Semicalibrated Relative Pose from an Affine Correspondence and Monodepth

Petr Hruby[1], Marc Pollefeys[1,2], and Daniel Barath[1]

[1] ETH Zürich, Switzerland
[2] Microsoft

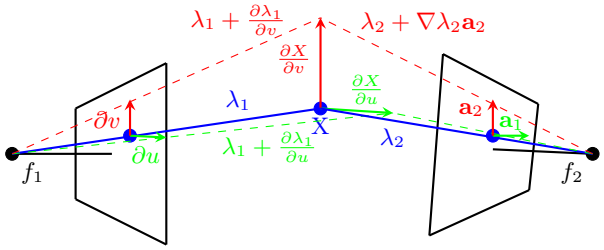

Fig. 1: Illustration of the problem. The notation is given in Sec. 2.1. Perturbing the observation $\mathbf{x}$ in the first camera in direction $\mathbf{m}$ will lead to the same change of 3D point $\mathbf{X}$, as perturbing the observation in the second camera in direction $\mathbf{Am}$. This relationship enables us to estimate the focal length and the relative pose of the cameras.

Abstract. We address the semi-calibrated relative pose estimation problem where we assume the principal point to be located in the center of the image and estimate the focal lengths, relative rotation, and translation of two cameras. We introduce the *first* minimal solver that requires only a single affine correspondence in conjunction with predicted monocular depth. Recognizing its degeneracy when the correspondence stems from a fronto-parallel plane, we present an alternative solver adept at automatically recovering the correct solution under such circumstances. By integrating these methods within the GC-RANSAC framework, we show they surpass standard approaches, delivering more accurate poses and focal lengths at comparable runtimes across large-scale, publicly available indoor and outdoor datasets. The code is available at `https://github.com/petrhruby97/semicalibrated_1AC_D`.

Keywords: Relative Pose · Affine Correspondence · Monodepth

1 Introduction

Estimating the relative pose between two cameras is a fundamental task in computer vision [32, 38, 43, 54, 68, 69] and robotics, with applications in 3D recon-

struction [1,9,33,66,67,71], visual localization [20,21,24,53], simultaneous localization and mapping (SLAM) [45,57,64,65], multi-view stereo [16,26,27,36], and visual odometry [55, 56]. In this paper, we focus on the semi-calibrated relative pose, when the focal lengths of the cameras are unknown, while the principal point is considered to be in the center of the images. This is an important task, since estimating focal length is challenging and tends to be inaccurate even when available in the EXIF tag [66]. On the other hand, assuming the principal point at the image center usually works well, even for cropped images, since the principal point can be recovered with bundle adjustment (BA). Note that the semi-calibrated setting is important in various applications such as absolute pose [14,39], relative pose [41,42], and vanishing point estimation [58]. Following prior work [7, 17, 31], we assume no camera distortion. In practice, this usually is a reasonable assumption, and a subsequent numerical optimization procedure estimates the distortion parameters. We use an affine correspondence [50] with a pre-trained relative (non-metric) monocular depth predictor [60, 61] to estimate the semi-calibrated pose from a single correspondence.

Solvers for estimating relative pose, commonly deployed within a RANSAC framework [25], are indispensable for achieving accurate results, especially when tackling the inherent noise and outliers in real-world data. The size of the minimal sample used for estimating the pose directly correlates with the problem complexity. In RANSAC, the runtime grows exponentially with the sample size, making methods that use fewer data points highly preferred. A special category of solvers is those that require a single correspondence. Their main advantage is that they can replace random sampling with exhaustive search, rendering the process deterministic. However, to solve a problem from a single data point, it must be equipped with rich information about the underlying scene geometry, or we need to make assumptions about the camera motion to reduce its degrees of freedom (DoF). Recognizing this, various single-point solvers for calibrated relative pose have been proposed, combining an affine correspondence (AC) with additional information, such as planar motion assumption [30], known vertical direction [29], and monocular depth [22].

Affine correspondences (AC) are a potent tool for relative pose estimation [48, 50–52]. Their appeal stems from the ability to reduce the sample size by imposing additional constraints. Early works like [17,59] introduced approximate solutions for uncalibrated and calibrated scenarios, leveraging 3 and 2 ACs, respectively. The constraints allowing direct use of ACs for relative pose estimation were studied in [10, 13], leading to exact solvers [6, 62]. In [5, 37], 2 ACs are used to find the unknown homography. A solution for relative pose between central cameras from 2 ACs was given in [23]. While there are solvers that marry a single AC to constraints such as the planar motion assumption [30], known vertical direction [29], or monocular depth [22], a notable gap persists: the absence of a single-correspondence solver tailored for scenarios with unknown focal lengths.

In this paper, we introduce two solvers tailored for semi-calibrated relative pose estimation, both of which harness a single affine correspondence and predicted monocular depth. The first method estimates the relative pose from a

single AC. Recognizing its degeneracy when the correspondence stems from a fronto-parallel plane, the other solver recovers the correct solution. Incorporating these solvers within the state-of-the-art Graph-Cut RANSAC framework [8] leads to more accurate results than the widely used methods both on outdoor [28, 34] and indoor [19] datasets.
To summarize, our contributions are the following:

- We propose the *first* minimal solver for semi-calibrated relative pose estimation from a single affine correspondence and monocular depth.
- We propose another solver applied to the same input as the first one when it can only estimate the camera rotation due to degeneracies in the data.

2 Minimal Solvers

In this section, we will first discuss the theoretical background. Then two minimal solvers will be proposed. The first one estimates the relative pose and focal lengths from a single affine correspondence and monocular depth. The second one is run when the first solver fails due to degeneracies in the data. It samples an additional correspondence via an exhaustive search approach to recover the model.

