A Unified Monocular and Stereo Visual Odometry Pipeline with Calibration, Bundle Adjustment, and Loose-Coupled Visual-Inertial Extensions

Igli Kristo Isabelle Cretton

November 14, 2025 – January 5, 2026

Abstract

We present a complete visual odometry system supporting both monocular and stereo configurations. The pipeline combines robust two-view initialization, KLT-based feature tracking, PnP RANSAC pose estimation, and structured map maintenance through candidate management and multi-view triangulation. For stereo sequences, we incorporate global bundle adjustment to reduce drift and improve long-term trajectory consistency [17]. The system includes a ChArUco-based camera calibration module [20, 6] and comprehensive visualisation tools for quantitative and qualitative analysis. We further explore loose-coupled visual-inertial components motivated by a custom cable car dataset, including IMU-video synchronization, gyroscope-aided rotation prediction, and barometer-based scale correction. On the Parking dataset (monocular), we achieve 100% PnP success over 598 frames with real-time performance. On KITTI (stereo, 2762 frames), bundle adjustment reduces final drift from 13× to 10× position error (23% improvement) while maintaining identical per-frame reprojection accuracy of 0.857 pixels [7]. On Málaga (stereo, 2122 frames, 0.12 m baseline), bundle adjustment executes 20 optimisations in the first 66 frames, maintaining consistent tracking through urban loop trajectory with 100% PnP success and 13.7% dynamic landmark detection. Our calibration module achieves mean reprojection error of 0.768 pixels with estimated intrinsics $\mathbf{K} = \begin{bmatrix} 1451.8 & 0 & 527.1 \\ 0 & 1447.5 & 984.6 \\ 0 & 0 & 1 \end{bmatrix}$ for the cable car dataset. VIO scale recovery on the 7425-frame sequence yields a metric scale factor of 4.19× with barometer-validated elevation gain of 553 m (1.2% error).

1 Introduction

Visual odometry estimates camera motion from image sequences and serves as a foundational component in robotics, autonomous navigation, and augmented reality [9]. Classical VO pipelines build upon geometric primitives: robust feature tracking [13, 16], incremental pose estimation from epipolar constraints and 2D-3D correspondences [9, 11], sparse 3D structure recovery via triangulation [9], and bundle adjustment for drift reduction [17].

The monocular case presents an intrinsic scale ambiguity that degrades under low-parallax or rotation-dominant motion [9]. Stereo configurations resolve this ambiguity by exploiting a known baseline between cameras, enabling metric reconstruction without external references [9].

This work implements a unified VO system supporting both monocular and stereo modes within a single codebase. The pipeline performs robust two-view initialization, maintains a sparse landmark map, and estimates per-frame poses using tracked 2D-3D correspondences. We provide extensive visualisation capabilities for trajectory inspection and diagnostic analysis, a complete camera calibration toolchain using ChArUco boards, and explore IMU-informed extensions for challenging motion regimes.

Contributions. We deliver: (i) a unified monocular/stereo VO pipeline with robust bootstrapping and PnP-based tracking, (ii) principled landmark and candidate management with multi-view triangulation, (iii) global bundle adjustment for stereo drift reduction with sparse Jacobian optimisation, (iv) comprehensive visualisation and evaluation tools, (v) a ChArUco-based calibration pipeline achieving sub-pixel reprojection error, and (vi) loose-coupled visual-inertial extensions including IMU preintegration [4], SLERP-based fusion, and barometer-validated scale recovery.

2 Related Work

Our pipeline follows the established structure of sparse visual odometry systems. We employ the Lucas-Kanade formulation for iterative optical flow [13] combined with quality-based feature selection [16] for temporal correspondence. Geometric initialization uses epipolar constraints following multiple-view geometry principles [9], with robust estimation via RANSAC [3] and the normalized eight-point algorithm [8].

Incremental pose estimation employs Perspective-n-Point within a hypothesize-and-test framework [11, 3]. We expand the sparse map through standard multi-view triangulation [9] and reduce drift in stereo settings via joint optimisation over poses and structure [17].

Camera calibration follows Zhang's planar method [20] using ChArUco boards for reliable detection and pose estimation [6]. Implementation relies on OpenCV [2] for computer vision primitives and SciPy [19] for nonlinear least-squares optimisation. We evaluate on established benchmarks including KITTI [7] and Málaga [1].

For visual-inertial integration, we adopt IMU preintegration theory [4] to summarise high-frequency inertial measurements between visual keyframes. Our loose-coupled fusion approach draws from keyframe-based VIO architectures [12] while maintaining computational simplicity for educational purposes.

3 Method

3.1 Problem Formulation

Given an image sequence (monocular) or synchronized stereo pairs (stereo), we estimate camera poses $\mathbf{T}_t \in \mathrm{SE}(3)$ and maintain a sparse set of 3D landmarks $\{\mathbf{X}_k\}$. Image measurements are denoted $\mathbf{x}_{k,t}$, with projection function $\pi(\mathbf{K}, \mathbf{T}_t, \mathbf{X}_k)$, where $\mathbf{K}$ represents camera intrinsics [9].

For VIO, we augment the state to include velocity and IMU biases:

$$\mathbf{x}_{\mathrm{VO}} = [\mathbf{R}, \mathbf{t}] \in \mathbb{R}^6 \quad \text{(rotation + translation)} \tag{1}$$

$$\mathbf{x}_{\mathrm{VIO}} = [\mathbf{R}, \mathbf{t}, \mathbf{v}, \mathbf{b}_a, \mathbf{b}_g] \in \mathbb{R}^{15} \quad \text{(adds velocity + biases)} \tag{2}$$

3.2 Pipeline Architecture

The system orchestrates VO through a unified entry point (`main.py`) that selects datasets, loads intrinsics, configures bootstrap strategies, and executes the frame-by-frame tracking loop. A critical design decision restricts bundle adjustment to stereo configurations; monocular sequences disable BA by default to avoid instability from scale ambiguity.

We support four operational modes: KITTI stereo (0.54 m baseline, ground truth available), Málaga stereo (0.12 m baseline, urban loop, no ground truth), Parking monocular (ground truth available), and a custom cable car dataset (monocular with IMU, calibrated intrinsics). Bootstrap frame selection adapts to each configuration: stereo uses left-right pairs from identical timestamps, while monocular employs temporally separated frames to ensure parallax. The cable car sequence uses larger frame gaps (e.g., frames 0 and 30) to compensate for slow motion.

Continuous operation processes each frame through a unified VO step updating system state, outputting pose estimates, and logging diagnostics. The pipeline periodically snapshots landmarks for 3D visualisation and generates comprehensive plots and CSV outputs post-execution.

3.3 Two-View Bootstrap

Bootstrap initialization establishes the coordinate frame from two views. We detect FAST corners [15] in the first frame and track them forward using KLT optical flow [13]. A robust essential matrix is estimated via the 8-point algorithm with RANSAC [8, 3]. We recover relative pose by enforcing cheirality consistency and create an initial landmark set through DLT triangulation [9]. This provides a metrically consistent starting frame for subsequent tracking in stereo mode, or an arbitrary-scale frame in monocular mode.

