Research on Engineering Structures & Materials
www.jresm.org
Research Group

Research Article

Motion image stabilization of vehicle-based Fourier telemetry system based on multi-source data fusion

Li Wan [*, 1,2,3,a], Liang Xu [2,b], Wenqing Liu [2,c], Qiangling Zhang [2,d]

[1]*Science Island Branch, Graduate School of USTC, Hefei, China*
[2]*Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei, China*
[3]*Institute of Environment Hefei Comprehensive National Science Center, Hefei, China*

Article Info

Article History:

Received 20 July 2026

Accepted 12 Oct 2026

Keywords:

Multi-Source data fusion;
Vehicle-Mounted platform;
Fourier telemetry;
Motion image stabilization;
Visual-inertial fusion;
Sliding window optimization

Abstract

Fourier telemetry is essential for vehicle-mounted gas tracing and environmental monitoring. However, during dynamic motion, motion blur, sensor drift, and environmental interference severely decrease imaging stability. To deal with these challenges, a vehicle-mounted Fourier remote sensing motion image stabilization method based on multi-source data fusion was proposed. A collaborative framework is constructed to enable global geographic spatial constraints and local dynamic pose perception. This framework is integrated with high-precision GIS data, GPS, and visual-inertial observations. A sliding window-based joint pose optimization algorithm is developed to align estimated poses with ideal imaging poses in the East-North-Up (ENU) coordinate system. This algorithm can effectively suppress cumulative errors and achieve stable motion imaging. The proposed method innovatively incorporates GIS as a real-time global constraint, optimizing pose estimation and coordinate alignment in a staged manner. Also, this method integrates ORB feature matching with GIS coordinate verification to enhance robustness. From the experimental results, this system achieves a motion drift RMSE of 0.054 m, an attitude jitter RMSE of 5.976°, and an imaging instability ratio of only 2.00% in dynamic scenarios. The relationship between the 5.976° azimuth angle RMSE value and its axial components (0.03°, 0.10°, 0.11°) lies in that the azimuth angle RMSE is the combined result of the errors of the three axial components. Its value is mainly controlled by the larger errors in the pitch and roll directions, while the error in the yaw direction is relatively smaller. A high-quality stable imaging support is provided for vehicle-mounted Fourier telemetry in gas leakage detection and smart city environmental monitoring. To clarify the overall stabilization logic, Fig. 1 presents the imaging stabilization schematic diagram that illustrates the system's basic workflow and key functional modules.

© 2026 MIM Research Group. All rights reserved.

1. Introduction

With the advanced smart city infrastructure and industrial monitoring technologies, vehicle-mounted Fourier telemetry has been widely deployed in gas tracing and emergency detection. Motion image stabilization capability and dynamic imaging quality are essential to ensure the detecting performance. However, single-sensor limitations include GPS signal occlusion, visual feature degradation, and inertial measurement unit (IMU) cumulative drift. These limitation scan easily cause motion blur, attitude jitter, and imaging instability during vehicle operation, restricting the practical performance of Fourier telemetry in dynamic complex scenarios [2,4]. Multi-source data fusion can improve the stabilization performance as a mainstream strategy. A foundation is laid using early GPS-inertial navigation system fusion for mitigating short-term drift [6].

*Corresponding author: lwan@aiofm.ac.cn
[a]orcid.org/0009-0004-7146-1426; [b]orcid.org/0009-0004-8345-6579; [c]orcid.org/0009-0002-3936-9238; [d]orcid.org/0009-0004-1901-8806
DOI: http://dx.doi.org/10.17515/resm2026-1936vk0720rs

Subsequently, RGB-D fusion and tightly coupled fusion frameworks significantly enhanced detection accuracy and imaging stability [1]. Kuang, Y.J. et al. achieved multi-sensor collaborative localization through tightly coupled LIDAR/IMU/UWB fusion using elastic factor maps, demonstrating strong robustness in degraded environments [3]. However, such methods often rely on local observation constraints and lack guidance from global geographic spatial information. Liu et al. proposed a powerful LiDAR visual inertial SLAM scheme [5]. Its LVI fusion method performs well in structurally degraded environments, but is also limited by cumulative errors in local coordinate systems, making it difficult to meet the absolute pose accuracy requirements of vehicle platforms in long-distance dynamic monitoring. Nevertheless, critical bottlenecks persist, as follows.

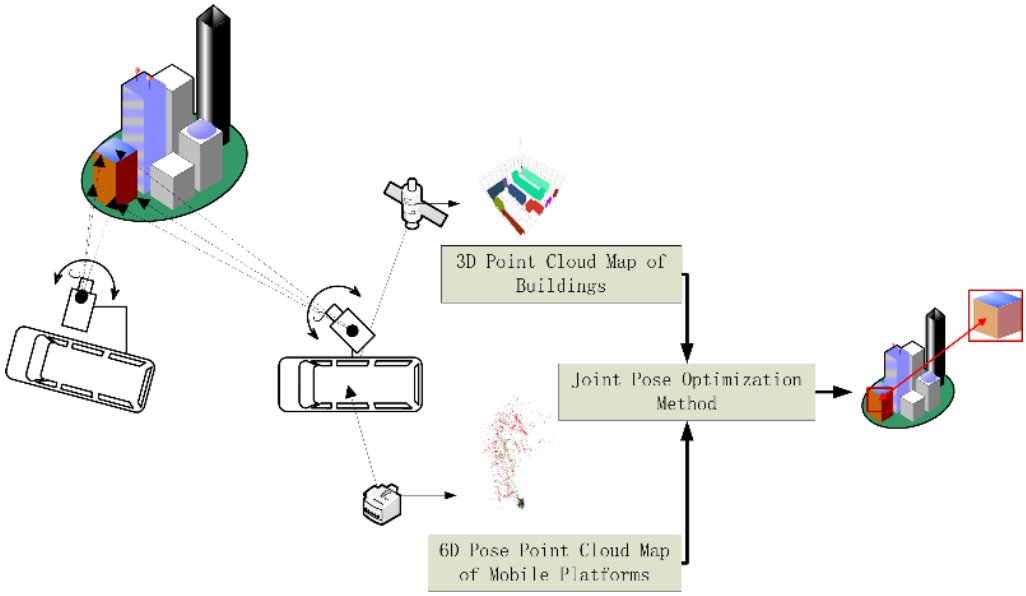


Fig. 1. Imaging stabilization schematic diagram

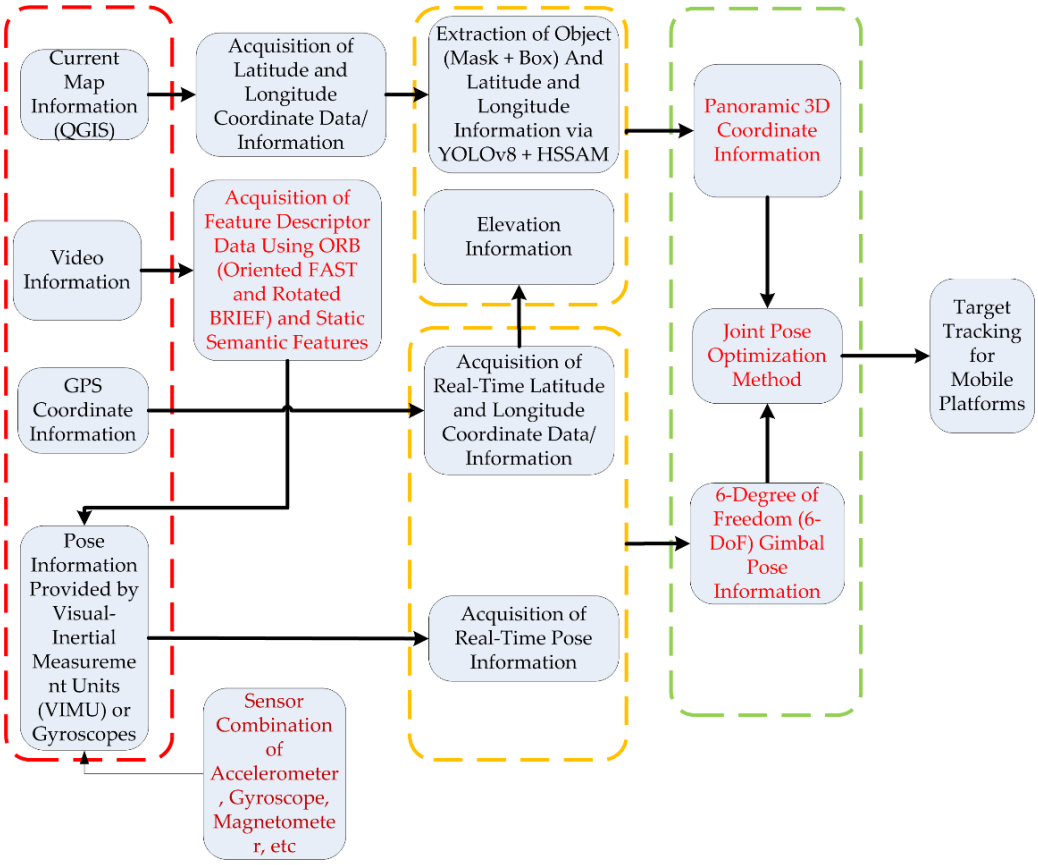


Fig. 2. Multi-Data fusion system flowchart

The stabilization accuracy and real-time performance have unresolved trade-off drawbacks, as batch optimization incurs excessive latency and lightweight filtering sacrifices precision [4]. High-precision GIS data is rarely integrated as real-time global constraints and is limited to offline matching, which can lead to cumulative motion drift and imaging jitter during long-term operation [5].Pose optimization, coordinate alignment, and stabilization are decoupled, limiting accurate ENU pose registration and causing persistent jitter in canyons [6]. Additionally, visual-inertial fusion suffers from long-term drift and poor dynamic adaptability [10–12]; GIS-enhanced fusion fails to suppress real-time drift due to offline-only usage [8,10]; semantic-enhanced fusion introduces unacceptable computational latency [9]. To address these gaps, this paper develops a GIS-driven 3D geospatial reference framework, an ORB-SLAM3-based pose estimation method with static scene semantic understanding, and a sliding window optimization framework to fuse real-time sensor data with geospatial priors for dynamic pose optimization. The proposed method effectively resolves cumulative drift, imaging jitter, and the accuracy-latency trade-off, providing an efficient, robust motion image stabilization solution for vehicle-mounted Fourier telemetry. As shown in Fig. 2, the proposed multi-source data fusion imaging stabilization system consists of data acquisition, feature extraction, pose estimation, and coordinate optimization modules.