2.1 Notation and concepts

Let $\mathbf{X} \in \mathbb{R}^3$ be a 3D point, and $\mathbf{P} \in \mathbb{R}^{3,4}$ a matrix representing a pinhole camera. We can decompose $\mathbf{P} = \mathbf{K}\,[\mathbf{R}\,\mathbf{t}]$, where $\mathbf{K} \in \mathbb{R}^3$ is the intrinsics matrix, $\mathbf{R} \in \mathrm{SO}(3), \mathbf{t} \in \mathbb{R}^3$ is the pose of the camera $\mathbf{P}$. Here, we assume the semi-calibrated setting with square pixels and a known principal point. Then, $\mathbf{K}$ has the following form:

$$\mathbf{K} = \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{1}$$

where f is the focal length of the camera.

Let $\mathbf{x} \in \mathbb{R}^3$, $\mathbf{x} = [u\ v\ 1]^{\mathrm{T}}$ be the homogeneous representation of the projection of $\mathbf{X}$ into $\mathbf{P}$. Then, there holds:

$$\mathbf{x} \sim \mathbf{P}[\mathbf{X}^{\mathrm{T}} 1]^{\mathrm{T}} = \mathbf{K}(\mathbf{R}\mathbf{X} + \mathbf{t}),\ \mathbf{K}^{-1}\mathbf{x} \sim \mathbf{R}\mathbf{X} + \mathbf{t}. \tag{2}$$

Let $\lambda \in \mathbb{R}$ be the distance between point $\mathbf{X}$, and the center of camera $\mathbf{P}$. Then, we can rewrite the constraint as

$$\lambda \frac{\mathbf{K}^{-1}\mathbf{x}}{\|\mathbf{K}^{-1}\mathbf{x}\|} = \mathbf{R}\mathbf{X} + \mathbf{t}. \tag{3}$$

The expression $\frac{\mathbf{K}^{-1}\mathbf{x}}{\|\mathbf{K}^{-1}\mathbf{x}\|}$ is the normalized bearing vector of a projection $\mathbf{x}$. We define a function $q(u, v, f) = \frac{\mathbf{K}^{-1}\mathbf{x}}{\|\mathbf{K}^{-1}\mathbf{x}\|}$, which maps the image coordinates and

focal length bearing vectors as follows:

$$q(u, v, f) = \frac{1}{\sqrt{u^2 + v^2 + f^2}} \begin{bmatrix} u \\ v \\ f \end{bmatrix}. \tag{4}$$

Using this, we obtain the 3D point $\mathbf{X}$ as:

$$\mathbf{X} = \mathbf{R}^{\mathrm{T}}(\lambda q(u, v, f) - \mathbf{t}). \tag{5}$$

Derivatives of the 3D point (5) with respect to the image coordinates describe how the 3D point changes if we apply an infinitesimal perturbation to the image coordinates. We calculate the derivatives as:

$$\begin{aligned} \frac{\partial \mathbf{X}}{\partial u} &= \mathbf{R}^{\mathrm{T}} \left(\frac{\partial \lambda}{\partial u} q(u, v, f) + \lambda \frac{\partial q(u, v, f)}{\partial u} \right), \\ \frac{\partial \mathbf{X}}{\partial v} &= \mathbf{R}^{\mathrm{T}} \left(\frac{\partial \lambda}{\partial v} q(u, v, f) + \lambda \frac{\partial q(u, v, f)}{\partial v} \right), \end{aligned} \tag{6}$$

where

$$\begin{aligned} \frac{\partial q(u, v, f)}{\partial u} &= \frac{1}{(u^2 + v^2 + f^2)^{\frac{3}{2}}} \begin{bmatrix} v^2 + f^2 \\ -uv \\ -uf \end{bmatrix}, \\ \frac{\partial q(u, v, f)}{\partial v} &= \frac{1}{(u^2 + v^2 + f^2)^{\frac{3}{2}}} \begin{bmatrix} -uv \\ u^2 + f^2 \\ -vf \end{bmatrix}. \end{aligned}$$

We use the compact notation as follows:

$$\nabla \mathbf{X} = \begin{bmatrix} \frac{\partial \mathbf{X}}{\partial u} \frac{\partial \mathbf{X}}{\partial v} \end{bmatrix}, \ \nabla \lambda = \begin{bmatrix} \frac{\partial \lambda}{\partial u} \frac{\partial \lambda}{\partial v} \end{bmatrix}, \ \nabla q = \begin{bmatrix} \frac{\partial q(u, v, f)}{\partial u} \frac{\partial q(u, v, f)}{\partial v} \end{bmatrix}.$$

In order to get the derivative of point $\mathbf{X}$ in a general direction $\mathbf{m} \in \mathbb{R}^2$, we have the following formula:

$$\nabla \mathbf{X}\mathbf{m} = \mathbf{R}^{\mathrm{T}} \left(q(u, v, f) \nabla \lambda \mathbf{m} + \lambda \nabla q \mathbf{m} \right). \tag{7}$$

Note, that the derivative $\nabla \mathbf{X}\mathbf{m}$ does not depend on $\mathbf{t}$.

Two cameras. Let us have two cameras $\mathbf{P}_1 = [\mathbf{I}\ 0]$, $\mathbf{P}_2 = [\mathbf{R}\ \mathbf{t}]$, and a 3D point $\mathbf{X}$. Let $\mathbf{x}_1$, $\mathbf{x}_2$ be the projections of $\mathbf{X}$ into both cameras, and let λ_1, λ_2 be the relative depths of $\mathbf{X}$ in each camera. Then, there holds:

$$\lambda_1 \mathbf{R} q(u_1, v_1, f_1) + \mathbf{t} = \lambda_2 q(u_2, v_2, f_2). \tag{8}$$

Local affine frame (LAF) $(\mathbf{x}, \mathbf{A})$ consists of a projection $\mathbf{x} \in \mathbb{R}^3$, and a linear transformation $\mathbf{A} \in \mathbb{R}^{2 \times 2}$ describing the local coordinate system of the image region. Two LAFs $(\mathbf{x}_1, \mathbf{A}_1)$, $(\mathbf{x}_2, \mathbf{A}_2)$ form an affine correspondence.

