Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

BPC-SLAM: Part-Level Dynamic Suppression and Structure-Constrained RGB-D SLAM for Human-Centric Dynamic Environments

Wang Yang
Kunming University of Science and Technology
Jiupeng Chen
183144902250163.com

Kunming University of Science and Technology
Hongjun San
Kunming University of Science and Technology
Fan Zhang
Kunming University of Science and Technology
Wunyu Xu
Kunming University of Science and Technology

Research Article

Keywords: RGB-D SLAM, Dynamic scenes, Human Motion Suppression, Dense Optical Flow, YOLOv8, Graph Optimization

Posted Date: April 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9197331/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

BPC-SLAM: Part-Level Dynamic Suppression and Structure-Constrained RGB-D SLAM for Human-Centric Dynamic Environments

Wang Yang[1], Jiupeng Chen[1,2,*],Hongjun San[1,2], Fan Zhang[1,2],Wunyu Xu[1]
1 Faculty of Mechanical and Electrical Engineering, Kunming University of Science and Technology, Kunming 650500, China
2 Key Laboratory of Advanced Equipment Intelligent Manufacturing Technology of Yunnan Province, Kunming 650500, China
* Corresponding author. Jiupeng Chen
E-mail address: 18314490225@163.com

ABSTRACT

To address the tracking degradation caused by global semantic-mask removal in human-dynamic scenes—where stable observations are mistakenly discarded, leading to insufficient geometric constraints—we propose a part-level motion determination strategy that suppresses truly dynamic interference while preserving valid observations from relatively static human parts, and further improves the robustness and accuracy of pose estimation by incorporating human structural constraints. Accordingly, we present BPC-SLAM, an RGB-D SLAM system featuring part-level dynamic suppression and structure-constrained optimization for dynamic human environments. Built upon ORB-SLAM3, the proposed system introduces human instance segmentation and pose keypoint estimation in the front end to partition each human instance into semantic parts such as the head, torso, and limbs. We then perform part-level robust statistics on the residual of dense optical flow after camera-motion compensation, so that only genuinely moving parts are suppressed while geometrically informative observations from relatively static parts are retained, thereby alleviating over-rejection in low-dynamic or locally moving scenarios. Furthermore, in the back end, constraints including the human–camera relative pose, keypoint reprojection, temporal motion priors, and bone-length consistency are integrated into a unified graph optimization framework to enhance observability and robustness under dynamics. Experiments on the TUM RGB-D dataset show that, compared with ORB-SLAM3, our method reduces ATE-RMSE by 97.38% on the highly dynamic sequence fr3/walking and by 46.79% on the low-dynamic sequence fr3/sitting. Additional comparisons, ablation studies, and real indoor-scene tests further demonstrate the effectiveness and stability of the proposed approach.

Keywords: RGB-D SLAM, Dynamic scenes, Human Motion Suppression, Dense Optical Flow, YOLOv8, Graph Optimization

1. Introduction

Visual simultaneous localization and mapping (Visual SLAM) is a core enabling technology for mobile robots, autonomous navigation, and augmented reality systems [1]. It leverages visual sensors to estimate the camera pose online while constructing a map of an unknown environment in real time. Compared with LiDAR-based SLAM, Visual SLAM offers advantages such as lower cost and richer information modalities, and has been widely applied to scenarios including indoor service robots and autonomous agricultural machinery [2], [3].

Classic Visual SLAM approaches, represented by ORB-SLAM2 [4], ORB-SLAM3 [5], DSO [6], and LSD-SLAM [7], can achieve high accuracy in static environments; however, their designs typically assume the scene to be approximately static. In real-world applications, dynamic objects such as pedestrians and vehicles are ubiquitous. Their motion regions introduce erroneous data associations and undermine map consistency, causing accumulated pose estimation errors and even tracking failure [8], [9]. A system-level comparison study targeting resource-constrained platforms has reported differences among multiple Visual SLAM pipelines in terms of accuracy, real-time performance, and resource consumption under embedded deployment, providing practical benchmarking evidence for engineering adoption [10].

Research on SLAM in dynamic environments broadly follows two lines of investigation. Geometry-driven dynamic rejection methods typically detect and exclude moving features using RANSAC [11], optical flow estimation [12], or epipolar-geometry consistency [13]. Kundu *et al.* achieved motion–static separation via multiview geometry and motion boundaries [14]. Lin *et al.* estimated the dynamic-state of points by fusing multiview information through probabilistic filtering [15]. Long *et al.* proposed DGFlow-SLAM, which suppresses dynamic interference by scene-flow grid segmentation without relying on semantic priors [16]. Deep-learning-based semantic-prior methods construct dynamic masks using semantic or instance segmentation. DynaSLAM employs Mask R-CNN to detect potentially dynamic objects and removes the corresponding observations [17], [18]. DS-SLAM combines semantic segmentation with geometric constraints to identify dynamic regions [19], [20]. Blitz-SLAM fuses semantic masks with depth information to generate dynamic occlusion masks for improved robustness [21]. The real-time instance segmentation network YOLACT supports higher frame-rate segmentation and reduces the latency introduced by the semantic module [22]. The fusion of semantics and geometry has gradually become the dominant paradigm. SG-SLAM jointly exploits semantic cues and geometric consistency to suppress dynamic points and improve trajectory accuracy in dynamic scenes [23]. Wang *et al.* further proposed an RGB-D SLAM method that removes dynamic feature points using combined semantic information and geometric constraints, and demonstrated improved trajectory accuracy on dynamic indoor data [24]. YOLO-SLAM integrates instance segmentation results with motion consistency to filter dynamic regions [25]. Detect-SLAM incorporates object detection into front-end association and map-state updates [26]. Ji *et al.* discussed a real-time semantic RGB-D SLAM implementation framework and engineering strategies for dynamic environments [27]. DP-SLAM provides continuous discrimination for dynamic points via motion-probability modeling, thereby enhancing robustness [28]. ARD-SLAM improves localization accuracy in dynamic scenes by integrating dynamic-object recognition with multiview geometric constraints [29]. SSF-SLAM combines semantic information with scene-flow geometry to handle complex dynamic environments [30]. At the level of system coupling, DynaSLAM II tightly integrates multi-object tracking with

SLAM and jointly optimizes the states of the camera and dynamic objects [31]. ReFusion models residuals and introduces tailored strategies for dynamic RGB-D reconstruction [32]. PR-SLAM improves real-time performance by running semantic segmentation in a parallel thread [33]. DC-SLAM designs category-specific suppression mechanisms for different types of dynamic features [34]. Under emerging representation-learning paradigms, Orbeez-SLAM introduces implicit representations to enhance map expressiveness [35], while SplaTAM leverages 3D Gaussian splatting for dense mapping with integrated tracking [36].

In indoor environments where humans constitute the dominant source of dynamics, existing methods still face challenges that are more specific and harder to address. Instance-mask-driven "all-or-nothing" rejection strategies (e.g., DynaSLAM [17] and DS-SLAM [19]) treat a person as a single dynamic object. However, due to non-rigid articulation and local motions, truly moving parts and relatively static parts often coexist within the same human instance. Consequently, instance-level one-shot removal mistakenly discards stable observations, resulting in insufficient geometric constraints; tracking degradation is particularly likely in low-dynamic scenarios with subtle pose changes. Moreover, instantaneous dynamic decisions based on pixel-level residuals or optical flow are sensitive to camera egomotion, depth noise, and registration errors. Without structured aggregation and adaptive thresholds, they may exhibit temporal jitter and allow residual dynamics to persistently contaminate pose estimation [37]. In most systems, human-related information is used only for front-end rejection; human structural and motion priors are rarely incorporated into back-end optimization, which leads to limited observability and accumulated drift when static constraints are scarce. Zhao *et al.* proposed Depth-aided RGBD-inertial odometry, introducing depth-reliability intervals and a sliding-window optimization mechanism in indoor scenarios to improve localization stability, highlighting the practical value of integrating multi-source constraints into the optimization framework [38]. The general graph-optimization framework g2o provides a standardized solver foundation for multi-constraint back ends [39]. Evaluation of dynamic SLAM commonly relies on public datasets for reproducible comparison, and the TUM RGB-D benchmark offers widely used protocols and data support for indoor RGB-D trajectory evaluation [40].

To address these issues, we propose a SLAM method tailored to typical indoor human-dynamic scenarios by combining part-level dynamic suppression with human structural constraints. In the front end, we use human instance segmentation together with a pose-skeleton prior to generate part-level masks, and conduct part-level robust statistical analysis on the residual optical flow after camera-motion compensation. This enables reliable identification of genuinely moving parts to suppress dynamic interference, while retaining valid observations from relatively static parts. In the back end, structural cues—such as the human–camera relative pose, keypoint reprojection, temporal motion priors, and bone-length consistency—are incorporated into a unified graph-optimization framework and jointly optimized with RGB-D geometric constraints, thereby improving the accuracy and robustness of pose estimation in dynamic scenes.