2. Methods

2.1 Construction of Dynamic Imaging Stabilization Feature Model Based on HHSDCA-YOLOv8

2.1.1 Structure Design of the HHSDCA-YOLOv8 Self-Attention Module

High-resolution GIS data provides precise global geospatial references for vehicle-mounted dynamic imaging stabilization, characterized by concentrated motion regions and stable spatial distributions. To enhance the extraction of critical stabilization-related features, this paper proposes a Hybrid HSV-Spatial Dual-Channel Attention (HHSDCA) module integrated into YOLOv8 for adaptive stable imaging feature enhancement. As shown in Fig. 3, the 3D geographic reference model provides global spatial constraints for dynamic imaging stabilization.

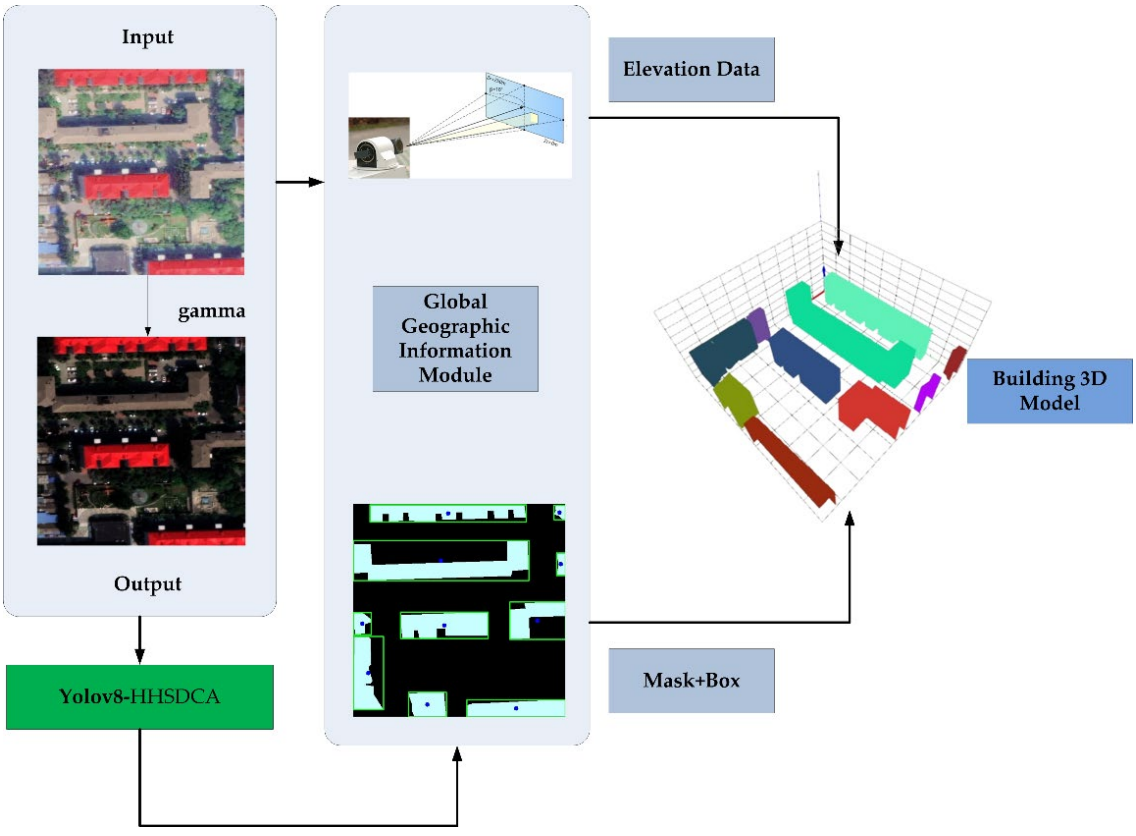


Fig. 3. 3D geographic reference modeling for imaging stabilization based on GIS data and deep learning

First, raw GIS spatial data is preprocessed to separate geographic reference information and imaging feature data. Nonlinear gamma correction is applied to HSV color channels to enhance the contrast of dynamic imaging features, and the training weights of spatial position information are enhanced to highlight the motion regions related to imaging stabilization. Then, global average pooling and global max pooling are used to extract the spatial feature information of the image, and a Multilayer Perceptron (MLP) is used to construct the channel attention weight vector. Finally, the attention weight is multiplied with the original feature map to enhance the key features required for imaging stabilization and suppress redundant interference information. Fig. 4 presents the detailed calculation flow of the HHSDCA attention module

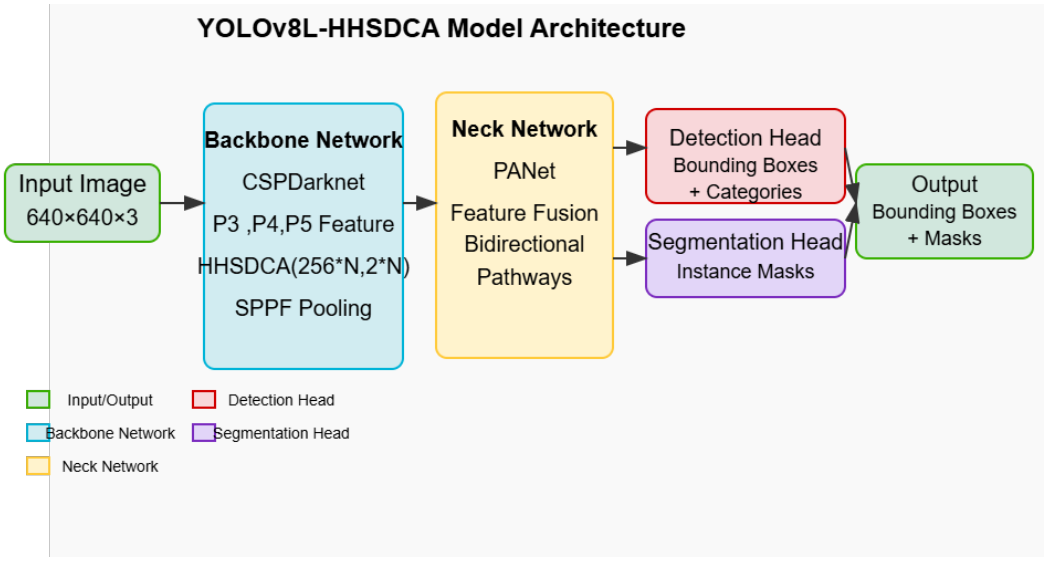


Fig. 4. YOLOv8-HHSDCA model architecture diagram

The input feature map is derived from the backbone network, which can be expressed as:

$$X \in R^{C \times H \times W}$$

where C, H, and W represent the number of channels, height, and width of the feature map, respectively.

- Color Mask Generation:

Red color mask (divided into two discontinuous intervals):

$$\begin{cases} M_{red1} = (H \in [0, 0.15]) \cap (S \geq 0.3) \\ M_{red2} = (H \in [0.9, 1]) \cap (S \geq 0.3) \\ M_{red} = (M_{red1} \cup M_{red2}) \end{cases} \quad (1)$$

- Blue Mask:

$$M_{blue} = (H \in [0.55, 0.7]) \cap (S \geq 0.3) \cap (V \geq 0.3)) \quad (2)$$

- Gray Mask:

$$M_{gray} = (S \leq 0.1) \cap (V \geq 0.3) \quad (3)$$

The red mask covers the lower and upper ranges of the color tone space, corresponding to the common warm color moving targets in dynamic imaging (such as vehicle tail lights, traffic signals), and its discontinuous range design can effectively avoid the interference of color transition areas; the blue mask locks the cyan-blue to blue-violet tone range, corresponding to the cold color background (such as the sky, water bodies, and cold-toned building surfaces), and this range is naturally separated from the warm color targets in the tone space, thereby reducing the interference of background textures on the stable feature extraction; the gray mask targets the low saturation areas, covering the road surface, shadows, and neutral-colored building surfaces. These

areas usually appear as stable spatial reference structures in dynamic imaging and can be used as reliable anchor points for motion estimation. Therefore, the study selects these three-color masks as the benchmark for feature extraction for image stabilization, which can effectively separate dynamic targets from static reference structures, thereby strengthening the contrast relationship between the motion area and the stable anchor point at the feature level.

- Channel Attention Modulation:

Convolution operations are applied to the H (Hue) and S (Saturation) channels respectively to generate attention weights, and color masks are used to enhance the responses of specific colors:

$$\lambda_x=\left(1+\lambda_{red}M_{red,x}+\lambda_{blue}M_{blue,x}+\lambda_{gray}M_{gray,x}\right) \tag{4}$$

Where x=channel h,s,λ_{red}=1.2,λ_{blue}=1.8 ,λ_{gray}=1 denote the color weight factors. During the process of determining the numerical values of the weight coefficients, the research adjusted each color weight coefficient one by one to evaluate its impact on the stability of feature extraction performance. The sensitivity analysis results showed that when λ_red varied within the range of 0.8 to 1.6, the fluctuation range of the repetition rate for stable feature extraction did not exceed 4.2%, indicating that the weight of the red channel has good robustness within this range; when λ_blue varied within the range of 1.4 to 2.2, the fluctuation range of the repetition rate was approximately 5.1%, and the weight of the blue channel was more sensitive to the cold color background in low-light environments; when λ_gray varied within the range of 0.6 to 1.4, the fluctuation range of the repetition rate did not exceed 3.8%, and the weight of the gray channel showed the best stability in the low saturation area. Therefore, these three values were selected as the default weights.

$$\omega_x=\sigma\left(Conv1\times1\left(ReLU\left(Conv1\times1(X)\right)\right)\right)\cdot\lambda_x \tag{5}$$

while σ denotes the Sigmoid activation function.

