# Information-Theoretic Online Multi-Camera Extrinsic Calibration

Eric Dexheimer[1], Patrick Peluse[2], Jianhui Chen[2], James Pritts[2,3], and Michael Kaess[1]

***Abstract*—Calibration of multi-camera systems is essential for lifelong use of vision-based headsets and autonomous robots. In this work, we present an information-based framework for online extrinsic calibration of multi-camera systems. While previous work largely focuses on monocular, stereo, or strictly non-overlapping field-of-view (FoV) setups, we allow arbitrary configurations while also exploiting overlapping pairwise FoV when possible. In order to efficiently solve for the extrinsic calibration parameters, which increase linearly with the number of cameras, we propose a novel entropy-based keyframe measure and bound the backend optimization complexity by selecting informative motion segments that minimize the maximum entropy across all extrinsic parameter partitions. We validate the pipeline on three distinct platforms to demonstrate the generality of the method for resolving the extrinsics and performing downstream tasks. Our code is available at https://github.com/edexheim/info_ext_calib.**

***Index Terms*—SLAM, Calibration and Identification**

## I. Introduction

MULTI-SENSOR calibration is an essential task as increasingly complex intelligent systems are deployed in the world. Cameras are low-cost, lightweight, and low-power, which makes them suitable for robotics and consumer headsets. Compared to monocular setups, multi-camera systems allow for increased FoV, which in turn improves robustness and facilitates richer scene understanding. However, in order for these vision systems to operate continuously in the real-world, sensor calibration is required. While factory calibration using targets [1] is repeatable and accurate, it is also time-consuming and not possible for systems deployed in the wild. Ideally, platforms should be able to passively correct for changes, such as from physical shock and thermal deformation, during regular operation.

Accurate camera extrinsics are required for fundamental building blocks of autonomous systems, such as visual odometry (VO) and stereo matching. Typical online calibration systems focus on monocular, stereo, or multiple cameras with non-overlapping FoV. However, configurations vary greatly across platforms, and may contain non-traditional FoV overlap. Although treating cameras independently is general, accuracy will be limited as compared to leveraging potential inter-camera observations. Thus, the calibration framework should incorporate all information while remaining flexible.

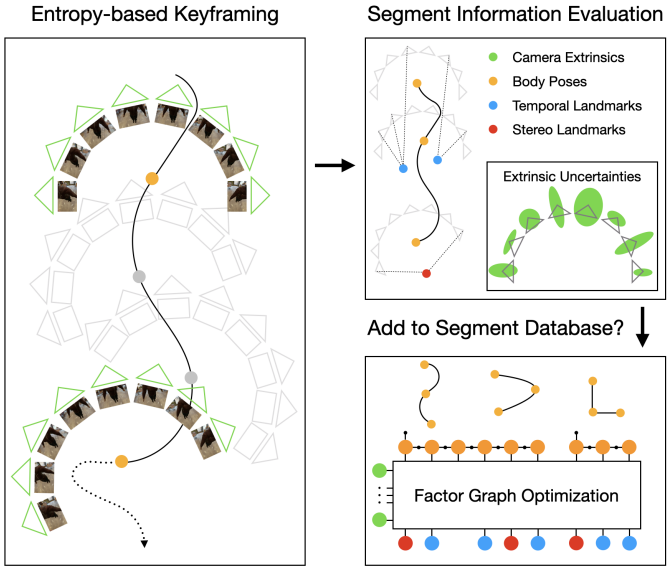



Fig. 1: Overview of proposed calibration pipeline. Keyframes are selected based on the extrinsic calibration entropy, the information content of fixed-length motion segments are evaluated, and a bounded database of segments is optimized to improve the extrinsics estimates.

In this work, we develop a general information-theoretic framework for online multi-camera extrinsic calibration. The frontend tracks intra-camera features temporally and matches inter-camera features on select keyframes, while the backend performs factor graph optimization of the extrinsics and auxiliary variables. Since the complexity can greatly increase with a large number of cameras, we first propose a novel entropy-based multi-camera keyframe selection method to sparsify the set of body poses. After a number of keyframes, a motion segment is generated, and its information content is checked against a database of previous segments. Compared to previous methods, the database scales independently of the number of cameras by minimizing the maximum entropy across all extrinsic parameter partitions. A high-level diagram of the proposed method is shown in Fig. 1. We demonstrate the performance of our pipeline on three distinct configurations: a stereo camera on-board a micro-aerial vehicle (MAV) [2], an 8-camera human-facing headset rig over realistic simulation data, and a 5-stereo platform on a ground vehicle.

## II. Related Work

### A. Multi-Camera Extrinsic Calibration

Since multi-camera calibration is a fundamental requirement for many autonomous systems, a wide variety of online methods across different platforms have been proposed. Stereo

Manuscript received: September, 9, 2021; Revised December, 1, 2021; Accepted January, 3, 2022.

This paper was recommended for publication by Editor Javier Civera upon evaluation of the Associate Editor and Reviewers' comments.

Please see Section VIII for support and grants.

[1]Eric Dexheimer and Michael Kaess are with the Robotics Institute, Carnegie Mellon University (CMU), Pittsburgh, PA 15213, USA. {edexheim, kaess}@andrew.cmu.edu.

[2]Patrick Peluse and Jianhui Chen are with Facebook Reality Labs (FRL), Pittsburgh, PA, USA. James Pritts contributed to this work while at FRL.

[3]James Pritts is with Chalmers University of Technology, Gothenburg, Sweden.

Digital Object Identifier (DOI): see top of this page.

extrinsic calibration methods minimize epipolar error or re-projection error [3]. Beyond stereo setups, multi-camera rigs have become very popular due to greater robustness for multi-camera VO [4], [5]. However, deployment of these setups is still a challenge, as resolving accurate calibration parameters during operation is required to achieve suitable performance. Online methods for multi-camera extrinsic calibration focus on independent extrinsic rotation estimation [6], monocular map matching, [7], non-overlapping extrinsic estimation for a car with odometry [8], independent multi-camera visual-inertial calibration [9], or non-overlapping stereo setups [10]. Treating cameras independently when they have overlapping FoV can allow significant relative extrinsic errors, which will hinder downstream dense correspondence algorithms. An efficient, general framework of calibration as a pose alignment problem with an application to two cameras is presented in [11]. Therefore, all of these works either focus on specific use cases or generalize the problem such that accuracy may be limited.