The main contributions of this work are as follows:

(1) We propose a part-level human mask generation method based on the YOLOv8 skeletal topology, and adopt a capsule-distance competition mechanism to achieve stable body-part segmentation, providing spatial support for fine-grained dynamic reasoning;

(2) We design a dense-optical-flow-driven part-level motion determination strategy that combines camera-motion compensation with robust threshold statistics, effectively distinguishing truly moving parts from locally spurious motion observations;

(3) We develop a joint optimization framework that fuses RGB-D geometric constraints with human structural constraints, where multi-dimensional constraints—including human-pose vertices, keypoint reprojection, and bone-length consistency—are leveraged in the back end to improve pose estimation accuracy, and we validate the stability and robustness of the proposed method in real indoor-scene experiments.

2. BPC-SLAM System

This paper targets RGB-D SLAM under dynamic human interference and proposes a localization framework that couples part-level dynamic suppression with joint optimization under human structural constraints. The system is built upon ORB-SLAM3 and preserves its three-thread architecture (tracking, local mapping, and loop closure), while adding an additional pipeline for human perception and motion reasoning. Human-related constraints are further integrated into the back-end local optimization stage, thereby reducing the contamination of pose estimation by inconsistent observations in highly dynamic scenes, and avoiding over-rejection induced by treating an entire human instance as dynamic in low-dynamic or locally moving scenarios.

As illustrated in Figure_1, the system takes RGB images as input. We first apply YOLOv8-seg for human instance segmentation to obtain pixel-level human priors, and then use YOLO v8-Pose for pose keypoint estimation to recover skeletal structure. Based on the person mask and the geometric constraints induced by skeletal line segments, we generate six categories of part-level masks. Next, dense optical flow is computed between adjacent frames and compensated for camera motion. We then perform part-level statistics of the residual motion magnitude to determine which parts are truly dynamic, producing the final dynamic mask. This dynamic mask is used for front-end feature filtering: feature observations falling inside the dynamic mask are rejected, and only static/reliable observations are retained for geometric tracking. During local optimization, in addition to the standard RGB-D reprojection constraints, we introduce constraints on the human–camera relative pose, keypoint reprojection, temporal motion priors, and bone-length consistency to form a unified graph-optimization objective, thereby improving observability and robustness in dynamic environments.

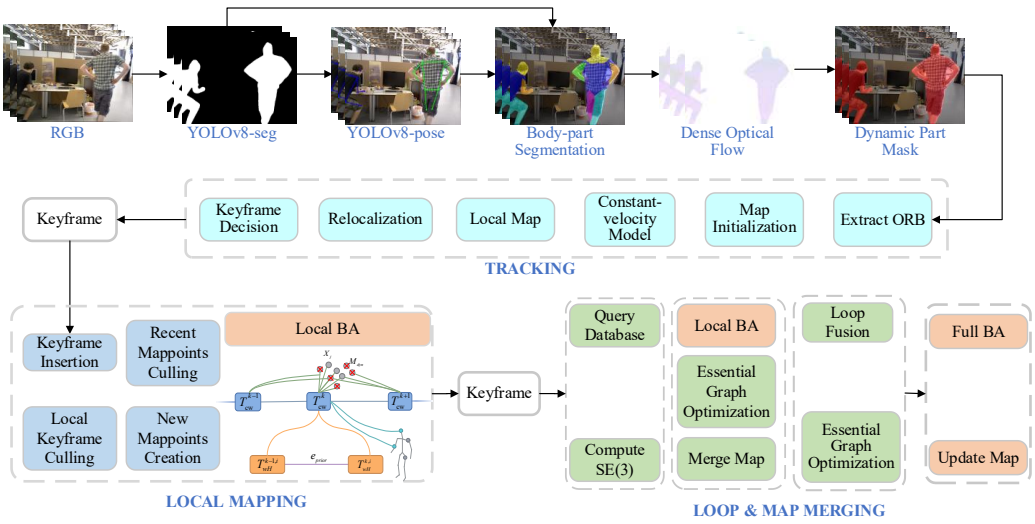


Figure_1 Schematic Illustration of the BPC-SLAM System Framework

2.1. Human Perception: Instance Segmentation and Pose Estimation

YOLOv8-seg is an instance segmentation network designed for real-time applications. It outputs, for each input image, both object detection results and the corresponding pixel-wise instance masks, thereby providing direct semantic region constraints for dynamic scenes. Owing to its real-time inference capability, the model can be seamlessly integrated into the SLAM tracking thread, supplying human instance masks to the front end without introducing significant additional latency. As shown in Figure_2, we apply YOLOv8-seg to each frame I_t to perform human instance segmentation, obtaining the human instance mask M_t^i and its associated bounding box B_t^i .

$$(B_t^i, M_t^i), \quad i = 1, \ldots, N_t \tag{1}$$

Here, $M_t^i \subset \Omega$ denotes the set of human pixels, and Ω denotes the image domain. The mask M_t^i explicitly separates the human region from the background, enabling subsequent part partitioning and motion statistics to be performed within the human pixel domain, thereby reducing interference from irrelevant background areas.

After obtaining the human pixel-domain prior, relying solely on the full-body mask is still insufficient to capture the locality of human motion. In practice, human motion is often dominated by local limb movements rather than global rigid-body motion. If the entire human region is naively treated as dynamic and all features within it are removed, over-rejection may occur and the available constraints for SLAM can be substantially reduced. To further incorporate human structural information, we employ YOLOv8-pose to estimate human keypoints and the skeletal topology. For the i-th pedestrian instance, YOLOv8-pose outputs a set of keypoints denoted as

$$K_t^i = \{(\mathbf{z}_{t,i,k}, c_{t,i,k})\}_{k=1}^{K} \tag{2}$$

Here, $\mathbf{z}_{t,i,k} \in \mathrm{R}^2$ denotes the pixel coordinates of the k-th keypoint and $c_{t,i,k} \in [0,1]$ is its confidence score. We associate the keypoints and the corresponding skeletal connections with the instance mask, thereby obtaining, for the same human instance, a joint representation consisting of the pixel-domain mask and the structure-domain skeleton. This provides the necessary input for the part-level mask generation described in the next section.



(a) (b) (c)

Figure_2 Illustration of Instance Segmentation and Pose Estimation (a) is the original RGB image, (b) is the human instance mask produced by YOLOv8-seg, and (c) shows the human

keypoints and skeletal connections estimated by YOLOv8-pose.

2.2. Skeleton-Guided Part-Level Mask Generation

Treating the entire human instance as dynamic based solely on an instance mask leads to over-suppression, particularly in low-dynamic or locally moving scenarios, where a large number of stable features may be unnecessarily removed. To explicitly model the locality of human motion, we introduce the skeletal structural constraints from YOLOv8-pose within the human instance mask produced by YOLOv8-seg. Specifically, we partition human pixels into semantic parts such as the head, torso, and limbs, and generate a set of part-level masks $M_t^{i,p}$. The key insight of this step is that the instance mask provides a pixel-domain boundary, while the keypoint skeleton offers a structural prior; combining the two enables stable and reliable part partitioning.

2.2.1. Part Set and Skeletal-Segment Representation

For the i-th pedestrian instance, we define the set of body parts as

$$\mathrm{P} = (\text{Head}、\ \text{Torso}、\ \text{LArm}、\ \text{RArm}、\ \text{LLeg}、\ \text{RLeg})$$

Following the pose-keypoint topology, we construct for each part a set of skeletal segments $\mathrm{S}_t^{i,p}$. Each segment is defined by connecting a pair of endpoint keypoints $(\mathbf{z}_a, \mathbf{z}_b)$, and we assign the segment a confidence score c_s, computed as the average confidence of the two connected keypoints. To accommodate scale variations and changes in human pose, we further introduce an adaptive radius r_p for each part and expand each segment into a capsule-shaped support region, which is used for pixel-to-part assignment.

2.2.2. Capsule Distance and Competitive Assignment from Pixels to Body Parts

Let $\mathbf{X}_j$ be an arbitrary pixel within the human instance mask M_t^i. We define the Euclidean distance from a point to a line segment as $\mathrm{dist}(\mathbf{X}_j, s)$. Then, the minimum distance from pixel $\mathbf{X}_j$ to body part P is:

$$d_p(\mathbf{X}_j) = \min_{s \in S_t^{i,p}} dist(\mathbf{X}_j, s) \tag{3}$$

To mitigate skeletal drift caused by low-confidence keypoints, we further introduce a confidence-weighted distance:

$$\tilde{d}_p(\mathbf{X}_j) = \min_{s \in S_t^{i,p}} \frac{\mathrm{dist}(\mathbf{X}_j, s)}{c_s + \varepsilon} \tag{4}$$

Here, ò is a numerical stabilization term. The distance is normalized by the part-wise scale, yielding the competitive score of pixel $\mathbf{X}_j$ with respect to part P :

$$\text{score}_p(\mathbf{X}_j) = \frac{\tilde{d}_p(\mathbf{X}_j)}{r_p} \tag{5}$$

Finally, the part label of a pixel is determined by the minimum competitive score:

$$\text{label}(\mathbf{X}_j) = \arg\min_{p \in \mathrm{P}} \text{score}_p(\mathbf{X}_j) \tag{6}$$

