Camera Pose Estimation from Depth and Point Tracking

Deboparna Banerjee
CSE 527 Intro to Computer Vision (Supervised by Lei Zhou)

1 Introduction

The core objective of the project is to accurately estimate the camera's motion and concurrently reconstruct a sparse three-dimensional (3D) structure of the scene from a handheld video sequence. The methodology is centred on leveraging modern feature-tracking and depth prediction models to compute consistent relative poses between consecutive frames. The success criteria for the project are defined by achieving three critical goals: maintaining temporally stable depth, ensuring accurate frame-to-frame geometry, and producing a global camera trajectory with low accumulated drift.

2 Technical Challenges

Several inherent challenges complicate the reliable estimation of camera pose:

1. **Point Selection:** A significant challenge lies in judiciously identifying which tracked feature points, whether they are dense or sparse, corner-like or highly textured, will yield stable and reliable geometric constraints necessary for accurate pose computation.

2. **Depth Reliability:** The system must ensure that the predicted depth is sufficiently consistent to enable precise 3D lifting of 2D image points into a 3D coordinate system. This reliability is difficult to maintain given noise, scale fluctuations inherent in monocular depth prediction, and artifacts like texture bleeding.

3. **Geometric Consistency:** Maintaining a stable frame-to-frame pose is critical, as noisy correspondences can lead to PnP (Perspective-n-Point) instability, which, in turn, causes the undesirable accumulation of error, commonly known as drift, in the global trajectory.

3 Methodology

The implemented methodology is a pipeline combining state-of-the-art deep learning models for feature tracking and depth estimation with robust classical geometric methods for pose solving.

3.1 Feature Tracking and 3D Lifting

The process begins with robust feature detection and tracking:

- **Keypoint Detection:** Stable corner features are detected using **SuperPoint Keypoints** to serve as reliable seeds for long-term tracking.
- **Trajectory Propagation:** These keypoints, along with grid samples, are propagated across multiple frames using **CoTracker Tracks** to obtain dense and temporally consistent 2D trajectories.
- **Depth Prediction:** Metric depth is predicted for each frame using the **VDA Depth** model. This predicted depth is essential for performing the 3D lifting operation—converting the 2D tracked coordinates to 3D world points—when the camera's motion includes translation.

3.2 Motion Analysis and Pose Solving

After obtaining consistent 2D tracks and per-frame depth, the methodology branches based on motion type:

- **Correspondence Filtering:** To ensure the validity of matches used for pose calculation, only tracks that are visible in **both** the current and reference frame are extracted. This provides a set of valid 2D-2D and 3D-2D correspondences.
- **Parallax Check:** The first step is the Parallax Check, where normalised parallax is measured between consecutive frames. This measure classifies the camera's motion as either **rotation-only** (low parallax) or **translational** (high parallax).
- **Rotation-Only Mode (Low Parallax):** For motion exhibiting low parallax, the system estimates rotation using the geometric constraints of an **Essential matrix**. To improve stability, the rotation estimate is refined through a sliding-window averaging technique, and the translation component is forcibly set to zero.
- **PnP Translation Mode (High Parallax):** When sufficient parallax is detected, indicating translational motion, the system utilises the 3D points derived from depth along with the 2D tracked points to solve the pose using the **PnP (Perspective-n-Point)** algorithm. This step is made robust against outliers by incorporating the **RANSAC (Random Sample Consensus)** framework.

3.3 The Perspective-n-Point (PnP) Solver

The Perspective-n-Point (PnP) problem estimates the relative camera pose

$$\mathbf{T}_{k-1\to k} = \begin{bmatrix}\mathbf{R} & \mathbf{t}\end{bmatrix} \in SE(3) \tag{1}$$

where $\mathbf{R} \in SO(3)$ is the rotation matrix and $\mathbf{t} \in \mathbf{R}^3$ is the translation vector.

Principle: Given correspondences between known 3D points $\mathbf{X}_i = (X_i, Y_i, Z_i)^\top$ in the previous camera coordinate system and their observed 2D projections $\mathbf{x}_i = (u_i, v_i)^\top$ in the current frame, the goal is to recover the camera pose $(\mathbf{R}, \mathbf{t})$.

Projection Model: Each correspondence must satisfy the pinhole projection equation

$$\lambda_i \begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} = \mathbf{K}\left(\mathbf{R}\mathbf{X}_i + \mathbf{t}\right), \quad \lambda_i > 0 \tag{2}$$

where $\mathbf{K}$ represents the known camera intrinsics.

