Fundamental Matrix Estimation Using Relative Depths

Yaqing Ding[1], Václav Vávra[1], Snehal Bhayani[2], Qianliang Wu[3], Jian Yang[3], and Zuzana Kukelova[1]

[1] Visual Recognition Group, Faculty of Electrical Engineering, Czech Technical University in Prague, Prague, Czech Republic
[2] Faculty of Information Technology and Electrical Engineering, Center for Machine Vision and Signal Analysis, University of Oulu, Oulu, Finland
[3] PCA Lab, Nanjing University of Science and Technology, Nanjing, China

Abstract. We propose a novel approach to estimate the fundamental matrix from point correspondences and their relative depths. Relative depths can be approximated from the scales of local features, which are commonly available or can be obtained from non-metric monocular depth estimates provided by popular deep learning-based methods. This makes the considered problem very relevant. To derive efficient solutions, we explore new geometric constraints on the fundamental matrix with known relative depths and present new algebraic constraints between the fundamental matrix and the translation vector. Using point correspondences and their relative depths, we derive novel efficient minimal solvers for two fully uncalibrated cameras, two cameras with different unknown focal lengths, and two cameras with equal unknown focal lengths, respectively. We propose different variants of these solvers based on the source of the relative depth information. We present detailed analyses and comparisons with state-of-the-art solvers, including results with $86,306$ image pairs from three large-scale datasets.

Keywords: Fundamental Matrix · Relative Depth

1 Introduction

Estimation of the relative pose of two uncalibrated cameras from image correspondences, also known as epipolar geometry or fundamental matrix estimation, is a fundamental problem with many applications, *e.g.*, in Structure-from-Motion (SfM) [44], localization [42, 47, 52], and augmented reality [43]. In these applications, robust optimization algorithms, such as random sample consensus (RANSAC) or its more modern variants [4, 41] are often used to find the relative pose of two cameras and potentially unknown internal calibration parameters, such as focal lengths.

By far, the most common approach for estimating the epipolar geometry between two cameras is based on point correspondences. For uncalibrated cameras, the well-known 7-point or the linear 8-point solvers [25] are widely used in practice. For cameras with common unknown equal focal length, the relative motion and common focal length can be estimated using six point correspondences [24, 29, 30, 46]. Although these algorithms, especially the 7- and 8-point algorithms, are quite efficient and simple to implement, they require a large number of point correspondences to be sampled in each iteration of RANSAC. It is known that the number of iterations required for RANSAC-based algorithms grows exponentially with the outlier ratio and the sample size, *i.e.*, the number of correspondences used for the estimation. Thus, especially in difficult image matching situations where the outlier ratio is high, reducing the sample size can lead to improvements in speed and accuracy of pose estimation.

Previous work has examined how to reduce the sample size for relative pose estimation either by reducing the estimated degrees of freedom (DoF) using additional information, *e.g.*, from an Inertial Measurement Unit (IMU) [13–15, 18, 40, 48], or by using additional information about correspondences such as information from local affine frames [5, 7].

Bentolila *et al.* [7] proposed a solution to estimate the fundamental matrix using three affine features. For cameras with a common unknown focal length, two affine correspondences are enough to estimate the epipolar geometry together with the unknown focal length [5]. Although affine correspondences significantly reduce the number of correspondences that need to be sampled, they are less commonly used in practice. The reason is that affine covariant features are much more expensive to compute compared to the most widely used feature detectors, which usually produce scale and orientation estimates for "free" [6, 36].

Thus, recently, solvers that estimate the relative pose of two cameras using scales and orientations of features have been proposed [3]. As the most closely related work, the main idea of [3] is to assume an additional constraint in which the orientations of local features in two images are related by the local affine transformation. By combining this constraint and the affine constraints, a new constraint on the epipolar geometry from the feature's scale and orientation has been proposed. Based on this assumption, scale- and orientation-covariant features can provide two linear constraints for the fundamental matrix or the essential matrix estimation. These constraints result in solvers with the same complexity as point-based relative pose solvers that, however, halve the number of correspondences required for the estimation. On the other hand, used scales and orientations of features introduce significantly higher noise than point matches.

Several solutions have recently been proposed for the relative pose of calibrated cameras. The scales and scale ratios of the features, together with the point correspondences, were used to estimate the relative pose of two calibrated cameras in [34] and of a multi-camera system in [22]. Feature rotations were used to estimate the relative pose of two calibrated cameras from four correspondences in [39] and of two uncalibrated cameras from five correspondences in [1]. Feature scales and rotations were also used to estimate homography [2, 11] and relative pose with known vertical direction [12]. A relative pose solver for calibrated cameras that uses a combination of a deep-learned non-metric monocular depth predictor with one affine correspondence was proposed in [17]. Most recently, Astermark *et al.* [27] showed a closed-form solution to the relative pose of two calibrated cameras with known relative depth from scale ratios. In addition, they proposed a neural network to estimate the relative depth.

In this paper, we also assume that the relative depth of the matched keypoints is known, and explore relative pose estimation using this additional information. However, in contrast to previous work, we do not assume that the cameras are calibrated. Information about relative depth is practical since it can be approximated using different inputs, *e.g.*, the scales of scale-covariant features [36] or the non-metric depth maps estimated by monocular depth estimation methods [20, 51]. The main contributions of the paper are:

- A new relative depth constraint on the fundamental matrix estimation is proposed. In order to find efficient solutions, we further derive new algebraic constraints between the fundamental matrix and the translation vector.
- Based on these constraints, we derive new efficient solutions to fundamental matrix estimation using point correspondences and their relative depths, including 4 point correspondences with their relative depths (**4p4d**) for general fundamental matrix estimation, 4 point correspondences with three relative depths (**4p3d**) for the case of different and unknown focal lengths, and 3 point correspondences with three relative depths (**3p3d**) for the case of equal and unknown focal length.
- We evaluate the proposed solvers on three large-scale datasets (`KITTI` , `Phototourism` , and `ETH3D`) using SIFT and Superpoint features with different matching strategies (mutual nearest neighbors and LightGlue). We evaluate three different strategies to obtain relative depths: depth maps from depth cameras or multi-view stereo, depth maps estimated by a monocular depth prediction algorithm [51], and approximate depths from feature scales.