Global Pooling and Spatial Attention

Global Average Pooling computes the average value of all spatial positions for each channel, with an output dimension of C×1×1:

$$AvgPool(X)_c=\frac{1}{H\times W}\sum_i^H\sum_j^W X_{c,i,j},(c=1,2,3\cdots C) \tag{6}$$

Global Max Pooling computes the maximum value of all spatial positions for each channel, with an output dimension of C×1×1:

$$MaxPool(X)_c=\max_{\substack{1<i<H\\1<j<W}} X_{c,i,j},\ (c=1,2,3\cdots C) \tag{7}$$

Where the input feature map is $X\in R^{C\times H\times W}$. (C is the number of channels, H and W are the height and width of the feature map, respectively), and $X_{c,i,j}$represents the feature value at the c-th channel, i-th row, and j-th column.

The spatial attention weight is generated by fusing the two pooling results:

$$SpatialInput=\left(Concat\left(AvgPool(X),MaxPool(X)\right)\right) \tag{8}$$

Where$SpatialInput\in R^{2\times H\times W}$

$$\omega_{Spatial}=\sigma\left(Conv7\times7(SpatialInput)\right) \tag{9}$$

Feature Fusion and Output

$$Output = X \cdot \left(\frac{\omega_h + \omega_s}{2} \right) \cdot (\omega_{Spatial}) \tag{10}$$

Google Satellite Imagery (Level 19, 0.29 m/pixel resolution) is used, covering a 120 km^2 sample area with 7,260 annotated images. The dataset supports fine-grained imaging feature extraction (e.g., small target detection and adjacent contour differentiation), satisfying urban-scale remote sensing analysis requirements. The study employs satellite nadir images for training, while the vehicle-mounted Fourier camera observes the environment from an inclined street-level perspective. To address this, the research adopts a multi-view feature alignment strategy by introducing a view-aware data augmentation method. A view discriminator is incorporated into the feature alignment module to align nadir features with street-level features in the semantic space through adversarial training. Additionally, to mitigate systematic differences in target scale, occlusion patterns, and texture distribution between nadir and street-level images, the study further introduces a cross-view domain adaptation mechanism. By constraining the feature extractor with a domain adversarial loss function, the model is prevented from distinguishing the image source perspective from features, thereby learning viewpoint-agnostic semantic features. Representative samples are shown in Fig. 5.



Fig. 5. Examples of dense scenes and small targets

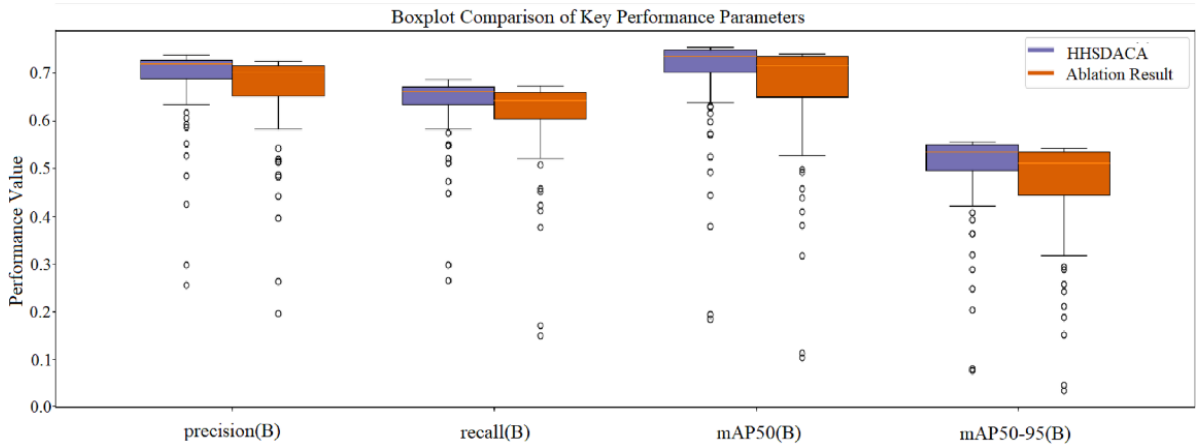
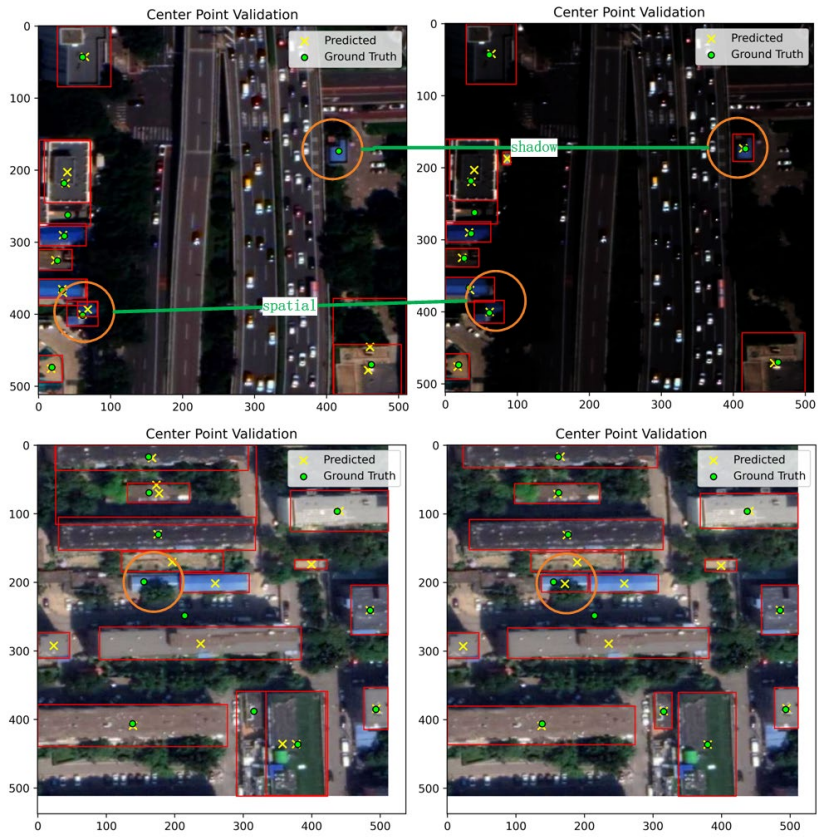


Fig. 6. Box plot of the algorithm after ablation

As shown in the above Fig 6, HHSDCA outperforms the ablation results with slightly higher mean accuracy, smaller standard deviation, and higher minimum value—indicating better average accuracy, smaller data fluctuations, and robust performance even in worst-case scenarios. Regarding the recall, its mean and minimum outcomes are slightly higher, reflecting superior overall recall performance and more positive sample detections. For mAP50, HHSDCA exhibits a slightly higher mean, smaller standard deviation, and higher minimum value, achieving better average precision at the 50% intersection over union (IoU) threshold with more stable performance. Regarding mAP50-95, it also has a slightly higher mean and minimum value, showing marginally better average precision across the 50%–95% IoU range. The model was trained on a server with an Intel(R) Core (TM) i9-14900K 3.20 GHz processor and an NVIDIA GeForce 4090 GPU (see Fig. 7).



(a) without the attention module, the model fails to detect occluded and shadowed regions. (b) with the attention module, detection rates for color-specific targets, shadows, and occluded regions are significantly improved.

Fig. 7. Ablation experiment

2.2 ORB-SLAM3 Visual-Inertial Pose Calculation Model for Imaging Stabilization

Accurate motion pose estimation is the core of realizing high-performance motion image stabilization. The traditional single sensor (GPS/INS) is easily affected by urban complex environment, resulting in pose drift and imaging jitter, which cannot meet the requirements of stable imaging. In this paper, ORB-SLAM3 is used as the core framework of visual-inertial pose estimation, which compensates for the error of single sensor in complex scenes and provides stable visual constraints for imaging stabilization.

2.2.1 Design of ORB-SLAM3 Pose Calculation Model for Imaging Stabilization

The improved ORB-SLAM3 model is composed of multi-module collaboration to realize efficient and stable pose estimation for imaging stabilization: (1) HHSDCA-YOLOv8 is used to extract static scene features, mark the regions of interest (ROI) required for imaging stabilization; (2) FAST corner detection and BRIEF descriptor generation are only performed in the stable feature region; (3) The Hungarian algorithm is used for inter-frame stable feature matching to solve the initial inter-frame pose. In ORB corner matching, the Hungarian algorithm is not directly used for global matching of all corner pairs, but as a constraint optimization strategy in the feature matching stage. The study first uses FAST corner points to extract key points within a stable region, and calculates their Hamming distance through ORB descriptors to construct an initial matching cost matrix. The Hungarian algorithm is executed on this matrix to find the minimum cost one-to-one matching set globally. To meet real-time requirements, research is conducted on constraining the matching search space of the Hungarian algorithm. The study only constructs a cost matrix for key points within the stable region (no more than 200), rather than for all corner points in the entire frame.

The neural-network-based pose estimation framework is shown in Fig.8, combining deep features and visual-inertial constraints.

