ms per frame—an overhead that is acceptable in practice.

4.5 Case Study

To further evaluate the system's performance under chal- lenging conditions characterized by the simultaneous presence of dynamic objects and motion degeneracy, we conducted case studies on the crowd1 and crowd2 sequences from the Bonn dataset. In these scenarios, three pedestrians traverse the scene with complex, intersecting trajectories while the camera undergoes combined translational and rotational ego- motion. Notably, the primary direction of camera movement aligns closely with that of the pedestrians, resulting in minimal relative motion between them. This situation leads to low parallax and near-degenerate motion patterns, which severely disrupt motion-based dynamic object detection methods. Con- sequently, due to insufficient motion cues, traditional methods such as ORB-SLAM2 and SG-SLAM fail to distinguish mov- ing pedestrians from static structures, erroneously incorporat- ing dynamic features into the background model, as illustrated in Fig. 7.

Particularly, in Fig. 7(a), it is evident that even though SG-SLAM integrates semantic segmentation and geometric constraints, it still performs poorly in these sequences because it does not account for motion degeneracy when handling person-labeled features. This highlights the limitations of overly relying on semantic information and geometric con- straints without considering the degeneration of geometric constraints under low-parallax conditions.

In contrast, our method utilizes semantic segmentation to identify human-shaped regions and introduces a degeneracy- aware constraint mechanism that suppresses feature matching within these regions when motion consistency is ambiguous. Even under conditions of minimal parallax and complex ego-motion, our approach mitigates the adverse impact of dynamic objects by computing the depth values of these features during pose optimization to predict their motion and thereby determine which points are most likely to be dynamic. This enables the system to adaptively remove potentially unstable features to the greatest extent. The dual mecha- nism—combining semantic awareness with motion-context- adaptive geometric constraints—enables robust camera local- ization in highly deceptive scenes where dynamic motion and degenerate camera motion coexist, as shown in Fig. 7(b)-(e). Our algorithm maximizes the suppression of dynamic feature points in degenerate scenarios compared to other strategies.

As summarized in Table 5, our method achieves the lowest Root Mean Square Error (RMSE): 0.0178 m on crowd1 and 0.0325 m on crowd2, outperforming ORB-SLAM2 (0.8632 m and 1.3573 m), SG-SLAM (0.0234 m and 0.0584 m), and RDS- SLAM (0.0180 m and 0.0363 m). Although SG-SLAM em- ploys a robust dynamic

filtering mechanism, its reliance on consistent motion cues makes it ineffective against slowly moving or co-directional pedestrians. Compared to existing methods, our approach significantly reduces localization er- rors, demonstrating its effectiveness and state-of-the-art ac- curacy under extreme motion degeneracy and high dynamic interference conditions. These results confirm that integrating semantic priors with degeneracy-aware geometric reasoning is crucial for robust SLAM in real-world crowded environments.

4.6 Discussion

The proposed SLAM framework demonstrates substantial advantages over existing dynamic SLAM methods across both the TUM RGB-D and Bonn datasets. Compared with ORB- SLAM2, SG-SLAM, YOLO-SLAM, DS-SLAM, and RDS- SLAM, our system consistently achieves the lowest Absolute Trajectory Error (ATE) and Relative Pose Error (RPE) in high-dynamic and motion-degenerate scenarios. On the TUM dataset, our method achieves up to 94-99% error reduction relative to ORB-SLAM2 across walking xyz, walking rpy, and walking half sequences, while achieving moderate improve- ments on low-dynamic sequences. On the Bonn dataset, our method attains an ATE of 0.0178 m on crowd1 and 0.0325 m on crowd2, outperforming the SG-SLAM by up to 58%, thus clearly evidencing its superior resilience to both dynamic disturbances and motion degeneracy.

The strong performance of our method can be attributed to two key factors: (1) tight integration of lightweight semantics with motion-consistency modeling, which reduces false-static classifications caused by degenerate constraints, and (2) multi- frame Bayesian fusion, which robustly accumulates evidence over time, improving discrimination accuracy beyond what is possible with single-frame checks.(3) an adaptive multi- metric keyframe insertion strategy, which jointly considers motion, image entropy, and static constraints to select highly informative keyframes.

Despite these significant advantages, some limitations re- main. First, the detection performance of YOLO-FASTEST may degrade under extreme illumination changes or for very rare object categories not present in the training data. Sec- ond, because the system currently relies solely on monocular visual information, tracking stability can still be challenged in extremely low-texture regions, large featureless surfaces, or under abrupt accelerations where parallax cues are minimal.