### B. Entropy-based Keyframe Selection

Selecting informative keyframes is an essential component of SLAM systems to bound computational complexity. [12] thresholds a ratio of direct image alignment entropies, specifically registering both the current frame and first frame after the last keyframe to the keyframe. [5] follows a similar ratio threshold, but instead measures the pose entropy based on Perspective-n-Point (PnP) optimization with respect to the current SLAM map. While these methods are suitable for SLAM, pose estimation entropy from either keyframe alignment or map estimation will largely be monotonic as the pose uncertainty increases with exploration. For extrinsic calibration, this is less clear, as the observability of parameters depends on the motion itself, not just the registration to a map. Furthermore, differential entropy can also be negative, so the ratio will not generalize in all cases. [13] computes an independent sum over entropy reduction for each map point. In this work for tractable multi-camera calibration, we wish to avoid maintaining a full 3D map for the frontend, and instead delay structure computation until the backend. In addition, each of these works lacks a probabilistic interpretation for the ratio of entropy and sum of entropy reduction heuristics, while we leverage a different measure in terms of conditional mutual information.

### C. Segment-based Self-Calibration

Calibration problems are expensive due to measurements depending on calibration variables. Including landmarks in the optimization can improve accuracy, but increases complexity. Since all portions of a trajectory are not equally informative about calibration parameters, some methods maintain a priority queue of useful segments. This ensures the optimization is tractable for real-time optimization of camera intrinsics [14] visual-inertial parameters [15], and slowly drifting camera extrinsics [11]. In this work, we develop a segment-based framework specifically for the multi-camera use case to maintain accuracy while limiting the complexity of an increasing number of cameras.

## III. Preliminaries

### A. Problem Formulation

For a platform with $K$ cameras, we denote each as $C_k, k \in 1, \ldots, K$. The extrinsic transformation from camera frame $k$ to the body (rig) frame $B$ is $\mathbf{T}_{BC_k} \in SE(3)$. In some cases, we will work directly with rotations $\mathbf{R} \in SO(3)$ and translations $\mathbf{t} \in \mathbb{R}^3$. As input, the system receives a stream of synchronized frames from each camera $C_k$, as well as an odometry estimate. The position estimate is the transformation from the body frame at time $t$ to the world frame, and is denoted as $\mathbf{T}_{WB_t}$. We focus on odometry information because it is applicable to a wide variety of platforms, such as those generating LiDAR or GPS-based state estimates, which can constrain the scale of the extrinsics. As output, we wish to optimize for the extrinsic calibrations $\mathbf{T}_{BC_k}, k \in 1, \ldots K$, as well as auxiliary variables. Specifically, we also optimize for body poses $\mathbf{T}_{WB_t}, t \in 1, \ldots, T$, temporally-tracked 3D landmarks, $\mathbf{l}_m, m \in 1, \ldots, M$, and stereo landmarks $\mathbf{s}_n, n \in 1, \ldots, N$.

### B. Nonlinear Least Squares

We wish to perform *maximum a posteriori* (MAP) estimation over a set of unknown variables $\mathbf{x}$ under the assumption of Gaussian measurement noise. The nonlinear least squares (NLLS) minimization is

$$\hat{\boldsymbol{x}} = \underset{\mathbf{x}}{\operatorname{argmin}} \frac{1}{2} \sum_i ||\mathbf{z}_i - \mathbf{f}_i(\mathbf{x}_i)||^2_{\boldsymbol{\Sigma}_i} \tag{1}$$

where $\mathbf{z}_i$ is the measurement, $\mathbf{f}_i(\mathbf{x}_i)$ is a nonlinear prediction function based on the current state, and $\boldsymbol{\Sigma}_i$ is the measurement covariance. Given an initial guess for the state $\mathbf{x}^0$ and after linearizing the constraints at the current variable estimates, we can solve for a state update vector

$$\delta\hat{\boldsymbol{x}} = \underset{\delta\mathbf{x}}{\operatorname{argmin}} \sum_i ||\mathbf{z}_i - \mathbf{f}_i(\mathbf{x}_i^0) - \mathbf{F}_i \delta\mathbf{x}_i||^2_{\boldsymbol{\Sigma}_i} \tag{2}$$

where $\mathbf{F}_i$ is the Jacobian of the measurement function $\mathbf{f}_i$ evaluated at the current linearization point. This can be solved by stacking the terms into the Gauss-Newton normal equations

$$\mathbf{J}^T \boldsymbol{\Sigma}_\mathbf{z} \mathbf{J} \delta\hat{\boldsymbol{x}} = \mathbf{J}^T \boldsymbol{\Sigma}_\mathbf{z} \mathbf{r} \tag{3}$$

where $\mathbf{J}$ stacks the $\mathbf{F}_i$ Jacobian terms, $\boldsymbol{\Sigma}_\mathbf{z}$ creates a block-diagonal matrix from the measurement covariances $\boldsymbol{\Sigma}_i$, and $\mathbf{r}$ stacks the residuals $\mathbf{z}_i - \mathbf{f}_i(\mathbf{x}_i^0)$.

### C. Posterior Information Content

We are often interested in approximating the uncertainty of the posterior instead of just the point estimate. We can efficiently recover a Gaussian approximation to the posterior via the Laplace approximation [16] with the Fisher information matrix $\mathbf{I}_\mathbf{x} = \mathbf{J}^T \boldsymbol{\Sigma}_\mathbf{z} \mathbf{J}$. The differential entropy of a $d$-dimensional multivariate Gaussian can be used to express the information content as a scalar value:

$$H(\mathbf{x}) = -\frac{1}{2} \ln |\mathbf{I}_\mathbf{x}| + \frac{d}{2}(1 + \ln(2\pi)). \tag{4}$$

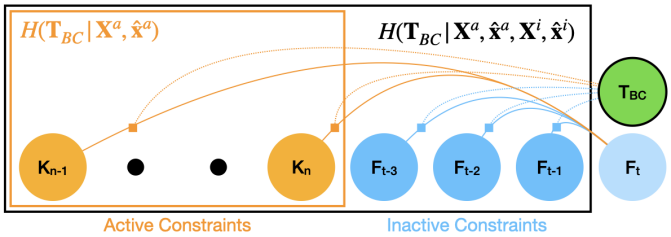



Fig. 2: Example of active and inactive constraints with current frame and corresponding entropy measures for keyframe selection.