2 Preliminaries

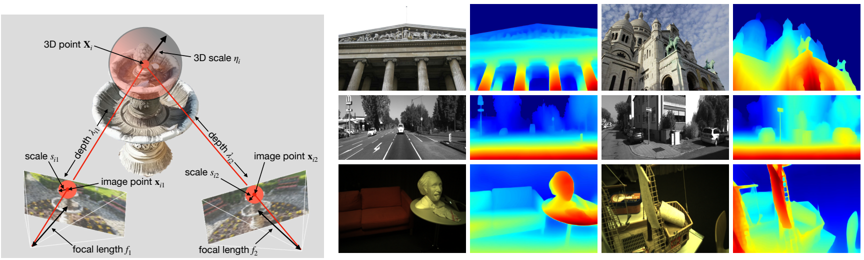
Assume that a set of 3D points $\{\mathbf{X}_i\}, i = 1, \ldots, n$ is observed by two cameras with projection matrices $\mathbf{K}_1[\mathbf{I} \mid \mathbf{0}]$ and $\mathbf{K}_2[\mathbf{R} \mid \mathbf{t}]$. Let $\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, i = 1, \ldots, n$ be a set of n 2D point correspondences, *i.e.*, the projections of 3D points $\{\mathbf{X}_i\}, i = 1, \ldots, n$ in the first and the second camera, respectively. Then we have

$$\lambda_{i1}\mathbf{K}_1^{-1}\mathbf{m}_{i1} = \mathbf{X}_i, \ \lambda_{i2}\mathbf{K}_2^{-1}\mathbf{m}_{i2} = \mathbf{R}\mathbf{X}_i + \mathbf{t}, \tag{1}$$

where λ_{i1} and λ_{i2} are the depths of the 3D point $\mathbf{X}_i$ in the first and the second camera, respectively. By eliminating the 3D point $\mathbf{X}_i$ from (1), we get

$$\sigma_i\mathbf{m}_{i2} = \mathbf{K}_2\mathbf{R}\mathbf{K}_1^{-1}\mathbf{m}_{i1} + \frac{1}{\lambda_{i1}}\mathbf{K}_2\mathbf{t}, \tag{2}$$

where $\sigma_i = \frac{\lambda_{i2}}{\lambda_{i1}}$ is the relative depth of the point $\mathbf{X}_i$, w.r.t. the second and the first camera. In this paper, we assume that the relative depth can be obtained from different geometric entities and discuss solutions to the problem of fundamental matrix estimation from point correspondences with known relative depths.



(a) **(b)**

Fig. 1: (a) Illustration of the derived constraint (4) relating feature scale, depths of the 3D point in two cameras, and their focal lengths. (b) Example RGB images from the three datasets, and their corresponding disparity images obtained using Depth Anything [51]. Top row: `Phototourism` , middle row: `KITTI` , bottom row: `ETH3D` dataset.

Relative Depth from Scales. First, we show that the relative depth can be approximated from features that provide scale information, such as SIFT [35] and SURF [6]. Assume a simplified scenario where a 3D sphere with radius η_i projects into circles with radii s_{i1} and s_{i2} in two cameras with focal lengths f_1 and f_2. From the law of similar triangles (*cf*. Fig. 1), it holds

$$\frac{s_{i1}}{\eta_i} = \frac{f_1}{\lambda_{i1}} \text{ and } \frac{s_{i2}}{\eta_i} = \frac{f_2}{\lambda_{i2}}. \tag{3}$$

Hence

$$\sigma_i = \frac{\lambda_{i2}}{\lambda_{i1}} = \frac{s_{i1} f_2}{s_{i2} f_1}, \tag{4}$$

where λ_{i1} and λ_{i2} are the depths of the center of the 3D sphere in the first and the second camera and σ_i is their ratio. In practice, the 3D sphere projects to ellipses in 2D, and the circles defined by the feature scales do not back-project to exactly the same 3D ellipsis. However, using feature scales (s_{i1}, s_{i2}) in (4) usually provides a good approximation of the relative depth σ_i. Note that in this case, the relative depth σ_i is parameterized using the focal lengths (f_1, f_2), which are unknown for uncalibrated cameras.

Relative Depth from Local Affine Transformation. Affine-covariance [38] is a desirable property of local features that provides strong constraints for camera geometry problems. For an affine correspondence, we have a triplet $(\mathbf{m}_{i1}, \mathbf{m}_{i2}, \mathcal{A}_i)$, where $\mathcal{A}_i$ is a 2×2 linear transformation that encodes the local affine frame. It is known [3] that the scales of the features are proportional to the area of the affine image region, *i.e.*, $\sqrt{\det(\mathcal{A}_i)} = \frac{s_{i2}}{s_{i1}}$. Therefore, we have the constraint $\sigma_i = 1/\sqrt{\det(\mathcal{A}_i)} \cdot \frac{f_2}{f_1}$.

Relative Depth from Learned Scales. Unlike SIFT and features that can provide scale information, state-of-the-art learned detectors such as Superpoint [9], DISK [49], D2-Net [16] do not provide such scales. However, there are several deep learning-based methods [32, 50] that can assign scales to arbitrary keypoints. Thus, it is possible to use modern features with scale information.

Relative Depth from Learned Depths. Recently, several learning-based methods [20, 51] that provide non-metric monocular depth estimates, *i.e.*, the depth estimates scaled with an unknown scale factor k, have been proposed. However, as we will show later, even depths that are only known up to an unknown scale k_i in the image i, can be used to obtain the relative depth that can be used for fundamental matrix estimation.

3 Fundamental Matrix Estimation From Four Points and Their Relative Depths

In this section, we first discuss the general case where we do not have any prior on the cameras' intrinsic parameters, *i.e.*, the calibration matrices $\mathbf{K}_1$ and $\mathbf{K}_2$ are unknown general triangular matrices. Let $\mathbf{T} = \mathbf{K}_2\mathbf{t}$. Based on (2), we have

$$\frac{1}{\lambda_{i1}}\mathbf{T} = \sigma_i \mathbf{m}_{i2} - \mathbf{K}_2\mathbf{R}\mathbf{K}_1^{-1}\mathbf{m}_{i1}, \tag{5}$$

Multiplying (5) by the skew-symmetric matrix $[\mathbf{T}]_\times$ of the vector $\mathbf{T}$ results in

$$[\mathbf{T}]_\times \sigma_i \mathbf{m}_{i2} - \mathbf{F}\mathbf{m}_{i1} = \mathbf{0}, \text{ with } \mathbf{F} = [\mathbf{T}]_\times \mathbf{K}_2\mathbf{R}\mathbf{K}_1^{-1}, \tag{6}$$

where $\mathbf{F}$ is the unknown fundamental matrix that contains information about the relative pose of two cameras and their intrinsic parameters. For the known relative depths σ_i, equation (6) provides three new linear constraints on 12 unknown elements of the fundamental matrix $\mathbf{F}$ and the translation vector $\mathbf{T}$. Given four point correspondences $\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, i = 1, \ldots, 4$, with known relative depths $\sigma_i,\ i = 1, \ldots 4$, we can obtain the following system of 12 linear equations

$$[\mathbf{T}]_\times \sigma_i \mathbf{m}_{i2} - \mathbf{F}\mathbf{m}_{i1} = \mathbf{0},\ i = 1, \ldots 4 \tag{7}$$

which can be rewritten as

$$\mathbf{B}[\mathbf{f}, \mathbf{T}]^\top = \mathbf{0}, \tag{8}$$

where $\mathbf{B}$ is a 12×12 matrix and $\mathbf{f}$ is a 9×1 vector containing the elements of the fundamental matrix $\mathbf{F}$, *i.e.*, $\mathbf{f} = vec(\mathbf{F})$. For non-collinear points $\mathbf{m}_{ij}, i = 1, \ldots, 4;\ j = 1, 2$, the matrix $\mathbf{B}$ has rank 11. In addition, the matrix $\mathbf{F}$ estimated using (8) is already singular and thus $\det(\mathbf{F}) = 0$ does not add any additional constraint. For the proof of the rank deficiency of the matrices $\mathbf{B}$ and $\mathbf{F}$, see the Supplementary Material (SM). Therefore, for general unknown calibration matrices, the minimal number of point correspondences, together with their relative depths, necessary for the fundamental matrix estimation is four. In this case, the solution can be efficiently obtained by solving the linear homogeneous system of 11 equations (8), *i.e.*, as the null space vector of the matrix $\mathbf{B}$ in (8).