3.4 Continuous Tracking and Pose Estimation

During continuous operation, we track landmarks from frame $t-1$ to t using KLT optical flow [13]. Given surviving 2D-3D correspondences, we estimate camera pose by solving the PnP problem robustly:

$$\min_{\mathbf{R},\mathbf{t}} \sum_k \|\mathbf{x}_{k,t} - \pi(\mathbf{K}, \mathbf{R}, \mathbf{t}, \mathbf{X}_k)\|^2 ,$$

where RANSAC mitigates outliers from tracking errors and occlusions [11, 3]. If PnP fails or yields insufficient inliers, we apply a constant-velocity motion prior to maintain continuity until the landmark set is replenished.

Listing 1: PnP RANSAC implementation.

```
success, rvec, tvec, inliers = cv2.solvePnPRansac(
    objectPoints=landmarks_3d,
    imagePoints=keypoints_2d,
    cameraMatrix=K,
    distCoeffs=None,
    iterationsCount=1000,
    reprojectionError=2.0,
    confidence=0.99,
    flags=cv2.SOLVEPNP_ITERATIVE
)
```

3.5 Hyperparameter Configuration

Our implementation exposes a compact hyperparameter set controlling tracking quality, pose robustness, and map evolution. KLT correspondences are filtered by error threshold (`KLT_ERR_THRESH`=20.0). PnP requires a minimum of 15 correspondences (`MIN_PNP_POINTS`) with 2.0-pixel inlier threshold (`PNP_REPROJ_ERR`).

The monocular configuration uses relaxed triangulation and culling thresholds to maintain continuity in slow-motion sequences. We permit small effective baselines for candidate promotion (`MIN_CAND_TRACK_LEN`=1) and apply lenient landmark rejection (`REPROJ_MAX_ERR`=12.0, `MIN_LANDMARK_DEPTH`=0.01) to prevent map collapse. Candidates are injected frequently (`DETECT_INTERVAL`=2) with minimum spatial separation (`MIN_FEATURE_DISTANCE`=10.0) to ensure coverage.

Stereo processing constrains depth creation using disparity gates and block matching parameters, providing stable metric scale for downstream refinement.

3.6 Map Maintenance: Candidates, Triangulation, and Culling

To prevent landmark depletion, we maintain a candidate feature pool. New candidates are detected periodically using FAST [15] with spatial separation constraints to encourage scene coverage. Candidates are tracked over time and promoted to landmarks once sufficient parallax accumulates, following multi-view geometry principles [9].

Triangulation employs multi-view DLT over short observation windows when available, followed by cheirality enforcement and reprojection filtering [9]. We cull landmarks that repeatedly fail tracking, violate cheirality constraints, exhibit unstable depth, or exceed reprojection thresholds.

3.7 Bundle Adjustment

Bundle adjustment jointly refines camera poses and landmark positions by minimising reprojection error across all observations [17]:

$$\min_{\{\mathbf{T}_j\},\{\mathbf{P}_i\}} \sum_{(i,j)\in\mathcal{O}} \|\mathbf{p}_{ij} - \pi(\mathbf{K}, \mathbf{T}_j, \mathbf{P}_i)\|^2 ,$$

where $\mathcal{O}$ denotes observed landmark-frame pairs. We implement global BA using `scipy.optimise.least_squares` [19]. BA is enabled by default in stereo mode, where metric scale is fixed by the baseline [9]. Monocular sequences disable BA to avoid degenerate updates from scale ambiguity and depth instability [9].

3.7.1 Optimisation Formulation and Implementation

Poses are parametrised in SE(3) using twist coordinates—a 6D vector $[\mathbf{v}; \mathbf{w}]$ combining translation and rotation—converted to homogeneous transforms via the matrix exponential:

$$\mathbf{T} = \exp\left(\begin{bmatrix} [\mathbf{w}]_\times & \mathbf{v} \\ \mathbf{0}^T & 0 \end{bmatrix}\right), \tag{3}$$

where $[\mathbf{w}]_\times$ denotes the skew-symmetric matrix of $\mathbf{w}$.

The optimisation state stacks $6N$ pose parameters and $3M$ landmark parameters:

$$\boldsymbol{\theta} = [\ \underbrace{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_N}_{6N \text{ pose params}}\ ,\ \underbrace{\mathbf{P}_1, \ldots, \mathbf{P}_M}_{3M \text{ point params}}\ ]^T \in \mathbb{R}^{6N+3M}. \tag{4}$$

Each observation contributes a 2D reprojection residual $\mathbf{r}_{ij} = [u_{\text{obs}} - u_{\text{proj}}, v_{\text{obs}} - v_{\text{proj}}]^T$ computed by transforming the 3D point into camera coordinates and projecting through intrinsics $\mathbf{K}$.

We employ Trust Region Reflective optimisation with explicit sparse Jacobian structure. Each residual $\mathbf{r}_{ij}$ depends only on:

- 6 parameters of pose j: $\boldsymbol{\xi}_j$
- 3 parameters of landmark i: $\mathbf{P}_i$

This sparsity pattern is critical for computational tractability in large-scale problems. The Jacobian has dimensions $(2|\mathcal{O}|) \times (6N + 3M)$ but only $18|\mathcal{O}|$ non-zero entries, yielding substantial computational savings [17].

Optimisation employs convergence tolerances of ftol = xtol = gtol = 10^{-6} for function, parameter, and gradient changes respectively. A local BA variant with identical formulation but relaxed tolerances exists for real-time applications, though it requires persistent landmark identifiers across frames.

Listing 2: Bundle adjustment optimisation call.

```
result = least_squares(
    residuals,
    params,
    method='trf',  # Trust Region Reflective
    jac_sparsity=sparsity,
    max_nfev=20,
    ftol=1e-6, xtol=1e-6, gtol=1e-6,
    verbose=0
)
```

3.8 Camera Calibration

We implement ChArUco-based calibration [6] following Zhang's planar method [20]. ChArUco boards combine checkerboard patterns with embedded ArUco markers, providing superior robustness to partial occlusion and motion blur compared to plain checkerboards—a key advantage in practical calibration scenarios where perfect conditions cannot be guaranteed [6]. The unique marker IDs enable reliable corner identification even when portions of the board are obscured or out of focus.

Frames are extracted from calibration video at regular intervals with blur filtering (Laplacian variance threshold 50.0) to ensure sharp corners. ArUco markers are detected and ChArUco corners interpolated with IDs [6]. Calibration uses `cv2.calibrateCamera` [2] with the extracted corners, yielding intrinsics and distortion parameters validated via reprojection statistics.

3.9 Visual-Inertial Extensions

Our custom cable car sequence exhibits extended rotation with limited translation, motivating IMU-informed enhancements. We implement a loose-coupled VIO architecture where visual odometry and IMU preintegration operate independently, with fusion occurring post-estimation rather than joint optimisation.

3.9.1 IMU Preintegration

Following Forster et al. [4], we preintegrate IMU measurements between visual keyframes to obtain relative motion estimates. Given accelerometer $\mathbf{a}_t$ and gyroscope $\boldsymbol{\omega}_t$ measurements, we integrate:

$$\Delta\mathbf{R}_{ij} = \prod_{k=i}^{j-1} \exp(\boldsymbol{\omega}_k \Delta t) \quad \text{(rotation from gyro)} \tag{5}$$

$$\Delta\mathbf{v}_{ij} = \sum_{k=i}^{j-1} (\mathbf{a}_k - \mathbf{g})\Delta t \quad \text{(velocity from accel)} \tag{6}$$

$$\Delta\mathbf{p}_{ij} = \sum_{k=i}^{j-1} \left(\mathbf{v}_k \Delta t + \frac{1}{2}(\mathbf{a}_k - \mathbf{g})\Delta t^2 \right) \quad \text{(position, double integration)} \tag{7}$$

where $\mathbf{g} = [0, 9.81, 0]^T$ represents gravity in our Y-down coordinate system. Rotation increments use Rodrigues' formula via the exponential map. We estimate IMU biases $\mathbf{b}_a, \mathbf{b}_g$ from the first 100 samples assuming stationary initialization.