Since $\mathbf{A} = \mathbf{A}_2 \mathbf{A}_1^{-1}$ represents the map that maps the infinitesimal vicinity of $\mathbf{x}_1$ to that of $\mathbf{x}_2$, the infinitesimal perturbation of $\mathbf{x}_1$ in direction $\mathbf{m}$, and the

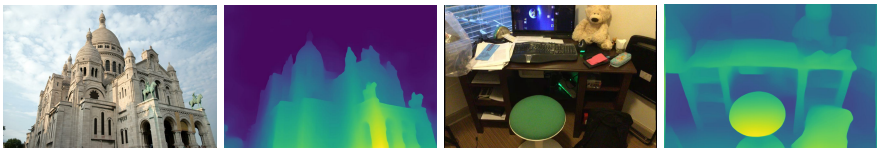

Fig. 2: Example images and their monocular depths by MiDaS-v3 [60, 61] from the PhotoTourism [35] and ScanNet [19] datasets used in the real experiments.

infinitesimal perturbation of $\mathbf{x}_2$ in direction $\mathbf{Am}$ lead to the same change in $\mathbf{X}$ [22]. For every $\mathbf{m}$, there holds:

$$\begin{aligned} q(u_1, v_1, f_1)\nabla\lambda_1\mathbf{m} + \lambda_1\nabla q(u_1, v_1, f_1)\mathbf{m} \\ = \mathbf{R}^{\mathrm{T}}(q(u_2, v_2, f_2)\nabla\lambda_2\mathbf{Am} + \lambda_2\nabla q(u_2, v_2, f_2)\mathbf{Am}). \end{aligned} \tag{9}$$

Since this holds for every $\mathbf{m}$, we get the constraints:

$$\begin{aligned} q(u_1, v_1, f_1)\nabla\lambda_1 + \lambda_1\nabla q(u_1, v_1, f_1) \\ = \mathbf{R}^{\mathrm{T}}(q(u_2, v_2, f_2)\nabla\lambda_2\mathbf{A} + \lambda_2\nabla q(u_2, v_2, f_2)\mathbf{A}). \end{aligned} \tag{10}$$

Relative depth. Usually, we only know the depth in each image up to unknown common scale α. Therefore, we obtain the following constraints:

$$\lambda_1\mathbf{R}q(u_1, v_1, f_1) + \mathbf{t} = \alpha\lambda_2 q(u_2, v_2, f_2), \tag{11}$$

$$\begin{aligned} q(u_1, v_1, f_1)\nabla\lambda_1 + \lambda_1\nabla q(u_1, v_1, f_1) = \\ \alpha\mathbf{R}^{\mathrm{T}}(q(u_2, v_2, f_2)\nabla\lambda_2\mathbf{A} + \lambda_2\nabla q(u_2, v_2, f_2)\mathbf{A}). \end{aligned} \tag{12}$$

The unknowns are the rotation $\mathbf{R}$, $\mathbf{t}$, focal lengths f_1, f_2, and the scale factor α. There are 9 variables: 3 for $\mathbf{R}$, 3 for $\mathbf{t}$, 2 for (f_1, f_2), and 1 for α. Equations (11), (12) together give 9 independent constraints. The problem is minimal.

2.2 Semi-calibrated Relative Pose (1AC+D)

Here, we describe how to solve for a semi-calibrated relative pose from a single affine correspondence and monocular relative depth. We know a single affine correspondence $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{A})$, relative depths λ_1 of $\mathbf{x}_1$, λ_2 of $\mathbf{x}_2$, and their derivatives $\nabla\lambda_1, \nabla\lambda_2$. First, we find focal lengths f_1, f_2, scale factor α, and rotation $\mathbf{R}$ from equation (12). Then, we use constraint (11) to find the translation.

Constraint (12) can be split into two constraints in the form $d_i(f_1) = \alpha\mathbf{R}^{\mathrm{T}}e_i(f_2), i \in \{1, 2\}$, as follows:

$$\begin{aligned} d_1(f_1) &= q(u_1, v_1, f_1)\frac{\partial\lambda_1}{\partial u_1} + \lambda_1\frac{\partial q(u_1, v_1, f_1)}{\partial u_1}, \\ d_2(f_1) &= q(u_1, v_1, f_1)\frac{\partial\lambda_1}{\partial v_1} + \lambda_1\frac{\partial q(u_1, v_1, f_1)}{\partial v_1}, \\ e_1(f_2) &= q(u_2, v_2, f_2)\nabla\lambda_2\mathbf{a}_1 + \lambda_2\nabla q\mathbf{a}_1, \\ e_2(f_2) &= q(u_2, v_2, f_2)\nabla\lambda_2\mathbf{a}_2 + \lambda_2\nabla q\mathbf{a}_2, \end{aligned} \tag{13}$$

where $\mathbf{a}_1$, and $\mathbf{a}_2$ are the columns of matrix $\mathbf{A}$.

Vectors $d_i(f_1)$, $e_i(f_2)$ are only related by rotation and common scale. As a result, the angles between $d_1(f_1)$ and $d_2(f_1)$, and between $e_1(f_2)$ and $e_2(f_2)$ are equal. Additionally, the ratios of the norms of $d_1(f_1)$ and $d_2(f_1)$, and of $e_1(f_2)$ and $e_2(f_2)$ are equal. These observations lead to the following equations:

$$\frac{d_1(f_1)^{\mathrm{T}} d_2(f_1)}{\|d_1(f_1)\|\|d_2(f_1)\|} = \frac{e_1(f_2)^{\mathrm{T}} e_2(f_2)}{\|e_1(f_2)\|\|e_2(f_2)\|}, \; \frac{\|d_1(f_1)\|}{\|d_2(f_1)\|} = \frac{\|e_1(f_2)\|}{\|e_2(f_2)\|}. \tag{14}$$