Derivation with feature scales. If the relative depths σ_i are not directly known but are approximated using scales of features, *i.e.*, $\sigma_i \approx (s_{i1}f_2)/(s_{i2}f_1)$, then the equations that have to be solved have a slightly different structure since the expression for σ_i contains unknown f_1, f_2. Substituting (4) into (6), we have

$$\frac{s_{i1}}{s_{i2}} \frac{f_2}{f_1} [\mathbf{T}]_\times \mathbf{m}_{i2} - \mathbf{F}\mathbf{m}_{i1} = 0. \tag{9}$$

For a pair of images, $\frac{f_2}{f_1}$ is an unknown constant, which can be absorbed into the translation $\mathbf{T}$. Let $\tilde{\mathbf{T}} = \frac{f_2}{f_1}\mathbf{T}$ and $s_i = \frac{s_{i1}}{s_{i2}}$. For four correspondences, we have

$$[\tilde{\mathbf{T}}]_\times s_i \mathbf{m}_{i2} - \mathbf{F}\mathbf{m}_{i1} = 0,\ i = 1, \ldots, 4. \tag{10}$$

Equations (10) have the same structure as (7) and can again be solved by computing the one-dimensional null space of a 11×12 coefficient matrix. The same solver also works for the case where the relative depth is approximated from a local affine transformation as $\sigma_i \approx 1/\sqrt{\det(\mathcal{A}_i)} \cdot \frac{f_2}{f_1}$

A similar situation arises when the depths λ_{ij} of the points $\mathbf{X}_i$ in the j^{th} image are known only up to an unknown scale factor k_j. Such depths can be obtained using a monocular depth estimation method [20, 51]. In this case

$$\sigma_i = \frac{\lambda_{i2}}{\lambda_{i1}} = \frac{k_2 \tilde{\lambda}_{i2}}{k_1 \tilde{\lambda}_{i1}}, \tag{11}$$

where $\tilde{\lambda}_{i1}, \tilde{\lambda}_{i2}$ are known depths from a monocular depth estimation, and k_1, k_2 are unknown scales of these depths. For a pair of images, k_1, k_2 are constant numbers. Substituting (11) into (6), the scale factor $\frac{k_2}{k_1}$ can be absorbed into translation $\mathbf{T}$ and we obtain a linear system with the same structure as (10).

Thus, the proposed linear solver works for general unknown calibration matrices $\mathbf{K}_i$, *i.e.*, $\mathbf{K}_i$ of a general triangular form with unknown focal length, skew, aspect ratio, and principal point, and for cases when the scales of the depths are unknown, *i.e.*, they are known up to an unknown scale factor in each image.

4 Focal Length Problems

In many practical scenarios, additional assumptions about the structure of calibration matrices can be used. For modern CCD or CMOS cameras, it is often reasonable to assume that the cameras have square-shaped pixels, and the principal point coincides with the image center [24]. This is a widely used assumption in many camera geometry solvers, where the only unknown intrinsic parameters are focal lengths, and the calibration matrices have the form $\mathbf{K}_i = diag(f_i, f_i, 1)$.

4.1 Different and Unknown Focal Lengths

First, let us assume that $\mathbf{K}_i = diag(f_i, f_i, 1),\ i = 1, 2$ and, in general, $f_1 \neq f_2$. By solving the system of equations (6), we simultaneously solve for $\mathbf{T}$ and $\mathbf{F}$. However, $\mathbf{T}$ and $\mathbf{F}$ are not independent and are related by

$$\mathbf{F} = [\mathbf{T}]_\times \mathbf{K}_2 \mathbf{R} \mathbf{K}_1^{-1}. \tag{12}$$

Constraints (12) can not be simply combined with (6) since they introduce additional unknowns from $\mathbf{R}$ and f_1, f_2, and would have resulted in a very complex solver. However, these constraints can be used to derive new constraints that contain only elements of $\mathbf{T}$ and $\mathbf{F}$. To do this, we can use the elimination ideal method [8], a method known from algebraic geometry that was recently used to derive many efficient camera geometry solvers [30]. In this case, we first create an ideal J generated by the 9 polynomials (12) [8]. Then the unknown elements of the rotation matrix $\mathbf{R}$ and the focal lengths f_1, f_2 are eliminated from the generators of J by computing the generators of the elimination ideal $J_1 = J \cap \mathbb{C}[f_{11}, \ldots, f_{33}, T_x, T_y, T_z]$. Here, f_{ij} are the entries of $\mathbf{F}$. The elimination ideal J_1 can be computed offline using some algebraic geometry software like Macaulay 2 [21]. In our case, the elimination ideal J_1 is generated by three quartic polynomials, one of which has the form

$$\begin{gathered}
-f_{11}^2 f_{13} f_{23} + f_{11} f_{13}^2 f_{21} - f_{11} f_{21} f_{23}^2 - f_{12}^2 f_{13} f_{23} + f_{12} f_{13}^2 f_{22} - f_{12} f_{22} f_{23}^2 \cdots \\
+ f_{13} f_{21}^2 f_{23} + f_{13} f_{22}^2 f_{23} + f_{12} f_{32} T_y T_z + f_{11} f_{31} T_y T_z - f_{21} f_{31} T_x T_z - f_{22} f_{32} T_x T_z = 0.
\end{gathered} \tag{13}$$

The form of the remaining two polynomials together with the input code for Macaulay2 is provided in the SM. Although there are three generators, it can be shown that they provide only one algebraic constraint on the elements of $\mathbf{T}$ and $\mathbf{F}$, for details see the SM. Note that similar additional constraints on $\mathbf{T}$ and $\mathbf{F}$ cannot be derived for general triangular matrices $\mathbf{K}_1$ and $\mathbf{K}_2$.