## IV. Frontend

The framework consists of a frontend, which tracks monocular features, selects keyframes, and finds stereo correspondences, and a backend, which incorporates this information to optimize for the extrinsics, as shown in Fig. 1. Within the frontend, the proposed entropy-based keyframe selection method ensures a minimal number of poses are selected for backend optimization, while providing sufficient information to resolve the extrinsics. Since there may be potentially many pairs of cameras with FoV overlap, we conduct stereo matching after a fixed number of keyframes. Furthermore, since stereo matching is performed across cameras at a single timestep, it can operate at a lower frequency than keyframing.

### A. Entropy-based Keyframing

Each camera performs independent feature detection and tracking as to not bias the calibration. First, features are detected using FAST corners [17] with grid bucketing to ensure an even distribution of features. Next, features are tracked temporally using KLT [18], and each camera performs 5-point essential matrix RANSAC to prune outliers [19].

While every frame could be passed to the backend, this would introduce redundancy, as little information about calibration parameters is gained without significant motion. Common keyframe heuristics include motion or feature thresholds, but these methods do not generalize across camera rigs and environments. Entropy-based keyframe selection has been leveraged for visual odometry [12], [13], [5], but pose estimation is not equivalent to calibration. For example, motion could be well-constrained for image alignment or PnP in a straight-line trajectory, but the extrinsic translations would not be. The uncertainty of the current pose will increase since the last keyframe, while the calibration uncertainty is less clear.

We introduce an approximate entropy-based keyframe selection method. We do not require a sparse map for pose estimation, and since explicitly maintaining one as in [13], [5] is expensive, we defer the use of triangulation and bundle adjustment to the backend. Given only feature tracks, estimates of the extrinsics, and the locally-accurate odometry, we wish to determine when a keyframe provides sufficiently new information. By formulating a NLLS problem with the extrinsics as unknowns, we can measure the entropy of the extrinsics $H(\mathbf{T}_{BC}) \triangleq H(\mathbf{T}_{BC_1}, \ldots, \mathbf{T}_{BC_K})$. Therefore, a residual function dependent on the extrinsics is required.

The multi-camera rig can be viewed as a generalized camera rig without a single center of projection, which follows the generalized epipolar constraint (GEC) [20]. From the notation in [21], the jth Plücker line in camera $k$ is denoted as

$$\boldsymbol{\ell}_{kj} = \begin{bmatrix} (\mathbf{R}_{BC_k}\hat{\mathbf{x}}_{kj})^T & ([\mathbf{t}_{BC_k}]_\times \mathbf{R}_{BC_k}\hat{\mathbf{x}}_{kj})^T \end{bmatrix}^T \tag{5}$$

where $\hat{\mathbf{x}}_{kj}$ is the normalized image coordinates, while $\mathbf{R}_{BC_k}$ and $\mathbf{t}_{BC_k}$ are the extrinsic rotation and translation, respectively, of camera $k$. Then, each Plücker line correspondence is related by the GEC:

$$r_{kj} = \boldsymbol{\ell}'^T_{kj} \begin{bmatrix} [\mathbf{t}]_\times \mathbf{R} & \mathbf{R} \\ \mathbf{R} & 0 \end{bmatrix} \boldsymbol{\ell}_{kj} \approx 0 \tag{6}$$

where $\mathbf{R}$ and $\mathbf{t}$ are the relative rotation and translation, respectively, between the two body frames, while $\boldsymbol{\ell}_{kj}$ and $\boldsymbol{\ell}'_{kj}$ are Plücker line correspondences between two time steps. In reality, these terms will be nonzero due to noise in the odometry, extrinsics, and data association. However, we are only interested in the information content of the extrinsics, and treat the body poses as locally accurate. We thus formulate these correspondences into the NLLS

$$\underset{\mathbf{R}_{BC_k}, \mathbf{t}_{BC_k}, k \in 1,\ldots,K}{\operatorname{argmin}} \sum_k^K \sum_j ||r_{kj}||^2 . \tag{7}$$

As mentioned in Section III, the entropy provides a scalar value measuring how well the extrinsics variables are constrained. Since each camera's correspondences are independent given fixed body poses, this can be efficiently calculated using the block diagonal determinant rule:

$$H(\mathbf{T}_{BC}) = -\frac{1}{2}\ln(|\mathbf{I}_{\mathbf{R}_1}| \cdot |\mathbf{I}_{\mathbf{t}_1}| \cdot \cdots \cdot |\mathbf{I}_{\mathbf{R}_K}| \cdot |\mathbf{I}_{\mathbf{t}_K}|) + C \tag{8}$$

where $C$ is the constant term from Eq. 4. As mentioned in [12], [13], [5], absolute thresholds on entropy do not generalize. We are also specifically interested in whether significant information about the calibration parameters is gained by inserting a new keyframe. Therefore, we evaluate the entropy using features with a reference keyframe, denoted *active*, as well as the entropy with both these features and features with no keyframe reference, denoted *inactive*. As shown in Fig. 2, active features have a correspondence with their first keyframe observation, while inactive features have one with their first observation in a pose buffer since the last keyframe.

The ratio heuristics proposed in [12], [5] do not generalize, as differential entropy can be negative, so sign flips are possible. Instead, we propose a principled threshold by leveraging the conditional mutual information:

$$\begin{aligned} &I(\mathbf{T}_{BC}; \mathbf{X}^i, \hat{\mathbf{x}}^i | \mathbf{X}^a, \hat{\mathbf{x}}^a) = \\ &H(\mathbf{T}_{BC}|\mathbf{X}^a, \hat{\mathbf{x}}^a) - H(\mathbf{T}_{BC}|\mathbf{X}^a, \hat{\mathbf{x}}^a, \mathbf{X}^i, \hat{\mathbf{x}}^i) \end{aligned} \tag{9}$$

which measures the information that we can observe about the calibration unknowns by incorporating the inactive poses $\mathbf{X}^i$ and normalized image coordinate correspondences $\hat{\mathbf{x}}^i$, when the active poses $\mathbf{X}^a$ and correspondences $\hat{\mathbf{x}}^a$ are already observed. Since mutual information is in the range $[0, \infty)$, we select a threshold $p \in (0, 1]$ such that a new keyframe is chosen when

$$I(\mathbf{T}_{BC}; \mathbf{X}^i, \hat{\mathbf{x}}^i | \mathbf{X}^a, \hat{\mathbf{x}}^a) > -\ln(p). \tag{10}$$

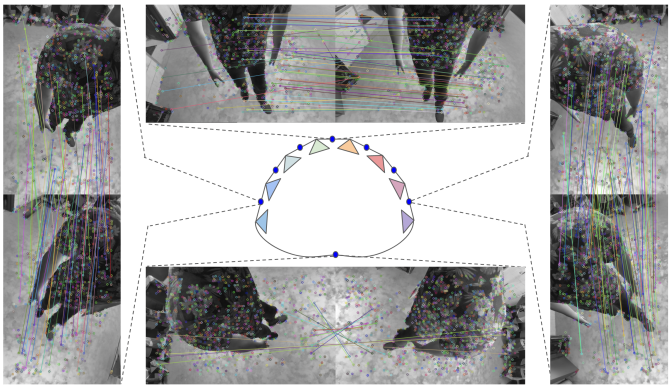


Fig. 3: Example of stereo constraints for an 8-camera human facing rig. Both static and dynamic observations are included, which improves robustness of matching against the large viewpoint changes.