2.2.3. Two-Stage Assignment: Head-First Locking

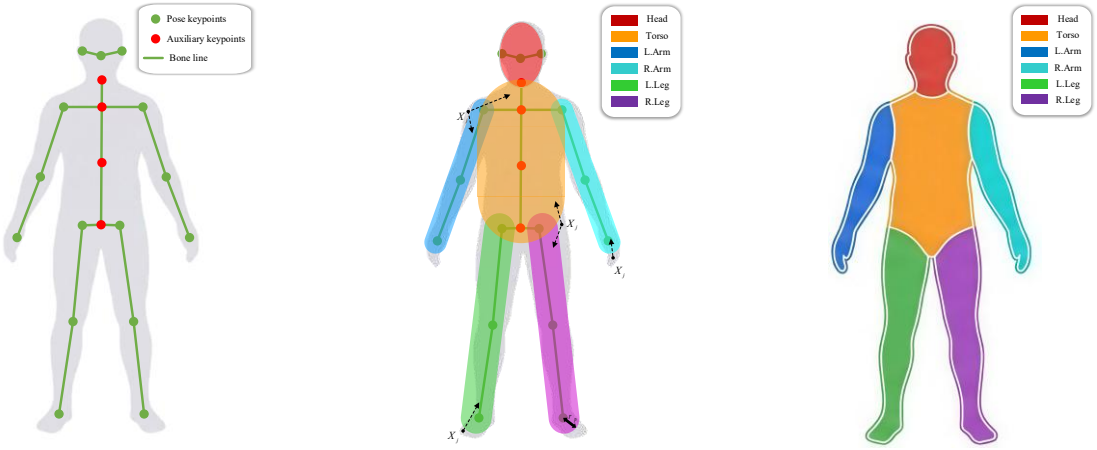
Since the head region is relatively small and its boundary can be easily confused with the torso and upper limbs, we adopt a two-stage assignment strategy to improve the stability of part partitioning. In the first stage, head pixels are prioritized and locked. Specifically, if

$$d_{\text{Head}}(\mathbf{X}_j) \le r_{\text{Head}} \tag{7}$$

we directly assign $\text{label}(\mathbf{X}_j) = \text{Head}$. In the second stage, the remaining unassigned pixels follow the competitive rule described above, completing the assignment to the torso and four limbs. Accordingly, the part-level masks are obtained as

$$M_t^{i,p} = \{\mathbf{X}_j \in M_t^i \mid \text{label}(\mathbf{X}_j) = p\}, \quad p \in \mathrm{P} \tag{8}$$

As shown in Figure_3, a structured set of part-level masks is generated within each human instance mask via keypoint-guided capsule-distance competition. These masks provide a spatial substrate for subsequent part-level motion reasoning based on dense optical flow, enabling a dynamic feature suppression strategy that suppresses only truly moving parts while preserving constraints from relatively static parts.



(a) (b) (c)

Figure_3 Illustration of Skeleton-Guided Part-Level Mask Generation. (a) Pixel-domain boundary and skeletal-structure priors. (b) Part partitioning results. (c) Visualization of the generated body parts.

2.2.4. Part-Level Motion Determination via Dense Optical Flow and Dynamic Mask Generation

To determine the true motion state of each human part, we estimate dense optical flow between adjacent frames and, after compensating for camera egomotion, perform dynamic reasoning based on motion statistics within each part mask. Let the grayscale images of two consecutive frames be I_{t-1} and I_t. Dense optical flow aims to solve for the per-pixel displacement field:

$$\mathbf{u}(\mathbf{X}_j) = (u(\mathbf{X}_j), v(\mathbf{X}_j)), \quad \mathbf{X}_j \in \Omega \tag{9}$$

Here, $u(\mathbf{X}_j), v(\mathbf{X}_j)$ denote the displacement of pixel $\mathbf{X}_j$ between the two frames. Optical-flow estimation is commonly derived from the brightness-constancy assumption and can be obtained by minimizing the following energy:

$$\min_{\mathbf{u}} \sum_{\mathbf{x}\in\Omega} | I_{t-1}(\mathbf{X}_j) - I_t(\mathbf{X}_j + \mathbf{u}(\mathbf{X}_j)) | + \lambda R(\mathbf{u}) \tag{10}$$

where $R(\mathbf{u})$ is a smoothness regularizer and λ is a weighting coefficient.

(1) Camera-Motion Compensation and Residual Optical Flow

Optical flow computed directly from consecutive frames contains both camera motion and object motion. To better approximate object-induced motion, we first estimate the global inter-frame motion in the background region and use it for alignment. Specifically, we remove all human instance masks and aggregate them into the human region M_t^{person}, and define the background mask as $M_t^{\text{bg}} = \neg M_t^{\text{person}}$. We then estimate the global affine transformation W_t over the background region and warp the previous frame to the current one:

$$\hat{G}_{t-1} = W_t(G_{t-1}) \tag{11}$$

Subsequently, we compute dense optical flow on $(\hat{G}_{t-1}, G_t)$ to obtain the residual flow field $\mathbf{u}(\mathbf{X}_j)$, and define the flow magnitude map as

$$m(\mathbf{x}) = \| \mathbf{u}(\mathbf{X}_j) \|_2 \tag{12}$$

(2) Robust Thresholding and Part-Level Dynamic Decision

Because noise levels and registration errors vary across scenes, fixed thresholds are prone to failure. We therefore perform robust statistics on the magnitude map $m(\mathbf{x})$ within the background region M_t^{bg}, and construct an adaptive dynamic threshold τ based on the median and the median absolute deviation (MAD), so as to characterize the typical range of background residual motion.

After obtaining the part masks $M_t^{i,p}$, we compute, for each part, a high-percentile statistic of the flow magnitude within its region as the motion strength:

$$s_{i,p} = \text{Percentile}_q(m(\mathbf{X}_j) | \ \mathbf{X}_j \in M_t^{i,p}) \tag{13}$$

Here, q is a fixed percentile. We choose a relatively high percentile to reduce the influence of occlusion boundaries and a small number of outlier flow vectors. A part is then classified as dynamic if its motion strength exceeds a certain ratio of the threshold τ:

$$\text{Mov}(i, p) = \text{I}(si, p > \alpha_p \cdot \tau) \tag{14}$$

Here, α_p is a part-dependent sensitivity coefficient used to accommodate differences in typical motion magnitudes across limbs.

(3) Dynamic Part Mask

Based on the part-level motion decision, we construct the dynamic mask as the union of all masks corresponding to parts classified as dynamic:

$$M_t^{\mathrm{dyn}}(\mathbf{X}_j) = \bigvee_i \bigvee_{p\in \mathrm{P}} \left(M_t^{i,p}(\mathbf{X}_j) \wedge \mathrm{Mov}(i,p) \right) \tag{15}$$

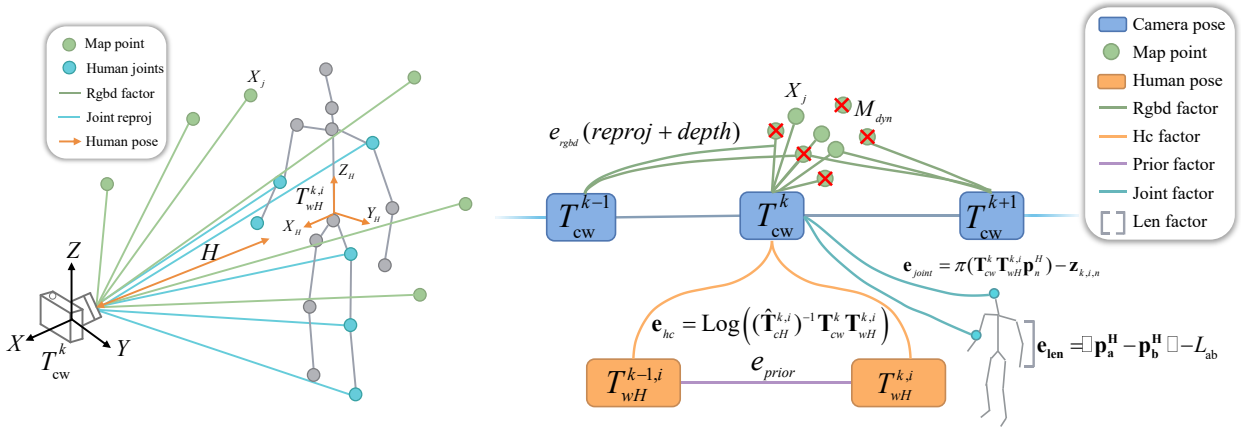
This mask covers only those body-part regions that are identified as truly moving. It is used in the SLAM front end to reject features affected by dynamic interference, thereby substantially reducing inconsistent constraints in highly dynamic scenes. Meanwhile, in low-dynamic scenarios, it avoids over-removal of the entire human region and preserves more stable geometric constraints.

2.2.5. Graph-Optimization Back End with Human Structural Constraints

In RGB-D SLAM, camera pose estimation relies on the assumption that the static scene satisfies multi-view geometric consistency. However, dynamic humans introduce observations in the front end that are inconsistent with camera egomotion, leading to incorrect reprojection constraints and subsequent pose drift. Based on the proposed part-level dynamic mask, we first reject feature observations within dynamic regions in the front end. We then incorporate the remaining static observations together with human structural/motion constraints into a unified graph-optimization framework in the back end, thereby jointly improving the accuracy and robustness of pose estimation.

