PRISM-VO: Scale-Aware Visual Odometry Using Photometric Plenoptic Bundle Adjustment

Aymeric Fleith[1,2], Julian Zirbel[1,2], Daniel Cremers[1], and Niclas Zeller[2]

[1] Technical University of Munich, Munich, Germany
{aymeric.fleith, julian.zirbel, cremers}@tum.de
[2] Karlsruhe University of Applied Sciences, Karlsruhe, Germany
niclas.zeller@h-ka.de

Abstract. We introduce PRISM-VO, a novel pure optimization-based sparse photometric visual odometry framework for focused plenoptic cameras. The core of PRISM-VO is a novel photometric plenoptic bundle adjustment which jointly optimizes camera poses and inverse depth values of points in a sliding window. By combining geometric depth from a single plenoptic image with temporal multi-view constraints, PRISM-VO achieves accurate and drift-resilient motion estimation. Through explicit modeling of the plenoptic projection, PRISM-VO provides reliable metric-scale reconstructions, overcoming the scale ambiguity of monocular SLAM algorithms. Importantly, our approach relies solely on a single plenoptic sensor and avoids complex initialization, as depth priors are computed directly from plenoptic imaging.
Experiments show that PRISM-VO outperforms the current state-of-the-art plenoptic visual odometry method on indoor and outdoor scenes. The proposed approach rivals other optimization- and learning-based methods while accurately and reliably recovering a metric scale of the scene.
Project page: https://prism-vo.github.io/.

Keywords: Plenoptic camera · Light field · Micro-lens array · Visual odometry · SLAM · Scale

1 Introduction

3D reconstruction of the environment and motion estimation are essential for real-time localization in robotics, autonomous vehicles, drones, virtual reality, and augmented reality. While lidars, radars, GPS, and inertial sensors are widely used, camera-based simultaneous localization and mapping (SLAM) and visual odometry (VO) have become widely adopted for mapping and localization. Passive cameras offer high resolution, rich visual information, versatility, and a cost-effective, lightweight, and compact sensing solution. However, monocular cameras still cannot recover absolute scale without additional information.

Alternatives such as stereo, RGB-D, and ToF cameras address the scale ambiguity of monocular vision but remain constrained by the stereo baseline, the structured light range, or the trade-off between range and accuracy. By placing

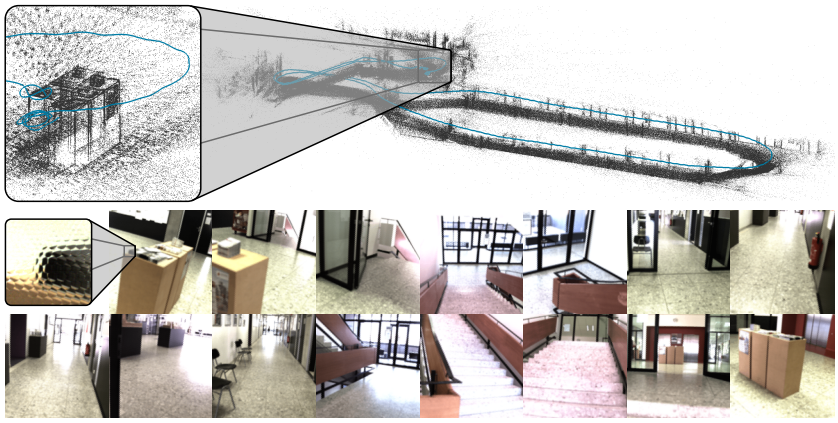


Fig. 1: Metric 3D reconstruction of a 150 m long sequence (seq_007) from the dataset [39] using PRISM-VO. The zoom shows the accumulated drift over the whole sequence. The images below are examples of raw plenoptic images from the sequence.

a micro-lens array (MLA) between the main lens and the sensor of a camera, a plenoptic camera extends conventional imaging by simultaneously capturing spatial and angular information of the scene. This produces multiple viewpoints in the form of micro-images, also offering a very wide depth of field. This makes them well suited for true-to-scale robust SLAM in compact systems.

To exploit this advantage, we propose PRISM (**P**lenoptic **R**econstruction via **I**nverse-depth **S**parse **M**apping) visual odometry, a new pure optimization-based photometric VO approach for focused plenoptic cameras. It combines disparity between micro-images for absolute scale estimation and larger baseline from temporal disparity to improve tracking accuracy and robustness. It enables reliable sparse metric reconstruction (see Fig. 1), which is not possible with a standard monocular camera. Our main contributions are:

- A sparse and photometric plenoptic bundle adjustment formulation that jointly optimizes camera poses and environment points.
- Tight integration of the plenoptic camera model into front-end tracking and back-end bundle adjustment of the VO pipeline.
- Depth refinement using variance-weighted plenoptic depth residuals combined with temporal optimization.

We extensively evaluate the method on several datasets containing various types of scenes and different cameras. We demonstrate excellent performance compared to state-of-the-art methods by reducing the errors accumulated along the trajectory while providing a metric scale.

2 Related Work

This section reviews classical SLAM, plenoptic and light-field methods, and prior structure from motion (SfM) and VO approaches using plenoptic cameras.

2.1 Classical SLAM Approaches

Early visual SLAM systems relied on feature-based approaches (also called indirect methods), such as PTAM [19] and ORB-SLAM [25], which estimate camera motion by tracking sparse keypoints across frames. Later, direct methods such as DTAM [26] and LSD-SLAM [11] bypassed feature extraction by optimizing over image intensities. DSO [10] further improved robustness due to photometric bundle adjustment. Methods such as DROID-SLAM [30] rely on deep learning and achieve state-of-the-art accuracy through end-to-end optimization of camera poses and dense depth. More recently, visual–inertial SLAM frameworks such as DM-VIO [32] have shown real-time performance in challenging environments.

Despite their success, these methods rely on pinhole or fisheye camera models. When applied to plenoptic cameras, their geometric and radiance assumptions no longer hold, motivating the need for adapted formulations.

2.2 Plenoptic and Light-Field Vision

We distinguish two main configurations of plenoptic cameras: unfocused (plenoptic camera 1.0) and focused (plenoptic camera 2.0).

Prior works studied unfocused plenoptic cameras, where the main lens is focused on the MLA. The MLA is itself focused at infinity. Calibration methods typically reconstruct sub-aperture images to detect features [5, 6, 42] or detect features directly in the micro-images [2,28,41]. However, they are less used today due to low spatial resolution.

Focused plenoptic cameras improve the trade-off between spatial and angular resolution by focusing micro-lenses on the main lens image plane [23, 24]. Subsequent work has established new light-field models to estimate all intrinsic, extrinsic, and scene parameters from reconstructed images [35,37,38] or directly from raw images [20,21,27]. To eliminate the need for a calibration target, [13,14] use sub-aperture views, and LiFCal [15] allows recalibration on any scene.

Several methods exploit light-field data for depth estimation within a single shot [22,33]. In addition, applications such as post-capture refocusing [17], super resolution [34] or view synthesis [12] highlight the versatility of light-field data.