Each point correspondence together with the known relative depth gives us three linear homogeneous constraints of the form (6) on the 12 unknown elements of $\mathbf{T}$ and $\mathbf{F}$. Therefore, three point correspondences with three relative depths result in nine linear homogeneous equations. To solve this problem, we thus need, in addition to the quartic constraint (13) from the elimination ideal, one more constraint. This constraint can be obtained from an additional point correspondence, *i.e.*, we can use the standard linear epipolar constraint on $\mathbf{F}$. Four points with three relative depths give 10 linear homogeneous equations

$$\begin{gathered}
[\mathbf{T}]_\times \sigma_i \mathbf{m}_{i2} - \mathbf{F}\mathbf{m}_{i1} = \mathbf{0}, \ i = 1, 2, 3 \\
\mathbf{m}_{42}^\top \mathbf{F} \mathbf{m}_{41} = 0,
\end{gathered} \tag{14}$$

which can be rewritten as

$$\mathbf{B}[\mathbf{f}, \mathbf{T}]^\top = \mathbf{0}, \tag{15}$$

where the matrix $\mathbf{B}$ is a 10×12 matrix. The solution to the vector $[\mathbf{f}, \mathbf{T}]^\top$ can thus be written as a linear combination of the two basis vectors from the 2-dimensional null space of the matrix $\mathbf{B}$ as

$$[\mathbf{f}, \mathbf{T}]^\top = \alpha_1 \mathbf{N}_1 + \alpha_2 \mathbf{N}_2, \tag{16}$$

where α_1, α_2 are new unknowns. Since (12) is homogeneous, we can set $\alpha_2 = 1$.

By substituting parameterization (16) into (13), we obtain a quartic equation in the unknown α_1. In practice, this equation has one trivial solution. Thus, we only need to find the roots of a cubic univariate polynomial. In general, there are up to three possible solutions to the fundamental matrix $\mathbf{F}$.

Derivation using local feature scales. If scales of features are used to approximate relative depths, the linear equations have the form (10). Therefore, instead of solving for $\mathbf{F}$ and $\mathbf{T}$, we solve for $\mathbf{F}$ and $\tilde{\mathbf{T}} = \frac{f_2}{f_1}\mathbf{T}$. The constraint that relates elements of $\mathbf{F}$ and $\tilde{\mathbf{T}}$ can be obtained using a similar elimination ideal technique, *i.e.*, by eliminating the rotation and focal length parameters from the ideal generated by the nine equations $\mathbf{F} = \frac{f_1}{f_2}[\tilde{\mathbf{T}}]_\times \mathbf{K}_2\mathbf{R}\mathbf{K}_1^{-1}$. The elimination ideal J_2 is generated by three cubic polynomials, one of which has the form

$$f_{13}f_{21}f_{31} - f_{11}f_{23}f_{31} - f_{12}f_{23}f_{32} + f_{13}f_{22}f_{32} + f_{13}\tilde{T}_y\tilde{T}_z - f_{23}\tilde{T}_x\tilde{T}_z = 0. \quad (17)$$

The final solver performs the same steps as for known relative depths, however, in this case it solves the univariate cubic polynomial (17). More details about this solver are given in the SM. Note that a similar 4p3d (4-points-3-depths) solver cannot be derived for the case with an unknown scale factor $k = \frac{k_2}{k_1}$, since in this case, in contrast to the scale factor $\frac{f_2}{f_1}$, we introduce a new unknown k and thus also increase the DoF that must be estimated.

4.2 Equal and Unknown Focal Length

By assuming unknown equal focal lengths, *i.e.*, $\mathbf{K}_1 = \mathbf{K}_2 = diag(f, f, 1)$, which is a useful assumption, *e.g.*, when estimating the motion of a single uncalibrated camera, we decrease the DoF by one compared to the different unknown focal length case described in Section 4.1. Thus, we need at least three points with three relative depths to solve this problem. The problem can be solved using the null-space parameterization and the elimination ideal method, similarly to the 4p3d solver for different focal length case. In this case, three points with three relative depths give us nine homogeneous equations of the form (6). Therefore, the solution to the vector $[\mathbf{f}, \mathbf{T}]^\top$ can be written as a linear combination of the three basis vectors of a 3-dimensional null space of the coefficient matrix with two new unknowns β_1 and β_2. This parameterization can be plugged into the generators of the elimination ideal that we obtain by eliminating elements of the rotation matrix $\mathbf{R}$ and the focal length f from the generators of the ideal generated by (12). In this case this results in solving one cubic and one quartic equation in two unknowns. These equations can be solved using the Gröbner basis method [31], where the final solver performs Gauss-Jordan elimination of a 6×10 matrix and extracts solutions from the eigenvectors of a 4×4 matrix. Hence, the same solver works for the case where scales of features are used to approximate relative depths. For the case of an unknown scale factor $k = \frac{k_2}{k_1}$, a similar solver that, however, uses four points and three relative depths can be derived. More details on these solvers can be found in the SM.

In this section, we describe an alternative solution, which is based on homography parameterization and which results in a more efficient 3p3d solver for the case of equal unknown focal lengths. Three points define a plane in the 3D space.

Let $\mathbf{n}$ be the unit normal vector of the plane with respect to the first camera frame, and let d denote the distance from the plane to the optical center of the first camera. Then we have

$$\mathbf{n}^\top \mathbf{X}_i = d \Rightarrow \frac{1}{d}\mathbf{n}^\top \mathbf{K}^{-1}\mathbf{m}_i = \frac{1}{\lambda_i}. \tag{18}$$

Substituting (18) into (2) we have

$$\sigma_i \mathbf{m}_{i2} = \mathbf{G}\mathbf{m}_{i1}, \text{ with } \mathbf{G} = \mathbf{K}\mathbf{H}\mathbf{K}^{-1}, \text{ and } \mathbf{H} = \mathbf{R} + \frac{\mathbf{t}}{\mathrm{d}}\mathbf{n}^\top, \tag{19}$$

where $\mathbf{G}$ is 2D homography, and $\mathbf{H}$ is Euclidean homography. Given 3 point correspondences, we have

$$[\sigma_i \mathbf{m}_{i1}, \sigma_i \mathbf{m}_{i2}, \sigma_i \mathbf{m}_{i3}] = \mathbf{G}[\mathbf{m}_{i1}, \mathbf{m}_{i2}, \mathbf{m}_{i3}] \tag{20}$$

and $\mathbf{G}$ is given by $[\sigma_i \mathbf{m}_{i1}, \sigma_i \mathbf{m}_{i2}, \sigma_i \mathbf{m}_{i3}] \cdot [\mathbf{m}_{i1}, \mathbf{m}_{i2}, \mathbf{m}_{i3}]^{-1}$. The Euclidean homography matrix $\mathbf{H}$ can be formulated as $\mathbf{H} = \mathbf{K}^{-1}\mathbf{G}\mathbf{K}$. As shown in [37,53], a Euclidean homography matrix should satisfy the singular value constraint

$$\mathrm{median}(\mathrm{svd}(\mathbf{H})) = 1, \tag{21}$$

where the second largest singular value of $\mathbf{H}$ should be 1. Hence, we have