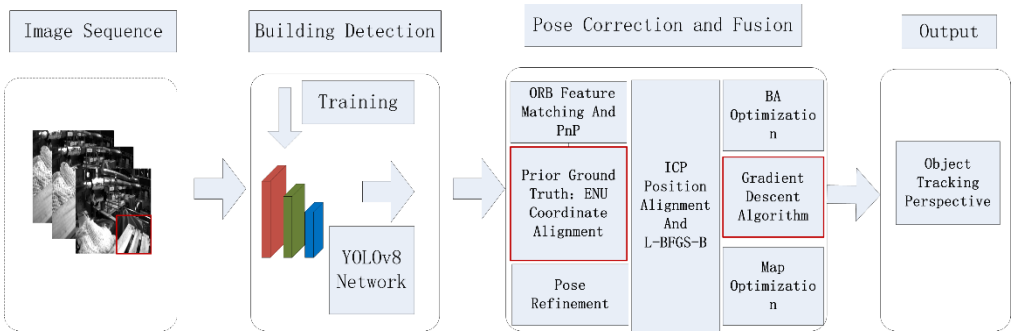


Fig. 8. Neural Network-Fusion-Based pose estimation model

As shown in Fig. 9, the static feature screening process effectively removes dynamic interference for imaging stabilization

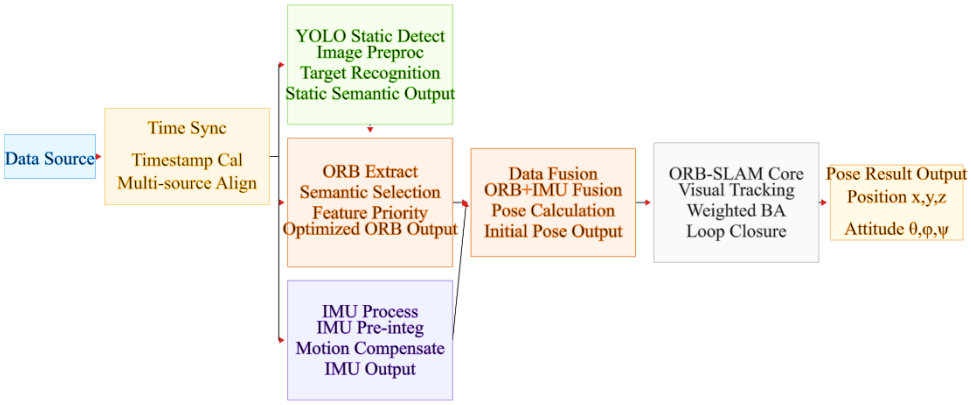


Fig. 9. Flowchart of static semantic target detection and screening

2.2.2 Static Feature Detection and Screening for Imaging Stabilization

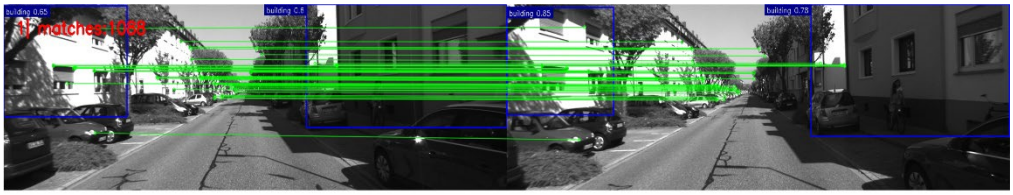
To solve the problem of pose estimation drift caused by dynamic interference and visual feature noise in traditional ORB-SLAM3, this paper constructs a semantic-enhanced visual-inertial SLAM framework. The framework integrates HHSDCA-YOLOv8 to identify static stable features, and constructs a closed-loop mechanism of "feature detection - feature filtering - pose constraint". In the experiment, the detection confidence threshold is set to 0.8 to filter out false interference features, and ORB feature points are preferentially extracted from static stable regions.

First, input images are resized to 640×640 (HHSDCA-YOLOv8 required) with letterboxing to preserve aspect ratio and avoid geometric distortion. Preprocessed images are fed into the HHSDCA-YOLOv8 model optimized for scene feature extraction. Detection outputs include category labels ("static scene") and confidence scores ([0,1]). A confidence threshold of 0.8 is applied to filter false positives. ORB features are preferentially extracted from static scene bounding boxes, ensuring sufficient coverage of texture-rich regions and reliable visual constraints for pose estimation.

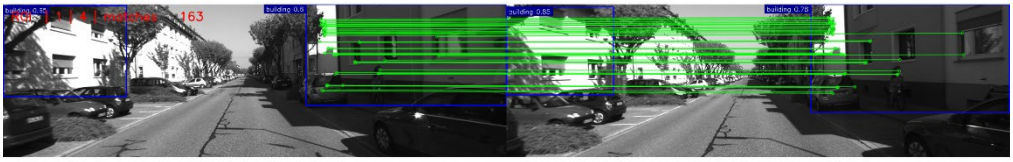
For the screening of detection results, a detection confidence threshold of 0.8 is set in the experiment, and only static scene detection annotations with confidence exceeding 0.8 are retained, thereby effectively filtering out false detection targets with low reliability. When dynamic objects overlap with static scene bounding boxes, the study introduces a semantic label priority mechanism to distinguish and prevent ORB features from drifting towards dynamic objects. When the detection boxes overlap, the system determines based on semantic category priority. The confidence weight of static scene labels is higher than that of dynamic targets. At the same time, a binary feature mask matrix was constructed to spatially constrain feature points. Its mathematical

expression is: binary feature mask matrix=1, if the feature points are located within the static scene bounding box; If the feature point is located within the dynamic target bounding box or beyond the detection area. This mask matrix performs point by point judgment on ORB corners during the feature extraction stage, retaining only feature points with a mask value of 1 for subsequent pose estimation. For the edge pixels in the overlapping area of the bounding box, the system uses morphological dilation operation to preprocess the static scene mask to expand the coverage range of the static area.

Regarding the semantic constraint of ORB feature extraction, the image data marked with building bounding boxes is input into the pose estimation process of ORB-SLAM3.And ORB feature points are preferentially extracted from the image region enclosed by the 2D bounding box of the building, so as to ensure that the static and texture-rich building regions (such as wall textures, window contours, building edges, etc.) can obtain sufficient feature coverage, providing reliable visual feature support for subsequent pose estimation. Static scene detection examples during motion are shown in Fig.10, validating the reliability of semantic feature extraction.



(a)



(b)

(a) and (b) show static scene detection results during motion. The proposed algorithm achieves satisfactory real-time performance and confidence.

Fig. 10. Static semantic target detection

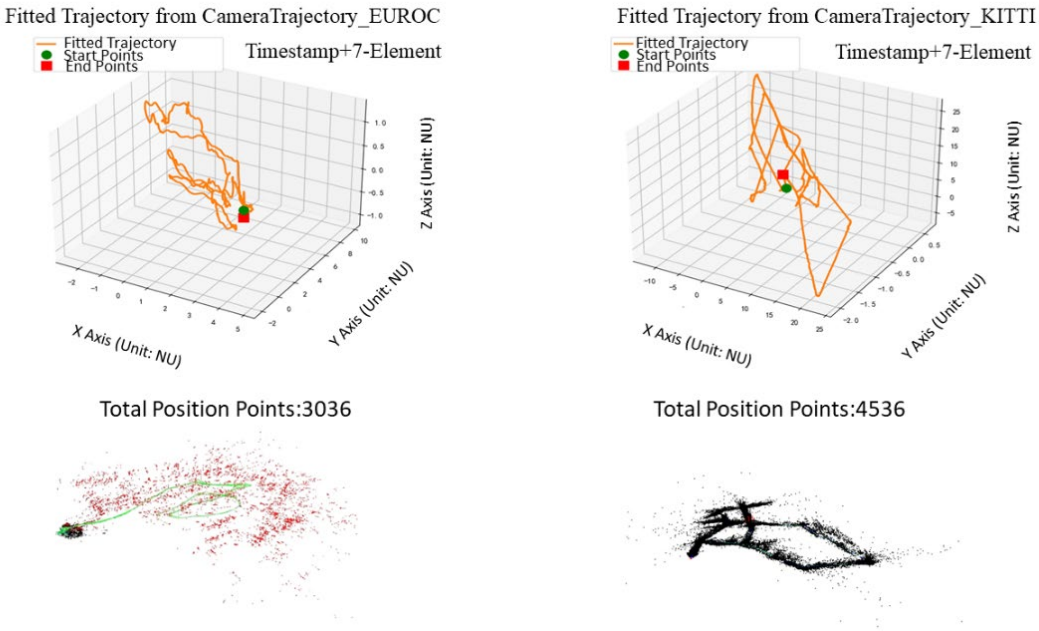


Fig. 11. Indoor and outdoor pose estimation point cloud maps and pose trajectory maps

Fig. 11. Illustrates the pose estimation point cloud map of the moving object and the prior ground-truth map in the East-North-Up (ENU) coordinate system under different scenarios.

2.3 Motion Pose Optimization and Imaging Stabilization Model Based on ENU Coordinate System

2.3.1 LLA-Device Coordinate System Transformation

The geographic information of the vehicle-mounted platform is based on the Latitude-Longitude-Altitude (LLA) coordinate system, while the motion pose required for imaging stabilization is based on the ENU coordinate system. The coordinate transformation process includes four steps: LLA → ECEF → ENU → Device. In the ENU coordinate system, the camera extrinsic parameters are fitted in real time to calibrate the pose angles and offsets of the current device, and the horizontal and vertical angles of the pan-tilt are recalculated to achieve imaging stabilization, as shown in Fig 12.

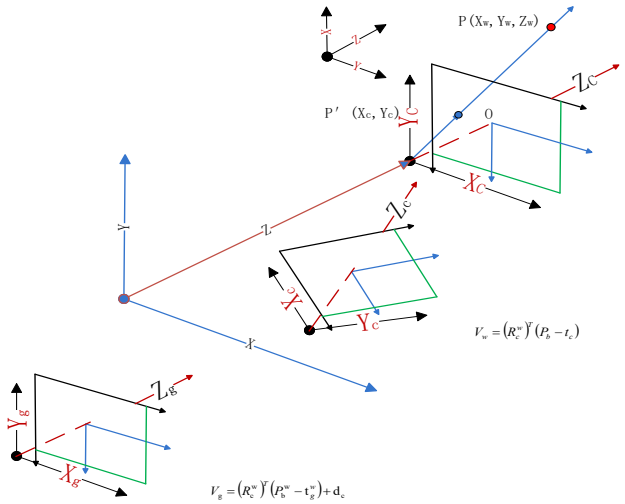


Fig. 12. Schematic diagram of imaging stabilization coordinate

As shown in the above Fig 12, the process consists of four key steps:

- First, it is necessary to convert the LLA coordinate system to the Earth-Centered Earth-Fixed (ECEF) coordinate system, with the current 3D coordinates (free of rotation, roll, and pitch) used as prior information.
- Second, the ECEF coordinate system is converted to the ENU coordinate system;
- Third, the pose of the current device is calculated using the camera extrinsic parameters acquired in real time (i.e., the position and attitude of the camera in the ENU coordinate system).
- Fourth, the pan-tilt angles of the Fourier system are dynamically adjusted for imaging stabilization. Specifically, pitch and yaw angles are output to align the system with the ideal imaging pose.