2.3 SfM and VO with Plenoptic Cameras

Few works directly integrate plenoptic cameras into VO/SLAM. Early approaches proposed closed-form 6-DOF VO [4] or multi-scale motion estimation for low-resolution plenoptic vision [8], while [18] introduced a light-field front-end for indirect SLAM. However, these rely on custom setups rather than true plenoptic cameras. SPO, A semi-dense direct plenoptic odometry method, introduced in [38] and improved in [39], operates on micro-images and recovers metric scale from the light-field geometry. Nevertheless, this method is limited to projecting points from the penultimate image onto the last image without performing a bundle adjustment across multiple frames, which makes it less robust and more sensitive to drift. [7] relies on unsupervised learning for depth estimation and VO

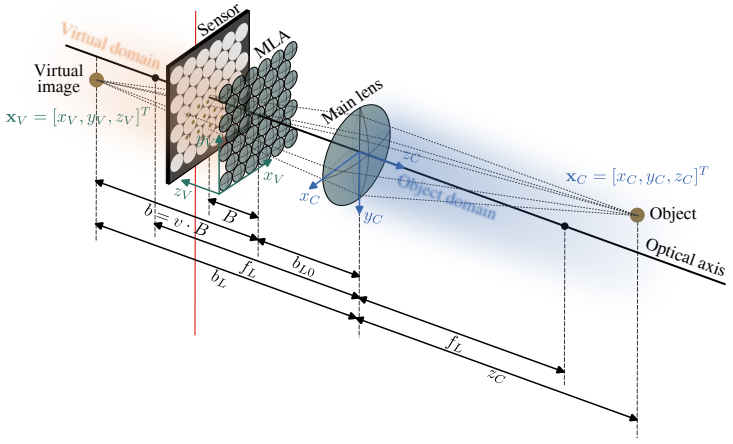


Fig. 2: Representation of the model of the focused plenoptic camera in Galilean mode.

but is tailored to sparse light-field cameras. An indirect VO framework leveraging light-field cameras is introduced in [1], which enables metric-scale translation estimation and simplifies calibration from a novel keypoint extraction. However, this approach requires an additional step of feature detection depending on textured scenes. The methods in [9] and [15] also perform reconstruction but operate as SfM pipelines without considering temporal links between images.

Overall, light-field sensing shows strong potential for robust tracking and scale estimation. However, existing approaches target different camera setups, require training, or rely on highly textured environments.

3 Preliminaries

We first provide an overview of the plenoptic camera model used in the VO pipeline (Sec. 3.1). Then, we introduce the notation used in the paper (Sec. 3.2).

3.1 The Focused Plenoptic Camera

A plenoptic camera extends the conventional imaging model by capturing not only the intensity of light rays on the sensor but also their directions within the optical system by placing an MLA between the main lens and the sensor.

Throughout the paper, we consider a focused plenoptic camera in Galilean mode and base our method on the plenoptic camera model introduced in [15], modeling the main lens as a thin lens and the micro-lenses as pinholes. The plenoptic camera model is defined by key geometric parameters (see Fig. 2): f_L (main lens focal length), b_{L0} (distance from main lens to MLA), and B (distance from MLA to sensor). The main lens forms a virtual image of the scene at a distance b_L from the main lens and b from the MLA. The virtual depth v, introduced in [29] as $v = \frac{b}{B}$, maps the distance z_C between the real 3D

(a) Raw image **(b)** Totally focused image **(c)** Virtual depth map **(d)** Virtual depth uncertainty map

Fig. 3: Results of depth estimation on plenoptic images based on [36] and [15].

point in camera metric coordinates $\mathbf{x}_C = [x_C, y_C, z_C]^T$ in the 3D virtual image space $\mathbf{x}_V = [x_V, y_V, z_V]^T$. This projection is denoted as $\Pi_{\text{pl}}(\cdot)$. The virtual depth allows a depth map to be generated without prior metric calibration of the camera. The main lens principal point is denoted as $\mathbf{c}_L = [c_x, c_y]^T$ in pixels.

Each micro-lens forms a small image on the sensor (see Fig. 3a). Neighboring micro-images represent the same points from slightly different perspectives. We rely on the depth-estimation pipelines of [36] and [15] to compute virtual depth maps (Fig. 3c), the so-called totally focused image (Fig. 3b) created by the main lens and the virtual depth uncertainty maps (Fig. 3d) from a single raw plenoptic image. More details about the camera model are provided in [15] and in the supplementary material.

3.2 Notation

In the remainder of this document, vectors are denoted in bold lowercase letters ($\mathbf{x}$), matrices in bold uppercase letters ($\mathbf{J}$) and scalars in non-bold letters (v). Indexed quantities use subscript i ($i \in \{1, \dots, n\}$) for all indexed variables, including scalars (v_i) and vectors ($\mathbf{x}_{C,i}$).

4 PRISM-VO: Plenoptic Direct Visual Odometry

We first give an overview of the pipeline in Sec. 4.1, followed by the main components: the multi-scale image representation (Sec. 4.2), the point selection (Sec. 4.3), and the plenoptic bundle adjustment formulation (Sec. 4.4).

4.1 Method Overview

PRISM-VO uses a novel sparse and photometric plenoptic bundle adjustment formulation within a keyframe-based front-end/back-end architecture. It integrates the plenoptic camera model into both tracking and mapping (Fig. 4). This tight coupling is required because the plenoptic camera model differs fundamentally from pinhole stereo/RGB-D models.

Each input raw image is processed into a totally focused image, virtual depth map, and virtual depth uncertainty map in a similar way as in [36] and [15]. The

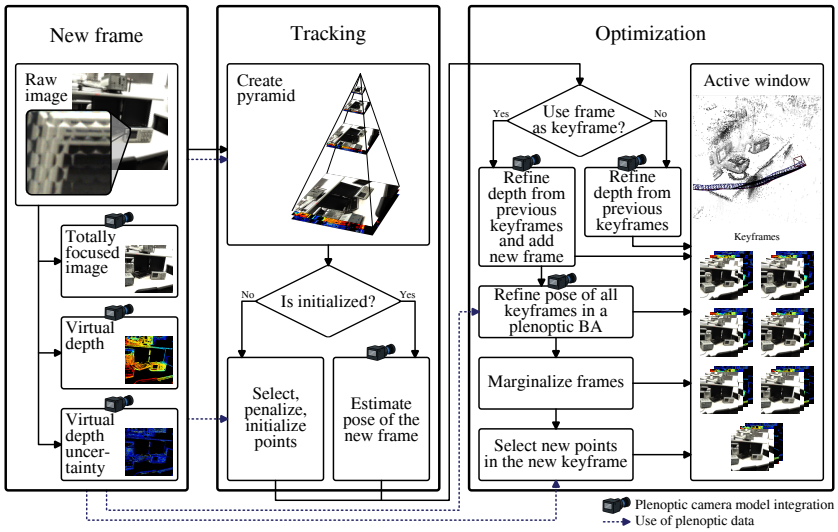


Fig. 4: Overview of the PRISM-VO algorithm pipeline: image processing, tracking in the front-end, and optimization by plenoptic bundle adjustment in the back-end.

totally focused image drives photometric tracking, while depth and uncertainty provide geometric priors for both front-end tracking and back-end optimization.