(1) Optimization Variables

Assume that the camera pose of the k-th keyframe is $\mathrm{T}_{cw}^{k} \in SE(3)$, and the map points are $\mathbf{X}_j \in \mathbb{R}^3$. Meanwhile, to model the temporal consistency and structural constraints of the human body, we introduce a human pose vertex $\mathbf{T}_{wH}^{k,i} \in SE(3)$ for the i-th pedestrian in the k-th keyframe, where HHH denotes the transformation from the human coordinate frame to the world coordinate frame. As shown in Figure_4(a), we provide a geometric illustration of the joint optimization in BPC-SLAM.



(a) (b)

Figure_4 (a) Geometric Illustration of the Joint Optimization in BPC-SLAM. (b) Illustration of Joint Graph Optimization Integrating RGB-D Geometric Constraints and Human Structural Constraints

(2) RGB-D Geometric Residuals (Static Feature Constraints)

For all static feature observations retained after dynamic-mask filtering, we formulate the RGB-D reprojection error as

$$\mathbf{e}_{rgbd} = \begin{bmatrix} \pi(\mathbf{T}_{cw}^{k}\mathbf{X}_{j}) - \mathbf{z}_{k,j} \\ d(\mathbf{T}_{cw}^{k}\mathbf{X}_{j}) - \hat{d}_{k,j} \end{bmatrix} \tag{16}$$

Here, $\mathbf{z}_{k,j}$ denotes the image observation and $\hat{d}_{k,j}$ denotes the depth measurement.

This term enforces static geometric consistency between the pose and the map point, and by using only observations from static regions it avoids contamination from dynamics.

(3) Human-Related Constraint Terms

To further exploit human observations to improve the system's observability and stability in dynamic environments, we introduce the following human-related constraints in the back end.

1）Human–Camera Relative Pose Observation Constraint

For each keyframe k, we construct a relative pose observation between the camera and the human, denoted as $\hat{\mathrm{T}}_{cH}^{k,i}$. This observation can be expressed as

$$\mathbf{T}_{cH}^{k,i} = \mathbf{T}_{cw}^{k}\mathbf{T}_{wH}^{k,i} \tag{17}$$

We then formulate the residual using the Lie-algebra logarithm map:

$$\mathbf{e}_{hc} = \mathrm{Log}\left((\hat{\mathbf{T}}_{cH}^{k,i})^{-1}\mathbf{T}_{cw}^{k}\mathbf{T}_{wH}^{k,i}\right) \tag{18}$$

This term couples human observations with camera pose estimation, providing additional constraint sources in dynamic environments.

2）Keypoint Reprojection Constraint (Structural Consistency)

Given n keypoint image observations $\mathbf{z}_{k,i,n} \in \mathrm{R}^2$ obtained from human pose estimation, together with the corresponding 3D joint points $\mathbf{p}_n^h \in \mathrm{R}^3$ in the human coordinate frame, we define the reprojection error as

$$\mathbf{e}_{joint} = \pi(\mathbf{T}_{cw}^{k}\mathbf{T}_{wH}^{k,i}\mathbf{p}_n^H) - \mathbf{z}_{k,i,n} \tag{19}$$

This term enforces consistency between the estimated human pose and the camera pose in the image domain, thereby enhancing the geometric interpretability of the human constraints.

3）Human Motion Prior Constraint (Temporal Consistency)

To avoid unreasonable jumps of the human pose between adjacent keyframes, we impose a relative-motion prior on consecutive human poses. Let the measured relative motion be $\mathbf{T}_h^{k-1\to k,i}$. The residual is defined as

$$\mathbf{e}_{prior} = \mathrm{Log}\left((\Delta\hat{\mathbf{T}}_H^{k-1\to k})^{-1}(\mathbf{T}_{wH}^{k-1,i})^{-1}\mathbf{T}_{wH}^{k,i}\right) \tag{20}$$

4）Bone-Length Consistency Constraint (Shape Consistency)

Let the human skeleton contain mmm bone segments (a,b), with endpoints $(\mathbf{p}_a^H, \mathbf{p}_b^H)$ in the human coordinate frame. Given the expected bone length L_{ab}, we define the length-consistency residual as

$$e_{len} = \| \mathbf{p}_a^H - \mathbf{p}_b^H \| - L_{ab} \tag{21}$$

This term suppresses structural distortions caused by unstable keypoints, improving the

stability of human constraints during optimization.

(4) Unified Objective Function and Robust Kernel

In summary, the back-end joint graph optimization is depicted in Figure_4 (b), and the overall objective function constructed in this work is given by

$$\min \sum \rho\left(\left\| \mathbf{e}_{rgbd} \right\|^2\right) + \sum \rho\left(\left\| \mathbf{e}_{hc} \right\|^2\right) + \sum \rho\left(\left\| \mathbf{e}_{joint} \right\|^2\right) + \sum \rho\left(\left\| \mathbf{e}_{prior} \right\|^2\right) + \sum \rho\left(\left\| e_{len} \right\|^2\right) \quad (22)$$

To reduce the impact of outlier observations (e.g., mismatches and occluded keypoints), we apply a robust kernel to reweight a subset of the residual terms, thereby improving the robustness of the optimization.

3. Experiments and Results

3.1. Experimental Setup and Evaluation Metrics

We evaluate the proposed system on both standard benchmarks and real indoor scenes to verify its localization robustness and accuracy under dynamic human interference. All experiments are conducted on the same workstation: an Intel i5-12600KF CPU (16 threads), 32 GB DDR4 RAM, and an NVIDIA GeForce RTX 4060 GPU, running Ubuntu 18.04 LTS.

For quantitative evaluation, we use Absolute Trajectory Error (ATE) and Relative Pose Error (RPE). ATE measures global trajectory accuracy by computing the overall deviation between the estimated and ground-truth poses along the entire trajectory. RPE is reported in both translation and rotation components: the translational error quantifies position drift over a fixed time interval, while the rotational error evaluates angular discrepancies. We report multiple statistics, including root mean square error (RMSE), mean, median, and standard deviation (S.D.), to ensure a comprehensive and reliable evaluation.

3.2. Standard Datasets

Experiments are conducted on the TUM RGB-D benchmark dataset. It provides synchronized RGB images, depth maps, and ground-truth camera poses, and is widely used for quantitative evaluation of visual SLAM. Focusing on dynamic human interference, we select representative sequences from two typical scenarios: fr3/walking and fr3/sitting. The fr3/walking sequences are dominated by pronounced dynamic occlusions and large-amplitude human motion, whereas fr3/sitting mainly involves local hand motions of a seated person near a desk, resulting in relatively weak overall dynamics. Specifically, the evaluated sequences include five clips: fr3/sitting_static, fr3/walking_halfsphere, fr3/walking_rpy, fr3/walking_static, and fr3/walking_xyz. For brevity, we refer to them as fr3_s_static, fr3_w_half, fr3_w_rpy, fr3_w_static, and fr3_w_xyz, respectively, in the remainder of this paper.

3.3. Comparison with the ORB-SLAM3 Baseline

To quantitatively evaluate the localization performance of our method in dynamic environments, we analyze the errors of the estimated trajectories and conduct comparative experiments against the ORB-SLAM3 baseline. The improvement ratio is defined as

$$\gamma = \frac{e - f}{e} \times 100\% \tag{23}$$

where e and f denote the error values of ORB-SLAM3 and our method on the same sequence, respectively. This definition is used to compute the improvement ratio for ATE, translational RPE, and rotational RPE.

As reported in Tables 1–3, our method consistently outperforms ORB-SLAM3 in both highly dynamic scenarios (fr3_w) and moderate/low-dynamic scenarios (fr3_s). The most significant gains are observed on fr3_w_xyz: compared with ORB-SLAM3, our method reduces the RMSE of ATE, translational RPE, and rotational RPE by 97.38%, 50.45%, and 24.89%, respectively. On fr3_w_static, the ATE RMSE decreases from 0.1535 to 0.0082 (a 94.66% reduction), indicating that under strong dynamic interference induced by walking pedestrians, the proposed part-level motion determination and dynamic-feature suppression strategy can substantially mitigate the impact of the dynamic foreground on camera pose estimation. Even in low-dynamic environments, our method remains superior on most sitting sequences; for instance, on fr3_s_static it reduces the ATE RMSE from 0.0109 to 0.0058 (an average reduction of approximately 46.79%), demonstrating that the part-level strategy avoids treating the entire human region as dynamic and thus preserves more stable constraints in locally moving scenarios.