$$\det(\mathbf{H}^\top \mathbf{H} - \mathbf{I}_3) = 0. \tag{22}$$

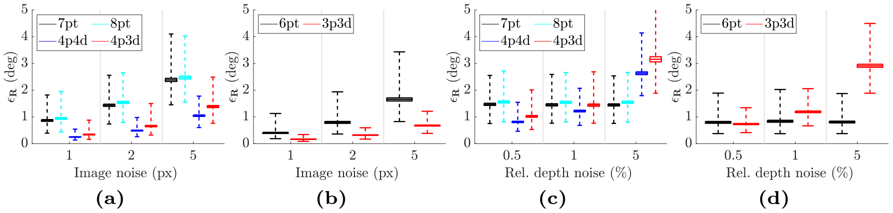
After substituting $\mathbf{H} = \mathbf{K}^{-1}\mathbf{G}\mathbf{K}$ into (22), we obtain a quadratic equation in f^2. There are up to two possible solutions since the focal length should be positive. Note that, we still need to verify if the focal length satisfies (21), as (22) is only a *necessary condition* for (21). Once the focal length is known, the Euclidean homography matrix can be decomposed into rotation and translation.

5 Experiments

We evaluate the proposed solvers, *i.e.*, the 4p4d, 4p3d, and 3p3d solvers, on both synthetic data and real-world images. For general fundamental matrix estimation, *i.e.*, in general, different focal lengths, we compare our 4p4d and 4p3d solvers to the most closely related 4SIFT solver from [3], the standard 7pt algorithm [25], and the 5ORI solver from [1] (using 5 point correspondences with known feature rotations). For the equal and unknown focal length case, we compare the proposed 3p3d solver with the 3SIFT solver [3] and the 6pt solver [30].

5.1 Synthetic Evaluation

We generate 200 random 3D points in the cube $[-10, 10] \times [-10, 10] \times [2, 22]$ and 5K camera pairs with random relative poses. In the case of two equal focal lengths, we set them to 1000 px, and for the case of different focal lengths, we



(a) **(b)** **(c)** **(d)**

Fig. 2: A comparison of the performance of point-based and the proposed solvers in the presence of image noise (**a,b**), and relative depth noise (**c,d**) and 2px image noise.

set them to 1000 px and 2000 px. For the case of different focal lengths, we compare our 4p4d and 4p3d solvers with the 4SIFT solver [3] and the standard 7pt algorithm [25]. For the case of equal focal lengths, we compare our 3p3d solver with the 3SIFT solver [3] and the 6pt solver [30]. We evaluate solvers performance by studying the error in the estimated rotation w.r.t. the ground truth, defined as $\epsilon_{\mathbf{R}} = 2\arcsin\left(\frac{\|\mathbf{R}_{gt}-\mathbf{R}_e\|}{2\sqrt{2}}\right)$. We use the arcsin formulation for rotation [10], since the arccos metric suffers from precision issues with noise-free data. Errors in translation and focal length estimates are studied in the SM.

First, we test the performance of all solvers in the presence of Gaussian noise with standard deviation σ, added to the image points in both cameras. Fig. 2(a,b) shows the rotation error (in degrees) for different focal lengths (a) and equal focal lengths (b). Here, we depict the results as box plots that show the 25% and 75% quantile values as boxes with a horizontal line for the median. Observe that our 4p4d (followed by 4p3d solver) and 3p3d solvers have the best accuracy in the presence of image pixel noise.

We also test the performance of all the solvers in the presence of noise in the relative depths. We add Gaussian noise to the relative depths and vary the standard deviation σ as a (%) of the relative depths. To simulate real-world scenarios, we also add 2px Gaussian noise to the images. Fig. 2(c,d) show the rotation error (in degrees) for different focal lengths (c) and equal focal lengths (d). We note that for 0.5% noise in the relative depths, our solvers, 4p4d, 4p3d, and 3p3d have better or comparable accuracy to the SOTA solvers. For larger noise levels, our solvers return larger errors than the point-based solvers. However, as shown next, our solvers perform at least comparable to the point-based solvers under depth noise levels observed in the real world.

5.2 Real-world Experiments

Datasets. In order to test the proposed solvers [1] on real-world data, we choose the `Phototourism` [26], the `KITTI` odometry [19], and the `ETH3D` [45] datasets. They cover three situations: unordered images for Structure-from-Motion, sequential images in an outdoor environment, and sequential images in an indoor

[1] `https://github.com/yaqding/FMRD`

environment. Fig. 1 shows example images and their corresponding disparity maps. The images of the Phototourism dataset were collected from multiple cameras obtained at different times, from different viewpoints, and with occlusions. It is a challenging dataset that is commonly used as a benchmark dataset [26] and can be used to evaluate the performance of methods in a wide range of situations. We used nine test scenes with $43,678$ image pairs from this dataset, including images, ground-truth poses, and sparse depth maps. The KITTI odometry benchmark consists of 22 stereo sequences. Only 11 sequences (00–10) are provided with ground truth trajectories. We thus used these 11 sequences for evaluation. In total, $23,190$ image pairs were used. The ETH3D SLAM dataset covers a variety of indoor and outdoor scenes. A DSLR camera as well as a synchronized multi-camera rig with varying field-of-view was used to capture images. In total, $19,438$ image pairs were used from ETH3D dataset.

Robust estimation. For testing minimal solvers on real-world data, we use Graph-Cut RANSAC [4] (GC-RANSAC). In GC-RANSAC (and other RANSAC variants with local optimization), two different solvers are used: (a) one for estimating the pose from a minimal sample and (b) one for fitting to a larger sample when doing final pose polishing on all inliers or in the local optimization step. For (a) the main goal is to solve the problem using as few correspondences as possible since the number of RANSAC iterations depends exponentially on the number of correspondences required for the pose estimation. All the proposed solvers, as well as state-of-the-art solvers that are used for comparison, are applied in this step of the GC-RANSAC. The goal of (b) is to generate a pose hypothesis that minimizes the error on all detected inliers. Similarly to previous works that used GC-RANSAC for relative pose estimation [3], in (b) we use the normalized 8-point algorithm [23] for linear least squares fitting.

Feature detection and matching. There are mainly two types of feature detectors, handcrafted features and learning-based features. For handcrafted features, we choose SIFT. The scale and orientation information from the SIFT features can be used for our solver, the related SIFT-based solver [3], and the 5ORI solver [1]. For deep learning-based features, we choose the popular Superpoint [9] (SP) features, which are used in SfM and localization pipelines [42]. For feature matching, we also evaluated two methods. One is the classical mutual nearest neighbors (NN) combined with the standard ratio test [36]. The other is the state-of-the-art deep learning-based feature matching method LightGlue [33] (LG), which shows significant improvement over classical matching methods. In general, we use four different combinations of feature detection and matching: SIFT+NN, SIFT+LG, SP+NN, and SP+LG. Due to the space limitation, the results of SP+NN are provided in the SM.