The front-end estimates camera pose via direct image alignment against reference keyframes. To increase robustness and convergence, it uses a coarse-to-fine approach. During initialization, points are selected according to their gradient and their depth information. Keyframes are added based on field of view changes, occlusions, and exposure variations in a manner inspired by [10].

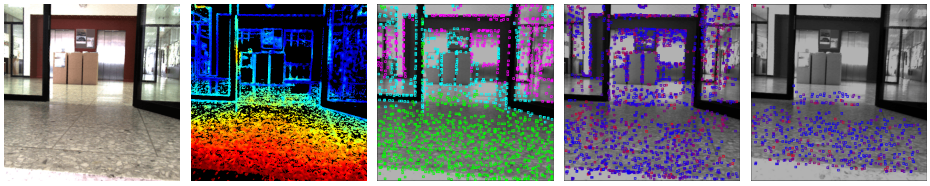
The back-end maintains an active window of keyframes. As new keyframes are added, depth hypotheses are refined by combining multi-view photometric consistency with plenoptic depth cues. Camera poses and scene geometry are jointly optimized through a novel sparse photometric plenoptic bundle adjustment. Older keyframes are then marginalized, with their information retained in the optimization via prior constraints.

4.2 Multi-Scale Image Representation

To enable robust multi-scale processing, we construct image pyramids for the totally focused image, virtual depth map, and virtual depth uncertainty map. Each pyramid level is half the resolution of the previous level in both dimensions.

Totally Focused Image Pyramid. The totally focused image pyramid is built by averaging 2×2 pixel blocks from the previous level.

Virtual Depth Map Pyramid. Since the inverse virtual depth is proportional to the micro-image disparity, one can assume it to be normally distributed. Virtual depth maps are downsampled via variance-weighted averaging, giving more weight to reliable measurements and ignoring invalid pixels. The



(a) Totally focused image **(b)** Virtual depth map **(c)** Point selection **(d)** Photometric residuals **(e)** Inverse virtual depth residuals

Fig. 5: Point selection and residual definitions. (5a) Totally focused image. (5b) Corresponding virtual depth map. (5c) Selection of points when adding a new frame. *Green*: valid depth used in the bundle adjustment; *cyan*: selected points with depth information available but unused due to a high uncertainty; *magenta*: selected points without depth. Representation of photometric (5d) and depth (5e) residuals, where *red* indicates high energy and *blue* low energy.

inverse virtual depth $\bar{\rho}_{v,i}$ at the coarser level is computed with Eq. (1), where $N_x^{(i)}$ denotes the neighborhood pixels with valid inverse virtual depth $\rho_{v,i}$ and the corresponding variance $\sigma^2_{\rho_{v,i}}$.

$$\bar{\rho}_{v,i+1} = \frac{\sum_{k\in N_x^{(i)}} \rho_{v,k} \cdot \left(\sigma^2_{\rho_{v,k}}\right)^{-1}}{\sum_{k\in N_x^{(i)}} \left(\sigma^2_{\rho_{v,k}}\right)^{-1}}, \qquad \bar{\sigma}^2_{\rho_{v,i}} = \frac{\left|N_x^{(i)}\right|}{\sum_{k\in N_x^{(i)}} \left(\sigma^2_{\rho_{v,k}}\right)^{-1}} \tag{1}$$

Virtual Depth Uncertainty Map Pyramid. The inverse virtual depth variance $\bar{\sigma}_{\rho_{v,i}}$ at coarser levels is computed using a harmonic mean over valid child pixels with Eq. (1), where $\left|N_x^{(i)}\right|$ is the cardinal of $N_x^{(i)}$. This approach ensures that regions with high uncertainty are appropriately reflected in coarser levels, *i.e.* assuming a strong correlation between depths of neighboring pixels.

4.3 Point Selection

Adaptive Depth Cutoff. We compute an adaptive inverse depth cutoff based on the mean inverse virtual depth $\rho_{v,\text{av}}$ across valid points to favor points near the camera with lower variance. Using a normalized weighting factor $w_{\text{cut}} \in [0,1]$, we define a percentile p_a with Eq. (2) between $p_{\min}$ and $p_{\max}$ that reduces the influence of distant outliers for reliable depth selection (see Fig. 5). The values $\rho_{v,\text{n}}$ (near) and $\rho_{v,\text{f}}$ (far) represent reference inverse virtual depth bounds.

$$p_{\text{a}} = p_{\min} + w_{\text{cut}} \cdot (p_{\max} - p_{\min}), \qquad \text{with} \qquad w_{\text{cut}} = \frac{\rho_{v,\text{av}} - \rho_{v,\text{f}}}{\rho_{v,\text{n}} - \rho_{v,\text{f}}} \tag{2}$$

Point Penalization. The gradient-based score $s_{\text{grad}}(\mathbf{x}_V)$ of a point $\mathbf{x}_V$ is defined from the totally focused image gradient $\nabla I(\mathbf{x}_V)$ and a randomly sampled unit direction $\mathbf{d}$ to encourage orientation diversity. The score of a pixel $\mathbf{x}_V$

containing valid depth is weighted more heavily by a constant factor $w_d(\mathbf{x}_V)$ than those without depth information. The final selection score used for ranking candidate pixels is therefore computed with Eq. (3). Points without depth data are also selected (but with a lower weight $w_d(\mathbf{x}_V)$) to ensure good distribution of points across the image. Fig. 5 illustrates the resulting point selection. This approach favors reliable nearby points while ensuring coverage in regions with noisy, missing, or distant depth.

$$s(\mathbf{x}_V) = \left|\nabla I(\mathbf{x}_V)^\top \mathbf{d}\right| \; w_d(\mathbf{x}_V) \tag{3}$$

4.4 Optimization Formulation

The state vector, including camera poses, inverse depths, and brightness parameters (to account for exposure changes), is estimated by minimizing a joint nonlinear least-squares problem. A point $\mathbf{x}_{V,i}$ in the host frame i is projected with Eq. (4) to a point $\mathbf{x}_{V,j}$ in a target frame j using the plenoptic camera model introduced in Sec. 3.1. $\mathbf{R}, \mathbf{t} \in SE(3)$ denote the relative pose and $\Pi_{\text{pl}}(\cdot)$ denotes the plenoptic projection (see supplementary material for details).

$$\mathbf{x}_{V,j} = \Pi_{\text{pl}} \left(\mathbf{R} \cdot \Pi_{\text{pl}}^{-1} \left(\mathbf{x}_{V,i} \right) + \mathbf{t} \right) \tag{4}$$

The photometric residual r^{photo} for a point $\mathbf{x}_{V,i}$ in a reference frame I_i observed in a target frame I_j is defined in Eq. (5). The affine brightness parameters a_i, a_j, b_i and b_j are inspired by [10].

$$r^{\text{photo}} = \left(I_j \left[\Pi_{\text{pl}}(\mathbf{x}_{V,i})\right] - b_j\right) - \frac{e^{a_j}}{e^{a_i}} \left(I_i[\mathbf{x}_{V,i}] - b_i\right) \tag{5}$$

The depth residual is formulated in the inverse virtual depth space as defined in Eq. (6). Here, ρ_v is the inverse virtual depth resulting from the projection and ρ_v^{meas} the corresponding measured value from the depth map.