Table1 Comparison of the Absolute Trajectory Error (ATE) (M).

sequence	**ORB-SLAM3**				**BPC-SLAM**				**Improvements**			
	RMSE	**Mean**	**Median**	**S.D**	**RMSE**	**Mean**	**Median**	**S.D**	**RMSE**	**Mean**	**Median**	**S.D**
fr3_s_static	0.0109	0.0098	0.0094	0.0047	0.0058	0.0050	0.0045	0.0029	46.79%	48.98%	52.13%	38.30%
fr3_w_half	0.3576	0.3413	0.3148	0.1067	0.0239	0.0121	0.0202	0.0144	95.36%	93.23%	93.58%	86.50%
fr3_w_rpy	0.7855	0.6542	0.6553	0.4348	0.0336	0.0278	0.0227	0.0189	95.72%	95.75%	96.54%	95.65%
fr3_w_static	0.1535	0.14029	0.1265	0.0623	0.0082	0.0068	0.0059	0.0046	94.66%	94.58%	95.26%	85.07%
fr3_w_xyz	0.5389	0.4039	0.2619	0.3568	0.0141	0.0123	0.0112	0.0068	97.38%	96.95%	95.73%	98.10%

Table2 Comparison of the metric translational drift (RPE) (M/S).

sequence	**ORB-SLAM3**				**BPC-SLAM**				**Improvements**			
	RMSE	**Mean**	**Median**	**S.D.**	**RMSE**	**Mean**	**Median**	**S.D.**	**RMSE**	**Mean**	**Median**	**S.D.**
fr3_s_static	0.0055	0.0047	0.0040	0.0029	0.0045	0.0039	0.0034	0.0023	18.67%	17.66%	14.24%	21.41%
fr3_w_half	0.0213	0.0153	0.0115	0.0149	0.0149	0.0112	0.0092	0.0098	30.16%	26.70%	20.44%	34.06%
fr3_w_rpy	0.0275	0.0214	0.0167	0.0173	0.0193	0.0150	0.0114	0.0123	29.64%	29.96%	31.64%	29.14%
fr3_w_static	0.0164	0.0101	0.0069	0.0129	0.0062	0.0049	0.0039	0.0039	49.83%	48.86%	43.25%	50.44%
fr3_w_xyz	0.0218	0.0166	0.0120	0.0142	0.0108	0.0090	0.0076	0.0060	50.45%	45.63%	37.06%	57.84%

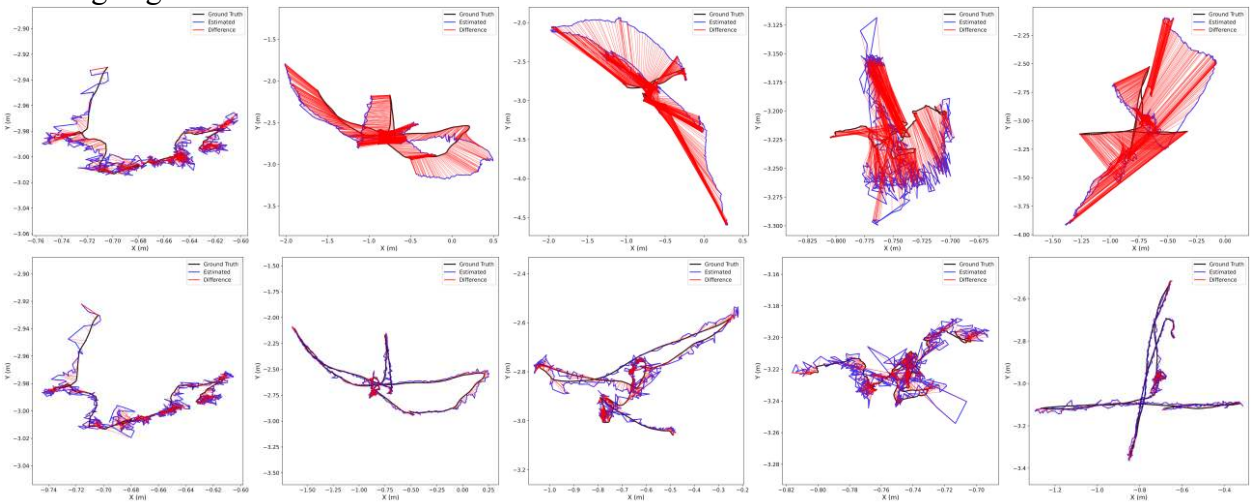
Table3 Comparison of the metric rotational drift (RPE) (◦/S).

	ORB-SLAM3				**BPC-SLAM**				**Improvements**			
	RMSE	Mean	Median	S.D.	RMSE	Mean	Median	S.D.	RMSE	Mean	Median	S.D.
fr3_s_static	0.1672	0.1423	0.1199	0.0879	0.1491	0.1252	0.1042	0.0810	10.85%	12.01%	13.14%	7.87%
fr3_w_half	0.5233	0.4111	0.3406	0.3238	0.4133	0.3468	0.2975	0.2247	21.02%	15.62%	12.65%	30.60%
fr3_w_rpy	0.6362	0.5191	0.4401	0.3679	0.4743	0.3778	0.3035	0.2868	25.45%	27.22%	31.06%	22.04%
fr3_w_static	0.3282	0.2374	0.1897	0.2266	0.2042	0.1529	0.1244	0.1353	37.79%	35.58%	34.45%	40.31%
fr3_w_xyz	0.5109	0.3872	0.3085	0.3333	0.3837	0.2719	0.2235	0.2708	24.89%	29.80%	27.56%	18.74%

As shown in Figure_6, compared with ORB-SLAM3, our method yields lower and

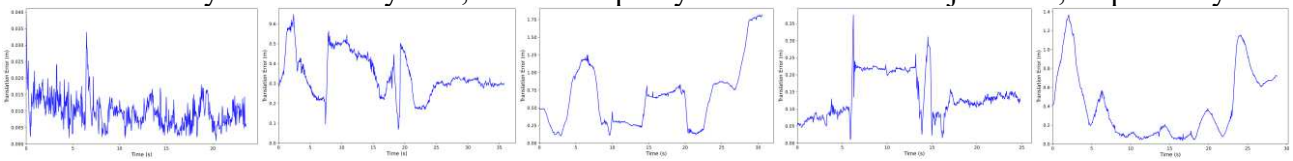
smoother translational RPE curves across all test sequences. In particular, on the highly dynamic *walking* sequences, ORB-SLAM3 exhibits pronounced error peaks and severe fluctuations during periods of human occlusion and rapid motion. In contrast, our method significantly reduces the peak magnitude and narrows the fluctuation range, indicating that the proposed part-level dynamic suppression effectively removes inconsistent constraints introduced by dynamic interference, thereby alleviating drift accumulation within short time windows. On the low-dynamic sequence fr3/sitting_static, our method likewise maintains a more stable error profile, suggesting that the strategy of "suppressing only truly moving parts while retaining observations from relatively static parts" avoids the over-rejection and insufficient geometric constraints caused by whole-body masking.

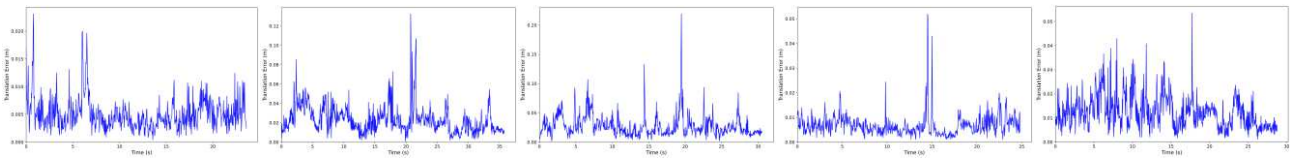
Furthermore, Figure_5 provides a geometric comparison between the estimated and ground-truth trajectories. On the *walking* sequences, the ORB-SLAM3 trajectory shows noticeable offsets and deformations, and the error correspondences diverge substantially, whereas our estimated trajectory aligns more closely with the ground truth, with overall shorter error links. This demonstrates that erroneous matches and dynamic-constraint contamination induced by human motion are effectively suppressed, improving global trajectory consistency. Figure_7 further visualizes the system workflow by presenting the outputs of each module. In low-dynamic scenarios such as sitting_static, our method also reduces local trajectory deviations, further confirming that the part-level strategy preserves stable observations and mitigates tracking degradation.



(a) (b) (c) (d) (e)

Figure_5 ATE results on five sequences for ORB-SLAM3 (first row) and the proposed method (second row): (a) fr3_s_staic, (b) fr3_w_half, (c) fr3_w_rpy, (d) fr3_w_staic, and (e) fr3_w_xyz. The black, blue, and red curves denote the ground-truth trajectory, the camera trajectory estimated by the SLAM system, and the disparity between the two trajectories, respectively.





(a) (b) (c) (d) (e)

Figure_6 RTE results on five sequences for ORB-SLAM3 (first row) and the proposed method (second row): (a) fr3_s_staic, (b) fr3_w_half, (c) fr3_w_rpy, (d) fr3_w_staic, and (e) fr3_w_xyz



(a) (b) (c) (d) (e) (f)

Figure_7 Visualization of the System Pipeline on Standard Benchmark Datasets. The first row corresponds to fr3_w_half, the second row to fr3_w_rpy, and the third row to fr3_w_xyz. From left to right, the columns show: (a) the RGB image; (b) the human instance segmentation mask; (c) keypoints and skeleton; (d) the part-level segmentation mask; (e) dense optical flow; and (f) the dynamic mask. The visualizations clearly highlight the distinct contribution of each module to the overall system.