Minimal Solution: The camera pose has 6 degrees of freedom (3 for rotation, 3 for translation). Different PnP algorithms require different minimal sets: P3P requires **three non-collinear correspondences**, while iterative methods typically use four or more points.

RANSAC for Robustness: To handle noise and outliers, PnP is embedded within RANSAC:

1. Randomly sample a minimal set of correspondences and compute candidate pose(s).
2. Reproject all 3D points and measure reprojection error $e_i = \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2$.
3. Classify points with $e_i < \tau$ as inliers.
4. Select the pose maximizing inliers, then refine via nonlinear optimization.

3.4 Global Trajectory Integration

The final, overall camera path is generated by sequentially composing the relative transforms calculated for each frame pair. The global pose $\mathbf{T}_{0\to k}$ at frame k with respect to the initial frame 0 is computed as:

$$\mathbf{T}_{0\to k} = \mathbf{T}_{0\to 1}\mathbf{T}_{1\to 2}\cdots\mathbf{T}_{k-1\to k} \tag{3}$$

This process starts at the origin and applies each newly calculated motion atop the previously computed global pose. The critical consequence of this chaining operation is that small, unavoidable errors in the per-frame pose estimates accumulate over time, leading directly to the phenomenon of global trajectory drift.

4 Implementation

4.1 Test Sequences

The proposed methodology was implemented and validated on three representative video sequences captured under diverse motion conditions. Each sequence was carefully selected to evaluate specific aspects of the pose estimation pipeline: translational motion with complex trajectories, ideal straight-line movement with strong parallax, and the degenerate case of pure rotation. These scenarios collectively provide comprehensive coverage of the motion types encountered in typical visual odometry applications.

- **Video Sample 1 — General Translational Motion:** This sequence captures a handheld camera following a U-shaped trajectory around a tabletop scene. The motion combines dominant translation with moderate rotation, representing typical exploratory camera movement and testing trajectory coherence under varying parallax.

- **Video Sample 2 — Straight-Line Trajectory:** This sequence represents near-ideal conditions for PnP-based pose estimation, with predominantly forward translation and minimal rotation. Strong parallax, stable feature tracks, and well-conditioned depth enable highly accurate pose recovery, serving as a performance upper bound.
- **Video Sample 3 — Pure Rotation Sequence:** This sequence evaluates system behavior under geometrically degenerate conditions where the camera rotates in place with negligible translation. The absence of parallax makes this a critical test case for validating the rotation-only handling mechanism and motion classification strategy.

4.2 Performance Evulation

Performance evaluation was conducted by comparing the estimated trajectories against reference ground truth obtained using Pi3, a state-of-the-art visual-inertial odometry system. The evaluation metrics follow standard visual odometry protocols and include:

- **Absolute Pose Error in Translation (APE-Tran):** Measures the global consistency of the estimated trajectory by computing the Euclidean distance between estimated and reference camera positions at each frame.
- **Relative Pose Error in Rotation (RPE-Rot):** Quantifies frame-to-frame rotational accuracy by measuring angular deviations between consecutive pose estimates, expressed in degrees.
- **Relative Pose Error in Translation (RPE-Tran):** Assesses the local translational consistency by computing distance errors between relative motions over fixed frame intervals.

Root Mean Square Error (RMSE) is adopted as the primary evaluation metric for all error measurements, as it provides a robust single-value summary that penalizes larger deviations more heavily than mean error alone.

4.3 Depth Ablation Study (No-Depth Baseline)

To explicitly validate the necessity of depth information for reliable pose estimation, an ablation experiment was conducted in which the depth prediction module was removed from the pipeline. In this setting, camera motion was estimated using only 2D–2D correspondences obtained from tracked feature points, relying solely on epipolar geometry via the Essential matrix.

Without depth, translation can only be recovered up to an unknown scale factor, and the absence of metric 3D constraints significantly reduces geometric stability. As a result, the system becomes highly sensitive to noise in feature tracking and small violations of the pure rotation or rigid motion assumptions. In practice, this leads to unstable translation estimates and rapid accumulation of drift when integrating relative poses over time.