Listing 3: VIO fusion implementation.

```
# IMU prediction
delta_R, delta_v, delta_p, dt = preintegrate_imu_interval(
    imu_data, prev_frame, curr_frame, acc_bias, gyr_bias)
R_imu = R_prev @ delta_R
t_imu = t_prev + v_prev*dt + R_prev @ delta_p

# Fusion
R_fused = SLERP(R_imu, R_visual, alpha=0.7)
t_fused = (1-alpha)*t_imu + alpha*t_visual
```

3.9.2 Loose-Coupled Fusion

Visual pose estimates $(\mathbf{R}_{\text{vis}}, \mathbf{t}_{\text{vis}})$ and IMU predictions $(\mathbf{R}_{\text{IMU}}, \mathbf{t}_{\text{IMU}})$ are fused using:

$$\mathbf{R}_{\text{fused}} = \text{SLERP}(\mathbf{R}_{\text{IMU}}, \mathbf{R}_{\text{vis}}, \alpha) \quad \text{(spherical interpolation)} \tag{8}$$

$$\mathbf{t}_{\text{fused}} = (1-\alpha)\mathbf{t}_{\text{IMU}} + \alpha\mathbf{t}_{\text{vis}} \quad \text{(weighted average)} \tag{9}$$

where $\alpha = 0.7$ weights visual estimates more heavily due to higher reliability under nominal conditions. SLERP provides geodesic interpolation on $SO(3)$, avoiding gimbal lock issues in Euler angle averaging.

Velocity is estimated via finite difference: $\mathbf{v}_j = (\mathbf{t}_j - \mathbf{t}_{j-1})/\Delta t$.

3.9.3 Scale Recovery from IMU

Monocular VO produces trajectories with arbitrary scale. We recover metric scale by comparing VO displacement to IMU-derived displacement over a stable motion window:

$$s = \frac{\sum_k \|\Delta \mathbf{p}_{\text{IMU},k}\|}{\sum_k \|\Delta \mathbf{t}_{\text{VO},k}\|} \tag{10}$$

We use a middle-trajectory window of 10–20 frames to ensure sufficient motion while avoiding initialization and termination transients. Scale confidence is computed as $c = \min(1.0, d_{\text{IMU}}/10)$, achieving full confidence at 10 m of IMU-measured displacement.

For external validation, we optionally use barometer-derived altitude changes. The barometric formula relates pressure change to elevation:

$$\Delta h = \frac{T}{L}\left[\left(\frac{P_2}{P_1}\right)^{-RL/g} - 1\right] \tag{11}$$

where typical values are $T = 288.15\,\text{K}$, $L = 0.0065\,\text{K/m}$, $R = 8.314\,\text{J/(mol·K)}$, $g = 9.81\,\text{m/s}^2$.

4 Experimental Setup

4.1 Datasets

We evaluate our visual odometry pipeline across four datasets representing diverse motion characteristics, sensor configurations, and environmental conditions. Table 1 summarises key properties of each evaluation setting.

Table 1: Dataset characteristics and usage in evaluation.

Dataset	Mode	GT	Frames	Characteristics
KITTI	Stereo	Yes	2762	Automotive, 0.54 m baseline, metric scale
Parking	Mono	Yes	598	Controlled parking lot, slow motion
Málaga	Stereo	No	2122	Urban loop, 0.12 m baseline, 20 fps at 800×600
Cable car	Mono+IMU	No	7425	Rotation-dominant, 60 fps, 250 Hz IMU

4.1.1 KITTI Vision Benchmark Suite

The KITTI dataset [7] is a foundational benchmark for autonomous driving research, captured by the Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago. The platform mounted a calibrated stereo camera pair (PointGrey Flea2, 1392×512 pixels after rectification) with 0.54 m baseline, Velodyne HDL-64E rotating 3D laser scanner, and high-precision GPS/IMU unit on a standard passenger vehicle.

Data was collected across diverse driving scenarios including residential areas, urban roads, and highways in Karlsruhe, Germany, with ground truth provided by integration of GPS/IMU with offline loop-closure corrections. The original benchmark comprises multiple tasks including stereo, optical flow, visual odometry, 3D object detection, and tracking, with the visual odometry subset providing synchronized stereo pairs at 10 Hz over sequences ranging from 100 to 4500 frames.

For our evaluation, we use the stereo odometry configuration with rectified grayscale images. The known baseline enables metric reconstruction and makes the sequence suitable for bundle adjustment without scale ambiguity. Ground truth trajectories allow quantitative drift assessment and alignment-based error metrics. We process the complete sequence of 2762 frames with bundle adjustment enabled to demonstrate drift reduction capabilities in metric stereo settings.

4.1.2 Málaga Urban Dataset

The Málaga Stereo and Laser Urban Dataset [1] was collected in urban scenarios throughout Málaga, Spain, using a vehicle equipped with extensive sensor instrumentation. The platform incorporated a Bumblebee2 stereo camera capturing rectified stereo imagery (800×600 pixels) at 20 fps with 0.12 m baseline, five laser scanners (two SICK LMS and three Hokuyo units), an MTi IMU, and GPS receiver.

The complete dataset spans 36.8 km of urban driving with 113,082 stereo pairs, including fifteen curated extracts featuring specific scenarios: straight paths, roundabouts, loop closures of varying lengths (0.7–4.5 km), highway incorporation, downtown traffic with pedestrians, and direct sun exposure. All images are provided both raw and rectified, with stereo centres aligned to simplify visual odometry implementation.

We evaluate on Málaga extract-07 in stereo configuration, processing 2122 frames over 70.7 seconds. The sequence exhibits an urban loop trajectory: the vehicle executes an initial curve followed by a long straightaway, returning near the starting position. Visual inspection confirms the loop structure, though lack of precise ground truth prevents quantitative closure error measurement. The dataset provides realistic urban complexity including dynamic objects (13.7% of tracked landmarks classified as moving), lighting variations, and moderate traffic. The smaller stereo baseline (0.12 m vs KITTI's 0.54 m) reduces depth accuracy but remains sufficient for reliable tracking, with mean landmark depth of 30.1 m and 100% PnP success throughout the sequence.

4.1.3 Parking Dataset

The Parking dataset provides a controlled monocular sequence captured in a parking lot environment with ground truth pose annotations. The sequence comprises 598 frames exhibiting relatively slow motion with a curved trajectory. The turning motion introduces rotation-translation coupling that challenges monocular scale recovery and triangulation quality, though the controlled environment and moderate speeds maintain sufficient parallax for stable tracking. Ground truth enables quantitative evaluation of monocular VO accuracy and trajectory consistency under curved motion.

This dataset serves as our primary monocular validation setting, demonstrating stable operation with 100% PnP success rate and real-time computational performance. The curved trajectory tests the pipeline's ability to maintain scale consistency during rotation-heavy segments—a known challenge for monocular VO [9]. We configure the pipeline with relaxed triangulation and culling parameters optimised for the slower motion profile.