3.4. Comparison with Representative Dynamic SLAM Methods

To further validate the effectiveness of the proposed method in human-dynamic scenarios, we compare it with several representative approaches, including ORB-SLAM3, DS-SLAM, DynaSLAM, YOLO-SLAM, SSF-SLAM, and Orbeez-SLAM. The quantitative results are reported in Table 4. All methods are evaluated on the same sequences under identical settings, and we consistently report ATE-RMSE and standard deviation (S.D.) to characterize both the mean localization error and its stability.

As can be seen from Table 4, our method achieves overall lower errors and better stability on the evaluated sequences, with particularly pronounced advantages on the highly dynamic walking sequences. Specifically, it obtains an ATE-RMSE of 0.0141 on fr3_w_xyz and 0.0239 on fr3_w_half, both clearly outperforming the ORB-SLAM3 baseline and reaching the best or near-best performance among the compared methods. On sequences such as fr3_w_rpy and fr3_w_static, our method is on par with DS-SLAM, DynaSLAM, YOLO-SLAM, and SSF-SLAM in terms of error magnitude, while exhibiting smaller fluctuations (S.D.), indicating strong robustness. Moreover, on the low-dynamic sequence fr3_s_static, our method still

achieves a low error (ATE-RMSE of 0.0058), outperforming ORB-SLAM3.

These improvements stem from two main factors. First, unlike strategies that treat an entire human instance as dynamic and remove it wholesale, our part-level motion reasoning suppresses only genuinely moving parts while preserving stable observations from relatively static body regions to maintain geometric constraints. Second, the back end incorporates human structural constraints and temporal consistency priors, further improving observability and estimation stability under dynamic interference. Overall, the proposed method significantly enhances localization accuracy in highly dynamic scenes and avoids over-suppression in low-dynamic or locally moving scenarios, yielding more stable overall performance.

Table4 Comparison with the state-of-the-art dynamic SLAM system on ATE (M/S).

sequence	ORB-SLAM3		Ds-slam		Dyna-slam		Yolo-slam		Ssf-slam		Orbeez-slam		BPC-SLAM	
	RMSE	S.D.	RMSE	S.D.	RMSE	S.D.	RMSE	S.D.	RMSE	S.D.	RMSE	S.D.	RMSE	S.D.
fr3_s_static	0.0109	0.0047	0.0065	0.0033	0.0064	0.0044	0.0066	0.0033	0.0060	0.0029	0.5245	0.2567	0.0058	0.0029
fr3_w_half	0.3576	0.1067	0.4442	0.2350	0.0417	0.0276	0.2164	0.1001	0.0324	0.0187	0.4733	0.2134	0.0239	0.0121
fr3_w_rpy	0.7855	0.4348	0.0303	0.0159	0.0302	0.0156	0.0283	0.0138	0.0268	0.0134	0.2966	0.1897	0.0336	0.0189
fr3_w_static	0.1535	0.0623	0.0081	0.0036	0.0072	0.0034	0.0073	0.0035	0.0070	0.0038	0.1979	0.0926	0.0082	0.0046
fr3_w_xyz	0.5389	0.3568	0.0247	0.0161	0.0156	0.0079	0.0146	0.0070	0.0148	0.0067	0.5245	0.2567	0.0141	0.0068

3.5. Ablation Studies

To quantitatively validate the contribution of each module in BPC-SLAM, we conduct a set of ablation experiments while keeping the ORB-SLAM3 (RGB-D) backbone pipeline and all other parameter settings unchanged. BPC-SLAM suppresses dynamic observations in the front end through an "instance segmentation–(optional) part partitioning–dense optical-flow-based motion determination" pipeline, and introduces human structural and temporal-consistency constraints in the back end for joint graph optimization. Considering the different functionalities of these components, we design the following variants: ORB-SLAM3 as the baseline; BPC-SLAM (S+F), which generates a dynamic mask using instance segmentation and dense optical flow and removes features within dynamic regions; BPC-SLAM (S+P+F), which further introduces part segmentation to refine instance-level dynamic rejection into part-level dynamic suppression; BPC-SLAM (S+F+C), which adopts the same front-end dynamic rejection as (S+F) but only adds human structural and temporal constraints in the back end; and the full model BPC-SLAM, which enables both part-level dynamic suppression and joint optimization with human constraints.

Table 5 reports the ATE-RMSE and standard deviation (S.D.) over five sequences. The baseline ORB-SLAM3 exhibits notably large errors on the highly dynamic walking sequences (e.g., an RMSE of 0.7855 on fr3_w_rpy and 0.5389 on fr3_w_xyz), mainly because dynamic humans introduce numerous inconsistent feature correspondences and contaminate the optimization with dynamic constraints, leading to accumulated pose drift. After introducing BPC-SLAM (S+F), the system can remove salient dynamic regions based on instance segmentation and optical flow, yielding a substantial reduction in ATE-RMSE across all walking sequences (e.g., from 0.5389 to 0.0152 on fr3_w_xyz, and from 0.3576 to 0.0273 on fr3_w_half), confirming the effectiveness of instance-level dynamic suppression under strong dynamics.

Furthermore, BPC-SLAM (S+P+F), which incorporates part segmentation, shows a clearer

advantage in low-dynamic or locally moving scenarios. On fr3_s_static, the RMSE is further reduced from 0.0075 to 0.0059, indicating that the part-level strategy avoids over-removal of the entire human region and preserves stable observations from relatively static parts, thereby mitigating tracking degradation caused by insufficient geometric constraints. This strategy remains beneficial under high dynamics as well (e.g., reducing fr3_w_rpy from 0.0630 to 0.0369). In addition, introducing back-end human optimization in BPC-SLAM (S+F+C) further decreases error fluctuations and improves stability on multiple sequences (e.g., reducing the S.D. on fr3_w_static from 0.0106 to 0.0088), suggesting that human structural and temporal-consistency constraints enhance observability and estimation stability under dynamic interference. Finally, the full BPC-SLAM combines the complementary benefits of part-level dynamic suppression and joint optimization with human constraints, achieving the lowest or near-lowest errors across all five sequences, and attaining the best results on fr3_s_static, fr3_w_half, and fr3_w_xyz (RMSEs of 0.0058, 0.0239, and 0.0141, respectively), demonstrating improved accuracy and robustness.

Table5 Ablation experiment ATE results (M).

sequence	**ORB-SLAM3**		**BPC-SLAM(S+F)**		**BPC-SLAM(S+P+F)**		**BPC-SLAM(S+F+C)**		**BPC-SLAM**	
	Rmse	**S.D.**	**Rmse**	**S.D.**	**Rmse**	**S.D.**	**Rmse**	**S.D.**	**Rmse**	**S.D.**
fr3_s_static	0.0109	0.0047	0.0075	0.0035	0.0059	0.0030	0.0063	0.0032	0.0058	0.0029
fr3_w_half	0.3576	0.1067	0.0273	0.0144	0.0242	0.0118	0.0263	0.0132	0.0239	0.0121
fr3_w_rpy	0.7855	0.4348	0.0630	0.0502	0.0369	0.0242	0.0591	0.0458	0.0336	0.0189
fr3_w_static	0.1535	0.0623	0.0136	0.0106	0.0093	0.0062	0.0115	0.0088	0.0082	0.0046
fr3_w_xyz	0.5389	0.3568	0.0152	0.0074	0.0144	0.0070	0.0149	0.0075	0.0141	0.0068
Average	0.3693	0.1931	0.0253	0.0172	0.0181	0.0104	0.0236	0.0157	0.0171	0.0091
Improvement			93.14%	91.08%	95.09%	94.59%	93.60%	91.87%	95.36%	95.31%

3.6. Runtime Analysis

In terms of computational efficiency, we profile the additional human-perception modules introduced in this work. Specifically, we measure the average runtime of the human instance segmentation module (YOLOv8-seg), the human pose keypoint detection module (YOLOv8-pose), and the mask-and-skeleton-constrained body-part segmentation module (Body-part segmentation). Timing is performed by recording timestamps at the entry and exit of each module; the runtimes are accumulated over an entire sequence and then averaged. As reported in Table 6, the average runtimes of YOLOv8-seg, YOLOv8-pose, and Body-part segmentation are 19 ms, 9 ms, and 15 ms, respectively, resulting in an overall overhead of approximately 43 ms per frame for the added modules. Note that this section profiles only the human-perception components newly introduced relative to the baseline; the remaining SLAM backbone follows the original ORB-SLAM3 implementation and is therefore not detailed here.