Relative Depth Estimation. We use three methods to obtain the relative depth: *i) Relative scale.* Given the SIFT features, the relative depth can be directly approximated from the scales. For Superpoint, we employ the Self-scale-ori [32] scale estimator to obtain the scale and orientation (the orientation is used for the originally SIFT-based solvers [1, 3]). *ii) Depth map.* The Phototourism and ETH3D datasets provide depth images that can be used to obtain the relative depth. For the KITTI dataset, we use ground truth poses for triangulation to get the relative depth. *iii) Monocular depth estimation.* Depth Anything [51] is one of the state-of-the-art techniques for estimating intra-image relative depth (depth of the image up to a scale factor). We use this intra-image relative depth for inter-image relative depth estimation between image pairs. We used a pre-trained model not trained on our datasets. While using depth maps or SIFT features does not incur additional computation, incorporating learning-based methods, such as Depth Anything [51] or Self-scale-ori [32], introduces additional computational costs. For example, on a single RTX4090, the inference time for Depth Anything [51] (small model) is $\sim 3ms$. Note that in some applications, depth estimation methods are also used for other tasks, and thus this computational overhead is not purely introduced by the pose estimation method.

Experimental results. *i) Fundamental matrix estimation for general intrinsics.* Table 1 shows the rotation, translation (in degrees) errors, and run-times (in milliseconds) on the Phototourism , KITTI , and ETH3D datasets for fundamental matrix estimation. Since the Phototourism and ETH3D datasets are very challenging, all methods fail on some pairs. For each dataset, we first obtain the median of the rotation and translation errors for each sequence / scene, and then report the mean over all the sequences / scenes. Given accurate relative depth, the proposed methods outperform the existing methods on all the datasets with different features in terms of speed and accuracy. However, for Phototourism and ETH3D datasets, when using relative depth approximated from SIFT scales or monocular depth estimation, there is a slight gap compared to point-based solvers. There are mainly two reasons: first, relative depth from SIFT scales may not work well for images collected from significantly different viewpoints; second, we used the pre-trained Depth Anything model which can give imprecise depth estimates for our datasets. For the KITTI dataset, the proposed methods are comparable or slightly better than point-based solvers. Based on the results, SIFT+NN is still a good choice for sequential datasets with textured scenes, *e.g.*, KITTI. SIFT can provide good enough features, and NN is good enough to give good matches under small motion. However, for challenging datasets (Phototourism , ETH3D), SP+LG can significantly improve performance, especially translation estimation. On the other hand, LightGlue usually gives more matches, and sometimes more outliers. Thus, SIFT+LG and SP+LG are more time consuming than SIFT+NN. Note that the 4SIFT and 5ORI solvers cannot make use of relative depth or Depth Anything for uncalibrated cameras.

ii) Equal and unknown focal length. To use the Phototourism dataset for testing equal and unknown focal length solvers, we resize the image pairs to have the same focal length. The KITTI and ETH3D datasets are captured from a single camera, and we did not do any preprocessing. We used the 6pt solver for non-minimal fitting. Table 2 shows the rotation, translation (in degrees), and focal length errors, and run-times (in ms). Here, we only show the results of SP+LG, remaining results are in the SM. With the equal focal length constraint, the overall performance is much better. The proposed 3p3d solver is always among the top-performing methods in terms of speed and accuracy.

Table 1: Rotation and translation errors (in degrees), and run-times (in milliseconds) on the `Phototourism` , `KITTI` , and `ETH3D` datasets for fundamental matrix estimation. The **best** and the $\underline{\text{second best}}$ methods are highlighted.

Feature	Depth	Method	Phototourism			KITTI			ETH3D		
			$\epsilon_{\mathbf{R}}(^\circ)$	$\epsilon_{\mathbf{t}}(^\circ)$	$\tau(ms)$	$\epsilon_{\mathbf{R}}(^\circ)$	$\epsilon_{\mathbf{t}}(^\circ)$	$\tau(ms)$	$\epsilon_{\mathbf{R}}(^\circ)$	$\epsilon_{\mathbf{t}}(^\circ)$	$\tau(ms)$
SIFT+NN	-	7pt	2.58	14.7	2.74	0.11	0.59	3.78	2.34	45.7	2.12
	Relative depth	4p4d	$\underline{1.90}$	$\underline{11.2}$	1.59	**0.09**	**0.56**	$\underline{2.63}$	**1.61**	**27.5**	**1.54**
		4p3d	**1.90**	**10.4**	2.94	$\underline{0.10}$	$\underline{0.57}$	4.62	$\underline{1.67}$	$\underline{28.4}$	2.59
	SIFT	4SIFT	2.69	16.7	3.29	0.12	0.60	3.61	2.52	50.9	2.30
		4p4d	2.69	16.8	**1.47**	0.11	0.59	**2.46**	2.60	51.5	1.89
		4p3d	2.75	18.0	2.01	0.11	0.60	3.29	2.62	52.0	1.99
		5ORI	2.82	17.8	2.84	0.12	0.60	4.08	2.65	53.4	2.67
	DepthAny	4p4d	2.73	17.4	$\underline{1.50}$	0.11	0.60	2.69	2.67	51.2	$\underline{1.68}$
		4p3d	2.76	17.9	2.95	0.11	0.60	5.25	2.62	50.3	3.09
SIFT+LG	-	7pt	3.33	13.9	4.84	0.16	0.69	5.19	2.36	46.2	2.57
	Relative depth	4p4d	$\underline{2.22}$	$\underline{9.23}$	$\underline{3.27}$	**0.10**	**0.56**	**4.94**	**1.53**	**24.9**	**1.70**
		4p3d	**2.20**	**8.13**	4.81	$\underline{0.11}$	$\underline{0.59}$	6.61	$\underline{1.65}$	$\underline{27.2}$	2.92
	SIFT	4SIFT	4.08	17.1	5.04	0.17	0.72	6.29	2.57	52.2	2.58
		4p4d	4.14	17.0	**2.78**	0.16	0.71	$\underline{4.96}$	2.63	53.8	2.12
		4p3d	4.17	17.2	3.80	0.16	0.70	6.18	2.65	53.7	2.25
		5ORI	4.41	18.1	4.66	0.16	0.71	6.55	2.67	54.3	3.14
	DepthAny	4p4d	4.10	16.9	3.32	0.17	0.72	5.29	2.61	52.2	$\underline{1.86}$
		4p3d	4.19	17.5	3.69	0.16	0.70	7.53	2.56	50.8	3.53
SP+LG	-	7pt	2.62	10.6	7.05	0.18	0.67	3.81	1.47	27.1	3.89
	Relative depth	4p4d	$\underline{1.90}$	$\underline{7.24}$	**4.66**	**0.13**	**0.59**	**2.63**	**0.68**	$\underline{11.5}$	**2.00**
		4p3d	**1.83**	**6.24**	6.84	$\underline{0.14}$	$\underline{0.62}$	3.71	$\underline{0.75}$	**11.2**	3.42
	Self-sca-ori	4SIFT	3.32	13.3	5.98	0.17	0.67	3.43	1.79	39.8	2.75
		4p4d	3.31	13.1	$\underline{4.70}$	0.17	0.66	$\underline{2.73}$	1.74	37.6	2.34
		4p3d	3.29	13.3	5.76	0.17	0.66	3.35	1.77	37.2	4.01
		5ORI	3.57	14.2	6.56	0.17	0.66	3.68	1.91	43.0	3.58
	DepthAny	4p4d	3.32	13.3	4.98	0.17	0.67	2.78	1.77	38.8	$\underline{2.33}$
		4p3d	3.33	13.3	8.87	0.17	0.67	5.01	1.74	37.1	4.17