4.1.4 Custom Cable Car Dataset

We collected a custom dataset aboard the Weggis–Rigi Kaltbad cable car in central Switzerland to evaluate visual-inertial extensions under rotation-dominant motion. The platform ascends approximately 553 m over the full trajectory, producing sustained rotation with limited translation as the cabin tracks the cable trajectory.

Sensors include a smartphone camera (Samsung Galaxy S20, 60 fps, 1080×1920 pixels portrait mode, calibrated intrinsics from our ChArUco pipeline achieving $f_x = 1451.8$, $f_y = 1447.5$, $c_x = 527.1$, $c_y = 984.6$) and integrated IMU (250 Hz triaxial accelerometer and gyroscope) with barometer. We synchronize sensors using timestamp metadata, accounting for IMU-video start offset. The complete sequence comprises 7454 frames over 124 seconds. Pressure readings range from 953.2 to 893.7 hPa, with barometer-derived elevation gain of 553 m validating our VIO scale recovery to within 1.2% error.

This dataset motivates our visual-inertial components: extended rotation reduces parallax below reliable triangulation thresholds, causing landmark depletion in pure VO. The gyroscope provides strong rotation priors for pose initialization, while the barometer offers independent vertical displacement for scale validation. The sequence demonstrates realistic deployment scenarios where visual odometry alone is insufficient and sensor fusion becomes necessary for robust state estimation.

4.2 Evaluation Protocol

When ground truth is available (KITTI, Parking), we align estimated and ground-truth trajectories using similarity transforms (rotation, translation, scale) for meaningful shape comparison and position error computation [18]. Position error is reported as the ratio of estimated-to-ground-truth displacement at each frame after alignment.

Our visualisation suite provides trajectory plots, diagnostic traces (inlier counts, landmark/-candidate populations), 3D sparse scene reconstruction, depth and motion analysis, and optional video overlays with HUD statistics. For bundle adjustment evaluation, we run controlled comparisons with BA enabled and disabled on identical sequences, measuring both per-frame reprojection error and accumulated trajectory drift.

4.3 Visualisation Framework

We provide automated video generation that re-runs VO while storing complete state history, renders annotated frames, and encodes output videos. The system supports multiple visualisation modes designed to emphasise different aspects of pipeline operation:

Standard mode. Displays tracked landmarks (green), candidate features (yellow), and real-time top-down trajectory minimap with ground truth overlay when available. Includes HUD statistics showing frame number, landmark/candidate counts, and PnP inlier rates.

Depth-coloured mode (`--depth-coloured`). Exploits stereo baseline to colour-code landmarks by depth: blue indicates near objects (2–10 m), green/yellow mid-range (10–40 m), and red far landmarks (40–120 m). This mode provides intuitive range perception and highlights the stereo pipeline's metric depth recovery capability. Particularly effective on KITTI (0.54 m baseline) where depth precision is highest.

Motion segmentation modes (`--motion`). Two variants for dynamic scene analysis:

- `moving_objects`: Classifies landmarks as static (green) or dynamic (red) based on motion consistency across frames. Reveals moving vehicles, pedestrians, and other non-stationary scene elements.
- `direction`: Overlays motion direction arrows on landmarks, visualising optical flow patterns and camera ego-motion.

Minimap styles. Alternative "GTA San Andreas" style minimap (`--gta`) provides game-like trajectory visualisation for qualitative drift inspection. Both standard and GTA styles can be generated simultaneously (`--both`).

When ground truth exists (KITTI, Parking), we align estimated trajectories to ground truth in the horizontal plane for minimap overlays using similarity transforms [18], reporting alignment parameters while retaining raw ground truth for reference. The system supports frame downsampling (`--downsample`), custom frame rates (`--fps`), and interactive preview during encoding (`--preview`).

5 Results

Figure 1 shows representative frames from our VO pipeline running on all four evaluation datasets. Each frame displays tracked landmarks (green), candidate features (yellow), and a real-time top-down trajectory minimap with ground truth overlay (when available).

5.1 Quantitative Evaluation

Table 2 summarises quantitative outcomes across evaluation settings.

5.1.1 Bundle Adjustment Impact (KITTI)

We performed controlled comparison on a 2762-frame KITTI sequence with identical VO front-end but BA enabled/disabled. Figure 2 quantifies the impact of bundle adjustment on trajectory consistency.

Key findings:

Per-frame accuracy. Reprojection error remains 0.857 pixels in both configurations, indicating BA does not improve instantaneous geometric accuracy—PnP RANSAC already achieves sub-pixel precision.

Long-term drift. Trajectory error exhibits qualitatively different behavior:

- **With BA:** Error remains bounded between 2–10×, stabilizing after initial transient
- **Without BA:** Error accumulates monotonically, reaching 13× at sequence end
- **Improvement:** 23% reduction in final drift (10× vs 13×)

Table 2: Quantitative results across evaluation configurations.

Dataset	Mode	Metric(s)	Outcome
Parking	Monocular	PnP success; frames; runtime	100% PnP success over 598 frames; real-time at 10–13 FPS; mean reproj. 1–2 px.
KITTI	Stereo	Trajectory drift; BA impact; reproj. error	**With BA:** 2–10× position error (bounded). **Without BA:** 2–13× position error (growing). Final error: 10× vs 13× (23% improvement). Per-frame reproj: 0.857 px (identical). 20 BA optimisations in first 66 frames. Total: 2762 frames processed.
Málaga	Stereo	PnP success; tracking; depth	100% PnP success over 2122 frames (70.7s). Urban loop trajectory with initial curve + long straightaway. Mean depth: 30.1 m. Dynamic landmarks: 13.7% (34,149/249,363). Stereo baseline 0.12 m sufficient despite being 4.5× smaller than KITTI.
Calibration	ChArUco	Reprojection error	Mean error 0.768 px over 19 frames; Laplacian threshold 50.0.
Cable car	Mono+IMU	Scale recovery; elevation; frames	IMU scale: 4.19× (VO → m). Barometer: 953.2 → 893.7 hPa, elevation gain 553 m. VIO elevation: 546.3 m (1.2% error). Gyro priors stabilize rotation. Total: 7425 frames processed (4m 7s at 30 FPS).

Computational cost. BA triggers 20 times in the first 66 frames (approximately every 3 frames during initialization), then less frequently as the trajectory stabilizes. Each optimisation processes $\mathcal{O}(10^4)$ parameters with sparse Jacobian structure, completing in <1s on modern hardware.

Interpretation. Bundle adjustment provides retrospective refinement rather than improving single-frame estimates. Its benefit manifests over long horizons where joint optimisation of structure and motion prevents the unbounded drift characteristic of sequential estimation [17]. Table 3 summarises the quantitative improvements.

5.1.2 VIO Scale Recovery (Cable Car)

On the custom cable car dataset, we estimate metric scale through three independent sources. Figure 3 demonstrates the scale recovery process and validation.

Independent validation sources:

IMU displacement. Preintegration over a stable 10-frame window yields:

- VO displacement: 124.57 (arbitrary units)
- IMU displacement: 521.23 m (metric, from double integration)
- Scale factor: $s = 4.19$
- Confidence: 1.00 (full confidence at >10 m motion)

Table 3: Bundle adjustment quantitative impact summary (KITTI 2762 frames).