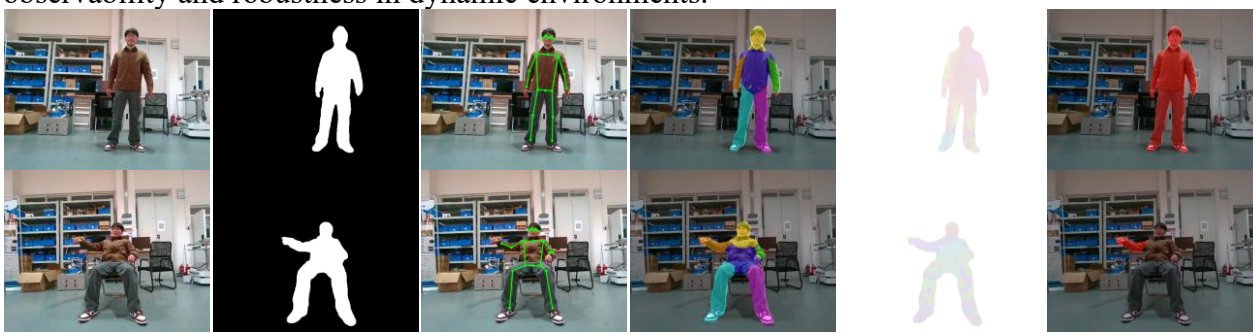
Table6 Runtime of the Added Modules(ms)

	YOLOv8-seg	YOLOv8-pose	Body part seg
Time (ms)	19	9	15

3.7. Real Indoor-Scene Experiments

To further validate the applicability and robustness of the proposed method in real-world settings, we collect an indoor RGB-D sequence using an Intel RealSense D455 and conduct comparative experiments. The captured scene includes dynamic interference such as walking pedestrians and local motions from a seated person, along with moderate viewpoint changes and occlusions, which can easily corrupt front-end tracking through dynamic feature contamination. After data acquisition, we replay the same sequence using ORB-SLAM3, DynaSLAM, and our method under identical settings, except for the modules related to dynamic suppression and human constraints, to ensure a fair comparison.

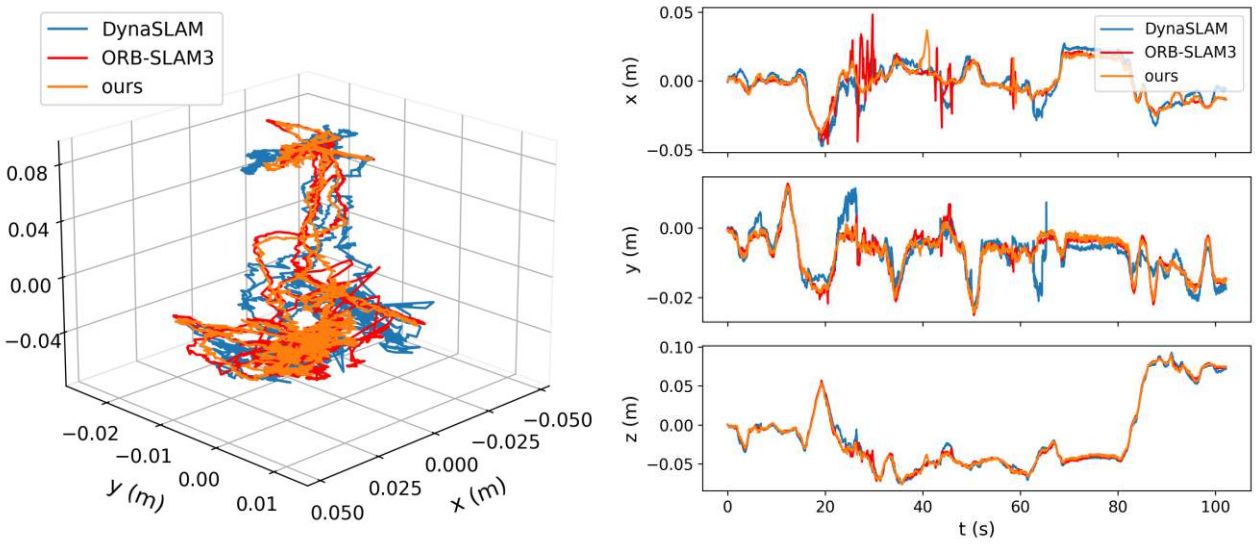
The overall processing pipeline of our method follows the proposed framework. Given an input RGB image, we first perform human instance segmentation and pose keypoint estimation, based on which part-level masks are generated. We then incorporate motion cues to obtain a dynamic mask, and reject feature observations within dynamic regions in the front end, retaining only static/reliable observations for geometric tracking. In the back end, during local optimization, we further introduce structural constraints such as human pose variables and keypoint reprojection to form a unified graph-optimization objective, thereby improving observability and robustness in dynamic environments.



(a) (b) (c) (d) (e) (f)

Figure_8 Qualitative visualization of the real indoor-scene experiments, showing the outputs of each module. the columns show: (a) The RGB image; (b) The human instance segmentation mask; (c) keypoints and skeleton; (d) The part-level segmentation mask; (e) dense optical flow; and (f) the dynamic mask. The visualizations clearly highlight the distinct contribution of each module to the overall system.

To qualitatively demonstrate the behavior of each module in real-world scenes, we visualize the online outputs during system execution, as shown in Figure_8. The proposed part-level motion reasoning can effectively distinguish *truly moving parts* from *relatively static parts*, thereby avoiding the "one-shot" over-removal of the entire human region and providing more stable static constraints for the front end. Furthermore, Figure_9 compares the trajectories produced by three methods on the same real sequence. Both the 3D trajectory plots and the component-wise error curves indicate that, during periods of strong dynamic interference, ORB-SLAM3 exhibits more pronounced trajectory fluctuations. While DynaSLAM mitigates dynamic effects to some extent, it still suffers from local instability. In contrast, our method yields a more continuous trajectory with reduced jitter, demonstrating superior overall stability and robustness.



(a) (b)

Figure_9 Analysis of the real indoor-scene experiment: (a) 3D trajectory plot; (b) component-wise error curves.

4. Conclusion

This paper addresses two key challenges of RGB-D SLAM in human-dynamic environments: (i) pose estimation being easily corrupted by dynamic-feature contamination, and (ii) stable observations being overly suppressed when semantic instance-level masking removes an entire person as a dynamic object. We propose BPC-SLAM, a part-level dynamic suppression and structure-constrained joint optimization approach tailored to dynamic human scenes. In the front end, we leverage human instance segmentation and pose-skeleton priors to generate part-level masks, and perform dense-optical-flow-based, part-level statistical motion reasoning to achieve the principle of "suppressing only truly moving parts while retaining observations from relatively static parts." This reduces dynamic-constraint contamination under strong dynamics and alleviates tracking degradation caused by insufficient geometric constraints in low-dynamic or locally moving scenarios. In the back end, we incorporate human structural and temporal-consistency constraints into a unified graph-optimization framework, further improving estimation consistency and robustness under dynamic interference.

Experiments on the TUM RGB-D benchmark and real indoor scenes show that the proposed method achieves superior localization accuracy and stability on both highly dynamic *walking* sequences and low-dynamic *sitting* sequences. Ablation studies further confirm the complementary contributions of part-level dynamic suppression and the human-aware back-end optimization.

Funding

This work is partially supported by The National Natural Science Foundation of China (Grant NO. 52565057), Yunnan Province Basic Research Program Project (grant NO. 202301AU070059)", Yunnan Provincial Major Science and Technology Special (grant NO. 202502AC080001) and Yunnan Province's "Caiyun Postdoctoral" Innovation Project

Author Contributions Statement

Wang Yang: Writing-review&editing, Writing-original draft, Visualization, Validation, Software, Methodology, Conceptualization, Data curation, Formal analysis. **Jiupeng Chen**: Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Data curation. **Hongjun San**: Resources, Project administration, Investigation, Funding acquisition. Fan Zhang: Validation, Methodology, Investigation, Funding acquisition. **Wenyu Xu**: Validation.

Declarations

We confirm that this manuscript has not been published and is not under consideration elsewhere. All authors have read and approved the final manuscript. The authors declare no competing interests.

References

[1] A.J. Davison, I.D. Reid, N.D. Molton, et al., MonoSLAM: Real-time single camera SLAM, IEEE Trans. Pattern Anal. Mach. Intell. 29 (6) (2007) 1052–1067. https://doi.org/10.1109/TPAMI.2007.1049.

[2] G. Klein, D. Murray, Parallel tracking and mapping for small AR workspaces, In: Proc. 2007 6th IEEE and ACM Int. Symp. Mixed and Augmented Reality (ISMAR), IEEE, 2007, pp. 225–234. https://doi.org/10.1109/ISMAR.2007.4538852.

[3] T. Qin, P. Li, S. Shen, VINS-Mono: A robust and versatile monocular visual-inertial state estimator, IEEE Trans. Robot. 34 (4) (2018) 1004–1020. https://doi.org/10.1109/TRO.2018.2853729.

[4] R. Mur-Artal, J.D. Tardós, ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras, IEEE Trans. Robot. 33 (5) (2017) 1255–1262. https://doi.org/10.1109/TRO.2017.2705103.

[5] C. Campos, R. Elvira, J.J.G. Rodríguez, et al., ORB-SLAM3: An accurate open-source library for visual, visual-inertial, and multimap SLAM, IEEE Trans. Robot. 37 (6) (2021) 1874–1890. https://doi.org/10.1109/TRO.2021.3075644.

[6] R. Wang, M. Schwörer, D. Cremers, Stereo DSO: Large-scale direct sparse visual odometry with stereo cameras, In: Proc. 2017 IEEE Int. Conf. Computer Vision (ICCV), IEEE, 2017, pp. 3923–3931. https://doi.org/10.1109/ICCV.2017.421.

[7] J. Engel, T. Schöps, D. Cremers, LSD-SLAM: Large-scale direct monocular SLAM, In: Proc.