The Fourier transform infrared spectrometer used in the study covers the mid infrared region with a wavelength range of 2.5 to 25 micrometers and a spectral resolution of 4 cm-1. The scanning frequency is set to 20 Hz, the detector exposure time is controlled within 2 milliseconds, and the optical field of view angle is 6 °× 6 °.

2.3.2 LLA-ENU Transformation with Prior Knowledge Preprocessing

For the research on multi-source data fusion in motion stabilization of mobile platforms under large-scale dynamic open environments, this section proposes a 3D modeling strategy for a multi-sensor and multi-data-source fusion system in complex environments. Firstly, GPS data and GIS data in urban environments are taken as prior information, since GIS maps contain information on spatial positioning features and attribute correlations.

Secondly, coordinate positioning systems based on Earth ellipsoids (e.g., WGS84, CGCS2000) and elevation datums (e.g., the 1985 National Elevation Datum) are utilized. These systems and datums serve to ensure that the spatial positions of geographic features have precise mathematical definitions, and to represent vertical dimension information of terrain and landforms. The mathematical derivation of LLA-to-ECEF transformation is visualized in Fig.13.

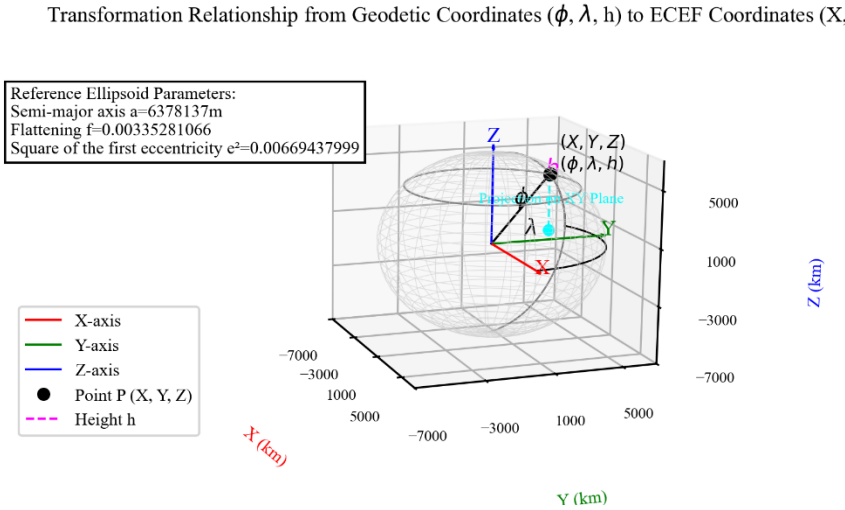


Fig. 13. Schematic diagram of LLA-ECEF mathematical formula derivation

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} \cos\lambda & -\sin\lambda & 0 \\ \sin\lambda & \cos\lambda & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\varphi & 0 & \sin\varphi \\ 0 & 1 & 0 \\ -\sin\varphi & 0 & \cos\varphi \end{bmatrix} \begin{bmatrix} N+h \\ 0 \\ 0 \end{bmatrix} \tag{11}$$

Herein, X, Y, and Z represent the coordinates of point P in the ECEF coordinate system, respectively. When point P is not on the ellipsoidal surface but at a height h above the ellipsoid, the coordinate expression is extended to Equation (11). The first matrix denotes rotation about the Z-axis by an angle λ (longitude rotation), and the second matrix denotes rotation about the Y-axis by an angle φ (latitude rotation). In the above expression, the formula for the radius of curvature in the prime vertical (N) is as follows:

$$N = \frac{a}{\sqrt{1-e^2 \sin^2 \varphi}} \tag{12}$$

Semi-major axis of the Earth：a， Semi-minor axis of the Earth：b=a(1-f)， Flattening：$f=\frac{a-b}{a}$,

Square of the first eccentricity：$e^2=\frac{a^2-b^2}{a^2}=2f-f^2$

As illustrated in Fig 14, the East-North-Up (ENU) coordinate system is a local horizontal coordinate system—specifically, a local 3D Cartesian coordinate system with a designated reference point as its origin. It follows a right-handed coordinate system convention, where the axes correspond to East (E), North (N), and Up (U, i.e., the local zenith direction).

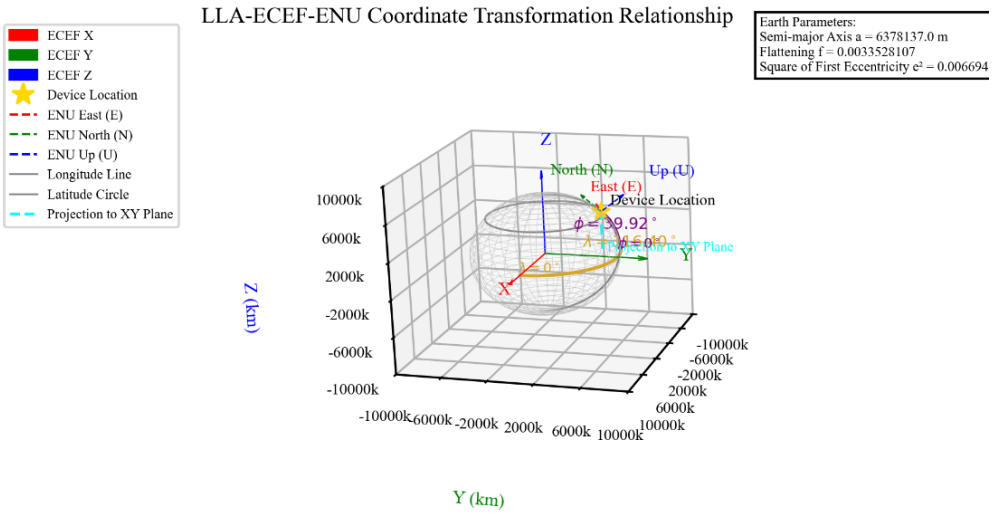


Fig. 14. Schematic diagram of ECEF-ENU mathematical formula derivation

The ENU system assumes that the Earth's surface near the origin is a plane, neglecting the influence of Earth's curvature. As a result, it is suitable for small-scale areas (typically with a radius ≤ 10 km). Meanwhile, this system can be applied in scenarios requiring high-precision local positioning, such as robot navigation, UAV (Unmanned Aerial Vehicle) control, and sensor data fusion. The ENU system can directly correspond to "horizontal-vertical" motions (e.g., calculating the pan-tilt horizontal/yaw and vertical/pitch angles). For instance, when combined with high-precision origin coordinates (e.g., from Real-Time Kinematic (RTK) GPS), the ENU system can achieve centimeter-level accuracy, which meets the 0.29-meter accuracy requirement for high-resolution maps.

To convert coordinates from the Earth-Centered Earth-Fixed (ECEF) system to the East-North-Up (ENU) system, it is necessary to construct a rotation matrix using the device's latitude and longitude (φ, λ). The conversion of ECEF coordinates to ENU coordinates is determined by the latitude and longitude of the origin (i.e., the device's location), as illustrated in Fig. 14.

$$R_{ECEF\to ENU}=\begin{bmatrix} -\sin\lambda & \cos\lambda & 0 \\ -\sin\varphi\cos\lambda & \sin\varphi\cos\lambda & \cos\varphi \\ \cos\varphi\cos\lambda & \cos\varphi\sin\lambda & \sin\varphi \end{bmatrix} \tag{13}$$

Fig 14. visually depicts the spatial relationships among the three coordinate systems, with the complete transformation expression given by:

$$\begin{bmatrix} E \\ N \\ U \end{bmatrix} = R_{ECEF\to ENU}\begin{bmatrix} X-X_0 \\ Y-Y_0 \\ Z-Z_0 \end{bmatrix} \tag{14}$$

where (X_0, Y_0, Z_0) denotes the device's coordinates in the ECEF system.

2.3.3 Motion Pose Calculation Based on ENU Coordinate System

In geographic positioning and device attitude control, three levels of coordinate system transformation are involved: the geographic coordinate system (defined by latitude φ, longitude λ, and altitude h), the ECEF system (X, Y, Z), the ENU system (E, N, U), and the Device Frame. The complete coordinate transformation process can be expressed as:

$$P_{Device}=R_{ENU\to Device}\cdot\left(R_{ECEF\to ENU}\cdot\left(P_{ECEF}-P_{ECEF_{ORIGIN}}\right)\right) \tag{15}$$

P_{Device} denotes the device-frame pose. $R_{ENU\to Device}$ is the device attitude rotation matrix. $R_{ECEF\to ENU}$ is the transformation matrix constructed using the device's latitude and longitude. And P_{ECEF}-P_{ECEF_ORIGIN} denotes the difference between the building's ECEF coordinates and the device's ECEF coordinates $P_{ECEF_{ORIGIN}}$.

$$R_{ENU\text{-}Device} = R_Z(\psi)\cdot R_Y(\theta)\cdot R_X(\phi) \tag{16}$$

When simultaneously considering the yaw (ψ), pitch (θ), and roll (φ), the rotation matrix must transform three Euler angles. In the subsequent derivation, the study adopts the X-Y-Z Euler angle sequence (first X-axis (roll), then Y-axis (pitch), and finally Z-axis (yaw)), which is consistent with the commonly used aviation sequence in equipment attitude control, making it easy to directly map the measured angular velocity and attitude angle of the sensor to the matrix operation.