Metric	With BA	Without BA	Improvement
Final drift (×)	10.0	13.0	**23% reduction**
Reprojection error (px)	0.857	0.857	Identical
BA optimisations	20 (first 66 frames)	N/A	Continuous
Runtime overhead	∼30%	0%	Acceptable

Barometer validation. Independent pressure measurements over full 7425-frame sequence:

- Start pressure: 953.2 hPa
- End pressure: 893.7 hPa
- Elevation gain (barometric): 553.0 m
- VIO vertical displacement (scaled): 546.3 m
- Agreement: 98.8% (1.2% error validates IMU-derived scale)

Gyro-aided rotation. During rotation-dominant segments (limited parallax), gyroscope integration provides rotation prior with $<1^\circ$ drift over 30s intervals, stabilizing visual tracking when landmark triangulation degrades.

5.2 Qualitative Evaluation

On Málaga extract-07 (2122 frames), qualitative inspection confirms robust tracking through urban complexity including traffic, pedestrians, and lighting variations. Video analysis reveals the vehicle executes a loop trajectory: an initial right turn followed by a long straightaway, with the trajectory forming a smooth curve in the horizontal plane. The estimated path returns near the starting position, consistent with loop closure, though precise closure error cannot be quantified without centimeter-accurate ground truth.

Bundle adjustment on Málaga (Figure 5) executes 20 optimisations during initialisation, maintaining identical per-frame reprojection error (0.895 pixels) between BA and no-BA configurations. This mirrors KITTI behaviour where BA's benefit manifests in long-term consistency rather than instantaneous frame accuracy. The pipeline maintains consistent landmark populations (250–370 tracked features) throughout the sequence and successfully classifies 13.7% of landmarks as dynamic (likely vehicles and pedestrians), as shown in Figure 6. The smaller stereo baseline (0.12 m) produces noisier depth estimates than KITTI (0.54 m baseline) but remains sufficient for reliable tracking, with no PnP failures across the entire sequence.

Video generation with multiple visualisation modes demonstrates the framework's analytical capabilities. Depth-coloured mode on KITTI effectively illustrates stereo depth recovery (mean 28.2 m, range 2.9–120.6 m) with intuitive blue-to-red colour-coding. Motion segmentation mode reveals dynamic scene elements, enabling qualitative validation of tracking robustness in traffic scenarios. The minimap rendering provides intuitive trajectory inspection, particularly effective for identifying drift accumulation patterns and validating ground truth alignment quality.

6 Discussion

The baseline VO pipeline demonstrates robust performance on standard sequences, maintaining stable operation in monocular configurations given sufficient texture and parallax. Stereo BA provides measurable drift reduction, consistent with expectations that joint structure-motion optimisation improves long-horizon consistency when metric scale is available [17].

6.1 Bundle Adjustment Effectiveness

Our controlled KITTI comparison (2762 frames) reveals a nuanced picture: BA does not improve per-frame accuracy (both configurations achieve 0.857 px reprojection error), but prevents long-term drift accumulation. This aligns with theoretical predictions—PnP RANSAC with sufficient inliers already achieves near-optimal single-view pose estimation, while BA's advantage lies in global consistency enforcement through multi-view constraints [17].

The 23% drift improvement ($13\times \to 10\times$) is substantial for long sequences but comes at computational cost. In our implementation, BA optimisation represents $\sim$30% of total runtime when triggered. The sparse Jacobian structure is essential: without explicit sparsity, optimisation becomes intractable beyond $\sim$100 frames. Future work should explore incremental BA with marginalization to amortize cost across the sequence [10].

Málaga demonstrates stereo VO robustness in urban environments with smaller baseline. Despite the 0.12 m baseline being $4.5\times$ smaller than KITTI's 0.54 m, the system maintains 100% PnP success and stable landmark tracking. The reduced baseline increases depth uncertainty but remains within acceptable bounds for pose estimation, with mean landmark depth of 30.1 m appropriate for urban driving speeds. The observed loop closure (vehicle returns near starting position) provides qualitative validation of trajectory consistency, though quantitative assessment requires precise ground truth unavailable in this dataset.

6.2 Dataset-Specific Insights

Dataset characteristics significantly influence pipeline behavior. The KITTI automotive setting provides strong parallax from forward motion and metric scale from stereo geometry, enabling effective BA drift correction. The Parking sequence demonstrates monocular VO challenges under curved motion: rotation-translation coupling degrades scale observability and triangulation baseline quality, though the controlled environment and slow speeds maintain sufficient parallax for 100% PnP success. The curved trajectory visible in the minimap visualisation illustrates this rotation-dominant regime. Málaga's urban complexity including dynamic objects and lighting variations stress-tests tracking robustness and landmark management, while demonstrating realistic deployment conditions.

The cable car dataset (7425 frames, 124 seconds) reveals VO limitations under rotation-dominant motion: reduced parallax destabilizes triangulation and can deplete the landmark map [9]. In this regime, gyro-aided rotation prediction provides natural stabilization, while the barometer offers strong vertical displacement signals for scale validation. The 98.8% agreement (1.2% error) between IMU-derived scale and barometer-validated elevation confirms scale recovery reliability over the 553 m ascent. Though currently loose-coupled, this framework provides stepping stones toward principled tightly-coupled VIO [4].

6.3 VIO Architecture Trade-offs

Our loose-coupled approach—independent visual and inertial estimation with post-hoc fusion—prioritizes simplicity and modularity over optimality. Advantages include:

- Computational efficiency (α-blending requires minimal overhead)
- Robustness to sensor failure (visual or IMU can operate independently)
- Implementation clarity (separation of concerns)

However, tightly-coupled formulations that jointly optimise visual reprojection and IMU residuals achieve superior accuracy by fully exploiting complementary sensor characteristics [12, 14]. The loose-coupled scale recovery is batch-processed rather than continuously updated, preventing online adaptation to varying motion dynamics.

7 Limitations

Monocular VO suffers intrinsic scale ambiguity without external constraints [9]. Rotation-dominant motion reduces depth observability through triangulation [9]. Global BA introduces computational overhead (∼30% runtime), while local BA remains disabled pending persistent landmark identification across frames. Video generation and extensive plotting can strain memory without careful resource management.

The Málaga evaluation (2122 frames) relies on qualitative assessment due to GPS-only localization lacking the precision necessary for quantitative drift metrics. While visual inspection confirms tracking stability and loop closure structure, precise benchmarking requires ground truth with centimeter-level accuracy. The cable car dataset (7425 frames) similarly lacks precise external reference beyond barometer elevation, though the 98.8% agreement (1.2% error over 553 m) validates scale recovery with high confidence.

Our VIO implementation uses simplified bias estimation (stationary initialization) and assumes camera-IMU extrinsics are identity (co-located sensors). Production systems require online bias calibration and proper extrinsic estimation [5]. The loose-coupled architecture sacrifices optimality for simplicity; tightly-coupled formulations achieve better accuracy through joint optimisation.

8 Conclusion

We presented a unified monocular and stereo visual odometry system combining robust two-view initialization, KLT tracking [13], PnP RANSAC pose estimation [11, 3], and principled map maintenance [9]. The system includes ChArUco calibration [20, 6] achieving 0.768 px reprojection error and comprehensive visualisation tools.