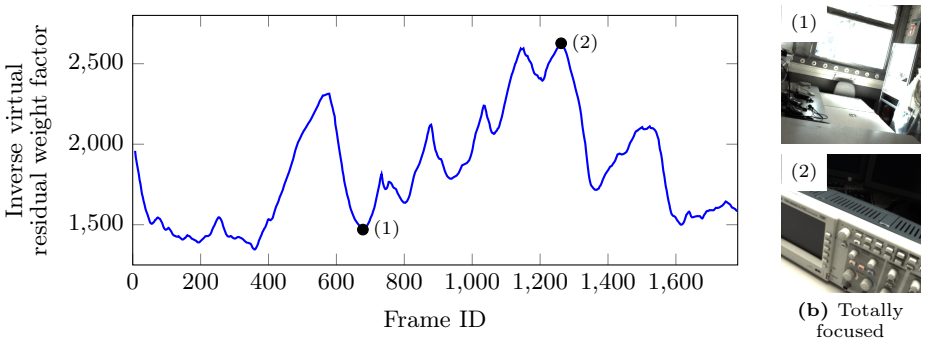
$$r^{\text{depth}} = \rho_v - \rho_v^{\text{meas}} \tag{6}$$

The total energy is given by

$$E = \sum_k w_k^{\text{photo}} \left\| r_k^{\text{photo}} \right\|_\gamma + \eta \sum_l w_l^{\text{depth}} \left\| r_l^{\text{depth}} \right\|_\gamma, \tag{7}$$

where w_k^{photo} and w_l^{depth} are per-residual weights. The quadratic error of the inverse virtual depth can be minimized as it is normally distributed. The inverse virtual depth residual is weighted by the inverse virtual depth variance, resulting in $w_l^{\text{depth}} = \left(\sigma^2_{\rho_{v,l}}\right)^{-1}$, so more reliable depth measurements have greater influence. The symbol $\|\cdot\|_\gamma$ denotes the Huber norm, and η balances the photometric and depth residual terms.

Adaptive Cross-Modal Weighting. To balance photometric and depth residuals in the joint optimization, we introduce a scalar weight η. It is adapted



(a) Adaptive weight factor η for the depth residual as a function of frame IDs.

(b) Totally focused images.

Fig. 6: Evolution of the inverse virtual depth residual weight factor η on seq_004 from the dataset [40]. Fig. 6a: Local maxima occur with nearby geometry, increasing depth influence, while local minima correspond to distant scenes dominated by photometric consistency. Fig. 6b: Corresponding representative frames are shown.

online by matching the average Gauss–Newton curvature, determined by $\mathbf{J}^T\mathbf{J}$, of both residual types with Eq. (8).

$$\eta = c \cdot \frac{\sum_i \|J_i^{\text{photo}}\|_2^2}{\sum_j \|J_j^{\text{depth}}\|_2^2} \quad \text{with} \quad J_i^{\text{photo}} = \frac{\partial r_i^{\text{photo}}}{\partial \rho_{C,i}} \quad \text{and} \quad J_i^{\text{depth}} = \frac{\partial r_j^{\text{depth}}}{\partial \rho_{C,j}} \tag{8}$$

Here, c is a constant factor and J_i^{photo} and J_j^{depth} are respectively the Jacobian contributions of the photometric and the inverse virtual depth residuals related to the inverse point depths.

For temporal stability, the weight η is updated with

$$\eta_{i+1} = \exp((1-\alpha)\log\eta_i + \alpha\log\eta)\,, \tag{9}$$

using a fixed sensitivity parameter α and exponential smoothing in the logarithmic domain to decrease the weight for older values. This results in a scene-adaptive balance between photometric and depth constraints without manual tuning as shown in Fig. 6. Since inverse depth is more sensitive to nearby points due to stronger parallax, scenes dominated by close objects increase the scaling factor, giving depth residuals greater influence (see Fig. 6).

Levenberg–Marquardt Optimization. The nonlinear problem is solved with a Levenberg–Marquardt optimization in a sliding window, jointly optimizing camera poses $\boldsymbol{\xi}$ in the tangent space, affine brightness parameters a, b, and inverse depths ρ_C. At each iteration, all residuals are linearized around the current state $\mathbf{s}$ as

$$r(\mathbf{s}+\delta\mathbf{s}) \approx r(\mathbf{s}) + \mathbf{J}\,\delta\mathbf{s}, \tag{10}$$

where $\mathbf{J}$ stacks the photometric and depth Jacobians (see the supplementary material for the complete derivation). The inverse virtual depth residual r^{depth}

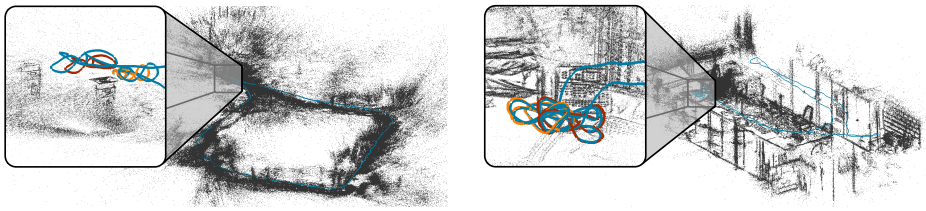


Fig. 7: Point clouds and trajectories estimated by PRISM-VO on sequences from the dataset [40]. On the *left* is the sequence seq_002 (200 m long / outdoor) and on the *right* is the seq_009 (30 m long / indoor). The zoomed views show the accumulated drift and the ground-truth trajectory (front part in *orange*, back part in *red*).

depends only on inverse depth and does not depend on a, b or the pose $\boldsymbol{\xi}$. The overall structure of the Hessian matrix is therefore preserved, making the integration very efficient to implement. Stacking all residuals yields the Eq. (11), where $\mathbf{H}$ is the approximate Gauss-Newton Hessian, $\mathbf{b}$ the gradient vector and $\mathbf{W}$ the weight matrix. The Schur complement is used to efficiently calculate the increment of the state vector by taking advantage of the sparsity of $\mathbf{H}$ and to marginalize keyframes that leave the optimization window, similar to [10].

$$(\mathbf{H} + \mu\mathbf{I})\,\delta\mathbf{s} = -\mathbf{b}, \qquad \text{with} \qquad \mathbf{H} = \mathbf{J}^\top\mathbf{W}\mathbf{J}, \qquad \mathbf{b} = \mathbf{J}^\top\mathbf{W}\mathbf{r} \tag{11}$$

5 Evaluation

We first qualitatively demonstrate the advantages of the plenoptic camera. Then, we evaluate PRISM-VO against plenoptic-specific methods and state-of-the-art monocular VO/SLAM pipelines and perform an ablation study of its key components. Our experiments rely on two plenoptic datasets with synchronized multi-sensor data and ground truth.