We get the first constraint by multiplying the left sides, and the right sides of the equations in (14):

$$\frac{d_1(f_1)^{\mathrm{T}} d_2(f_1)}{d_2(f_1)^{\mathrm{T}} d_2(f_1)} = \frac{e_1(f_2)^{\mathrm{T}} e_2(f_2)}{e_2(f_2)^{\mathrm{T}} e_2(f_2)}. \tag{15}$$

If we multiply this with both denominators, we obtain:

$$(d_1(f_1)^{\mathrm{T}} d_2(f_1))(e_2(f_2)^{\mathrm{T}} e_2(f_2)) = (e_1(f_2)^{\mathrm{T}} e_2(f_2))(d_2(f_1)^{\mathrm{T}} d_2(f_1)). \tag{16}$$

We get a second constraint by squaring the second equation of (14) and multiplying the result by both denominators as:

$$(d_1(f_1)^{\mathrm{T}} d_1(f_1))(e_2(f_2)^{\mathrm{T}} e_2(f_2)) = (e_1(f_2)^{\mathrm{T}} e_1(f_2))(d_2(f_1)^{\mathrm{T}} d_2(f_1)). \tag{17}$$

Eqs. (16), and (17) build together a system of two independent polynomial equations in f_1, f_2. Since the variables appear in the system only with even exponents, we first substitute $g_1 = f_1^2$ and $g_2 = f_2^2$. We build a solver for this system with an automatic Gröbner basis solver generator [40]. The system has 9 solutions in terms of g_1, g_2, and the elimination template has 36 rows. The running time is $29\mu s$. We use the solver to find g_1, g_2 and calculate

$$f_1 = \sqrt{g_1}, \; f_2 = \sqrt{g_2}, \; \alpha = \frac{\|d_1(f_1)\|}{\|e_1(f_2)\|}.$$

Then, we find $\mathbf{R}$ as $\mathbf{R} = \mathbf{E}\mathbf{D}^{-1}$, where:

$$\mathbf{E} = \left[\frac{e_1(f_2)}{\|e_1(f_2)\|} \; \frac{e_2(f_2)}{\|e_2(f_2)\|} \; \frac{e_1(f_2) \times e_2(f_2)}{\|e_1(f_2) \times e_2(f_2)\|} \right], \; \mathbf{D} = \left[\frac{d_1(f_1)}{\|d_1(f_1)\|} \; \frac{d_2(f_1)}{\|d_2(f_1)\|} \; \frac{d_1(f_1) \times d_2(f_1)}{\|d_1(f_1) \times d_2(f_1)\|} \right],$$

and refine rotation $\mathbf{R}$ using SVD [2]. Finally, we calculate $\mathbf{t}$ from (11). Note, that the overall scale is fixed by setting the scale of the first depth to one, instead of the common choice $|\mathbf{t}| = 1$. This solver works for generic camera configurations. However, it is degenerate when the optical axes of the cameras are parallel and orthogonal to the plane from which the used correspondence stems. We design a solver for this situation in Sec. 2.3. In Sec. 2.4, we show how the two solvers are used together in practice.

2.3 Fronto-parallel Planes Solver (1AC+1PC)

Here, we are going to discuss the case, when the optical axes of the cameras $\mathbf{P}_1$, $\mathbf{P}_2$ are parallel, and the plane containing point $\mathbf{X}$ is orthogonal to them. This is a degenerate case of the problem solved in Sec. 2.2. A calibrated version of this configuration was studied in [47] in the context of dense stereo reconstruction. Since it is a common case in real-world scenarios [34], we present a solution for it, which requires one affine correspondence $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{A})$, and one point correspondence $(\mathbf{y}_1, \mathbf{y}_2)$.

In this case, the views are related by homography $\mathbf{H} = \mathbf{K}_2(\mathbf{R} - \mathbf{t}\mathbf{n}^{\mathrm{T}})\mathbf{K}_1^{-1}$ [46], where $\mathbf{n} = [0\ 0\ 1]^{\mathrm{T}}$, and $\mathbf{R}$ rotates around the z-axis. This homography is proportional to:

$$\mathbf{H} \sim \begin{bmatrix} \cos\varphi & -\sin\varphi & -f_1 t_1 \\ \sin\varphi & \cos\varphi & -f_1 t_2 \\ 0 & 0 & \frac{f_1}{f_2}(1 - t_3) \end{bmatrix}, \tag{18}$$

where φ is the angle of $\mathbf{R}$, and $\mathbf{t} = [t_1\ t_2\ t_3]$. This homography is an affine transformation, and, therefore, it can be obtained from the AC $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{A})$ as:

$$\mathbf{H} = \begin{bmatrix} \mathbf{A} & \mathbf{x}_2 - \mathbf{A}\mathbf{x}_1 \\ \mathbf{o}^{\mathrm{T}} & 1 \end{bmatrix}. \tag{19}$$

Since matrices (18), (19) are proportional, we obtain the rotation as follows:

$$\mathbf{R} = \begin{bmatrix} \frac{\mathbf{A}}{\det \mathbf{A}} & \mathbf{o} \\ \mathbf{o}^{\mathrm{T}} & 1 \end{bmatrix}. \tag{20}$$

Let $\mathbf{b} = (\mathbf{x}_2 - \mathbf{A}\mathbf{x}_1)/\det \mathbf{A}$, and let b_1, b_2 be the elements of $\mathbf{b}$. Then, we express the elements of $\mathbf{t}$ as follows:

$$t_1 = \frac{b_1}{f_1},\ t_2 = \frac{b_2}{f_1},\ t_3 = \frac{f_2}{f_1}\frac{1}{\det \mathbf{A}} - 1. \tag{21}$$

The fundamental matrix is composed as $\mathbf{F} = \mathbf{K}_2^{-\mathrm{T}}[\mathbf{t}]_\times \mathbf{R}\mathbf{K}_1^{-1}$. There holds

$$\mathbf{F} \sim \begin{bmatrix} -\sin\varphi t_3 & -\cos\varphi t_3 & f_1 t_2 \\ \cos\varphi t_3 & -\sin\varphi t_3 & -f_1 t_1 \\ f_2(\sin\varphi t_1 - \cos\varphi t_2) & f_2(\cos\varphi t_1 + \sin\varphi t_2) & 0 \end{bmatrix}. \tag{22}$$