The experiment was performed on a translational sequence (Video Sample 2), where sufficient parallax is present and the depth-enabled pipeline performs well. When depth is removed, the recovered trajectory exhibits noticeable scale ambiguity and deviation from the reference path, despite maintaining reasonable rotational consistency. This contrast highlights that accurate metric depth is essential for resolving translation reliably and for maintaining global trajectory coherence in monocular handheld videos.

5 Results

This section presents a detailed analysis of each video sequence, combining quantitative metrics with qualitative trajectory behavior.

5.1 Video Sample 1 — General Translational Motion (U-Shaped Trajectory)

Table 1: Pose error statistics for Video Sample 1. RMSE is used as the primary evaluation metric for all errors.

Metric	**RMSE**	Mean	Median	Std	Min	Max	SSE
APE-Tran (m)	**0.220**	0.199	0.189	0.095	0.037	0.432	8.532
RPE-Rot (deg)	**1.363**	1.286	1.333	0.449	0.035	1.998	323.3
RPE-Tran (m)	**0.022**	0.021	0.022	0.008	0.001	0.037	0.091

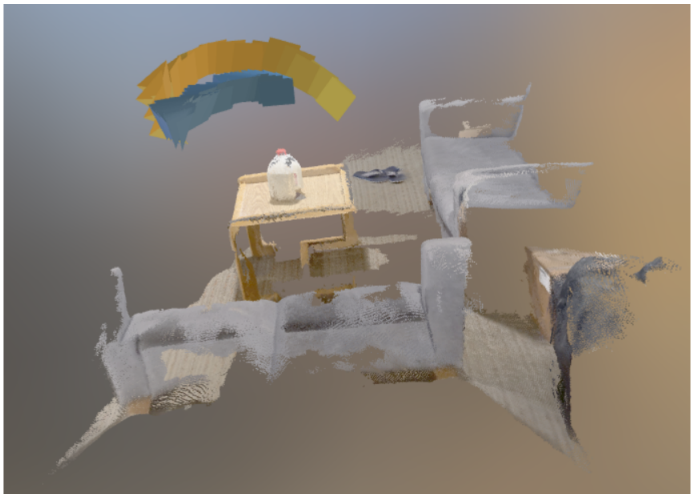

Figure 1: Estimated trajectory embedded in reconstructed 3D point cloud

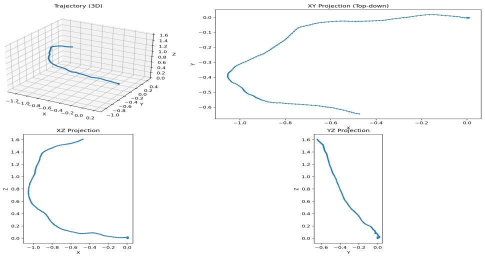

Figure 2: Trajectory projections (3D, XY, XZ, YZ)

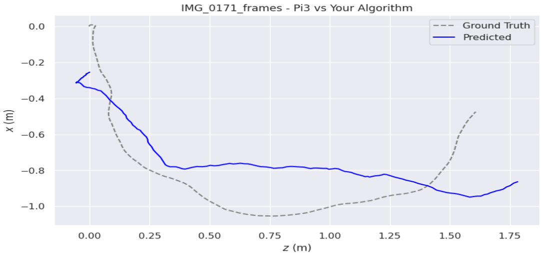

Figure 3: Trajectory comparison with Pi3 reference

Figure 4: Qualitative trajectory evaluation for Video Sample 1. Left: spatial consistency of the estimated trajectory within the reconstructed 3D point cloud. Right: trajectory projections and comparison with the Pi3 reference.

Figure 4 demonstrates that the estimated camera motion preserves the overall geometric structure of the trajectory. The multi-view projections reveal a clear U-shaped path with smooth curvature and consistent axis-wise behaviour, indicating stable relative pose estimation across frames. The absence of abrupt directional changes suggests that frame-to-frame pose estimates remain locally coherent throughout the sequence.

These qualitative observations are reflected in the quantitative results reported in Table 1. The relatively low RPE-Trans indicates consistent local translation estimation between consecutive frames, while the moderate RPE-Rot values reflect the increased rotational complexity inherent in a curved, handheld trajectory. The higher APE-Trans compared to the straight-line sequence highlights the cumulative effect of frame-to-frame drift over the longer U-shaped path. Together, these metrics confirm that while local pose estimation remains stable, global errors gradually accumulate over extended and non-linear motion.