- A Synchronized Stereo and Plenoptic Visual Odometry Dataset [40]. It consists of 11 indoor and outdoor sequences acquired with a Raytrix R5 plenoptic camera and a synchronized stereo camera pair, featuring large loops and depths up to several hundred meters for drift and scale evaluation.
- LiFMCR Dataset [16]. It contains 7 scenes recorded with two high-resolution Raytrix R32 plenoptic cameras, providing a precise 6-DoF ground truth from a Vicon system for precise pose accuracy assessment.

These datasets provide complementary scenarios for assessing long-term drift, scale consistency, and pose accuracy. Qualitative 3D reconstructions are shown in Fig. 7 and Fig. 8 for both datasets respectively.

5.1 Qualitative Results Against Other Sensors

To demonstrate the benefits of the plenoptic technology, we recorded a sequence using a synchronized plenoptic-monocular rig consisting of an R32 Raytrix cam-

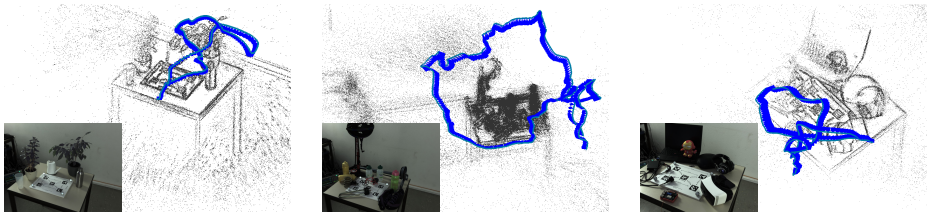


Fig. 8: Point clouds and trajectories estimated by PRISM-VO on sequences from the LiFMCR dataset [16] with the associated totally focused image (scenes from left to right: 01_Plants, 02_Bike, 04_Electronics). The camera poses are shown in *blue*.

era and a Basler acA1920-40gc camera. As shown in Fig. 9, DSO and ORB-SLAM3 fail in a challenging fan scene, despite a wide field of view, due to parallax at intersecting grilles, transparent fan blades, and reflective surfaces. DPVO remains stable but cannot recover metric scale, unlike PRISM-VO.

We also captured the same scene from the same viewpoint using the R32 Raytrix camera and an RGB-D camera (Intel RealSense D455). While the RGB-D camera produces unstable depth, the plenoptic system succeeds (see Fig. 10).

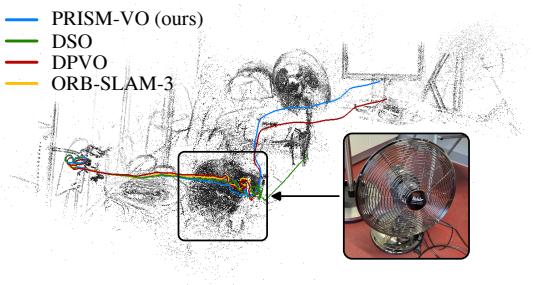
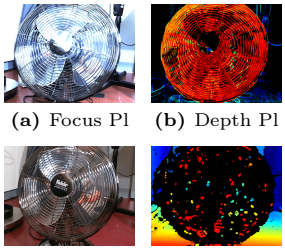


Fig. 9: Trajectories from a challenging sequence where PRISM-VO succeeds while other monocular methods fail.



(a) Focus Pl **(b)** Depth Pl **(c)** RGB RS **(d)** Depth RS

Fig. 10: Depth maps from plenoptic (*Pl*) and RealSense D455 (*RS*).

5.2 Quantitative Drift Study

To evaluate the accumulated drift over the complete VO pipeline, we rely on the dataset from [40]. Each sequence starts and ends at the same location, forming a large loop, allowing a quantitative assessment of accumulated drift compared to the ground truth obtained by loop closure. Following [40], we estimate two similarity transformations, $\mathbf{T}^{\text{gt}}_{\text{start}}$ (Eq. 12) and $\mathbf{T}^{\text{gt}}_{\text{end}}$ (Eq. 13), aligning estimated 3D points $\mathbf{x}_{C,i} \in \mathbb{R}^3$ and ground-truth points $\mathbf{x}^{\text{gt}}_{C,i} \in \mathbb{R}^3$ at the beginning and end of each sequence respectively. The start segment is defined by the index set S, and the end segment by the index set E.

$$\mathbf{T}_{\text{start}}^{\text{gt}} := \arg\min_{\mathbf{T}\in\text{Sim}(3)} \sum_{i\in S} \left\| \mathbf{T}\mathbf{x}_{C,i} - \mathbf{x}_{C,i}^{\text{gt}} \right\|_2^2 \tag{12}$$

$$\mathbf{T}_{\text{end}}^{\text{gt}} := \arg\min_{\mathbf{T}\in\text{Sim}(3)} \sum_{i\in E} \left\| \mathbf{T}\mathbf{x}_{C,i} - \mathbf{x}_{C,i}^{\text{gt}} \right\|_2^2 \tag{13}$$

From $\mathbf{T}_{\text{start}}^{\text{gt}}$ and $\mathbf{T}_{\text{end}}^{\text{gt}}$, we compute the accumulated drift transformation $\mathbf{T}_{\text{drift}} \in \text{Sim}(3)$ over the entire trajectory with Eq. (14).

$$\mathbf{T}_{\text{drift}} := \begin{bmatrix} e_s\mathbf{R} & \mathbf{t} \\ 0 & 1 \end{bmatrix} = \mathbf{T}_{\text{end}}^{\text{gt}} \left(\mathbf{T}_{\text{start}}^{\text{gt}}\right)^{-1} = \begin{bmatrix} s_e\mathbf{R}_e & \mathbf{t}_e \\ 0 & 1 \end{bmatrix} \begin{bmatrix} s_s\mathbf{R}_s & \mathbf{t}_s \\ 0 & 1 \end{bmatrix}^{-1} \tag{14}$$

For evaluation, we define the following metrics: absolute scale error $d_s = \sqrt{s_e \cdot s_s}$, scale drift $e_s = \frac{s_e}{s_s}$, rotation error e_r defined as the rotation angle around the Euler axis associated with the rotation matrix $\mathbf{R} \in \text{SO}(3)$ and a combined alignment error e_{align} captures overall trajectory inconsistency (Eq. (15)). To facilitate interpretation, we define $d_s' := \max\{d_s,\ d_s^{-1}\}$ and $e_s' := \max\{e_s,\ e_s^{-1}\}$.

$$e_{\text{align}} := \sqrt{\frac{1}{N}\sum_{i=1}^{N} \left\| \mathbf{T}_{\text{start}}^{\text{gt}}\mathbf{x}_{C,i} - \mathbf{T}_{\text{end}}^{\text{gt}}\mathbf{x}_{C,i} \right\|_2^2} \tag{15}$$

PRISM-VO is first compared against the plenoptic VO method SPO [39] using the dataset introduced in [40]. Tab. 1 reports the number of sequences that satisfy high, medium, and coarse precision thresholds for the different metrics. This evaluation protocol is adopted because only per-sequence results are available for SPO. Both methods recover the absolute scale accurately (scale close to 1), with PRISM-VO attaining slightly better results with $d_s' \leq 1.10$ on $8/11$ sequences and $d_s' \leq 1.30$ for $10/11$ sequences. More notably, PRISM-VO clearly reduces accumulated rotation error ($e_r < 4^\circ$ on $8/10$ sequences) and lowers worst-case rotation. The scale drift is comparable to SPO with $7/11$ sequences achieving $e_s' \leq 1.05$. The alignment error is also consistently smaller, especially at high and coarse precision. Additionally, PRISM-VO succeeds on one sequence where SPO fails, indicating greater robustness and stability beyond pure accuracy metrics. These results indicate that PRISM-VO provides more accurate and robust pose estimation, particularly in terms of alignment and rotational consistency, demonstrating the benefits of the proposed bundle-adjustment formulation for globally consistent VO. Therefore, PRISM-VO establishes a new state-of-the-art in plenoptic camera-based VO.