This becomes

$$\begin{bmatrix} -s_\varphi t_3 & -c_\varphi t_3 & f_1 t_2 \\ c_\varphi t_3 & -s_\varphi t_3 & -f_1 t_1 \\ f_2(s_\varphi t_1 - c_\varphi t_2) & f_2(c_\varphi t_1 + s_\varphi t_2) & 0 \end{bmatrix}, \tag{23}$$

and then

$$\begin{bmatrix} s_\varphi - \frac{f_2}{f_1}\frac{s_\varphi}{\det \mathbf{A}} & -c_\varphi + \frac{f_2}{f_1}\frac{c_\varphi}{\det \mathbf{A}} & b_2 \\ c_\varphi - \frac{f_2}{f_1}\frac{c_\varphi}{\det \mathbf{A}} & s_\varphi - \frac{f_2}{f_1}\frac{s_\varphi}{\det \mathbf{A}} & -b_1 \\ \frac{f_2}{f_1}(s_\varphi b_1 - c_\varphi b_2) & \frac{f_2}{f_1}(c_\varphi b_1 + s_\varphi b_2) & 0 \end{bmatrix}. \tag{24}$$

This can be split as:

$$\mathbf{F} = \begin{bmatrix} s_\varphi & -c_\varphi & b_2 \\ c_\varphi & s_\varphi & -b_1 \\ 0 & 0 & 0 \end{bmatrix} + \frac{f_2}{f_1} \begin{bmatrix} -\frac{s_\varphi}{\det \mathbf{A}} & \frac{c_\varphi}{\det \mathbf{A}} & 0 \\ -\frac{c_\varphi}{\det \mathbf{A}} & -\frac{s_\varphi}{\det \mathbf{A}} & 0 \\ s_\varphi b_1 - c_\varphi b_2 & c_\varphi b_1 + s_\varphi b_2 & 0 \end{bmatrix}. \tag{25}$$

This can be written in a compact form as $\mathbf{F} = \mathbf{F}_1 + \frac{f_2}{f_1}\mathbf{F}_2$, where $\mathbf{F}_1$, $\mathbf{F}_2$ can be fully determined from the affine correspondence $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{A})$.

Fundamental matrix $\mathbf{F}$ is therefore a function of the ratio $\frac{f_2}{f_1}$. In order to determine this ratio, we use an additional point correspondence $(\mathbf{y}_1, \mathbf{y}_2)$, and find the value of $\frac{f_2}{f_1}$ that satisfies $\mathbf{y}_2^\mathrm{T}\mathbf{F}\mathbf{y}_1 = 0$. This equation becomes $\mathbf{y}_2^\mathrm{T}\mathbf{F}_1\mathbf{y}_1 + \frac{f_2}{f_1}\mathbf{y}_2^\mathrm{T}\mathbf{F}_2\mathbf{y}_1 = 0$. Therefore, we can obtain the unique value of the ratio as:

$$\frac{f_2}{f_1} = -\frac{\mathbf{y}_2^\mathrm{T}\mathbf{F}_1\mathbf{y}_1}{\mathbf{y}_2^\mathrm{T}\mathbf{F}_2\mathbf{y}_1} \tag{26}$$

We substitute this value into (25) to obtain the parameters of fundamental matrix $\mathbf{F}$. However, in this case it is not possible to decompose $\mathbf{F}$ to get the translation and focal lengths. The reason for this is that, if $\mathbf{R}$, $\mathbf{t} = \begin{bmatrix} t_1 & t_2 & t_3 \end{bmatrix}^T$, f_1, f_2 build a fundamental matrix $\mathbf{F}$, then for every $\alpha \in \mathbb{R}$, $\mathbf{R}$, $\mathbf{t}' = \begin{bmatrix} t_1 & t_2 & \alpha t_3 \end{bmatrix}^T$, $f_1' = \alpha f_1$, $f_2' = \alpha f_2$ build the same fundamental matrix $\mathbf{F}$. Knowing the relative depth, we can find the focal lengths f_1, f_2. The procedure leads to solving a quadratic equation and it is described in the Supplementary material.

2.4 Combining the Solvers (1AC+D$^+$)

Here, we describe how to combine the single-point solver (Sec. 2.2) with the solver addressing the aforementioned degeneracy (Sec. 2.3). If the cameras and the observed plane are in degenerate configuration, matrix $\frac{1}{\det \mathbf{A}}\mathbf{A}$ is a rotation. In that case, the eigenvalues of this matrix are equal to 1. We employ the following procedure in order to detect degenerate cases after solver 1AC+D finished:

- Calculate eigenvalues λ_1, λ_2 of $\frac{1}{\det \mathbf{A}}\mathbf{A}$ $(\lambda_1 \geq \lambda_2)$.
- If $|\lambda_1| \leq \epsilon$, the configuration is labeled degenerate.

In case the configuration is not degenerate, the original 1AC+D solver returned the correct solutions, and we have nothing to do. We can verify the relative pose in the RANSAC procedure without additional steps. We use $\epsilon = 1.2$ in all experiments.

Otherwise, we run a model upgrade procedure similar to the DEGENSAC [18] algorithm. Knowing the special scene configuration that caused solver 1AC+D to fail, we can run an exhaustive search over the remaining correspondences and use the solver from Sec. 2.3 to estimate the model given the already selected AC and the new correspondence.

While the whole sampling process may be quadratic $\mathcal{O}(n^2)$ in the number of correspondences, our observations show that, in practice, the runtime is usually linear.

3 Experiments

In this section, we compare our 1AC+D (Sec. 2.2) and 1AC+1PC (Sec. 2.4) solvers with the 6PC [42], 7PC [32], 3AC [11], 3PC-to-AC [17], 4SIFT [7], 5ORB [4] methods both on synthetic and real-world data.