### B. Stereo Observations

While the temporal feature tracking and keyframing treats cameras independently, calibrating based only on monocular information is not sufficient for tasks such as dense correspondence and reconstruction when FoV overlap is present. These relative constraints may also improve robustness to odometry drift, as stereo constraints do not rely on temporal information. Furthermore, this allows for handling points that violate the static landmark assumption, such as points on the human body, which may provide essential information about the extrinsics. As a user-specified input, a set of pairs of cameras to attempt stereo matching is listed, which can be determined via rough FoV estimates or inspection of images.

For each pair, a matching procedure similar to [22] is performed. First, in addition to the current tracked features in each image, FAST features [17] are detected while ensuring distribution across the image via bucketing. While [22] uses BRIEF descriptors [23], we use ORB descriptors [24] since the cameras are not assumed to have near-identity rotation as in most stereo pairs. Features are undistorted to normalized image coordinates, and before being matched with cross-consistency and a uniqueness threshold, potential matches are pruned using the current extrinsics and a loose epipolar Sampson threshold [25]. Then, 5-point RANSAC [19] finds a single hypothesis, and if there are enough matches that triangulate in front of both cameras, features are passed to the backend. An example of stereo constraints for a human-facing 8-camera rig is shown in Fig. 3, where consecutive pairs are checked.

## V. Backend

The backend takes in temporally-tracked and stereo features, as well as initial keyframe pose estimates. A motion segment is generated after a fixed number of keyframes. The segment's information content with respect to the calibration parameters is evaluated based on the factor graph formulation in Section V-A. A segment is accepted by the segment database if it improves the total information content in the database, which is formulated specifically for the multi-camera use case as described in Section V-B. If a new segment is added, the database factor graph is modified, and optimization will recommence.

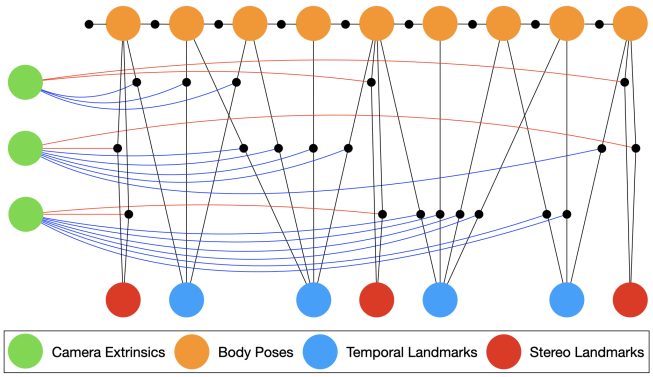



Fig. 4: Factor graph for 3-camera system consisting of body poses, extrinsics, and landmarks as the unknowns to be solved. In this example, stereo matching is conducted every 4 keyframes.

### A. Factor Graph Formulation

An example of a factor graph for a 3-camera rig is shown in Fig. 4. Extrinsic calibration variables are densely connected to projection factors, and stereo matching is run pairwise between cameras after a fixed number of keyframes. Given the set of landmarks observed by camera $C_k$ observed at time $t$ as $O(B_t, C_k)$, the NLLS problem is:

$$\begin{aligned}\min\ & ||\mathbf{p}(\mathbf{T}_{WB_1}, \mathbf{z_p})||^2_{\mathbf{\Sigma_p}} + \sum_{t=1}^{T-1} ||\mathbf{o}(\mathbf{T}_{WB_t}, \mathbf{T}_{WB_{t+1}}, \mathbf{z_o})||^2_{\mathbf{\Sigma}_{\mathbf{o},t}} \\ &+ \sum_{k}^{K}\sum_{t}^{T}\left(\sum_{\mathbf{l}_m}^{O(B_t,C_k)} ||\mathbf{r}^{t,k}(\mathbf{l}_m)||^2_{\mathbf{\Sigma_r}} + \sum_{\mathbf{s}_n}^{O(B_t,C_k)} ||\mathbf{r}^{t,k}(\mathbf{s}_n)||^2_{\mathbf{\Sigma_r}}\right) \end{aligned} \tag{11}$$

where the prior error $\mathbf{p}$ constrains the 6-DOF gauge freedom to a frame origin $\mathbf{z_p}$, the odometry error $\mathbf{o}$ enforces 6-DOF consistency with relative measurement $\mathbf{z_o}$, and $\mathbf{r}^{t,k}(\mathbf{l}) = \mathbf{z_r} - \pi(\mathbf{T}_{WB_t}, \mathbf{T}_{BC_k}, \mathbf{l})$ is the reprojection error for a landmark $\mathbf{l}$ with observation $\mathbf{z_r}$. Measurement covariances are assumed to be known *a priori*. Stereo landmarks lack temporal constraints, so they only enforce relative constraints between cameras, while monocular landmarks are temporally tracked. However, these may still incorporate observations from other cameras if they are matched via stereo.