Apart from PRISM-VO and SPO, no other plenoptic VO algorithms performing on this type of scene are available. Therefore, we also compare PRISM-VO with state-of-the-art monocular methods, namely DSO [10], ORB-SLAM3 [3] and DPVO [31]. For a fair comparison of the tracking performance, large-scale-loop-closure is disabled in ORB-SLAM3. For the comparison methods, we use the provided monocular images and crop them to match the field of view of the plenoptic images (while adapting the intrinsics to preserve the pinhole model)

Table 1: VO results for plenoptic methods on the dataset [40] over 11 sequences. The results are shown in terms of percentages of sequences reaching high/medium/coarse precision. The results for d'_s and e'_s are for a scale factor under 1.05/1.1/1.3, e_{align} for values under 1 %/2 %/4 %, e_{r} for angles under 1°/2°/4°. Best results are in bold.

Method	Abs. scale error d'_s	Scale drift e'_s	Alignment error e_{align}	Rotation error e_{r}
PRISM-VO (ours)	**5**/**8**/**10**	**7**/7/**10**	**4**/**6**/**10**	**6**/**8**/**10**
SPO [39]	**5**/7/9	**7**/**8**/9	2/**6**/9	3/7/8

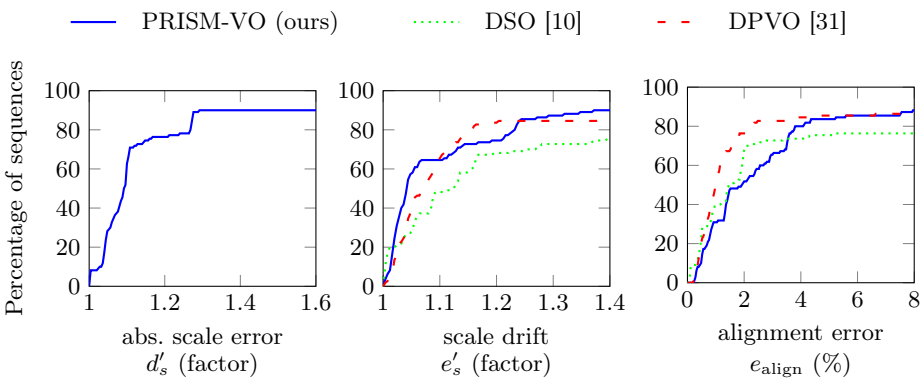


Fig. 11: Cumulative plots obtained on the dataset [40] for VO/SLAM.

to be able to compare the methods. To mitigate non-deterministic effects, each pipeline is run 10 times per sequence, with aggregated results shown in Fig. 11. ORB-SLAM3 fails on all the sequences of [40], likely due to the limited field of view. However, PRISM-VO is able to reliably recover the absolute scale of the scene. Furthermore, PRISM-VO exhibits a consistently lower scale drift over the optimization-based method DSO and performs on par with the deep-learning-based approach DPVO. With respect to the overall alignment error, PRISM-VO exhibits a performance similar to that of DSO and is only slightly worse than DPVO, which likely benefits from learned geometric priors.

5.3 Quantitative Pose Estimation Error

We further evaluate PRISM-VO on the LiFMCR dataset [16] to assess generalization to other plenoptic cameras. Raw images are calibrated with LiFCal [15]. The dataset only contains plenoptic data. For comparison with conventional methods, we generate totally focused images using a central perspective projection on a common image plane through the main lens center. This emulates a pinhole camera model and allows standard monocular methods to be applied (see supplementary material).

Table 2: Quantitative trajectory accuracy comparison on the LiFMCR dataset [16]. We report translational RMSE ($RMSE_t$, [mm]) and rotational RMSE ($RMSE_r$, [°]) for PRISM-VO and the reference methods. Best and second-best results per sequence are highlighted in bold and underlined, respectively.

	PRISM-VO (ours)		DSO		ORB-SLAM3 (mono)		DPVO	
Scene	$RMSE_t$	$RMSE_r$	$RMSE_t$	$RMSE_r$	$RMSE_t$	$RMSE_r$	$RMSE_t$	$RMSE_r$
01	**41.66**	**2.98**	<u>59.32</u>	48.56	271.24	98.79	607.44	<u>3.62</u>
02	**57.81**	<u>6.09</u>	<u>112.56</u>	21.24	316.78	25.66	356.60	**2.60**
03	**38.34**	2.86	326.70	**1.97**	<u>123.20</u>	10.71	169.23	<u>2.04</u>
04	**10.53**	**2.83**	<u>150.89</u>	87.20	261.02	45.97	268.47	<u>3.01</u>
05	**9.32**	**2.89**	239.16	9.15	<u>121.64</u>	56.93	219.66	<u>2.92</u>
06	**12.21**	**3.31**	<u>543.90</u>	164.22	773.65	<u>153.28</u>	704.67	163.09
07	<u>110.87</u>	<u>11.86</u>	253.36	162.17	225.55	21.12	**5.87**	**1.99**

Ground-truth poses for all camera views are provided by a Vicon motion capture system, enabling per-frame pose evaluation. We evaluate absolute pose errors over the full trajectory and report the root mean square error (RMSE) denoted as $RMSE_t$ and $RMSE_r$ for the translation and rotation errors respectively in Tab. 2. PRISM-VO achieves an error of a few millimeters and less than three degrees on most sequences. Although the camera model is not exactly a pinhole model for monocular methods, the comparison gives a rough idea. Overall, PRISM-VO consistently achieves the lowest translational error, obtaining the best $RMSE_t$ on six out of seven sequences. Rotational accuracy is also competitive, achieving the best $RMSE_r$ in four sequences and the second-best results in most remaining cases. The results demonstrate the advantages for plenoptic cameras in close range scenes, where reliable depth can be measured.

5.4 Ablation Study

We assess PRISM-VO components through incremental ablation, isolating each contribution and measuring its impact on accuracy and drift reduction.

- **(1) Pinhole camera model:** Baseline using a central perspective camera model without plenoptic geometry.
- **(2) Plenoptic depth initialization and model integration:** Metric scale from the plenoptic depth and deep integration of the plenoptic camera model for the tracking and the bundle adjustment.
- **(3) Full PRISM-VO algorithm:** Adds an inverse virtual depth residual term to the optimization. Automatically balances depth residuals by their uncertainty, assigning higher influence to points with lower variance.