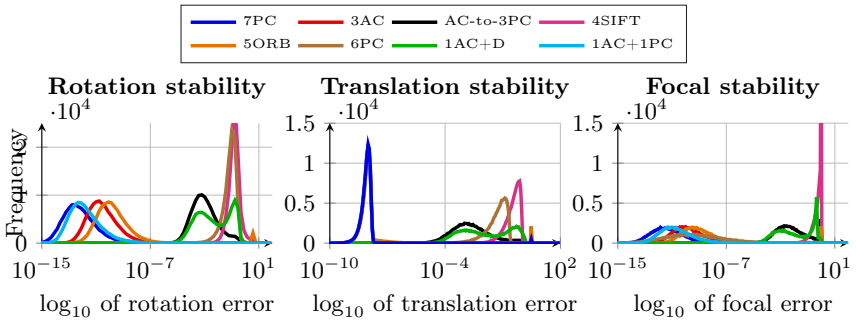

Fig. 3: Stability study. Histogram of $\log_{10}$ *left:* rotation, *middle:* translation errors in radians, and *right:* relative focal length errors of the poses estimated by the 6PC [42], 7PC [32], 3AC [11], AC-to-3PC [17], 4SIFT [7], 5ORB [4], the proposed 1AC+D (Sec. 2.2) and 1AC+1PC (Sec. 2.3) solvers, computed from 100k noiseless samples.

3.1 Synthetic Experiments

Numerical Stability. First, we generate a random rotation matrix $\mathbf{R}_{\text{gt}}$, a translation vector $\mathbf{t}_{\text{gt}}$, and focal lengths $f_{1,\text{gt}}$, $f_{2,\text{gt}}$ from uniform distribution $[500, 2000]$. To generate a PC, we sample a point $\mathbf{X} \in \mathbb{R}^3$ from a Gaussian distribution with mean $[0,\ 0,\ 5]^{\text{T}}$ and standard deviation 1. We project $\mathbf{X}$ into the first camera as $\mathbf{p}$ and into the second one as $\mathbf{q}$. To generate an AC, we sample four coplanar PCs, fit a homography onto them [32], and find the affine transformation $\mathbf{A}$ as the derivative of the homography [5] at the first PC. We combine the first PC and $\mathbf{A}$ to get the AC. We calculate the depths λ_1, λ_2 as the distance between $\mathbf{X}$ and the camera centers. To calculate the depth derivatives $\nabla\lambda_1$, we perturb the projection $\mathbf{p}$ in both u and v directions, intersect the perturbed projections with the plane defined by the four sampled points, measure the distance λ_1' between the camera center and the intersection, and calculate the derivative as $\lambda' - \lambda$. We find the derivative $\nabla\lambda_2$ in a similar way. Since the scale is relative, we generate a random scale factor $\sigma \in \mathbb{R}$ and multiply λ_2 and $\nabla\lambda_2$ by σ. Let $\mathbf{R}_{\text{est}}$, $\mathbf{t}_{\text{est}}$, f_{est}, respectively, the rotation, translation, and focal length estimated by the solver. For fundamental matrix solvers, we decompose the relative pose and the focal length from the estimated $\mathbf{F}_{\text{est}}$. We measure the rotation error as the angle of the rotation represented as ${\mathbf{R}_{\text{est}}}^{\text{T}}\mathbf{R}_{\text{gt}}$, and the translation error as the angle between vectors t_{est} and t_{GT}.

We generated $n = 100000$ random problem instances and ran the solvers on the noiseless samples. Fig. 3 shows histograms of rotation and translation errors in radians. The proposed solver 1AC+D has a peak close to 10^{-1} which is usually considered unstable. However, in our case, this case is detectable and we can run the other minimal solver (1AC+1PC) designed specifically for this particular

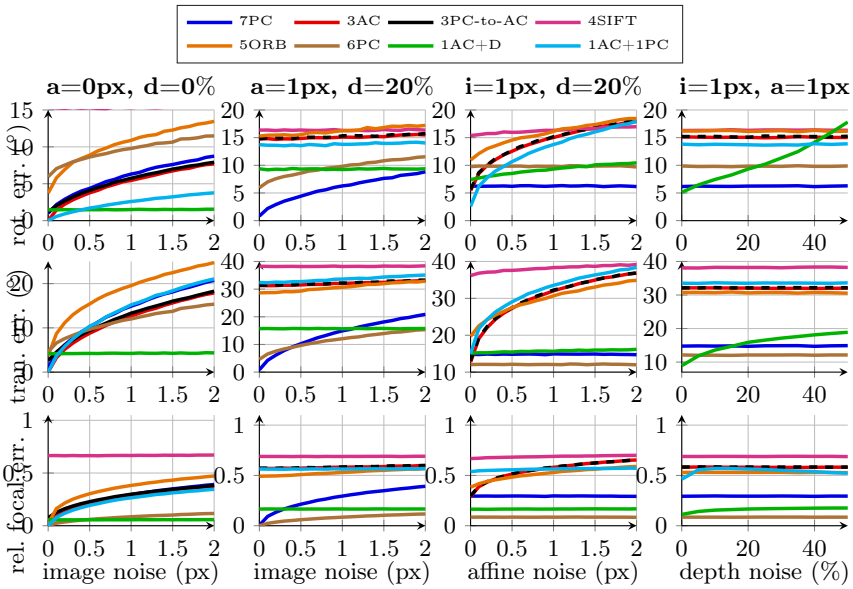

Fig. 4: Angular errors avg. over 10000 runs as a function of the image, affine, and depth noise (horizontal axis). The parameters fixed for a test are reported in the titles.

scenario. Note that the 7PC solver also has a degeneracy when observing close-to-planar scenes. This is visible by peaking at 10^0. The proposed 1AC+1PC solver that runs in the case the camera observes a fronto-parallel plane is particularly stable without any peak close to 10^0. In this case, decomposing $\mathbf{F}$ does not yield a correct solution, as described in Sec. 2.3. Therefore, we estimate the focal length with the procedure introduced in the Supplementary Material. We note that in practice, it is usually possible to retrieve the focal lengths for situations close to the fronto-parallel case by applying BA before decomposing $\mathbf{F}$.