### B. Segment Database for Multi-Camera Extrinsic Calibration

While all keyframes and landmarks could be continuously added to the optimization, the factor graph will grow unbounded. In addition, not all motions or environments are conducive for observing calibration parameters, so we follow the methodology of [26] and [27], in which a priority queue of the most informative segments is maintained. However, both [26] and [27] have a bounded number of parameters, as they only focus on monocular intrinsics and monocular visual-inertial odometry, respectively. In [27], calibration parameters are partitioned into three distinct partitions, each of which has a maximum number of segments. This ensures that all sets of calibration parameters are well-constrained, as a single priority queue could be dominated by only the most observable parameters. Following this technique for the multi-camera case, however, can create significant variation in the total number of segments. For example, 8 cameras each with rotation and translation partitions, along with a maximum of 4 segments

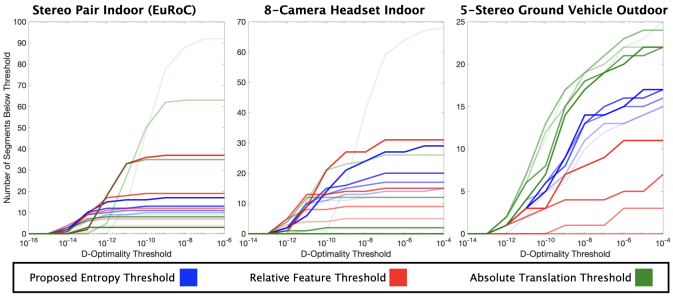
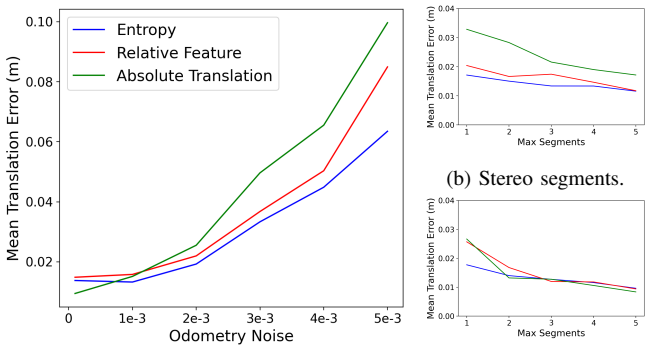



Fig. 5: Number of 10-pose segments that fall below D-optimality threshold for different keyframe measures. Five representative thresholds are tested per method, with the entropy and feature thresholds in $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ and translation threshold in $\{0.01, 0.05, 0.10, 0.50, 1.00\}$ meters. Greater transparency indicates a lower threshold.




(a) Stereo odometry noise. (b) Stereo segments. (c) Headset segments.

Fig. 6: Extrinsic errors across keyframe methods.

per partition, can result in a full database ranging from 4 to 64 segments. Instead, we propose to limit the absolute number of segments by minimizing the maximum entropy across all partitions.

To calculate the information content of a segment, landmarks are initialized via robust triangulation, and the segment factor graph is optimized so that the Laplace approximation can be utilized. For each camera $C_k$, we then obtain the marginal information in the extrinsic rotation $\mathbf{I}_{\mathbf{R}_k} = \mathbf{\Sigma}^{-1}_{\mathbf{R}_k}$ and translation $\mathbf{I}_{\mathbf{t}_k} = \mathbf{\Sigma}^{-1}_{\mathbf{t}_k}$ separately. Since different types of motion and observations may be useful for resolving rotation and translation, we keep them separate. To quantify information content with a scalar for new segment $\bar{S}$, we calculate the entropy as in Eq. 4 for each of the $2K$ parameter partitions $\boldsymbol{\theta}_j$, which is denoted as $H^{\bar{S}}(\boldsymbol{\theta}_j)$. Then, the partition with the maximum entropy in the current database $D$ is approximated by independently evaluating each segment $S_i$ and adding the marginal information matrices for a given partition $\mathbf{I}^{S_i}_{\boldsymbol{\theta}_j}$:

$$\mathbf{I}^{D}_{\boldsymbol{\theta}_j} = \sum_{S_i} \mathbf{I}^{S_i}_{\boldsymbol{\theta}_j} \tag{12}$$

$$H^{D}(\boldsymbol{\theta}_j) = -\frac{1}{2}\ln|\mathbf{I}^{D}_{\boldsymbol{\theta}_j}| + \frac{d}{2}(1+\ln(2\pi)) \tag{13}$$

$$j_{\max} = \underset{j}{\operatorname{argmax}}\, H^{D}(\boldsymbol{\theta}_j). \tag{14}$$

Note that the information content depends only on observations within a segment, which as mentioned in [15], is a conservative estimate, but is efficient and avoids biasing the estimation. If the entropy for partition $j_{\max}$ of the new segment $H^{\bar{S}}(\boldsymbol{\theta}_{j_{\max}})$ is less than that of the segment with maximum entropy in the same partition, then the new segment is a candidate. However, in order for the new segment to replace the old one, all entropy partitions must not increase beyond the current maximum entropy, $H^{D}(\boldsymbol{\theta}_{j_{\max}})$. This greedy selection strategy ensures that the maximum entropy across all partitions in the database never increases.

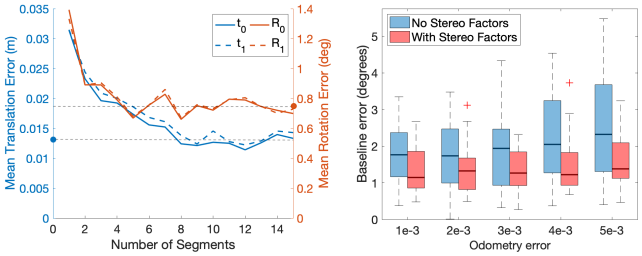
### C. Optimization

If a new segment is added, a new optimization will be triggered to determine the calibration variables. Segments are joined using odometry constraints or based on sufficient landmark co-visibility similar to [15]. For collections of segments that are disjoint, we need to place a pose prior on the first pose of each collection in order to constrain the gauge freedom. Observations of common features across segments are merged, and landmarks are initialized using robust triangulation. We use Levenberg-Marquardt to minimize the nonlinear objective. In practice, reprojection factors from Eq. 11 use the Huber cost for robustness against outliers.

## VI. Results

### A. Entropy-based Keyframe Selection

In Fig. 5, we compare the proposed entropy-based keyframe method against two baselines, a relative threshold on tracked features and an absolute translation threshold across three unique datasets. The plots look at the determinant of the extrinsic marginal covariances, known as the D-optimality, of generated 10-pose segments. In order to pass the threshold, every camera in the rig must have a D-optimality less than the threshold. An ideal threshold would induce a steep curve as far to the left as possible, indicating that many informative segments are being generated. Across all datasets, the proposed entropy threshold performs best, as it induces steep curves for a low number of segments, while the baselines have varying performance. Note that generating the largest number of segments is not the most desirable, as quality is sacrificed for quantity. For example, the translation threshold performs poorly in the indoor datasets, while the feature threshold generates the least informative segments in the outdoor dataset. The translation baseline performs the best in the outdoor experiment due to the fast vehicle motion and low frame rate, so almost all frames are selected. Since the proposed entropy threshold leverages the GEC, which does not directly correspond to the SfM problem used for marginal covariances, there can be some discrepancy in the objective for the outdoor dataset. The Plücker line correspondences may provide some information, but they do not account for all of a feature's observations, and if the points cannot be triangulated, which is often in the case of a forward-moving ground vehicle, then they provide no information for the extrinsic marginal covariance. The highest threshold of 0.9 for the proposed entropy method generalizes well across all datasets by generating highly-informative segments and a substantial number of useful ones, so this is used in the framework.