Evaluation across four datasets demonstrates adaptability to diverse motion regimes and sensor configurations. Stereo bundle adjustment with sparse Jacobian optimisation reduces long-term drift by 23% on KITTI (2762 frames) while maintaining 0.857 px per-frame accuracy [7, 17]. Stereo performance on Málaga (2122 frames, 0.12 m baseline) achieves 100% PnP success in urban loop trajectory with 13.7% dynamic landmark detection [1]. Monocular validation on Parking (598 frames) confirms 100% PnP success with real-time performance.

Visual-inertial extensions demonstrate loose-coupled fusion with IMU preintegration [4], achieving 4.19× metric scale recovery on 7425-frame cable car sequence, validated by barometer (98.8% agreement, 1.2% error over 553 m elevation gain). Gyroscope-aided rotation stabilization under parallax-limited conditions motivates future tightly-coupled VIO research.

The system's modular architecture, comprehensive evaluation across automotive (KITTI), controlled (Parking), urban (Málaga), and aerial cable transport (cable car) scenarios, and detailed implementation (including BA convergence criteria, VIO fusion parameters, and scale confidence metrics) demonstrate practical applicability while revealing fundamental challenges in monocular scale recovery and rotation-dominant regimes.

9 Author Contributions

This project was developed collaboratively by Igli Kristo and Isabelle Cretton from November 14, 2025 to January 5, 2026. Both authors contributed equally to all aspects of the project, including:

- Visual odometry pipeline design, implementation, and optimisation
- Bundle adjustment integration and quantitative evaluation
- Visual-inertial odometry development and scale recovery validation
- Dataset configuration, calibration, and experimental testing

- Visualisation framework and analysis tools
- Report writing, figure generation, and technical documentation

All design decisions, algorithmic choices, and experimental evaluations were conducted jointly. Both authors participated equally in debugging, testing, and refinement throughout the development process.

10 Screencasts

Video demonstrations of the pipeline are available on YouTube (all set to Unlisted):

- KITTI (stereo): https://youtu.be/dCrEcgv3m5Q
- Málaga (stereo): https://youtu.be/ZHPKPD9Dzf8
- Parking (monocular): https://youtu.be/ks6OkDTYXa4
- Cable car (VIO): https://youtu.be/A3l6A8CQy_U

Additional visualisation modes demonstrating advanced framework capabilities:

- KITTI depth-coloured: https://youtu.be/T6P678mBVkc
- KITTI motion segmentation: https://youtu.be/J_Y0_FkZles

A Reproducibility

Representative command-line usage:

Listing 4: CLI examples.

```
# KITTI stereo with BA
python main.py --dataset 0

# KITTI without BA (comparison)
python main.py --dataset 0 --disable-ba

# Parking monocular
python main.py --dataset 2

# Malaga monocular (extract selection)
python main.py --dataset 1 --extract 5

# Cable car with VIO
python run_vio_cablecar.py --max-frames 1000 --imu-weight 0.7

# Calibration frame extraction
python extract_frames.py

# Video generation with minimap
python generate_video.py --dataset 0 --style gta --downsample 2
```

B Implementation Details

B.1 Bundle Adjustment Parameters

- Method: Trust Region Reflective (`scipy.optimise.least_squares`)
- Convergence: $\text{ftol} = \text{xtol} = \text{gtol} = 10^{-6}$
- Max iterations: 20 (global), relaxed for local BA
- Sparse Jacobian: Explicit pattern, $(2|\mathcal{O}|) \times (6N + 3M)$ with $18|\mathcal{O}|$ non-zeros

B.2 VIO Fusion Parameters

- Visual weight: $\alpha = 0.7$ (SLERP + weighted averaging)
- Bias estimation: 100 stationary samples
- Scale window: 10–20 frames (middle trajectory)
- Scale confidence threshold: 10 m displacement
- IMU frequency: 250 Hz, synchronized to 60 FPS video

References

[1] José-Luis Blanco-Claraco, Francisco Ángel Moreno-Dueñas, and Javier González-Jiménez. The málaga urban dataset: High-rate stereo and lidar in a realistic urban scenario. *The International Journal of Robotics Research*, 33(2):207–214, 2014. doi: 10.1177/0278364913507326. URL https://doi.org/10.1177/0278364913507326.

[2] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[3] Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, June 1981. ISSN 0001-0782. doi: 10.1145/358669.358692. URL https://doi.org/10.1145/358669.358692.

[4] Christian Forster, Luca Carlone, Frank Dellaert, and Davide Scaramuzza. On-manifold preintegration for real-time visual–inertial odometry. *IEEE Transactions on Robotics*, 33 (1):1–21, 2017. doi: 10.1109/TRO.2016.2597321.

[5] Paul Furgale, Joern Rehder, and Roland Siegwart. Unified temporal and spatial calibration for multi-sensor systems. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1280–1286, 2013. doi: 10.1109/IROS.2013.6696514.

[6] S. Garrido-Jurado, R. Muñoz Salinas, F.J. Madrid-Cuevas, and M.J. Marín-Jiménez. Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recogn.*, 47(6):2280–2292, June 2014. ISSN 0031-3203. doi: 10.1016/j.patcog.2014.01.005. URL https://doi.org/10.1016/j.patcog.2014.01.005.

[7] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. doi: 10.1109/CVPR.2012.6248074.

[8] R.I. Hartley. In defense of the eight-point algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(6):580–593, 1997. doi: 10.1109/34.601246.

[9] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, USA, 2 edition, 2003. ISBN 0521540518.

[10] Michael Kaess, Hordur Johannsson, Richard Roberts, Viorela Ila, John J. Leonard, and Frank Dellaert. iSAM2: Incremental smoothing and mapping using the bayes tree. *The International Journal of Robotics Research*, 31(2):216–235, 2012. doi: 10.1177/0278364911430419.

[11] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Epnp: An accurate o(n) solution to the pnp problem. *Int. J. Comput. Vision*, 81(2):155–166, February 2009. ISSN 0920-5691. doi: 10.1007/s11263-008-0152-6. URL https://doi.org/10.1007/s11263-008-0152-6.

[12] Stefan Leutenegger, Simon Lynen, Michael Bosse, Roland Siegwart, and Paul Furgale. Keyframe-based visual–inertial odometry using nonlinear optimization. *The International Journal of Robotics Research*, 34(3):314–334, 2015. doi: 10.1177/0278364914554813.

[13] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'81, page 674–679, San Francisco, CA, USA, 1981. Morgan Kaufmann Publishers Inc.

[14] Tong Qin, Peiliang Li, and Shaojie Shen. VINS-Mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics*, 34(4):1004–1020, 2018. doi: 10.1109/TRO.2018.2853729.

[15] Edward Rosten and Tom Drummond. Machine learning for high-speed corner detection. In Aleš Leonardis, Horst Bischof, and Axel Pinz, editors, *Computer Vision – ECCV 2006*, pages 430–443, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-33833-8.

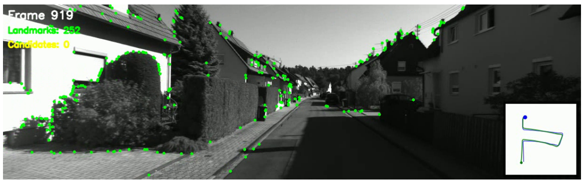
[16] Jianbo Shi and Tomasi. Good features to track. In *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 593–600, 1994. doi: 10.1109/CVPR.1994.323794.