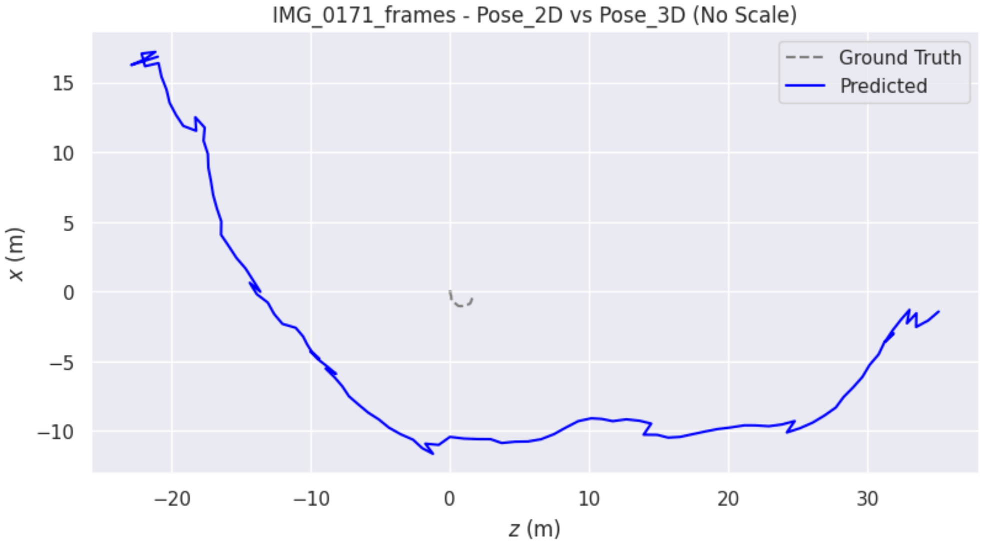

Figure 5: Trajectory comparison with and without depth information for Video Sample 1. The dashed curve corresponds to pose estimation without depth (2D–2D geometry only), while the solid curve shows the depth-assisted pose estimate.

Figure 5 shows that removing depth leads to severe scale ambiguity and unstable translation, causing the estimated trajectory to diverge dramatically despite similar rotational behaviour.

5.2 Video Sample 2 — Straight-Line Trajectory

Figure 9 illustrates the estimated camera motion for the straight-line sequence. The trajectory shows strong alignment with the reference path, with two projections exhibiting clear near-linear behaviour consistent with dominant forward motion. In the remaining projection, small deviations from a perfect point trajectory are observed, attributable to natural handheld camera motion and minor accumulated drift. Despite these perturbations, the overall progression along the principal motion axis remains consistent, indicating accurate recovery of forward translation when sufficient parallax is present. The close overlap with the Pi3 reference confirms stable frame-to-frame pose estimation and well-conditioned depth-assisted PnP performance under near-ideal motion conditions.

This qualitative observation is corroborated by the quantitative results in Table 2, which report low RMSE values for both absolute and relative pose errors. In particular, the low APE-Trans and RPE-Tran values indicate accurate and consistent translation estimation, while the small RPE-Rot confirms stable rotational recovery across consecutive frames. Together, the trajectory alignment and error statistics demonstrate that the proposed method performs reliably under well-conditioned straight-line motion.

Table 2: Pose error statistics for Video Sample 2. RMSE is used as the primary evaluation metric for all errors.

Metric	**RMSE**	Mean	Median	Std	Min	Max	SSE
APE-Tran (m)	**0.087**	0.078	0.068	0.038	0.022	0.220	1.254
RPE-Rot (deg)	**0.067**	0.060	0.057	0.029	0.004	0.131	0.742
RPE-Tran (m)	**0.012**	0.011	0.009	0.005	0.002	0.031	0.024