Figure 7. Comparison on the Bonn dataset: our SLAM removes dynamic features while preserving as many static ones as possible

5. Conclusion

This paper presents a robust dynamic SLAM system tai- lored for degenerate-motion scenarios, markedly improving localization accuracy and stability in highly dynamic envi- ronments while preserving real-time performance. By inte-grating a compact semantic detector, a degeneracy-consistent modeling mechanism, and

a multi-criteria-driven keyframe insertion strategy, the system effectively eliminates dynamic feature points in complex scenes and enhances map quality. Experimental results demonstrate that, compared with existing state-of-the-art methods, the proposed system achieves substantial improvements in both ATE and RPE. Meanwhile, computational overhead remains tightly controlled, ensuring strong real-time capability. Future work will enhance adapt- ability to extreme lighting and low-texture environments, and incorporate multi-sensor fusion to further improve robustness and generalization.

References

[1] Yang, Z., Song, S., & Shen, S. (2021). Deep factors: Learning discriminative and robust features for visual slam. In *2021 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 3158–3164). IEEE.

[2] Zhang, Y., Zhang, J., Prasad, D. K., Tan, D. S., & Liu, M. (2022). Analyzing and mitigating degeneracy in visual-inertial SLAM. *IEEE Transactions on Robotics*, *38*(4), 2337–2354.

[3] Zhang, Z., Wu, J., Zhang, C., Cheng, X., Li, X., Wu, X., & Liu, Y. (2021). Robust visual SLAM with dynamic object detection in highly dynamic environments. *IEEE Transactions on Robotics*, *37*(3), 588–602.

[4] Zhou, Z., Zhang, Z., Wu, J., Zhang, C., & Liu, Y. (2020). Degenerate motion handling in visual SLAM for complex environments. *IEEE Robotics and Automation Letters*, *5*(4), 6027–6034.

[5] Cherubini, A., De Souza, C., Burschka, A., & Oriolo, G. (2021). Dynamic object-aware visual SLAM for mobile robots in crowded environments. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 10345–10352). IEEE.

[6] Jocher, G., Stasi, B., & Chang, H. (2020). *YOLOv5: An open-source implementation of YOLO series*. GitHub. https://github.com/ultralytics/yolov5 (Accessed: 2025-08-29)

[7] Zhang, X., Li, P., Xu, C., & Zhang, Y. (2022). Modeling motion residuals in dynamic environments for robust SLAM. *IEEE Transactions on Robotics*, *38*(6), 1542–1556.

[8] Zhao, L., Wang, Y., Li, X., & Wu, T. (2023). Improved optical flow algorithms for SLAM in dynamic environments. *IEEE Transactions on Robotics*, *40*(2), 490–502.

[9] DefTruth. (2021). *YOLO-Fastest: A PyTorch implementation*. GitHub. Retrieve from 2025-08-29 https://github.com/DefTruth/yolo-fastest

[10] Zhang, L., Wang, X., Hou, Y., Ma, Z., Li, Y., & Liu, H. (2020). DS-SLAM: Dynamic semantic SLAM for real-time robotic perception. *IEEE Transactions on Robotics*, *36*(5), 1092–1105.

[11] Liu, Y., Zhang, X., Wu, C., & Li, M. (2021). Detect-SLAM: Real-time object detection for dynamic SLAM. *IEEE Robotics and Automation Letters*, *6*(3), 491–498.

[12] Li, M., Zhang, X., Wu, C., Liu, Y., & Wang, Q. (2021). YOLO-SLAM: A real-time SLAM system with YOLO-based object detection for dynamic environments. *IEEE Transactions on Robotics*, *38*(7), 1851–1865.

[13] Yang, C., Zhang, S., Liu, C., & Wang, Y. (2019). DynaSLAM: An open-source library for dynamic object removal in visual SLAM. *IEEE Transactions on Robotics*, *35*(3), 542–554.

[14] Huang, Z., Li, Y., Wu, X., & Liu, Y. (2020). RS-SLAM: Robust semantic SLAM with Bayesian optimization for dynamic environments. *IEEE Transactions on Robotics*, *39*(2), 309–323.

[15] Huang, S., Luo, Z., Peng, Z., Huang, T., Wang, J., Wu, X., Zhang, C., & Pan, X. (2021). YOLO-Fastest: A real-time object detection algorithm optimized for edge devices. *IEEE Sensors Journal*, *21*(12), 13867–13876.

[16] Wang, Z., Zhang, Y., Wu, J., Zhang, C., & Liu, Y. (2020). Dynamic feature point elimination for robust visual SLAM in dynamic environments. *IEEE Transactions on Robotics*, *39*(5), 1122–1134.

[17] Liu, Y., Zhang, X., Wu, C., & Li, M. (2020). ClusterSLAM: Motion consistency-based clustering for robust dynamic feature removal in SLAM. *IEEE Transactions on Robotics*, *37*(4), 958–970.

[18] Zhang, S., Wang, C., Liu, C., & Yang, C. (2021). StaticFusion: Using optical flow residuals for robust static feature detection in dynamic environments. *IEEE Robotics and Automation Letters*, *6*(4), 745–752.

[19] Wu, Q., Li, P., Xu, C., & Zhang, Y. (2020). DynaVINS: Dynamic VINS for visual-inertial navigation in dynamic environments. *IEEE Transactions on Robotics*, *36*(8), 1675–1689.