(a) (b)

Fig. 7: EuRoC results. 7a Mean error of extrinsics vs. max number of segments. Data points on y-axis and corresponding lines indicate average error from offline batch solution. 7b Baseline error vs. simulated odometry error with and without stereo factors.

To further evaluate the impact on the final extrinsic estimates, we conduct experiments on the two indoor environments, as we have a reliable comparison to the offline calibration of the stereo pair, and the simulated 8-camera dataset has known ground-truth extrinsics. We do not include the 5-stereo ground vehicle data, as there are no reliable known extrinsics since the pairs do not significantly overlap. In Fig. 6a, we limit the number of segments to 4 and run ten trials for each odometry noise parameter. The thresholds are selected according to the best curves from Fig. 5 for a low number of segments. Despite assuming locally accurate odometry in the formulation, the proposed entropy method is not severely affected, and is still able to perform well. In Fig. 6b and Fig. 6c, ten trials are run for an increasing number of segments, and the entropy method performs best, especially for a low number of segments. This demonstrates the ability of the method to provide informative segments. Once more segments are allowed, the methods all converge to similar errors as there is sufficient information present for all in the optimization. This result demonstrates that the proposed entropy-based keyframe procedure reduces the influence on the amount of data, and is useful for limiting the problem size in real-time operation.

### B. MAV Stereo Pair (EuRoC)

*1) Effect of Segment Database Size:* First, we evaluate the accuracy of the calibration against the maximum number of segments permitted in the database. We run 25 trials with simulated error on the pose estimates from the ground-truth motion capture and on the extrinsics provided via offline calibration. Since the database can be maintained over multiple sessions, we run on the datasets V1_02_medium and V2_02_medium, which provide reasonable motion for constraining the calibration parameters. The mean error from the offline calibration versus the number of segments is shown in Fig. 7a. In general, the error for the more easily observable rotation only decreases until 5 segments, while translation error stabilizes at 8 segments. We also plot the average error from a batch solution using all possible segments which does not run in real-time. The online method demonstrates asymptotic convergence to the batch solution, but also shows bias from the offline Kalibr calibration. As mentioned in [2], there are potential errors accumulated in the IMU-Vicon calibration due to deteriorated motion tracking and time offsets.

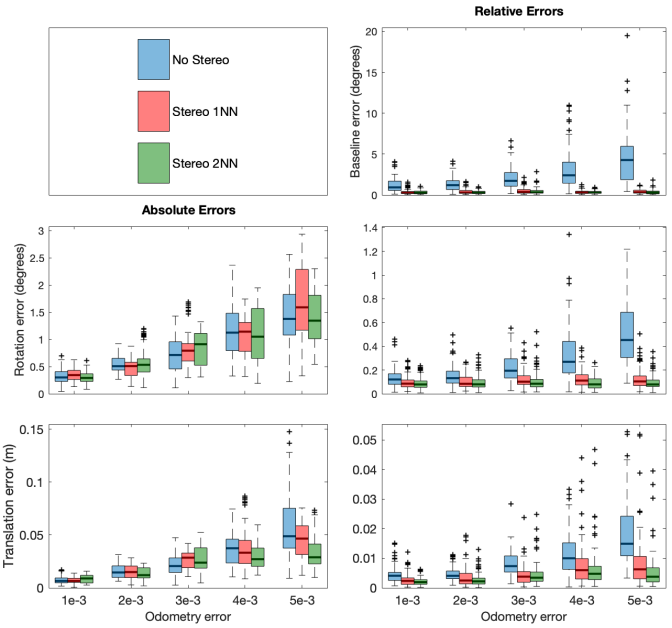



Fig. 8: Summary of errors across 12 datasets for 8-camera system with varying odometry noise. Each boxplot is for 96 data points.

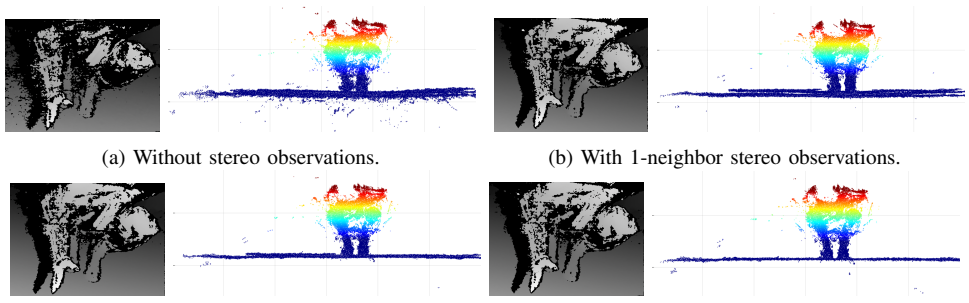
*2) Effect of Stereo Factors:* While optimizing for each of the extrinsics without relative constraints is often done, we show the effect of the estimated baseline direction vs. increasing odometry noise in Fig. 7b. Note that using stereo factors every 3 keyframes avoids any significant increase in the baseline error, which will improve disparity estimation.

### C. Simulated 8-Camera Human-Facing Headset Rig

We also conduct experiments on an 8-camera human-facing rig in a realistic simulation setup. Seven-thousand frames of human animation were captured using an Xsens MVN Link suit, and the motion was then retargeted to Mixamo bodies with varying size and appearance. A differentiable renderer, Unity HDRP DX12, provides a physically-based material and lighting setup for the human model and environment. Matterport environments with point lights and shadows were used to create realistic situations, as shown in Fig. 3.

A total of 12 datasets with varying indoor environments and characters were tested. Compared to standard stereo setups, there are significant viewpoint changes, wide-angle lenses, and potential overlap beyond neighboring pairs. In addition, a large portion of the image is dynamic, but these points can still be leveraged by stereo factors. We utilize a human segmentation mask by taking the ground-truth and roughly dilating it, which could alternatively be provided by learned methods.