Degeneracies. Similarly to standard point-based solvers, there are several degenerate cases to consider. Pure rotation and four coplanar points present degeneracy for both the 4p4d and 4p3d cases (the rank of matrix $\mathbf{B}$ in (8) is 9). In such instances, a homography should be utilized. Pure rotation does not result in degeneracy for the 3p3d scenario, as the homography formulation is employed in this case. Pure translation also generates some degeneracies for focal length recovery [28] (See Tab 2, the focal length estimation fails on the `KITTI` dataset).

Limitations. In this paper, we proposed new relative pose solvers for fundamental matrix estimation from point correspondences and relative depths. If precise relative depths are available, these solvers outperform the state-of-the-art point- [25], scale and orientation- [3], and orientation-based [1] solvers. However, if the relative depths are approximated by feature scales (either extracted from

Table 2: Rotation and translation errors (in degrees), focal length error, and run-times (in milliseconds) on the `Phototourism` , `KITTI` , and `ETH3D` datasets for the equal and unknown focal length problem. The **best** and the second best methods are highlighted.

Feature	Depth	Method	Phototourism				KITTI				ETH3D			
			$\epsilon_{\mathbf{R}}(^\circ)$	$\epsilon_{\mathbf{t}}(^\circ)$	ϵ_f	$\tau(ms)$	$\epsilon_{\mathbf{R}}(^\circ)$	$\epsilon_{\mathbf{t}}(^\circ)$	ϵ_f	$\tau(ms)$	$\epsilon_{\mathbf{R}}(^\circ)$	$\epsilon_{\mathbf{t}}(^\circ)$	ϵ_f	$\tau(ms)$
SP+LG	-	6pt	1.22	3.82	**0.09**	34.0	0.13	0.63	0.77	21.4	0.81	11.0	0.23	21.6
	Relative depth	3p3d	**1.11**	**3.47**	**0.09**	23.4	**0.10**	**0.56**	0.78	14.4	**0.70**	**9.91**	**0.22**	13.3
	Self-sca-ori	3SIFT	1.20	3.76	0.10	31.1	0.14	0.63	0.78	28.6	0.81	11.6	0.24	24.7
		3p3d	1.19	3.73	0.10	**20.1**	0.13	0.62	0.78	**13.5**	0.79	11.4	0.23	**12.2**
	DepthAny	3p3d	1.18	3.72	0.10	23.6	0.13	0.63	0.78	15.1	0.80	11.4	0.24	13.2

SIFT or learned scales [32]) or extracted from non-metric monocular depth estimates [51], the performance of the proposed solvers degrades and is sometimes even worse than the performance of the standard 7pt solver. The performance drop depends on the used features (SIFT vs. SP) and the matching method (NN vs. LG). Note that this performance drop is observed for all state-of-the-art solvers that use scales and orientations of features for the estimation. The drop is especially visible when using the LG matcher. In this scenario, it is still preferable to use the standard 7pt solver if no precise relative depths are available. Note that prior work [1,3] did not consider learned features and learned matchers and therefore did not notice this behavior. Learning-based monocular depth estimation [51] and estimation of scales and orientations of features [32] are active fields of research with rapid progress. Naturally, improved depth and scale estimates will automatically improve the performance of our solvers (as visible from their superior performance in the presence of precise relative depths).

6 Conclusion

We presented several different efficient minimal solvers for estimating the relative pose of uncalibrated or partially calibrated cameras from point correspondences and their relative depths. We derived different variants of these solvers based on the source of relative depth information. In extensive real experiments, we showed that in the presence of accurate relative depths, the proposed solvers outperform point- [25], scale and orientation- [3], and orientation-based [1] solvers. Although the combination of our solvers with learning-based monocular depth estimates [51] and scale estimates [32] does not, in general, outperform the standard 7pt solver when using LightGlue, we hope that with future progress in these learning-based techniques, our methods will be a very useful alternative to point-based solvers in the future. For now, the proposed algorithms can be used as complements to point-based algorithms to increase the reliability and speed for relative pose estimation. In addition, the new constraints derived between the fundamental matrix and the relative depth may also provide useful information for learning-based monocular depth estimation.

Acknowledgments. This work was supported by the Czech Science Foundation (GACR) JUNIOR STAR Grant No. 22-23183M, and the National Natural Science Foundation of China under Grant No. 62361166670.

References

1. Barath, D.: Five-point fundamental matrix estimation for uncalibrated cameras. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
2. Barath, D., Kukelova, Z.: Homography from two orientation-and scale-covariant features. In: Proceedings of the IEEE/CVF international conference on computer vision (2019)
3. Barath, D., Kukelova, Z.: Relative pose from sift features. In: European Conference on Computer Vision. pp. 454–469. Springer (2022)
4. Barath, D., Matas, J.: Graph-cut RANSAC. In: Computer Vision and Pattern Recognition (CVPR) (2018)
5. Barath, D., Toth, T., Hajder, L.: A minimal solution for two-view focal-length estimation using two affine correspondences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
6. Bay, H., Tuytelaars, T., Van Gool, L.: Surf: Speeded up robust features. In: European Conference on Computer Vision (ECCV) (2006)
7. Bentolila, J., Francos, J.M.: Conic epipolar constraints from affine correspondences. Computer Vision and Image Understanding (2014)
8. Cox, D.A., Little, J., O'shea, D.: Using algebraic geometry. Springer Science & Business Media (2006)
9. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops (2018)
10. Ding, Y., Chien, C., Larsson, V., Astrom, K., Kimia, B.: Minimal solutions to generalized three-view relative pose problem. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
11. Ding, Y., Astermark, J., Oskarsson, M., Larsson, V.: Noisy one-point homographies are surprisingly good. In: Computer Vision and Pattern Recognition (CVPR) (2024)
12. Ding, Y., Barath, D., Kukelova, Z.: Homography-based egomotion estimation using gravity and sift features. In: Proceedings of the Asian Conference on Computer Vision (2020)
13. Ding, Y., Yang, J., Ponce, J., Kong, H.: An efficient solution to the homography-based relative pose problem with a common reference direction. In: International Conference on Computer Vision (ICCV) (2019)
14. Ding, Y., Yang, J., Ponce, J., Kong, H.: Homography-based minimal-case relative pose estimation with known gravity direction. Trans. Pattern Analysis and Machine Intelligence (PAMI) (2020)
15. Ding, Y., Yang, J., Ponce, J., Kong, H.: Minimal solutions to relative pose estimation from two views sharing a common direction with unknown focal length. In: Computer Vision and Pattern Recognition (CVPR) (2020)
16. Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., Sattler, T.: D2-net: A trainable cnn for joint description and detection of local features. In: Computer Vision and Pattern Recognition (CVPR) (2019)