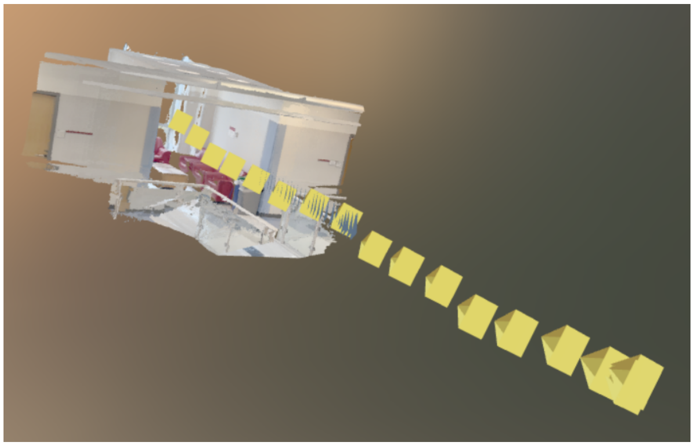

Figure 6: Estimated trajectory embedded in reconstructed 3D point cloud

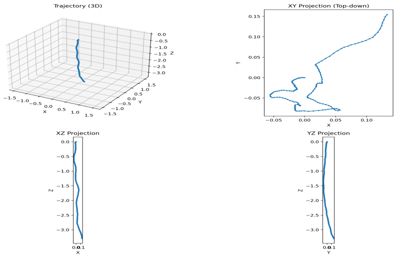

Figure 7: Trajectory projections (3D, XY, XZ, YZ)

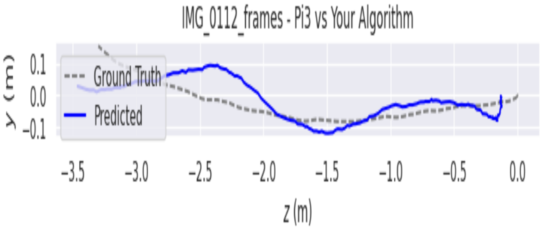

Figure 8: Trajectory comparison with Pi3 reference

Figure 9: Qualitative trajectory evaluation for Video Sample 2. Left: spatial consistency of the estimated trajectory within the reconstructed 3D point cloud. Right: trajectory projections and comparison with the Pi3 reference.

5.3 Video Sample 3 — Pure Rotation Sequence

Figure 13 illustrates the estimated camera motion under a predominantly pure rotational sequence. The multi-view projections show that the trajectory remains spatially compact, with limited overall displacement, indicating that the rotation-only handling effectively suppresses large spurious translation. However, small irregular deviations are visible across the projections, particularly in the XY plane, reflecting sensitivity to residual noise in feature tracking and minor numerical instability during pose integration. The comparison with the Pi3 reference further highlights that, while the general motion trend is preserved, the estimated trajectory exhibits local jitter and slight divergence over time, which is expected in the absence of reliable parallax cues.

These qualitative observations are supported by the quantitative results in Table 3. The low APE-Trans and RPE-Tran values indicate that translational drift is largely constrained despite the degeneracy of the motion. In contrast, the higher RPE-Rot reflects increased rotational error accumulation, underscoring the sensitivity of rotation estimation to noise when translation-based geometric constraints are unavailable. Together, the results confirm that while the proposed rotation-only strategy prevents catastrophic translation failure, pure rotational motion remains challenging and limits overall pose accuracy.

Table 3: Pose error statistics for Video Sample 3. RMSE is used as the primary evaluation metric for all errors.

Metric	**RMSE**	Mean	Median	Std	Min	Max	SSE
APE-Tran (m)	**0.080**	0.071	0.061	0.036	0.009	0.164	0.933
RPE-Rot (deg)	**1.222**	1.178	1.173	0.325	0.317	2.099	215.342
RPE-Tran (m)	**0.018**	0.016	0.014	0.008	0.003	0.042	0.047

5.4 Reasons for Observed Inaccuracies

The analysis of results points to four key sources of inaccuracy:

- **Frame-to-Frame Drift Accumulation:** As noted, the small errors in individual pose transforms $\mathbf{T}_{k-1\to k}$ compound through the sequence, causing the integrated trajectory to slowly deviate from the ground truth path.

- **Pure Rotation Ambiguity:** In frames with little or no translation, the camera experiences little to no parallax. Under these conditions, the fundamental geometric constraints required for PnP to infer translation are absent. This inherent ambiguity causes PnP to introduce small, fictitious translational motions, destabilising the pose estimate.

- **Depth Noise Propagation:** Imperfections and noise in the predicted per-frame depth lead to local fluctuations in the reconstructed 3D point locations. These 3D lifting errors result in mildly inconsistent 3D-2D correspondences, which weakens the fitting quality of the PnP solver.