Tests with noise. To evaluate the robustness of the solvers to the noise in the input data, we generated minimal problems like in the previous paragraph, and we perturbed them with artificial noise. Namely, we added zero-mean Gaussian noise with standard deviation i to the coordinates of each projected point. To perturb the affine matrices $\mathbf{A}$, we added zero-mean Gaussian noise with standard deviation d to each of the 4 PC used to calculate $\mathbf{A}$. To perturb the depths, we multiplied them with scalars sampled from Gaussian distribution with mean 1 and standard deviation d.

Errors of the solvers with artificial input noise are displayed in Fig. 4. The top row shows the average rotation in degrees, the middle row the translation errors in degrees, and the bottom row the relative error of the focal length. The main message of these synthetic experiments is that both the proposed solvers act reasonably w.r.t. to noise in the data. 1AC+D is almost always more accurate than the affine-based solver. In certain situations, it is more accurate than 7PC. 1AC+1PC shows similar trends to the 3AC solver. The image noise has

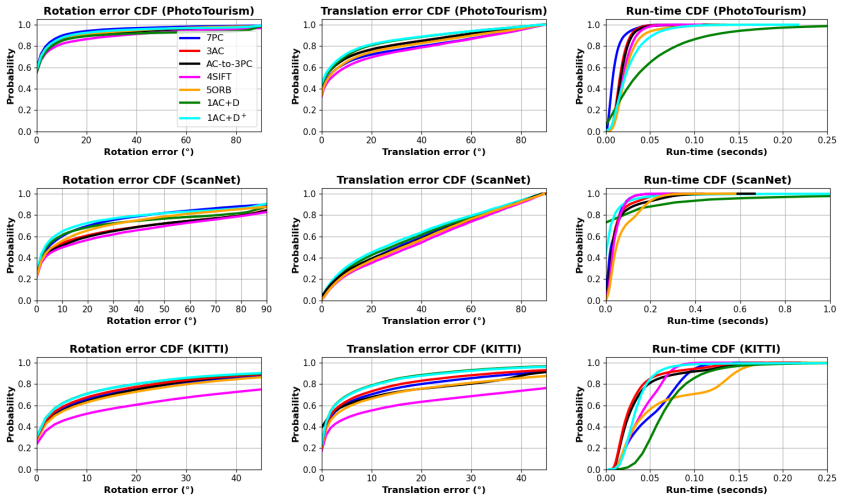

Fig. 5: The cumulative distribution functions (CDFs) of the rotation (left) and translation (middle) errors in degrees and the runtime (right) in secs of the 6PC [42], 7PC [32], 3AC [6], AC-to-3PC [17], 4SIFT [7], 5ORB [4], and the proposed 1AC+D (Sec. 2.3) and 1AC+D$^+$ (Sec. 2.4) solvers integrated into GC-RANSAC [8] on datasets PhotoTourism, ScanNet, and KITTI. A curve close to the top-left corner indicates accuracy.

a negligible effect on the 1AC+D solver, with depth noise having larger impact than affine noise. This is expected since only one constraint comes from the points and the rest from the depth or the affine elements; a similar behavior is seen in [70]. The average error of the 7PC solver is not zero when no noise is added since 7PC is degenerate when the points are close-to-planar. Therefore, the method fails in some cases. However, it is important to note that no work analyzes realistic noise levels in affine correspondences. Therefore, we cannot draw conclusions other than the proposed solver acting reasonably w.r.t. to increasing noise levels in the data. More synthetic tests, evaluating the solvers in specific scenarios, can be found in the Supplementary material.

3.2 Real-world Experiments

In this section, we test the proposed and other solvers on real-world data from public datasets. We obtain ACs by detecting DoG features [44], finding the affine shape by AffNet [50], and extracting HardNet [49] descriptors. This approach is among the leaders in the IMC 2020 benchmark [35]. We obtain relative depth by MiDaS-v3 [60, 61]. Note, that we do *not* use ground truth depth. The depth estimation takes about 12 ms, and it only needs to be calculated once per image, *i.e.*, if we perform exhaustive pose estimation on n images, which is a standard step of 3D reconstruction pipelines, we need to perform $\binom{n}{2}$ pose estimations, and only n depth estimations. As relative pose estimation is on average, 1-5 times slower than depth estimation, the latter becomes marginal when the image count

surpasses 20. Examples are in Fig. 2. The depth derivates are calculated from the depth images using bilinear sampling to achieve subpixel accuracy.

The tested solvers are integrated within the state-of-the-art GC-RANSAC [8] robust estimator, which is the standard approach today [11]. Although AC-based methods do not work without local optimization (LO) due to the noisy ACs, with LO, they can achieve SOTA accuracy as shown in [11], and in Tabs. 1, 2, 3. GC-RANSAC uses two types of solvers, one for estimating the model from a minimal sample and one for estimating from a larger-than-minimal sample. We set the parameters similarly as was proposed in [3]. We noticed that the focal lengths that solvers (such as 6PC, 3AC, 1AC+D, and 1AC+1PC) estimate lead to unstable results even when using weighted histogram voting [15]. Thus, we use the implied fundamental matrix inside GC-RANSAC. The reported relative poses and focal lengths are decomposed from the estimated **F**s by [12].

	7PC	6PC	3AC	AC-to-3PC	4SIFT	5ORB	1AC+D	1AC+D$^+$
AVG	19.5	17.7	17.4	18.6	20.7	16.3	<u>16.1</u>	**14.6**
MED	3.6	3.9	<u>3.1</u>	3.2	3.8	3.4	3.2	**3.0**
AUC@5°	41.2	39.3	<u>42.8</u>	42.4	39.1	40.2	41.9	**43.0**
AUC@10°	49.5	48.6	51.5	51.0	46.9	48.7	<u>51.6</u>	**52.8**
AUC@20°	58.1	58.7	60.5	59.9	55.1	58.0	<u>61.9</u>	**63.2**
t (ms)	15.4	**12.7**	<u>14.7</u>	15.3	15.8	19.9	45.9	24.2

Table 1: Avg. and median pose errors (in degrees; max. of the rotation and translation errors), the AUC score at 5°, 10°, and 20° and average runtime (in milliseconds) on the PhotoTourism dataset [35]. The best results are bold, the second bests are underlined.