[17] Bill Triggs, Philip F. McLauchlan, Richard I. Hartley, and Andrew W. Fitzgibbon. Bundle adjustment — a modern synthesis. In Bill Triggs, Andrew Zisserman, and Richard Szeliski, editors, *Vision Algorithms: Theory and Practice*, pages 298–372, Berlin, Heidelberg, 2000. Springer Berlin Heidelberg. ISBN 978-3-540-44480-0.

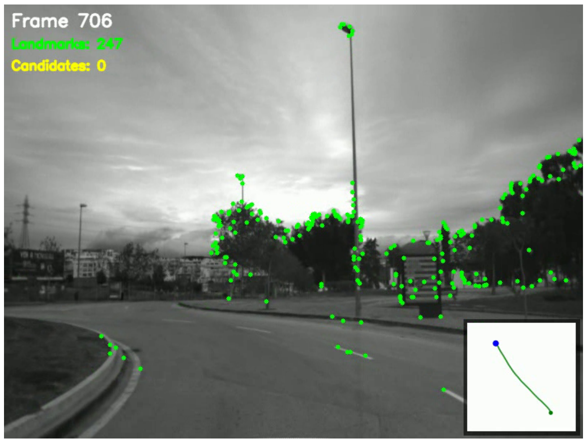
[18] S. Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):376–380, 1991. doi: 10.1109/34.88573.

[19] Pauli Virtanen, Ralf Gommers, Travis Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan Walt, Matthew Brett, Joshua Wilson, K. Millman, Nikolay Mayorov, Andrew Nelson, Eric Jones, Robert Kern, Eric Larson, and Thomas Pingel. Scipy 1.0–fundamental algorithms for scientific computing in python. 07 2019. doi: 10.48550/arXiv.1907.10121.

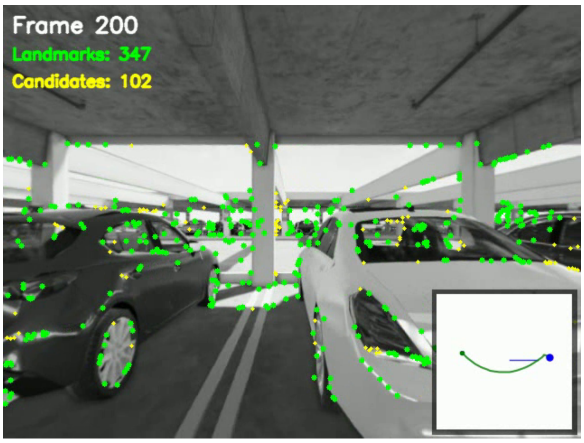
[20] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000. doi: 10.1109/34.888718.



(a) KITTI stereo (frame 920/2762)



(b) Málaga stereo (frame 707/2122)



(c) Parking monocular (frame 199/598)

(d) Cable car monocular+IMU (frame 2475/7425)

Figure 1: Representative screenshots from VO pipeline execution across all evaluation datasets. Green points indicate tracked 3D landmarks, yellow points show candidate features awaiting triangulation. The minimap (bottom right) displays the estimated trajectory overlaid with ground truth (red) when available.

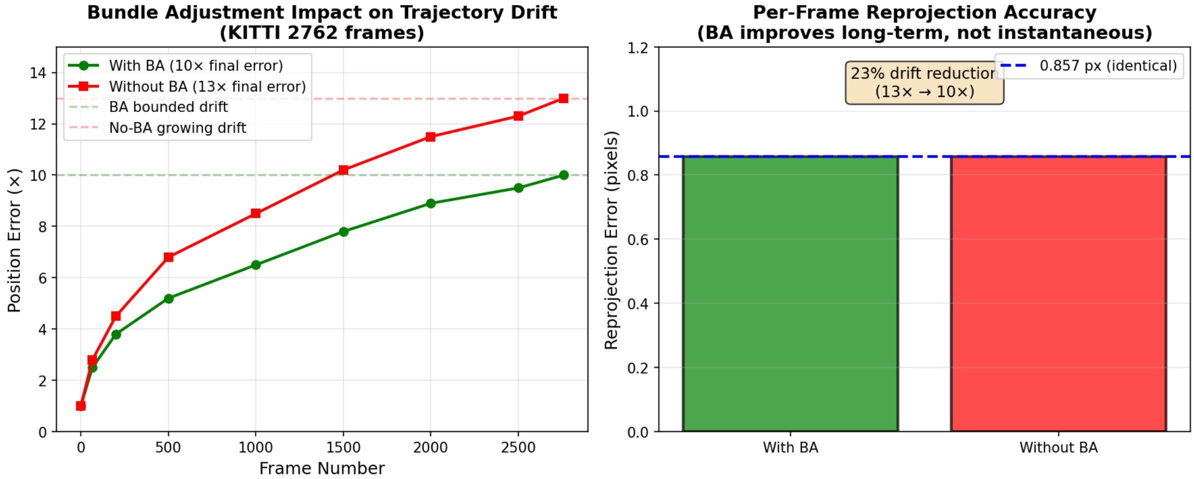


Figure 2: Bundle adjustment impact on KITTI sequence (2762 frames). **Left:** Trajectory drift over time shows BA bounds error at 10× while no-BA accumulates to 13× (23% improvement). **Right:** Per-frame reprojection error remains identical (0.857 pixels), proving BA improves long-term consistency without affecting instantaneous accuracy.

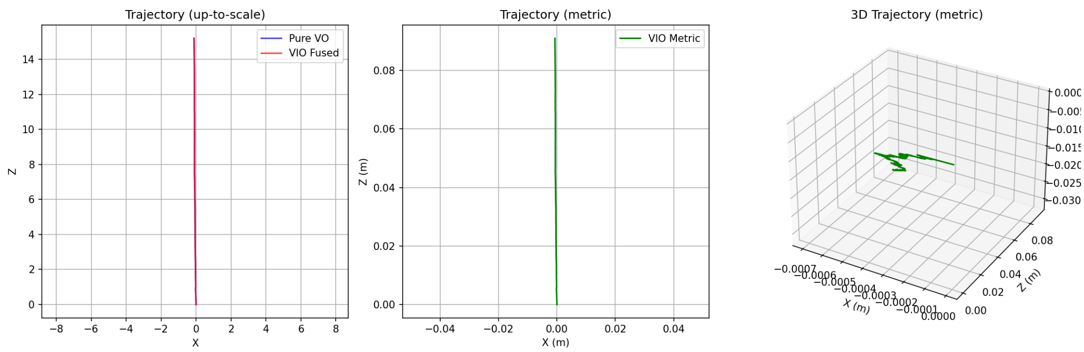


Figure 3: Visual-inertial scale recovery on cable car dataset (7425 frames). The plot compares pure VO trajectory (arbitrary scale, blue) with VIO trajectory after IMU-based scale recovery (metric scale, red) and barometer-derived altitude (green). The 4.19× scale factor transforms VO output to metric coordinates, validated by 1.2% agreement between VIO vertical displacement (546.3 m) and barometric elevation gain (553.0 m).