ECCV, Springer, 2014, pp. 834–849. https://doi.org/10.1007/978-3-319-10605-2_54.
[8] W. Dai, Y. Zhang, P. Li, et al., RGB-D SLAM in dynamic environments using point correlations, IEEE Trans. Pattern Anal. Mach. Intell. 44 (1) (2022) 373–389. https://doi.org/10.1109/TPAMI.2020.3010942.
[9] H. Yin, S. Li, Y. Tao, et al., DYNAM-SLAM: An accurate, robust stereo visual-inertial SLAM method in dynamic environments, IEEE Trans. Robot. 39 (1) (2023) 289–308. https://doi.org/10.1109/TRO.2022.3199087.
[10] R. Giubilato, S. Chiodini, M. Pertile, S. Debei, An evaluation of ROS-compatible stereo visual SLAM methods on a nVidia Jetson TX2, Measurement 140 (2019) 161–170. https://doi.org/10.1016/j.measurement.2019.03.038.
[11] M.A. Fischler, R.C. Bolles, Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography, Commun. ACM 24 (6) (1981) 381–395. https://doi.org/10.1145/358669.358692.
[12] A. Dosovitskiy, P. Fischer, E. Ilg, et al., FlowNet: Learning optical flow with convolutional networks, In: Proc. 2015 IEEE Int. Conf. Computer Vision (ICCV), IEEE, 2015, pp. 2758–2766. https://doi.org/10.1109/ICCV.2015.316.
[13] M.R.U. Saputra, A. Markham, N. Trigoni, Visual SLAM and structure from motion in dynamic environments: A survey, ACM Comput. Surv. 51 (2) (2018) 1–36. https://doi.org/10.1145/3177853.
[14] A. Kundu, K.M. Krishna, J. Sivaswamy, Moving object detection by multi-view geometric techniques from a single camera mounted robot, In: Proc. 2009 IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS), IEEE, 2009, pp. 4306–4312. https://doi.org/10.1109/IROS.2009.5354227.
[15] K.H. Lin, C.C. Wang, Stereo-based simultaneous localization, mapping and moving object tracking, In: Proc. 2010 IEEE Int. Conf. Robotics and Automation (ICRA), IEEE, 2010, pp. 3975–3980. https://doi.org/10.1109/IROS.2010.5649653.
[16] F. Long, L. Ding, J. Li, DGFlow-SLAM: A novel dynamic environment RGB-D SLAM without prior semantic knowledge based on grid segmentation of scene flow, Biomimetics 7 (4) (2022) 163. https://doi.org/10.3390/biomimetics7040163.
[17] B. Bescos, J.M. Fácil, J. Civera, et al., DynaSLAM: Tracking, mapping, and inpainting in dynamic scenes, IEEE Robot. Autom. Lett. 3 (4) (2018) 4076–4083. https://doi.org/10.1109/LRA.2018.2860039.
[18] K. He, G. Gkioxari, P. Dollár, et al., Mask R-CNN, In: Proc. 2017 IEEE Int. Conf. Computer Vision (ICCV), IEEE, 2017, pp. 2980–2988.
[19] C. Yu, Z. Liu, X.J. Liu, et al., DS-SLAM: A semantic visual SLAM towards dynamic environments, In: Proc. 2018 IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS), IEEE, 2018, pp. 1168–1174. https://doi.org/10.1109/IROS.2018.8593691.
[20] V. Badrinarayanan, A. Kendall, R. Cipolla, SegNet: A deep convolutional encoder–decoder architecture for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (12) (2017) 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615.
[21] Y. Fan, P. Jia, Z. Liu, et al., Blitz-SLAM: A semantic SLAM in dynamic environments, Pattern Recognit. 121 (2022) 108225. https://doi.org/10.1016/j.patcog.2021.108225.
[22] D. Bolya, C. Zhou, F. Xiao, Y.J. Lee, YOLACT: Real-time instance segmentation, In: Proc. 2019 IEEE/CVF Int. Conf. Computer Vision (ICCV), IEEE/CVF, 2019, pp. 9157–9166. https://doi.org/10.1109/ICCV.2019.00925.
[23] S. Cheng, C. Sun, S. Zhang, et al., SG-SLAM: A real-time RGB-D visual SLAM toward

dynamic scenes with semantic and geometric information, IEEE Trans. Instrum. Meas. 72 (2023) 1–12 (Article No. 5004812). https://doi.org/10.1109/TIM.2022.3228006.
[24] X. Wang, S. Zheng, X. Lin, F. Zhu, Improving RGB-D SLAM accuracy in dynamic environments based on semantic and geometric constraints, Measurement 217 (2023) 113084. https://doi.org/10.1016/j.measurement.2023.113084.
[25] W. Wu, L. Guo, H. Gao, et al., YOLO-SLAM: A semantic SLAM system towards dynamic environment with geometric constraint, Neural Comput. Appl. 34 (8) (2022) 6011–6026. https://doi.org/10.1007/s00521-021-06764-3.
[26] F. Zhong, S. Wang, Z. Zhang, et al., Detect-SLAM: Making object detection and SLAM mutually beneficial, In: Proc. 2018 IEEE Winter Conf. Applications of Computer Vision (WACV), IEEE, 2018, pp. 1001–1010. https://doi.org/10.1109/WACV.2018.00115.
[27] T. Ji, C. Wang, L. Xie, Towards real-time semantic RGB-D SLAM in dynamic environments, In: Proc. 2021 IEEE Int. Conf. Robotics and Automation (ICRA), IEEE, 2021, pp. 11175–11181. https://doi.org/10.1109/ICRA48506.2021.9561743.
[28] A. Li, C. Wang, S. Quan, et al., DP-SLAM: A visual SLAM with moving probability towards dynamic environments, Inf. Sci. 556 (2021) 128–142. https://doi.org/10.1016/j.ins.2020.12.019.
[29] Q. Ul Islam, H. Ibrahim, P. Kok Chin, et al., ARD-SLAM: Accurate and robust dynamic SLAM using dynamic object identification and improved multi-view geometrical approaches, Displays 82 (2024) 102654. https://doi.org/10.1016/j.displa.2024.102654.
[30] Z. Zhang, Y. Song, B. Pang, et al., SSF-SLAM: Real-time RGB-D visual SLAM for complex dynamic environments based on semantic and scene flow geometric information, IEEE Trans. Instrum. Meas. 74 (2025) 1–12 (Article No. 5043012). https://doi.org/10.1109/TIM.2025.3556918.
[31] B. Bescos, C. Campos, J.D. Tardós, et al., DynaSLAM II: Tightly-coupled multi-object tracking and SLAM, IEEE Robot. Autom. Lett. 6 (3) (2021) 5191–5198. https://doi.org/10.1109/LRA.2021.3068640.
[32] E. Palazzolo, J. Behley, P. Lottes, et al., ReFusion: 3D reconstruction in dynamic environments for RGB-D cameras exploiting residuals, In: Proc. 2019 IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS), IEEE, 2019, pp. 7855–7862. https://doi.org/10.1109/IROS40897.2019.8967590.
[33] J. Peng, H. Zhang, Q. Yang, PR-SLAM: Parallel real-time dynamic SLAM method based on semantic segmentation, IEEE Access 12 (2024) 45678–45692. https://doi.org/10.1109/ACCESS.2024.3373308.
[34] S. Ye, B. Xu, Y. Yang, et al., DC-SLAM: Dual-category dynamic feature suppression for RGB-D VSLAM, Expert Syst. Appl. 299 (2025) 129952. https://doi.org/10.1016/j.eswa.2025.129952.
[35] C.M. Chung, Y.C. Tseng, Y.C. Hsu, et al., Orbeez-SLAM: A real-time monocular visual SLAM with ORB features and NeRF-realized mapping, arXiv (2022) arXiv:2209.13274.
[36] N. Keetha, J. Karhade, K.M. Jatavallabhula, et al., SplaTAM: Splat, track & map 3D Gaussians for dense RGB-D SLAM, In: Proc. 2024 IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), IEEE/CVF, 2024, pp. 21357–21366.
[37] S. Jiang, D. Campbell, Y. Lu, et al., Learning to estimate hidden motions with global motion aggregation, In: Proc. 2021 IEEE/CVF Int. Conf. Computer Vision (ICCV), IEEE/CVF, 2021, pp. 9752–9761. https://doi.org/10.1109/ICCV48922.2021.00963.
[38] X. Zhao, Q. Li, C. Wang, H. Dou, B. Liu, Robust depth-aided RGBD-inertial odometry for

indoor localization, Measurement 209 (2023) 112487. https://doi.org/10.1016/j.measurement.2023.112487.
[39] G. Grisetti, R. Kümmerle, C. Stachniss, et al., g2o: A general framework for graph optimization, In: Proc. 2011 IEEE Int. Conf. Robotics and Automation (ICRA), IEEE, 2011, pp. 3607–3613. https://doi.org/10.1109/ICRA.2011.5979949.
[40] J. Sturm, N. Engelhard, F. Endres, et al., A benchmark for the evaluation of RGB-D SLAM systems, In: Proc. 2012 IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS), IEEE, 2012, pp. 573–580. https://doi.org/10.1109/IROS.2012.6385773.