17. Eichhardt, I., Barath, D.: Relative pose from deep learned depth and a single affine correspondence. In: European Conference on Computer Vision (ECCV) (2020)
18. Fraundorfer, F., Tanskanen, P., Pollefeys, M.: A minimal case solution to the calibrated relative pose problem for the case of two known orientation angles. In: European Conference on Computer Vision (ECCV) (2010)
19. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Computer Vision and Pattern Recognition (CVPR) (2012)
20. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: International Conference on Computer Vision (ICCV) (2019)
21. Grayson, D.R., Stillman, M.E.: Macaulay2, a software system for research in algebraic geometry. Available at `http://www2.macaulay2.com`
22. Guan, B., Zhao, J.: Relative pose estimation for multi-camera systems from point correspondences with scale ratio. In: Proceedings of the 30th ACM International Conference on Multimedia (2022)
23. Hartley, R.: In defence of the 8-point algorithm. In: International Conference on Computer Vision (ICCV) (1995)
24. Hartley, R., Li, H.: An efficient hidden variable approach to minimal-case camera motion estimation. Trans. Pattern Analysis and Machine Intelligence (PAMI) (2012)
25. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge university press (2003)
26. Jin, Y., Mishkin, D., Mishchuk, A., Matas, J., Fua, P., Yi, K.M., Trulls, E.: Image Matching across Wide Baselines: From Paper to Practice. International Journal of Computer Vision (2020)
27. Jonathan, A., Ding, Y., Larsson, V., Heyden, A.: Fast relative pose estimation using relative depth. In: International Conference on 3D Vision (3DV) (2024)
28. Kahl, F., Triggs, B.: Critical motions in euclidean structure from motion. In: Computer Vision and Pattern Recognition (CVPR) (1999)
29. Kukelova, Z., Bujnak, M., Pajdla, T.: Polynomial eigenvalue solutions to minimal problems in computer vision. Trans. Pattern Analysis and Machine Intelligence (PAMI) (2012)
30. Kukelova, Z., Kileel, J., Sturmfels, B., Pajdla, T.: A clever elimination strategy for efficient minimal solvers. In: Computer Vision and Pattern Recognition (CVPR) (2017)
31. Larsson, V., Åström, K., Oskarsson, M.: Efficient solvers for minimal problems by syzygy-based reduction. In: Computer Vision and Pattern Recognition (CVPR) (2017)
32. Lee, J., Jeong, Y., Cho, M.: Self-supervised learning of image scale and orientation. In: 31st British Machine Vision Conference 2021, BMVC 2021, Virtual Event, UK. BMVA Press (2021)
33. Lindenberger, P., Sarlin, P.E., Pollefeys, M.: Lightglue: Local feature matching at light speed. In: International Conference on Computer Vision (ICCV) (2023)
34. Liwicki, S., Zach, C.: Scale exploiting minimal solvers for relative pose with calibrated cameras. In: BMVC (2017)
35. Lowe, D.G.: Object recognition from local scale-invariant features. In: International Conference on Computer Vision (ICCV) (1999)
36. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**(2), 91–110 (2004)

37. Malis, E., Vargas Villanueva, M.: Deeper understanding of the homography decomposition for vision-based control. INRIA, Tech. Rep. (2007)
38. Mikolajczyk, K., Schmid, C.: Scale & affine invariant interest point detectors. International journal of computer vision (2004)
39. Mills, S.: Four-and seven-point relative camera pose from oriented features. In: 2018 International Conference on 3D Vision (3DV) (2018)
40. Naroditsky, O., Zhou, X.S., Gallier, J., Roumeliotis, S.I., Daniilidis, K.: Two efficient solutions for visual odometry using directional correspondence. Trans. Pattern Analysis and Machine Intelligence (PAMI) (2012)
41. Raguram, R., Chum, O., Pollefeys, M., Matas, J., Frahm, J.M.: USAC: A universal framework for random sample consensus. Trans. Pattern Analysis and Machine Intelligence (PAMI) **35**(8), 2022–2038 (2012)
42. Sarlin, P.E., Cadena, C., Siegwart, R., Dymczyk, M.: From coarse to fine: Robust hierarchical localization at large scale. In: Computer Vision and Pattern Recognition (CVPR) (2019)
43. Sattler, T., Leibe, B., Kobbelt, L.: Efficient & effective prioritized matching for large-scale image-based localization. IEEE transactions on pattern analysis and machine intelligence (2016)
44. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Computer Vision and Pattern Recognition (CVPR). pp. 4104–4113 (2016)
45. Schops, T., Sattler, T., Pollefeys, M.: Bad slam: Bundle adjusted direct rgb-d slam. In: Computer Vision and Pattern Recognition (CVPR) (2019)
46. Stewénius, H., Nistér, D., Kahl, F., Schaffalitzky, F.: A minimal solution for relative pose with unknown focal length. In: Computer Vision and Pattern Recognition (CVPR) (2005)
47. Svärm, L., Enqvist, O., Kahl, F., Oskarsson, M.: City-scale localization for cameras with known vertical direction. Trans. Pattern Analysis and Machine Intelligence (PAMI) (2016)
48. Sweeney, C., Flynn, J., Turk, M.: Solving for relative pose with a partially known rotation is a quadratic eigenvalue problem. International Conference on 3D Vision (3DV) (2014)
49. Tyszkiewicz, M., Fua, P., Trulls, E.: Disk: Learning local features with policy gradient. Advances in Neural Information Processing Systems (2020)
50. Yan, P., Tan, Y., Xiong, S., Tai, Y., Li, Y.: Learning soft estimator of keypoint scale and orientation with probabilistic covariant loss. In: Computer Vision and Pattern Recognition (CVPR) (2022)
51. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: Computer Vision and Pattern Recognition (CVPR) (2024)
52. Zeisl, B., Sattler, T., Pollefeys, M.: Camera pose voting for large-scale image-based localization. In: International Conference on Computer Vision (ICCV) (2015)
53. Zhang, Z., Hanson, A.R.: Scaled euclidean 3d reconstruction based on externally uncalibrated cameras. In: Proceedings of International Symposium on Computer Vision-ISCV. pp. 37–42. IEEE (1995)