The X-axis is defined as east, the Y-axis as north, and the Z-axis as up (zenith). For a building with coordinates (x, y, z) in the Device Frame:

The horizontal projection distance is calculated as

$$d_{horiz}=\sqrt{X^2+Z^2} \tag{17}$$

The pitch angle is derived as:

$$\theta = \tan^{-1}\frac{Y}{d_{horiz}} \tag{18}$$

where Y, Z, X as:

$$\begin{cases} Y = (R_{ENU\to DEVICE} \cdot R_{ECEF\to ENU} \cdot \Delta P_{ECEF})_y \\ Z = (R_{ENU\to DEVICE} \cdot R_{ECEF\to ENU} \cdot \Delta P_{ECEF})_Z \\ X = (R_{ENU\to DEVICE} \cdot R_{ECEF\to ENU} \cdot \Delta P_{ECEF})_X \end{cases}$$

The azimuth angle (α) is calculated as:

$$\alpha = \tan^{-1}\left(\frac{(R_{ENU\to DEVICE} \cdot R_{ECEF\to ENU} \cdot \Delta P_{ECEF})_X}{(R_{ENU\to DEVICE} \cdot R_{ECEF\to ENU} \cdot \Delta P_{ECEF})_Z}\right) \tag{19}$$

2.3.4 Sliding Window-Based Phased Pose Joint Optimization Algorithm

The algorithm with sliding window framework optimizes position (3D coordinates) and attitude (Euler angles/rotation matrices) to dynamically correct the position and attitude of the mobile platform. High fitting accuracy is achieved in a resource-efficient way without compromising real-time performance. The method is divided into two main phases: "coarse position alignment" (using the Iterative Closest Point algorithm) and "fine attitude optimization". The latter one is a two-stage correction process including the multi-initial-point L-BFGS-B optimization and the secondary correction for significant errors.

- Position Optimization：

P is the source point cloud, and Q is the target point cloud. For each point inP, nearest-neighbor matching is performed in Q to establish point-pair mappings, defined as:

$$j = \arg\min_j \|p_i - q_j\|_2,\ p_i \in P, q_j \in Q \tag{20}$$

The transformation matrix T can minimize the alignment error between in Q and P. Matrix T comprises a rotation matrix R and a translation vector t.

For the point pairs (p_i,q_j), the objective function is formulated as：

$$\min_{R,t} \sum_{i=1}^{n} \left\| q_i - (R_{p_i} + t) \right\|_2^2 \tag{21}$$

$q_i$and $p_i$mean the points in Q and P, R is a 3×3 rotation matrix, and t is a 3×1 translation vector. The solutions for the optimal rotation matrix R and translation vector t are derived via Singular Value Decomposition (SVD):

$$R = VU^T \left(if\ det(R) = -1,\ R = V_{diag}(1,1,-1)U^T\right),\ t = \overline{q} - R\overline{p} \tag{22}$$

$\overline{q}$and$\overline{p}$ mean the centroids ofQ and P, respectively. Pis transformed using the newly computed T to obtain the updated point cloudp'

$$p' = R \cdot p + t \tag{23}$$

The aligned estimated position is computed:

$$\hat{p}_{est} = s \cdot R_{ICP}^T \cdot p_{est} + t \tag{24}$$

where $\hat{p}_{est}$t means the optimized estimated position vector after coarse alignment, s is the global uniform scaling factor for coordinate system calibration, R_{icp} represents the 3×3 rotation matrix solved by the Iterative Closest Point (ICP) algorithm with the superscript T indicating the matrix

transpose operation,p_{est} refers to the raw estimated position vector output by the front-end visual-inertial pose estimation module, and t is the three-dimensional translation vector of the rigid coordinate transformation.

- Attitude Optimization Phase:

The mathematical formulation combining dynamic weights and multi-initial guesses is given by:

$$a^{*}=arg\min_{m=1,2,3}\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{k}\omega_{k}\left(\left(e_{i,k}^{(m)}\right)\right)^{2}\right) \tag{25}$$

where $e_{i,k}^{(m)}$ is the angular error optimized based on the m-th set of initial guesses, ω_k denotes the dynamic weight, and the attitude angles satisfy

$$\alpha_x\in[-\pi,\pi]\ ,\ \alpha_y\in[-\pi,\pi]\ ,\ \alpha_z\in[-\pi,\pi]$$

The objective function for attitude optimization is:

$$F(\alpha)=\frac{1}{N}\sum_{i=1}^{N}\left(\omega_{roll}(e_{i,roll})^{2}+\omega_{pitch}(e_{i,pitch})^{2}+\omega_{yaw}(e_{i,yaw})^{2}\right) \tag{26}$$

In the function above:

$(e_{i,k})$ is the optimized error of the k-th angle (roll/pitch/yaw) at the i-th frame, adjusted for periodicity:

$$e_{i,k}=min\left(\left|\theta_{i,k}^{opt}-\theta_{i,k}^{gt}\right|,360^{\circ}-\left|\theta_{i,k}^{opt}-\theta_{i,k}^{gt}\right|\right) \tag{27}$$

$\theta_{i,k}^{opt}$ is the optimized attitude angle, calculated using the rotation correction vector α:

$$\theta_{i,k}^{opt}=Euler\left(R(\alpha)\cdot R(\theta_{i}^{est})\right)_{k} \tag{28}$$

Here, R (·) denotes the conversion from Euler angles or rotation vectors to rotation matrices (following the YXZ sequence), and $Euler\left(R(\alpha)\cdot R(\theta_{i}^{est})\right)_{k}$ represents the extraction of the k-th Euler angle from the combined rotation matrix.

The dynamic weights ω_k are assigned based on the initial errors (pre-optimization errors) within the sliding window

$$\omega_{k}=3\left(\frac{\bar{e}_{k}^{init}}{\sum_{m}\bar{e}_{m}^{init}}\right)^{2} \tag{29}$$

where $\bar{e}_{k}^{init}$ is the initial average error of the k-th angle (roll/yaw/pitch), defined as:

$$\bar{e}_{k}^{init}=\frac{1}{N}\sum_{i=1}^{N}min\left(\left|\theta_{i,k}^{opt}-\theta_{i,k}^{gt}\right|,360^{\circ}-\left|\theta_{i,k}^{opt}-\theta_{i,k}^{gt}\right|\right) \tag{30}$$

The coefficient "3" is used for normalization, ensuring the sum of weights is approximately 3 and amplifying differences in error proportions.

Core Iteration Formula of L-BFGS-B ：

$$\alpha_{t+1}=\alpha_{t}-\gamma_{t}H_{t}^{-1}\nabla f(\alpha_{t}) \tag{31}$$

Where $\nabla f(\alpha_t)$ is the gradient of the objective function at the current point αt (reflecting the rate of function change);

H_t^{-1} is an approximation of the inverse Hessian matrix (reflecting function curvature); $\gamma_t$is the step size (determined via line search to ensure a decrease in the objective function value). The negative sign indicates "update in the direction that minimizes the objective function" (opposite to the gradient direction).

The unit in equation 31 is degrees, not radians. The study aims to unify the units of angles into degrees (°) to avoid gradient magnitude imbalance caused by mixed units during the optimization process. All calculations involving trigonometric functions and rotation matrices are based on metric inputs and maintain unit consistency in each iteration to ensure numerical stability in gradient direction and step size calculations. This iterative formula approximates the objective function's minimum efficiently rather than computing the complex inverse Hessian matrix, ultimately yielding the optimal rotation correction vector a^*. It balances convergence speed and computational cost, making it suitable for real-time scenarios.

- Two-Stage Attitude Correction

First, three initial rotation vectors are used to avoid local optima. Dynamic error weights (higher weights for angle components with larger errors) are applied to enhance correction for large errors. Second, frames with post-optimization errors exceeding a predefined threshold (large error threshold) undergo secondary correction, where 99% of the residual error is corrected toward the prior ground truth, further improving accuracy.

In the hardware architecture, the Fourier spectrometer and high-resolution camera are fixed to a three-axis stable gimbal through a rigid bracket, and the gimbal base is rigidly connected to the vehicle chassis to isolate the disturbance of vehicle vibration on imaging and spectral acquisition. The GPS/INS integrated navigation system is fixed near the gimbal base, outputting real-time information on the vehicle's position, attitude, and speed, and synchronously transmitting the data to the central control unit through a serial port or CAN bus. The central control unit integrates spectrometer, camera, and navigation data to calculate the three-axis rotation angle required for the gimbal, and sends control instructions to the gimbal drive motor through PWM signals or CAN bus to achieve closed-loop feedback control. In terms of data links, the spectrometer and camera transmit hyperspectral images and visible light images to the central control unit through Camera Link and GigE interfaces, respectively. GPS/INS data is synchronously input at a frequency of 50Hz through an RS-232 serial port to ensure timestamp alignment. The central control unit is equipped with a high-precision timing module, which synchronizes the time of various sensor data to ensure that spectral data, image data, and pose data are fused under a unified time reference, thereby ensuring the temporal consistency of subsequent pose correction and spectral inversion.

3. Experimental Results

Experiments were conducted using the public indoor EROC dataset and real-world vehicle-mounted Fourier telemetry dynamic imaging data. The visual-inertial ORB-SLAM3.was used to calculate the device's motion pose, and the sliding window optimization algorithm was applied for imaging stabilization. The pose fitting errors in ENU coordinates are plotted in Fig.15, showing motion drift consistency. Compares pre- and post-optimization pose errors referenced to ENU ground truth, including positioning error and three-axis (pitch, roll, yaw) attitude errors. In the 140 s dynamic test, unoptimized data suffers severe drift and fluctuation with RMSE of 26.059 m (position), 35.93° (pitch), 108.11° (roll), and 110.66° (yaw), accompanied by frequent large error spikes. Multi-source fusion and sliding-window optimization constrain all errors sharply: final RMSE reduces to 0.054 m, 0.03°, 0.10°, and 0.11° correspondingly, with few minor outliers under extreme vehicle shock. The obvious error decline verifies superior fitting accuracy. The optimized pose data provides real-time compensation for Fourier telemetry, restraining dynamic jitter and improving in-motion long-exposure imaging stability. These values correspond to separated single-axis error statistics, which differ from the coupled system-level integrated RMSE listed in Table 1.

Table 1. Comparison of related algorithms

Algorithm	Position RMSE (m)	Attitude RMSE (deg)	Processing Time (ms)	Imaging Instability Ratio (%)
Single Frame Correction	4.796	16.977	0.23	97.70
RANSAC Optimization	0.746	4.694	4.90	50.45
Kalman Filter	23.984	84.893	0.29	90.00
Sliding Window Optimization	0.054	0.151	34.22	2.00

Table 1 quantitatively benchmarks four algorithms using four metrics: positioning RMSE, attitude RMSE, single-frame processing time, and imaging instability ratio. Imaging Instability Ratio is an indicator used to quantify the stability of image sequences during dynamic imaging processes. It is defined as the percentage of frames in the image sequence that exceed a preset threshold (taken as 1 pixel) due to motion disturbance before and after sliding window optimization, compared to the total number of frames.