*1) Accuracy Across Datasets:* We test three configurations: without stereo constraints, using 1-nearest neighbor (1NN) stereo constraints, and up to 2-nearest neighbor (2NN) constraints. This results in 0, 8, and 16 pairs to be checked. Stereo matching is only run every 10 keyframes, or once per segment. The segment database is limited to 4 segments to achieve real-time performance. A summary of the simulated results are shown in Fig. 8. Each boxplot contains 8 cameras across 12 environments, for a total of 96 data points. As expected, both absolute rotation and translation errors increase with odometry



(a) Without stereo observations. (b) With 1-neighbor stereo observations.

(c) With 2-neighbor stereo observations. (d) With ground-truth extrinsics.

Fig. 9: Point clouds for 8-camera rig via pairwise stereo rectification, SGBM [28] disparity estimation, and triangulation.

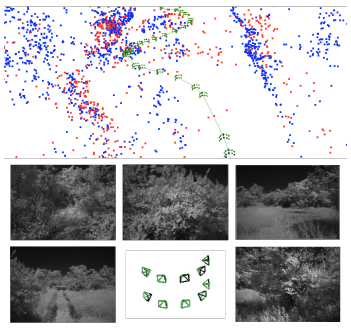


Fig. 10: 5-stereo NIR data with poses, landmarks, images, and rig.

noise. All methods are comparable for rotation, while 2NN performs best for absolute translation, followed by 1NN, since the additional long-range constraints are able to better resolve the extrinsics. For relative errors, not using stereo results in significant baseline errors with increasing noise, while both 1NN and 2NN stereo configurations are able to constrain these errors. In terms of relative calibration errors, 2NN performs slightly better than 1NN, while no stereo demonstrates that the lack of relative constraints permits significant errors.

*2) Qualitative Stereo:* A qualitative example example of dense correspondence after calibration is shown in Fig. 9 where semi-global block matching (SGBM) [28] is used. With no stereo observations, the quality of the disparity is not sufficient, as the epipolar lines are not accurate. Clearly, the 3D cloud accumulated by triangulating 3D points is inconsistent, especially in the ground plane. For 1NN, the disparity, or local consistency, is improved significantly, but the global consistency and flat ground is more evident for 2NN.

### *D. 5-Stereo Ground Vehicle*

We test calibration on a 5-stereo near-infrared (NIR) rig with a front-facing 190 degree FoV setup on a ground vehicle (GV). The datasets are challenging, as the vehicle can move up to a few meters per second in forest environments, yet the camera system only has 4 frames-per-second (FPS). One 650m trajectory is used for calibration purposes, while a second 2300m trajectory is used to evaluate VIO performance.

ORB matching [24] is included as a backup to KLT [18] since there is significant motion between frames. The database size is limited to 4 segments, and stereo matching is performed every 3 keyframes. A prior with 1cm standard deviation on the extrinsic translations from CAD is included due to the lack of observability while allowing the baseline to be optimized. The odometry estimates come from a GPS-IMU state estimate, which is modified with 5e-3 noise to create drift, and the extrinsics are initialized with perturbed values. An example visualization of the data is shown in Fig. 10.

The multi-stereo VIO pipeline uses the extrinsics to perform disparity estimation, and matching is conducted for 2D-3D ORB correspondences from disparity. Outlier rejection selects P3P models from the front stereo pair as it is the most reliable, while inliers are checked across all cameras. Stereo reprojection factors and IMU preintegration factors [29] are added to a fixed-lag smoother inspired by the backend of [30].

| | MAV 2 cam 1 stereo | Headset 8 cam 8 stereo | Headset 8 cam 16 stereo | GV 10 cam 5 stereo |
|---|---|---|---|---|
| Feature Tracking | 18.6 | 50.4 | 55.2 | 84.4 |
| Keyframe Selection | 0.2 | 1.3 | 1.3 | 0.5 |
| Stereo Matching | 82.4 | 274.6 | 470.1 | 59.0 |
| Segment Info Calculation | 128.3 | 910.9 | 1185.8 | 970.7 |
| Database Proposal | 58.8 | 77.6 | 108.2 | 268.3 |
| Optimization Step | 0.004 | 0.006 | 0.010 | 0.005 |

TABLE I: Timing in milliseconds for key steps of framework across datasets. All results are run on a 2.7 GHz Quad-Core Intel Core i7.

Four configurations are tested: offline calibration in conjunction with a CAD estimate of the IMU pose, online calibration without stereo constraints, calibration with stereo constrains, and a batch offline solution with stereo where the number of segments is not limited so that the entire trajectory is used. Qualitative VIO trajectories and relative pose error (RPE) statistics comparing the four are shown in Fig. 11. Without stereo constraints, the extrinsics are not suitable, and allow significant drift in the ground plane. Including stereo constraints, however, produces a significantly better trajectory. While there is some z-drift for stereo and batch as compared to the offline/CAD calibration, this is largely due to a single bad pose estimate, as seen by the maximum RPE, which also skews the standard deviation. As expected, the batch solution performs best overall in terms of the RPE statistics, but the stereo case is very competitive while also running online.

### *E. Timing*

Timing results in milliseconds for significant components of the calibration pipeline across each of the datasets are shown in Table I. Note that these steps operate at different frequencies. Frames in the 30 Hz headset datasets are dropped for feature tracking, but in the future, processing for each of the 8 cameras could be parallelized better. The keyframe selection method is very lightweight. Since the segment information calculation and database proposal only happens every 10 keyframes, and given the sparsity of frames selected as keyframes, one second is sufficient for real time operation. In the future, landmark triangulation could be reused between steps.