[20] Zhang, L., Wang, X., Hou, Y., & Liu, H. (2022). DOR-LINS: Lidar-based dynamic object removal for SLAM systems. *IEEE Transactions on Robotics*, *41*(2), 346–358.

[21] Qin, T., Li, P., Cao, S., & Shen, S. (2019). VINS-Fusion: A general optimization-based multi-sensor state

estimator. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 4004–4011). IEEE.

[22] Li, J., Zhang, Y., Wu, J., Zhang, C., & Liu, Y. (2020). Geometry-based dynamic point elimination in visual SLAM: Challenges and solutions. *IEEE Transactions on Robotics*, *38*(6), 1122–1135.

[23] Liu, X., Zhang, Y., Wu, J., Zhang, C., & Liu, Y. (2020). Motion consistency for dynamic object removal in visual SLAM. *IEEE Transactions on Robotics*, *39*(5), 1051–1063.

[24] Wang, L., Zhang, Y., Wu, J., Zhang, C., & Liu, Y. (2021). Multimodal fusion for robust dynamic object detection in SLAM systems. *IEEE Robotics and Automation Letters*, *6*(4), 567–573.

[25] Chen, L., Zhang, Y., Wu, J., Zhang, C., & Liu, Y. (2021). Threshold-driven keyframe insertion for efficient visual SLAM. *IEEE Transactions on Robotics*, *38*(6), 1194–1206.

[26] Liu, Z., Zhang, Y., Wu, J., Zhang, C., & Liu, Y. (2020). Feature tracking for robust keyframe selection in dynamic SLAM systems. *IEEE Transactions on Robotics*, *39*(4), 832–845.

[27] Zhang, X., Li, P., Xu, C., & Zhang, Y. (2022). Information gain-based keyframe selection for visual SLAM systems. *IEEE Sensors Journal*, *22*(1), 345–356.

[28] Li, M., Zhang, X., Wu, C., Liu, Y., & Wang, Q. (2021). SemSLAM: Semantic keyframe selection for robust visual SLAM. *IEEE Transactions on Robotics*, *40*(7), 1670–1683.

[29] Zhang, Y., Wu, J., Zhang, C., Liu, Y., & Wang, Z. (2020). A-VINS: A visual-inertial SLAM system with keyframe insertion based on multimodal fusion. *IEEE Transactions on Robotics*, *38*(8), 1956–1968.

[30] Wang, P., Zhang, Y., Wu, J., Zhang, C., & Liu, Y. (2020). Threshold-driven keyframe selection for visual SLAM in dynamic environments. *IEEE Transactions on Robotics*, *38*(4), 987–1001.

[31] Zhang, C., Zhang, Y., Wu, J., Liu, Y., & Wang, Z. (2020). Feature tracking for robust keyframe selection in weak texture environments. *IEEE Robotics and Automation Letters*, *6*(5), 625–634.

[32] Liu, S., Zhang, Y., Wu, J., Zhang, C., & Liu, Y. (2021). Semantic-based keyframe selection for real-time visual SLAM in dynamic scenes. *IEEE Transactions on Robotics*, *40*(6), 1539–1550.

[33] Wang, X., Zhang, L., Hou, Y., & Liu, H. (2020). Mitigating map drift in visual SLAM systems in dynamic environments. *IEEE Robotics and Automation Letters*, *5*(6), 3157–3163.

[34] Zhang, Y., Wu, J., Zhang, C., & Liu, Y. (2020). Geometric constraints failure in visual SLAM under degenerate motion. *IEEE Transactions on Robotics*, *37*(3), 472–483.

[35] Sturm, J., Engelhard, N., Endres, F., Rother, D., & Burgard, W. (2012). A benchmark for RGB-D visual odometry, 3D reconstruction, and SLAM. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 1–6). IEEE.

[36] Mur-Artal, R., & Tardós, J. D. (2017). ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras. *IEEE Transactions on Robotics*, *33*(5), 1255–1262.

[37] Wang, S., Yang, L., Sun, Y., & Zhang, T. (2021). SG-SLAM: A robust semantic graph SLAM system for dynamic environments. *IEEE Transactions on Robotics*, *37*(6), 1854–1868.

[38] Zhao, L., Zhang, Y., Li, X., & Wu, T. (2021). RDS-SLAM: Robust dynamic SLAM for real-time robotic perception in dynamic environments. *IEEE Transactions on Robotics*, *39*(5), 1379–1391.

[39] Zhang, X., Liu, H., Liu, Z., & Zhang, Y. (2020). CA-SLAM: A context-aware SLAM framework for dynamic environments. *IEEE Transactions on Robotics*, *36*(7), 1704–1716.

[40] Halber, M., Funk, H., Wandt, B., Rosenhahn, B., & Sturm, P. (2019). Bonn RGB-D dynamic dataset: Ground truth for dynamic RGB-D scenes. In *2019 International Conference on Robotics and Automation (ICRA)* (pp. 1–8). IEEE.