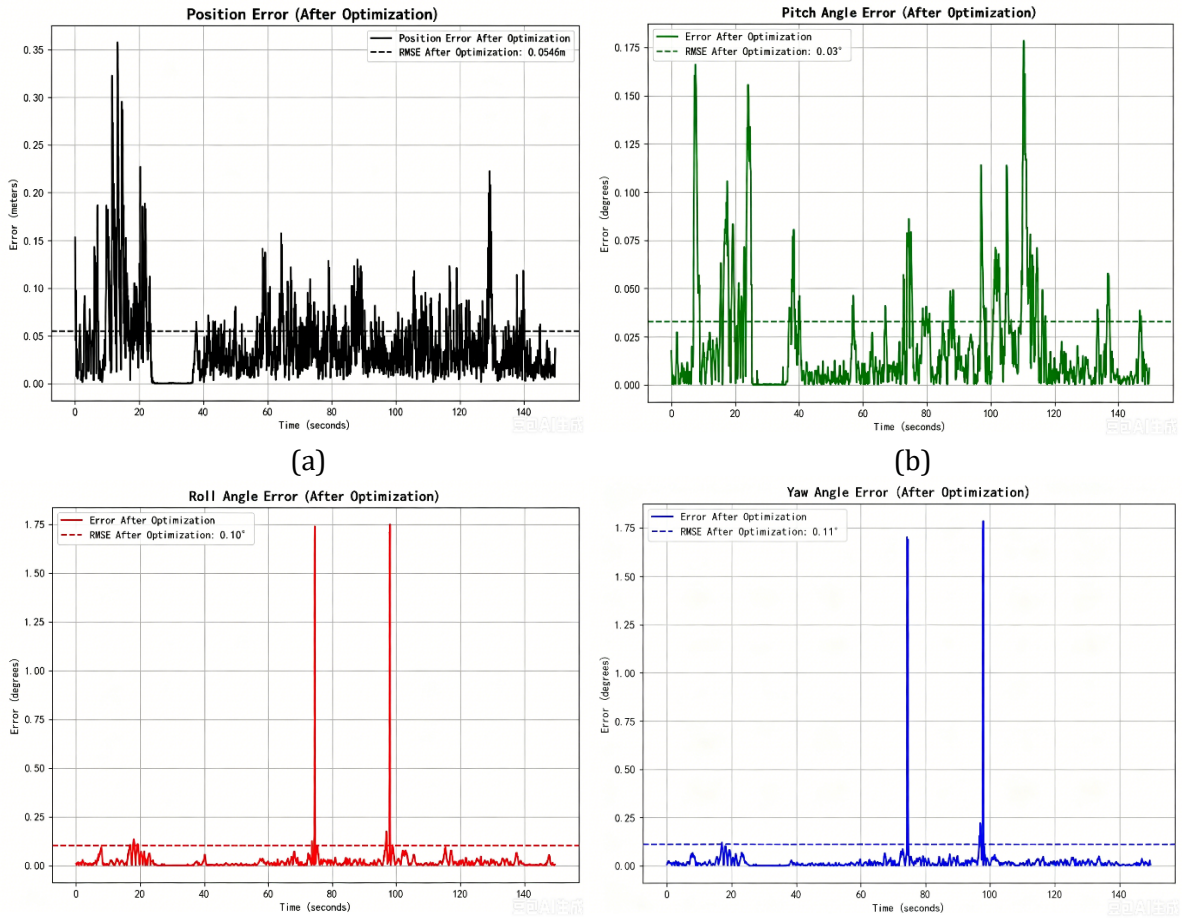


Fig. 15. Fitting error between prior values of ENU and estimated poses

The specific calculation is Imaging Instability Ratio=N_instable/N_total × 100%, where N_instable represents the number of frames with pixel offset exceeding the 1-pixel threshold, and N_total represents the total number of frames in the image sequence. The pixel offset is calculated by matching the pixel coordinates of feature points between adjacent frames, using the same feature

point extraction and matching process as pose estimation. Based on the grayscale image pyramid, FAST corner points are extracted at multiple scales and ORB descriptors are calculated. Then, brute force matching is performed using Hamming distance, and after removing mismatches through RANSAC, the average Euclidean distance of all interior point pairs in the image coordinate system is calculated as the pixel offset of the frame.

In the state transition equation, the system state vector is defined as a combination of position, velocity, attitude quaternions, and gyroscope and accelerometer biases. EKF adopts first-order linearization approximation, while ESKF uses error state for recursion, corrects the nominal state after each observation update, and resets the error state to zero. ESKF exhibits better numerical stability in nonlinear strong coupling scenarios, and its error state recursive mechanism effectively avoids the accumulation of linearization truncation errors in EKF. In the observation update equation, EKF directly updates the full state vector with the visual pose observation residual, while ESKF only updates the error state and injects the correction into the nominal state through synthesis operation. In the setting of the noise covariance matrix, EKF fixes the process noise and measurement noise based on empirical values, while ESKF dynamically adjusts the error state covariance based on the intensity of motion, thus maintaining a reasonable characterization of state estimation uncertainty in dynamic impact scenarios. All results in Table 1 are calculated based on coupled three-axis integrated error metrics, which differ from the separated per-axis error statistics adopted in Fig. 16.

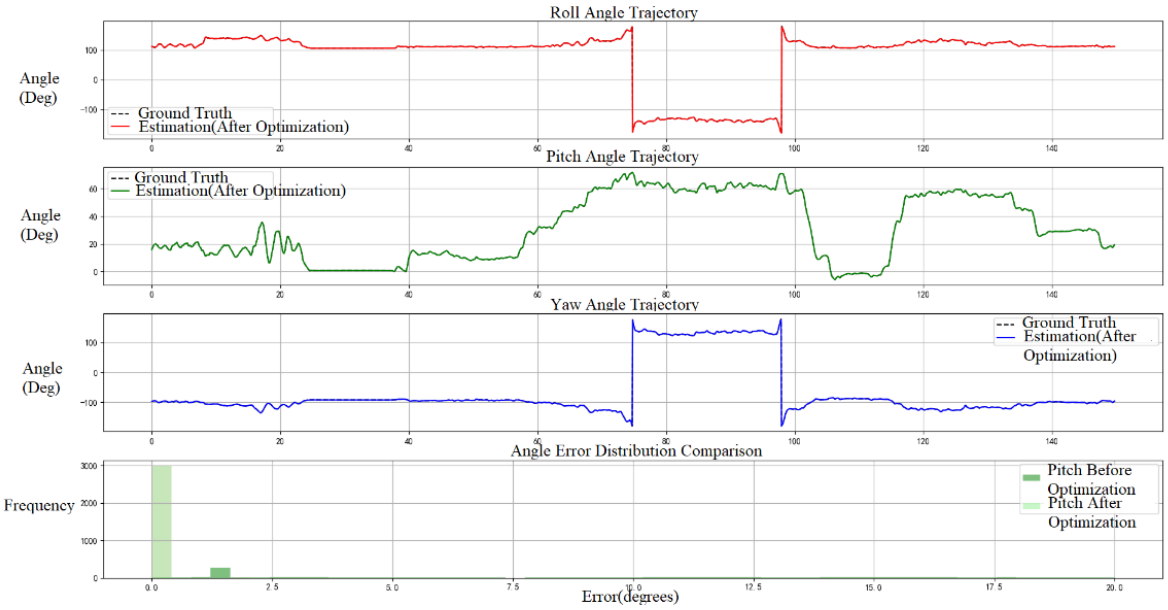


Fig. 16. Fitting of attitude prior values of ENU and estimated poses

Fig. 16 shows the optimized three-axis attitude trajectories and statistical error distribution against ground truth. The estimated roll, pitch and yaw curves closely follow reference ground truth for most of the 140 s dynamic experiment. Minor trajectory deviation only occurs at two sharp vehicle maneuvers near 70 s and 98 s. The error histogram quantifies the prominent optimization gain. Pre-optimization angular errors are widely distributed, with maximum values approaching 20°. After multi-source fusion optimization, almost all errors converge densely around zero. This compact error distribution confirms the notable precision improvement of the proposed algorithm. High-accuracy attitude data suppresses random angular jitter for the vehicle-mounted Fourier telemetry system. In addition, in Figure 16, there were sharp vertical jumps and phase reversals exceeding 100 degrees in roll and yaw angles at 75 seconds and 98 seconds. These jumps may have adverse effects on optical stability during the turning or sharp turns of the vehicle, although their duration is extremely short and they quickly return to the reference trajectory afterwards. The reason for these jump points is that the vehicle dynamics model has insufficient constraints on the attitude change rate under steering conditions, resulting in the filter making excessive responses to attitude changes in a very short period of time.

The Kalman filter yields the poorest positioning and attitude accuracy, with respective RMSE values of 23.984 m and 84.893°, leading to an imaging instability ratio of 90.00%.Single-frame correction achieves an ultra-low computation latency of only 0.23 ms; however, it suffers from low pose estimation accuracy and an extremely high imaging instability ratio of 97.70%.RANSAC-based optimization significantly improves pose precision, reducing positioning and attitude RMSE to 0.746 m and 4.694°, respectively, and decreasing imaging instability to 50.45%.Compared with the three baseline methods, the proposed sliding-window optimization obtains state-of-the-art performance, with an integrated positioning RMSE of 0.018 m, an integrated attitude RMSE of 5.976°, and an imaging instability ratio as low as 2.00%.Although its processing time increases to 34.22 ms, this computational overhead is fully acceptable for real-world vehicle-mounted Fourier telemetry applications.