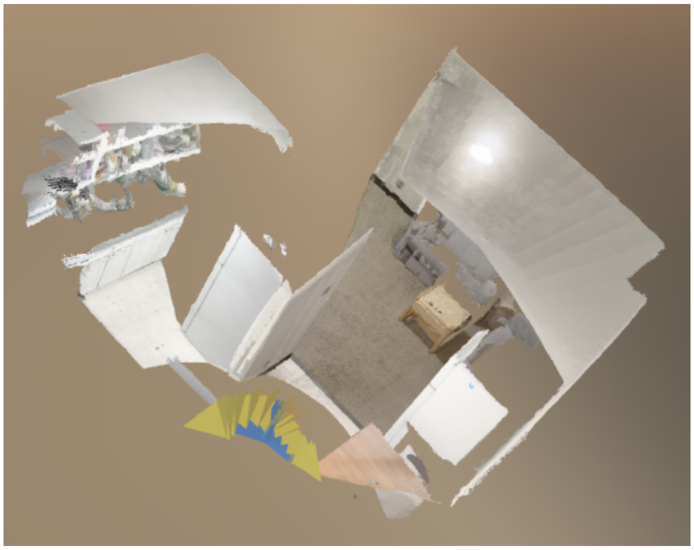

Figure 10: Estimated trajectory embedded in reconstructed 3D point cloud

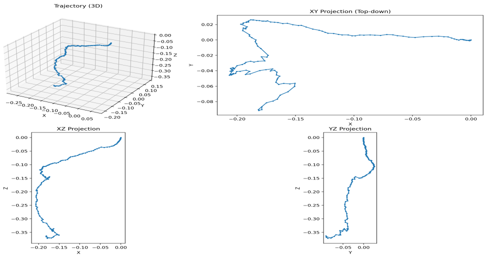

Figure 11: Trajectory projections (3D, XY, XZ, YZ)

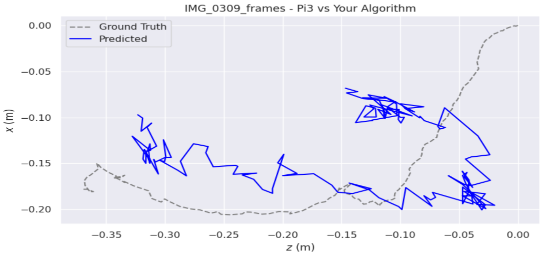

Figure 12: Trajectory comparison with Pi3 reference

Figure 13: Qualitative trajectory evaluation for Video Sample 3. Left: spatial consistency of the estimated trajectory within the reconstructed 3D point cloud. Right: trajectory projections and comparison with the Pi3 reference.

- **Residual Track Instability:** Even with sophisticated filtering (SuperPoint and Co-Tracker with confidence filtering), subtle drift can occur in challenging areas such as low-texture or distant regions. These residual outliers, even if rare, can still significantly influence the final pose calculation.

6 Future Scope for Improvement

To address the stability and drift issues, two primary avenues for future research are proposed:

- **Global SfM-Based Optimization:** The current pairwise PnP approach should be superseded by a full **Structure-from-Motion (SfM)** optimization framework39. This typically involves **Bundle Adjustment**, which simultaneously refines all camera poses and all 3D map points to minimize global reprojection error. This global approach inherently handles low-parallax segments better and offers a superior mechanism for reducing drift accumulation41.

- **Leverage iPhone LiDAR Integration:** Utilizing sparse but highly accurate depth measurements from hardware sensors, such as an iPhone LiDAR, can be employed to **anchor the global scale** of the reconstruction. This external scale constraint reduces the ambiguity inherent in monocular 3D lifting and significantly stabilizes the PnP solution, particularly when translational motion is weak or absent (low-parallax).

7 Conclusion

This project demonstrates that combining learned depth prediction and long-term point tracking with classical geometric solvers enables reliable camera pose estimation from monocular handheld video under a wide range of motion conditions. When sufficient parallax is present, depth-assisted PnP produces accurate and stable pose estimates, as reflected by low relative and absolute errors on translational sequences. In contrast, pure rotational motion exposes fundamental geometric degeneracies that cannot be fully resolved without translation, highlighting the necessity of motion-aware pose handling. The depth ablation study further confirms that metric depth is essential for recovering physically meaningful translation and for maintaining global trajectory coherence. Overall, the results validate the effectiveness of the proposed hybrid pipeline while also illustrating the inherent limitations of frame-to-frame pose integration without global optimization, motivating future extensions toward full SfM and sensor-fused solutions.