Fig. 12 reports the cumulative plots of the used metrics for the ablation stages. Compared to the pinhole baseline (1), adding plenoptic depth initialization and explicit plenoptic projection (2) reduces scale drift and makes absolute

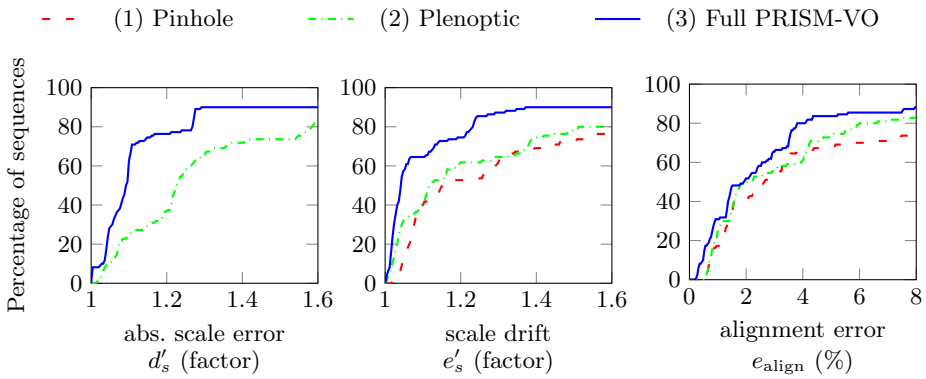


Fig. 12: Cumulative error distributions for the ablation study. Curves compare the pinhole baseline (1), plenoptic model with depth initialization (2), and the full PRISM-VO system with variance-weighted depth and adaptive balancing (3). Results show progressive improvement, highlighting the benefit of tightly integrating plenoptic data.

scale observable, confirming the benefit of exploiting plenoptic data. The full PRISM-VO algorithm (3) further improves all metrics, most notably scale consistency and alignment, through variance-weighted inverse virtual depth residuals and adaptive residual balancing. This uncertainty-aware integration of plenoptic depth within bundle adjustment enhances global consistency and reduces accumulated drift. Overall, each component contributes positively, with the largest gains from jointly optimizing and probabilistically weighting depth information.

6 Conclusion

We presented PRISM-VO, a novel VO method that tightly integrates plenoptic camera measurements into a photometric bundle adjustment. By explicitly modeling the plenoptic geometry, the method jointly leverages photometric and depth information via an inverse virtual depth residual. Experiments show improved trajectory consistency and accurate metric scale on challenging indoor and outdoor sequences. PRISM-VO outperforms the leading plenoptic VO pipeline and matches or surpasses monocular optimization and learning-based methods while reliably recovering metric scale. These results highlight the potential of tightly coupled plenoptic sensing and optimization for robust pose estimation and metric reconstruction with a single camera, despite limitations such as reduced resolution, a smaller field of view, and limited long-range depth accuracy.

Acknowledgements

This research was partially funded by the Federal Ministry of Research, Technology and Space of Germany in its program "FH-Kooperativ".

References

1. Al Assaad, M., Bazeille, S., Cudel, C.: Indirect visual odometry with a light-field camera. Intelligent Systems with Applications (ISWA) **28**, 200600 (2025). https://doi.org/10.1016/j.iswa.2025.200600
2. Bok, Y., Jeon, H.G., Kweon, I.S.: Geometric calibration of micro-lens-based light field cameras using line features. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **39**(2), 287–300 (2017). https://doi.org/10.1109/tpami.2016.2541145
3. Campos, C., Elvira, R., Rodríguez, J.J.G., Montiel, J.M., Tardós, J.D.: ORB-SLAM3: An accurate open-source library for visual, visual-inertial, and multi-map SLAM. Transactions on Robotics (T-RO) **37**(6), 1874–1890 (2021). https://doi.org/10.1109/TRO.2021.3075644
4. Dansereau, D.G., Mahon, I., Pizarro, O., Williams, S.B.: Plenoptic flow: Closed-form visual odometry for light field cameras. In: International Conference on Intelligent Robots and Systems (IROS). pp. 4455–4462. IEEE (2011). https://doi.org/10.1109/IROS.2011.6095080
5. Dansereau, D.G., Pizarro, O., Williams, S.B.: Decoding, calibration and rectification for lenslet-based plenoptic cameras. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1027–1034. IEEE (2013). https://doi.org/10.1109/cvpr.2013.137
6. Darwish, W., Bolsee, Q., Munteanu, A.: Plenoptic camera calibration based on sub-aperture images. In: International Conference on Image Processing (ICIP). pp. 3527–3531. IEEE (2019). https://doi.org/10.1109/icip.2019.8803473
7. Digumarti, S.T., Daniel, J., Ravendran, A., Griffiths, R., Dansereau, D.G.: Unsupervised learning of depth estimation and visual odometry for sparse light field cameras. In: International Conference on Intelligent Robots and Systems (IROS). pp. 278–285. IEEE (2021). https://doi.org/10.1109/IROS51168.2021.9636570
8. Dong, F., Ieng, S.H., Savatier, X., Etienne-Cummings, R., Benosman, R.: Plenoptic cameras in real-time robotics. International Journal of Robotics Research (IJRR) **32**(2), 206–217 (2013). https://doi.org/10.1177/0278364912469420
9. Dury, S., Bonatto, D., Sancho, J., Juarez, E., Teratani, M., Lafruit, G.: Structure-from-motion in the micro-image domain for uncalibrated plenoptic 2.0 cameras. International Journal of Computer Vision (IJCV) **134**(1), 34 (2026). https://doi.org/10.1007/s11263-025-02612-2
10. Engel, J., Koltun, V., Cremers, D.: Direct sparse odometry. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **40**(3), 611–625 (2018). https://doi.org/10.1109/TPAMI.2017.2658577
11. Engel, J., Schöps, T., Cremers, D.: LSD-SLAM: Large-scale direct monocular SLAM. In: European Conference on Computer Vision (ECCV). pp. 834–849. Springer (2014). https://doi.org/10.1007/978-3-319-10605-2_54
12. Fachada, S., Bonatto, D., Lafruit, G., Teratani, M.: Micro-image domain view synthesizer for free navigation with focused plenoptic cameras. Transactions on Multimedia **27**, 7179–7191 (2025). https://doi.org/10.1109/TMM.2025.3590906
13. Fachada, S., Bonatto, D., Losfeld, A., Lafruit, G., Teratani, M.: Pattern-free plenoptic 2.0 camera calibration. In: International Workshop on Multimedia Signal Processing (MMSP). pp. 1–6. IEEE (2022). https://doi.org/10.1109/mmsp55362.2022.9949312
14. Fachada, S., Losfeld, A., Senoh, T., Lafruit, G., Teratani, M.: A calibration method for sub-aperture views of plenoptic 2.0 camera arrays. In: International Workshop