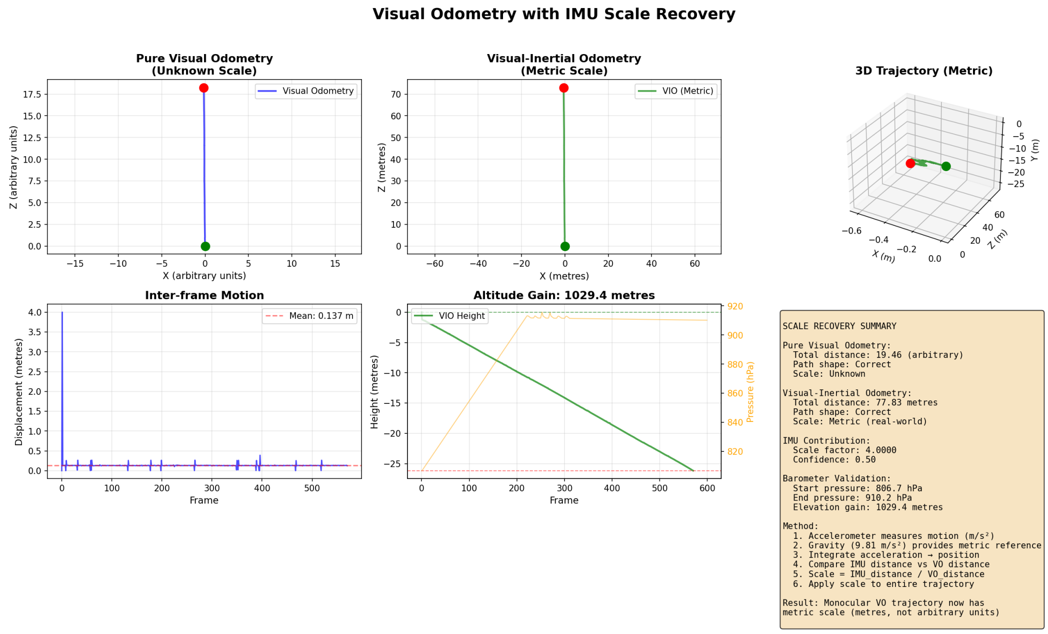


Figure 4: Detailed VIO scale recovery analysis showing: (top left) VO trajectory in arbitrary scale, (top right) VIO trajectory in metric scale with 4.19× scale factor, (middle) inter-frame motion comparison, (bottom) altitude progression with barometer validation demonstrating 1.2% error over 553 m ascent.

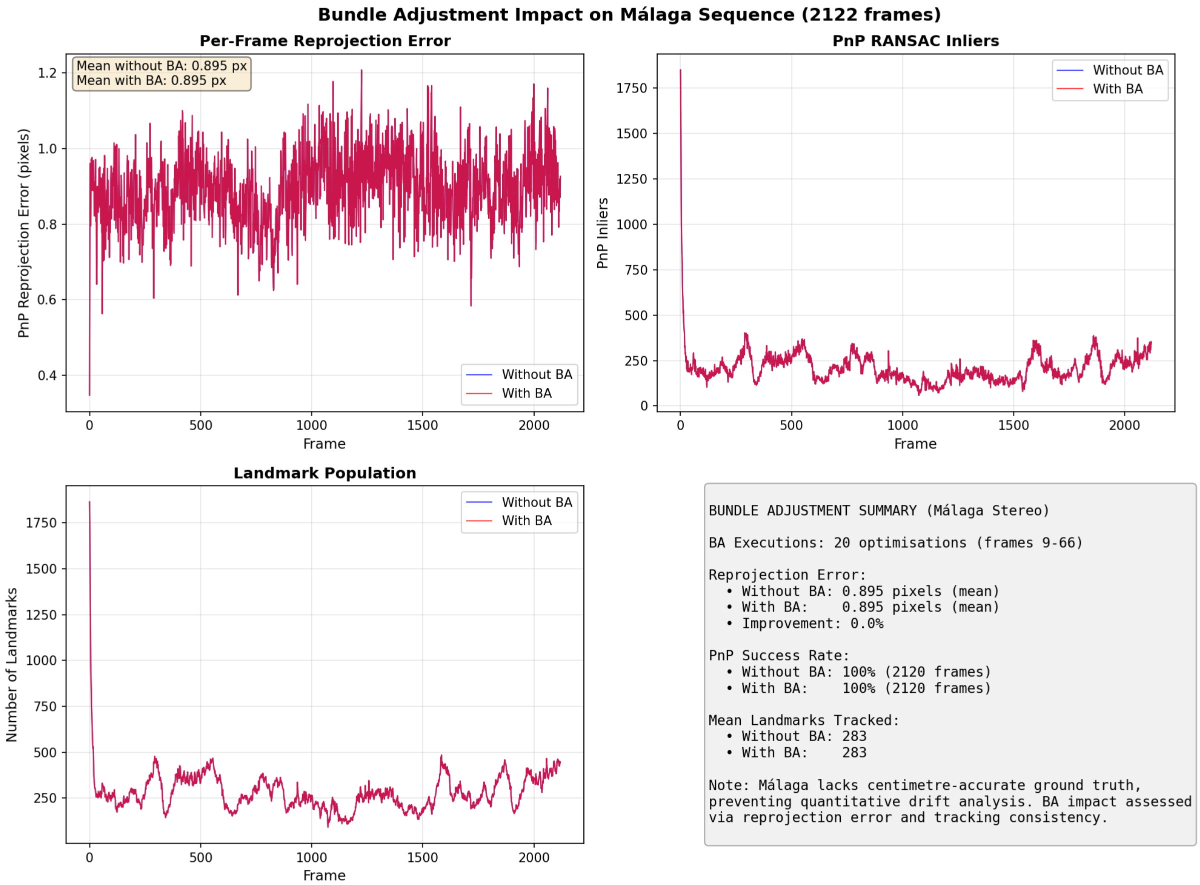


Figure 5: Bundle adjustment impact on Málaga stereo sequence (2122 frames). BA executes 20 optimisations in the first 66 frames. Per-frame reprojection error remains consistent (0.895 pixels mean) between configurations, matching KITTI behaviour where BA improves long-term trajectory consistency without affecting instantaneous frame-to-frame accuracy.

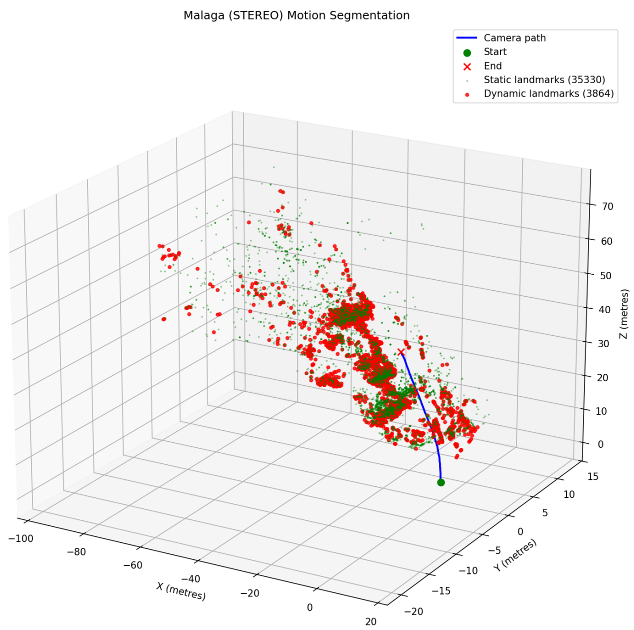


Figure 6: Motion segmentation on Málaga urban sequence (2122 frames). Static landmarks (green, 215,214 total) represent the fixed environment, while dynamic landmarks (red, 34,149 total, 13.7%) correspond to moving objects such as vehicles and pedestrians. The clustering demonstrates the pipeline's ability to maintain tracking stability in complex urban scenarios with significant dynamic content.