## VII. Conclusion

We have developed a general online multi-camera extrinsic calibration framework. In order to achieve efficient operation, we proposed a novel information-theoretic keyframe selection

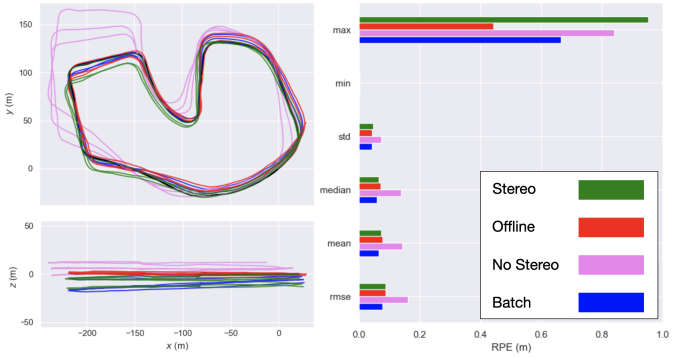



Fig. 11: 5-stereo VIO trajectory comparison and RPE statistics.

method and created a segment database approach specifically for multi-camera systems. We evaluated the pipeline on three distinct platforms to demonstrate the generality of the method, and showed that integrating inter-camera constraints improved results when possible. By ensuring accurate relative and absolute extrinsic transformations, we exhibited improved downstream tasks such as dense correspondence and VIO. For future work, it would be interesting to leverage the change detection from [31] into the segment-based calibration framework for true life-long operation. Lastly, performing degeneracy-aware optimization could alleviate issues with low observability.

## VIII. Acknowledgements

The authors would like to thank members of the Robot Perception Lab for insightful discussions. This work was supported by Facebook Reality Labs (Pittsburgh, PA). The authors would like to thank Matterport for providing the environments in the simulated 8-camera dataset. For the 5-stereo dataset, the CMU authors acknowledge partial support by the U.S. Army Research Office and the U.S. Army Futures Command under Contract No. W911NF-20-D-0002. The content of the information does not necessarily reflect the position or the policy of the government and no official endorsement should be inferred. James Pritts was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

## References

[1] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 11, pp. 1330–1334, 2000.

[2] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart, "The EuRoC micro aerial vehicle datasets," *Intl. J. of Robotics Research*, vol. 35, no. 10, pp. 1157–1163, 2016.

[3] T. Dang, C. Hoffmann, and C. Stiller, "Continuous stereo self-calibration by camera parameter tracking," *IEEE Trans. on Image Processing*, vol. 18, no. 7, pp. 1536–1550, 2009.

[4] P. Liu, M. Geppert, L. Heng, T. Sattler, A. Geiger, and M. Pollefeys, "Towards robust visual odometry with a multi-camera system," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2018, pp. 1154–1161.

[5] J. Kuo, M. Muglikar, Z. Zhang, and D. Scaramuzza, "Redesigning SLAM for arbitrary multi-camera systems," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2020, pp. 2116–2122.

[6] Z. Ouyang, L. Hu, Y. Lu, Z. Wang, X. Peng, and L. Kneip, "Online calibration of exterior orientations of a vehicle-mounted surround-view camera system," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2020, pp. 4990–4996.

[7] G. Carrera, A. Angeli, and A. J. Davison, "SLAM-based automatic extrinsic calibration of a multi-camera rig," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2011, pp. 2652–2659.

[8] L. Heng, B. Li, and M. Pollefeys, "CamOdoCal: Automatic intrinsic and extrinsic calibration of a rig with multiple generic cameras and odometry," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2013, pp. 1793–1800.

[9] K. Eckenhoff, P. Geneva, J. Bloecker, and G. Huang, "Multi-camera visual-inertial navigation with online intrinsic and extrinsic calibration," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2019, pp. 3158–3164.

[10] L. Heng, G. H. Lee, and M. Pollefeys, "Self-calibration and visual SLAM with a multi-camera system on a micro aerial vehicle," *Autonomous Robots*, vol. 39, no. 3, pp. 259–277, 2015.

[11] F. Nobre and C. R. Heckman, "FastCal: Robust online self-calibration for robotic systems," in *Intl. Sym. on Experimental Robotics (ISER)*, J. Xiao, T. Kröger, and O. Khatib, Eds. Cham: Springer International Publishing, 2018, pp. 737–747.

[12] C. Kerl, J. Sturm, and D. Cremers, "Dense visual SLAM for RGB-D cameras," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2013, pp. 2100–2106.

[13] A. Das and S. L. Waslander, "Entropy based keyframe selection for multi-camera visual SLAM," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2015, pp. 3676–3681.

[14] N. Keivan and G. Sibley, "Constant-time monocular self-calibration," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2014, pp. 1590–1595.

[15] T. Schneider, M. Li, C. Cadena, J. Nieto, and R. Siegwart, "Observability-aware self-calibration of visual and inertial sensors for ego-motion estimation," *IEEE Sensors Journal*, vol. 19, no. 10, pp. 3846–3860, 2019.

[16] C. M. Bishop, *Pattern Recognition and Machine Learning*. Berlin, Heidelberg: Springer-Verlag, 2006.

[17] E. Rosten, R. Porter, and T. Drummond, "Faster and better: A machine learning approach to corner detection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 32, no. 1, pp. 105–119, 2010.

[18] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Intl. Joint Conf. on Artificial Intelligence*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1981, p. 674–679.

[19] D. Nister, "An efficient solution to the five-point relative pose problem," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 26, no. 6, pp. 756–770, 2004.

[20] R. Pless, "Using many cameras as one," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, vol. 2, 2003, pp. II–587.

[21] G. H. Lee, M. Pollefeys, and F. Fraundorfer, "Relative pose estimation for a multi-camera system with known vertical direction," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2014, pp. 540–547.

[22] Y. Ling and S. Shen, "High-precision online markerless stereo extrinsic calibration," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2016, pp. 1771–1778.

[23] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "BRIEF: Binary robust independent elementary features," in *Eur. Conf. on Computer Vision (ECCV)*, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 778–792.

[24] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Intl. Conf. on Computer Vision (ICCV)*, 2011, pp. 2564–2571.

[25] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. USA: Cambridge University Press, 2003.

[26] N. Keivan and G. Sibley, "Online SLAM with any-time self-calibration and automatic change detection," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2015, pp. 5775–5782.

[27] T. Schneider, M. Li, M. Burri, J. Nieto, R. Siegwart, and I. Gilitschenski, "Visual-inertial self-calibration on informative motion segments," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2017, pp. 6487–6494.

[28] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 30, no. 2, pp. 328–341, 2008.

[29] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "On-manifold preintegration for real-time visual-inertial odometry," *IEEE Trans. Robotics*, vol. 33, no. 1, pp. 1–21, 2017.

[30] J. Jaekel, J. G. Mangelson, S. Scherer, and M. Kaess, "A robust multi-stereo visual-inertial odometry pipeline," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2020, pp. 4623–4630.

[31] F. Nobre, M. Kasper, and C. Heckman, "Drift-correcting self-calibration for visual-inertial SLAM," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2017, pp. 6525–6532.