on Multimedia Signal Processing (MMSP). pp. 1–6. IEEE (2021). https://doi.org/10.1109/mmsp53017.2021.9733556
15. Fleith, A., Ahmed, D., Cremers, D., Zeller, N.: LiFCal: Online light field camera calibration via bundle adjustment. In: DAGM German Conference on Pattern Recognition (GCPR). pp. 120–136. Springer (2024). https://doi.org/10.1007/978-3-031-85187-2_8
16. Fleith, A., Zirbel, J., Cremers, D., Zeller, N.: LiFMCR: Dataset and benchmark for light field multi-camera registration. In: International Symposium on Visual Computing (ISVC). Springer Nature Switzerland (2026). https://doi.org/10.1007/978-3-032-14492-8_35
17. Hahne, C., Aggoun, A., Velisavljevic, V., Fiebig, S., Pesch, M.: Refocusing distance of a standard plenoptic camera. Optics Express **24**(19), 21521–21540 (2016). https://doi.org/10.1364/OE.24.021521
18. Kaveti, P., Singh, H.: A light field front-end for robust SLAM in dynamic environments. arXiv preprint arXiv:2012.10714 (2020). https://doi.org/10.48550/arXiv.2012.10714
19. Klein, G., Murray, D.: Parallel tracking and mapping for small AR workspaces. In: International Symposium on Mixed and Augmented Reality (ISMAR). pp. 225–234. IEEE (2007). https://doi.org/10.1109/ISMAR.2007.4538852
20. Labussière, M., Teulière, C., Bernardin, F., Ait-Aider, O.: Blur-aware calibration of multi-focus plenoptic cameras. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2542–2551. IEEE (2020). https://doi.org/10.1109/cvpr42600.2020.00262
21. Labussière, M., Teulière, C., Bernardin, F., Ait-Aider, O.: Leveraging blur information for plenoptic camera calibration. International journal of computer vision (IJCV) **130**(7), 1655–1677 (2022). https://doi.org/10.1007/s11263-022-01582-z
22. Lasheras-Hernandez, B., Strobl, K.H., Izquierdo, S., Bodenmüller, T., Triebel, R., Civera, J.: Single-shot metric depth from focused plenoptic cameras. In: International Conference on Robotics and Automation (ICRA). pp. 9566–9573. IEEE (2025). https://doi.org/10.1109/ICRA55743.2025.11128276
23. Lumsdaine, A., Georgiev, T.: The focused plenoptic camera. In: International Conference on Computational Photography (ICCP). pp. 1–8. IEEE (2009). https://doi.org/10.1109/iccphot.2009.5559008
24. Lumsdaine, A., Georgiev, T., et al.: Full-resolution light field rendering. Indiana University and Adobe Systems, Tech. Rep **91**, 92 (2008)
25. Mur-Artal, R., Montiel, J.M.M., Tardos, J.D.: ORB-SLAM: A versatile and accurate monocular SLAM system. Transactions on Robotics (T-RO) **31**(5), 1147–1163 (2015). https://doi.org/10.1109/TRO.2015.2463671
26. Newcombe, R.A., Lovegrove, S.J., Davison, A.J.: DTAM: Dense tracking and mapping in real-time. In: International Conference on Computer Vision (ICCV). pp. 2320–2327. IEEE (2011). https://doi.org/10.1109/ICCV.2011.6126513
27. Noury, C.A., Teulière, C., Dhome, M.: Light-field camera calibration from raw images. In: International Conference on Digital Image Computing: Techniques and Applications (DICTA). pp. 1–8. IEEE (2017). https://doi.org/10.1109/DICTA.2017.8227459
28. O'brien, S., Trumpf, J., Ila, V., Mahony, R.: Calibrating light-field cameras using plenoptic disc features. In: International conference on 3D vision (3DV). pp. 286–294. IEEE (2018). https://doi.org/10.1109/3dv.2018.00041

29. Perwass, C., Wietzke, L.: Single-lens 3D camera with extended depth of field. In: Human Vision and Electronic Imaging (HVEI). vol. 8291, pp. 45–59. SPIE (2012). https://doi.org/10.1117/12.909882
30. Teed, Z., Deng, J.: DROID-SLAM: Deep visual SLAM for monocular, stereo, and RGB-D cameras. Advances in Neural Information Processing Systems (NeurIPS) **34**, 16558–16569 (2021)
31. Teed, Z., Lipson, L., Deng, J.: Deep patch visual odometry. Advances in Neural Information Processing Systems (NeurIPS) **36**, 39033–39051 (2023)
32. Von Stumberg, L., Cremers, D.: DM-VIO: Delayed marginalization visual-inertial odometry. Robotics and Automation Letters (RA-L) **7**(2), 1408–1415 (2022). https://doi.org/10.1109/LRA.2021.3140129
33. Wang, Y., Wang, L., Liang, Z., Yang, J., An, W., Guo, Y.: Occlusion-aware cost constructor for light field depth estimation. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19777–19786. IEEE (2022). https://doi.org/10.1109/cvpr52688.2022.01919
34. Xiao, Z., Liu, Y., Gao, R., Xiong, Z.: CutMIB: Boosting light field super-resolution via multi-view image blending. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1672–1682. IEEE (2023). https://doi.org/10.1109/cvpr52729.2023.00167
35. Zeller, N., Noury, C.A., Quint, F., Teulière, C., Stilla, U., Dhome, M.: Metric calibration of a focused plenoptic camera based on a 3D calibration target. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS Annals) **III–3**, 449–456 (2016). https://doi.org/10.5194/isprsannals-iii-3-449-2016
36. Zeller, N., Quint, F., Stilla, U.: Establishing a probabilistic depth map from focused plenoptic cameras. In: International Conference on 3D Vision (3DV) (2015). https://doi.org/10.1109/3DV.2015.18
37. Zeller, N., Quint, F., Stilla, U.: Depth estimation and camera calibration of a focused plenoptic camera for visual odometry. ISPRS Journal of Photogrammetry and Remote Sensing (P&RS) **118**, 83–100 (2016). https://doi.org/10.1016/j.isprsjprs.2016.04.010
38. Zeller, N., Quint, F., Stilla, U.: From the calibration of a light-field camera to direct plenoptic odometry. Journal of Selected Topics in Signal Processing (JSTSP) **11**(7), 1004–1019 (2017). https://doi.org/10.1109/jstsp.2017.2737965
39. Zeller, N., Quint, F., Stilla, U.: Scale-awareness of light-field camera-based visual odometry. In: European Conference on Computer Vision (ECCV). p. 732–747. Springer (2018). https://doi.org/10.1007/978-3-030-01237-3_44
40. Zeller, N., Quint, F., Stilla, U.: A synchronized stereo and plenoptic visual odometry dataset. arXiv preprint arXiv:1807.09372 (2018). https://doi.org/10.48550/arXiv.1807.09372
41. Zhao, Y., Li, H., Mei, D., Shi, S.: Metric calibration of unfocused plenoptic cameras for three-dimensional shape measurement. Optical Engineering **59**(7), 073104–073104 (2020). https://doi.org/10.1117/1.oe.59.7.073104
42. Zhou, P., Cai, W., Yu, Y., Zhang, Y., Zhou, G.: A two-step calibration method of lenslet-based light field cameras. Optics and Lasers in Engineering **115**, 190–196 (2019). https://doi.org/10.1016/j.optlaseng.2018.11.024