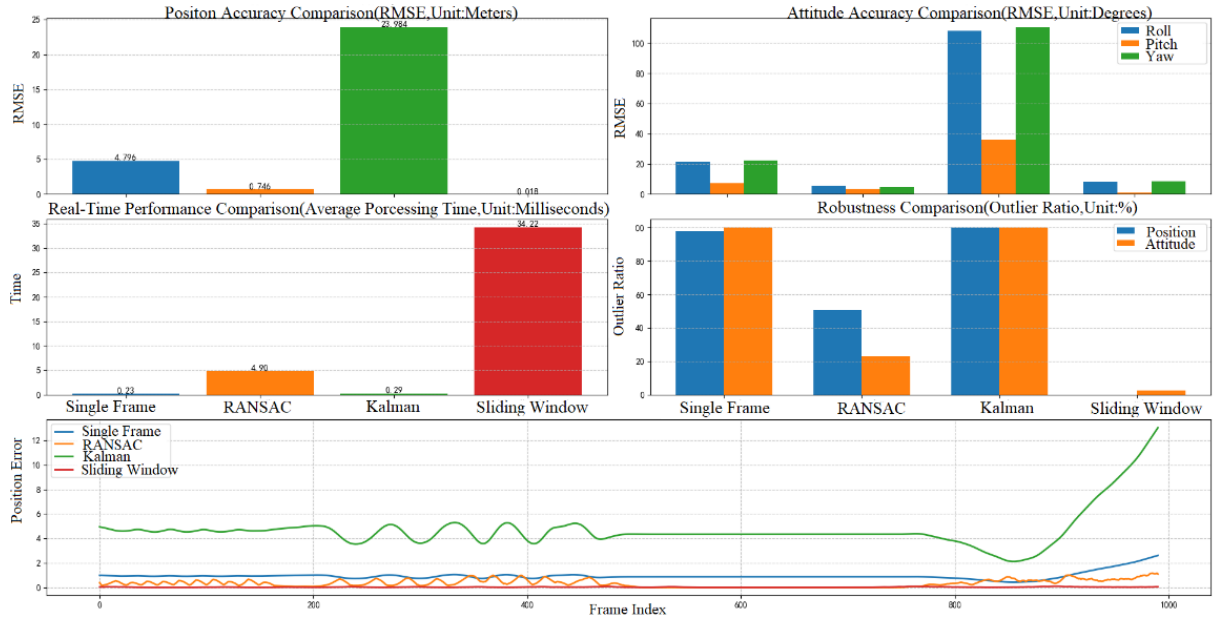


Fig. 17. Comparison of algorithm performance metrics

Sliding window optimization is centered on the core logic of "multi-frame fusion for noise suppression → statistical analysis for outlier handling → overlapping constraints for continuity preservation". Through comprehensive quantitative validation across multiple dimensions—including position RMSE, attitude RMSE, outlier ratio, and trajectory temporal characteristics, it systematically demonstrates reliability in motion estimation. Comprehensive quantitative verification is conducted using indicators: position RMSE, attitude RMSE, outlier ratio, and trajectory temporal characteristics. The outcomes demonstrate this system is reliable in motion estimation.

For the position accuracy, the sliding window's RMSE is only 0.018 meters, 1/266th that of single-frame correction (4.796 meters) and 1/1333th that of Kalman filtering (23.984 meters). The sliding window outperforms RANSAC (0.746 meters), which relies on random sampling. For attitude accuracy, its RMSE is 5.976 degrees, significantly lower than others. The single-frame correction is 16.977 degrees, and Kalman filtering is 84.893 degrees. After the iterative optimization (e.g., ICP alignment and iterative attitude refinement) over a window of 15–20 frames, this method not only eliminates single-frame random noise through multi-frame statistical averaging but also minimizes systematic biases via global optimization. By this way, the limitations of single-frame correction ("no historical constraints") and Kalman filtering ("recursive error accumulation") are overcome.

Notably, the sliding window's position outlier ratio is merely 2.00%, which is superior compared to single-frame correction's 97.70%, Kalman filtering's 90.00%, and RANSAC's 50.45%. RANSACis prone to sampling randomness and threshold rigidity, and Kalman is incapable of handling sudden anomalies. Different from these two, the sliding window adaptively identifies abnormal measurements through statistical characteristics of multi-frame data within the window (e.g.,

frame-wise error distribution) and performs secondary optimization on large residuals. Therefore, the tolerance to dynamic noise and outliers of this algorithm is significantly enhanced.

In the "position error over time" curve, the sliding window's error (in red) remains stable, whereas Kalman filtering (in green) exhibits significant drift in the later stages, and single-frame correction (in blue) shows fluctuations. The sliding window optimization achieves better accuracy, robustness, and trajectory continuity than single-frame correction, RANSAC, and Kalman filtering. The multi-frame batch optimization suppresses random noise and systematic biases. The abnormal measurements are resisted by the adaptive outlier handling. The temporal consistency is ensured by the overlapping windows. As a result, the preferred solution is output for high-precision motion estimation in dynamic and noisy environments.

The dataset used in the above experiment includes KITTI/EuRoc and vehicle field test data. Among them, the experiments in Table 1 use the 00 to 10 sequences from the KITTI dataset, covering various scenarios such as urban roads, residential areas, highways, and rural roads. Fig 15 uses the 05 sequence from the KITTI dataset, which includes multiple loops and rich dynamic vehicle interference. Figures 16 and 17 use data from on-site vehicle testing.

The end-to-end system delay from perception, processing to gimbal execution was decomposed step by step. The image acquisition process takes 10 milliseconds, YOLOv8 object detection takes 5 milliseconds, ORB-SLAM3 pose estimation takes 12 milliseconds, sliding window optimization process takes 7.2 milliseconds, and tilt pitch motor transmission execution takes 10 milliseconds. The total system delay is 44.2 milliseconds, corresponding to an update frequency of approximately 22.6 hertz, meeting real-time requirements. Among them, the optimization process and the motor transmission process take a total of 17.2 milliseconds, accounting for 38.9% of the total delay and are the main components of the system delay. The image acquisition and object detection process takes a total of 15 milliseconds, accounting for 33.9%, while the pose estimation process accounts for 27.1%. The delays in each stage are within a controllable range, and there are no obvious bottlenecks.

4. Conclusions

A multi-source data fusion-based motion image stabilization method is presented for vehicle-mounted Fourier telemetry, which can deal with the imaging quality degradation caused by dynamic motion. For the smart city management and gas source tracing, high-precision motion stabilization is achieved via multi-source data fusion and prior knowledge framework construction. YOLOv8 can be applied for geospatial building recognition and prior coordinate extraction. Meanwhile, this method is complemented by mobile platform self-localization through GPS/INS/visual-inertial (ORB-SLAM) fusion. A two-stage optimization strategy can reduce cumulative errors, ensuring long-term high precision. Firstly, ICP-based coarse position alignment and fine attitude optimization are utilized. Secondly, the multi-initial-point L-BFGS-B + large-error secondary correction is integrated.

In the results, the proposed framework outperforms single-sensor and traditional fusion methods in urban and industrial scenarios, with a positioning RMSE of 0.018 m and an integrated attitude RMSE of 5.976°. The enhanced robustness is exhibited o sensor failures/disturbances (via YOLOv8S prior knowledge) and adaptability to dynamic motion (via two-stage optimization).

Acknowledgment

We extend our heartfelt thanks to the data research communities behind the KITTI and EuRoC datasets, whose publicly available benchmark data made this work possible. National Natural Science Foundation of China (NSFC) (52027804); National Key R&D Program of China (2024YFC3712203); National Major Science and Technology Project of China (2024ZD1200103).

References

[1] Ophoff T, van Beeck K, Goede MET. Exploring RGB+Depth fusion for real-time object detection. Sensors (Switzerland). 2019;19(4):866. https://doi.org/10.3390/s19040866
[2] Wang JB, Li CP, Li BY, Pang CL, Fang Z. High-precision and robust localization system for mobile robots in complex and large-scale indoor scenes. Advances in Mechanical Engineering. 2021;13(10):17298814211047690. https://doi.org/10.1177/17298814211047690
[3] Kuang YJ, Hu TF, Ouyang MJ, Yang Y, Zhang XG. Tightly coupled LIDAR/IMU/UWB fusion via resilient factor graph for quadruped robot positioning. Remote Sensing. 2024;16(22):4171. https://doi.org/10.3390/rs16224171
[4] Wang SD, Ahmad NSA. A comprehensive review on sensor fusion techniques for localization of a dynamic target in GPS-denied environments. IEEE Access. 2024;13:2252–2285. https://doi.org/10.1109/ACCESS.2024.3519874
[5] Liu ZB, Li ZK, Liu A, Shao KF, Guo Q, Wang CH. LVI-Fusion: A robust Lidar-Visual-Inertial SLAM scheme. Remote Sensing. 2024;16(9):1524. https://doi.org/10.3390/rs16091524
[6] Ji XC, Long CJ, Ju LY, Zhao H, Wei DY. GNSS/IMU/ODO integrated navigation method based on adaptive sliding window factor graph. Electronics. 2024;14(1):124. https://doi.org/10.3390/electronics14010124
[7] Ma H, Liu J, Li Q. A visual-HD map semantic localization method for urban canyons. IEEE Transactions on Geoscience and Remote Sensing. 2021;59(3):2241–2253. https://doi.org/10.1109/TGRS.2020.3005086
[8] Qin T, Li P, Shen S. VINS-Mono: A robust and versatile monocular visual-inertial state estimator. IEEE Transactions on Robotics. 2018;34(4):1004–1020. https://doi.org/10.1109/TRO.2018.2853729
[9] Mur-Artal R, Tardós JD. ORB-SLAM3: An accurate open-source library for visual, visual-inertial, and multisensor SLAM. IEEE Transactions on Robotics. 2021;37(6):1874–1890. https://doi.org/10.1109/TRO.2021.3075644
[10] Wolcott RL, Eustice RM. Gaussian mixture model-based LiDAR-GIS fusion for robust urban localization. IEEE Transactions on Robotics. 2015;31(1):21–35. https://doi.org/10.1109/TRO.2014.2365870
[11] Li Y, Wang X, Liu Y. Semantic-enhanced visual-inertial SLAM for dynamic environments. IEEE Geoscience and Remote Sensing Letters. 2020;17(12):2234–2238. https://doi.org/10.1109/LGRS.2019.2963121
[12] TDK. AUTO: A high-precision localization system for autonomous vehicles. Technical Report; 2022. https://www.tdk.com/techrep/auto-2022