PhotoTourism. We use the data from the CVPR IMC 2020 PhotoTourism challenge [35]. It consists of 25 scenes (2 – validation; 12 – training; 11 – test sets) of landmarks with photos of varying sizes and focal lengths collected from the internet. We run the methods on the two scenes from the validation split with a total of 9900 image pairs.

The avg. and median pose errors in degrees, the Area Under the recall Curve (AUC) thresholded at 5°, 10°, and 20°, and the avg. runtime in milliseconds are reported in Table 1. The pose errors are calculated by taking the max. rotation and translation errors. The proposed approach (1AC+D$^+$) that runs the 1AC+D solver and, in case of degeneracy, runs an exhaustive search on the correspondences, leading to a significant improvement in terms of accuracy compared with other solvers. All methods run in real time.

The cumulative distribution functions (CDFs) of the rotation and translation errors and runtimes are shown in the top row of Fig. 5. The rotation errors of all methods show very similar trends. The translation errors of both 1AC+D and 1AC+D$^+$ are the best, with the curve of 1AC+D$^+$ being marginally higher than that of 1AC+D. The runtime curves show that 1AC+D is slow, due to the increased iteration number of GC-RANSAC to cope with the degeneracy. 1AC+D$^+$ runs at a similar speed as the other solvers, always terminating under $0.1s$.

ScanNet. The ScanNet dataset [19] contains 1613 monocular sequences with ground truth camera poses and depth maps. We evaluate the compared minimal solvers on the 1500 challenging pairs used in SuperGlue [63]. The results are

in Table 2. The proposed 1AC+D$^+$ solver achieves the best accuracy in all accuracy metrics. Despite the degeneracy check and model upgrade process, it is the second fastest, being only marginally slower than the 7PC solver (by 2ms). We can see that the degenerate solver without the upgrade requires significantly more iterations (thus the increased runtime) to find a good minimal sample.

The cumulative distribution functions (CDFs) of the rotation, translation errors, and processing times are in the middle row of Fig. 5. The proposed 1AC+D$^+$ solver leads to the best rotation accuracy, especially under 25°. The translation errors of the methods are similar, with 1AC+D$^+$ being marginally better than the rest of the methods. Similarly, as on PhotoTourism, the runtime curves show that 1AC+D as it requires more iterations due to the degeneracy. 1AC+D$^+$ runs at a similar speed as the other solvers.

	7PC	6PC	3AC	AC-to-3PC	4SIFT	5ORB	1AC+D	1AC+D$^+$
AVG	<u>43.5</u>	44.0	52.6	52.2	54.4	51.3	46.1	**42.9**
MED	34.8	<u>33.3</u>	43.8	43.0	46.8	45.3	35.8	**31.3**
AUC@5°	6.3	<u>6.5</u>	6.4	6.5	6.0	5.6	<u>6.5</u>	**6.6**
AUC@10°	13.7	14.0	13.5	13.6	12.4	12.0	<u>14.1</u>	**14.7**
AUC@20°	23.3	23.2	22.1	22.4	20.4	20.5	<u>24.0</u>	**25.2**
t (ms)	**39.9**	40.5	44.1	42.3	<u>41.3</u>	84.8	130.8	41.5

Table 2: Avg. and median pose errors (in degrees; max. of the rotation and translation errors), the AUC score at 5°, 10°, and 20° and average runtime (in milliseconds) on the ScanNet dataset [19].

KITTI. The KITTI dataset [28] is a real-world benchmark for tasks stereo, optical flow, visual odometry, 3D object detection, and tracking. It is captured by driving in the city of Karlsruhe with accurate ground truth from the laser scanner and GPS localization system. We tested our method on the 11 visual odometry sequences that are provided with ground truth. The avg. number of images in the sequences is 1826. We test the methods by using different frame distances d. We iterate through the frames in a sequence and, to form an image pair in the kth frame, we select the $(k+d)$th image. We run tests on $d \in \{5, 10, 25\}$.

The results are reported in Table 3. While the differences when $d = 5$ are small, they get more pronounced as d increases. The proposed solvers always lead to the best accuracy in all metrics. With large d, the improvements from the proposed 1AC+D$^+$ solver are significant compared to 7PC and 3AC. Also, the methods get faster due to finding fewer correspondences. For $d = 25$, the proposed 1AC+D$^+$ is the fastest, running twice as fast as the 7PC solver.

The cumulative distribution functions (CDFs) of the rotation and translation errors and processing times are shown in the bottom row of Fig. 5. For these plots, all $d \in \{5, 10, 25\}$ are considered. The rotation accuracy looks similar for all methods except for the 4SIFT solver, which is inaccurate on this dataset. The translations exhibit more significant differences, with the 1AC+D$^+$ and 7PC methods being the best. 1AC+D$^+$ is one of the fastest algorithms on this dataset, finishing under 0.1 seconds in all cases. Note, that KITTI contains surfaces *not* parallel to the camera (road, buildings on the side). Therefore, solver 1AC+D gets enough non-degenerate samples, although the rotation between views is often very small.

		7PC	6PC	3AC	AC-to-3PC	4SIFT	5ORB	1AC+D	1AC+D$^+$
$k, k+5$	AVG	6.3	6.7	**6.2**	6.7	9.8	7.3	**6.2**	**6.2**
	MED	**2.1**	2.2	**2.1**	2.2	2.4	2.2	**2.1**	**2.1**
	AUC@5°	46.3	45.4	46.3	45.3	42.8	44.9	46.4	**46.4**
	AUC@10°	61.7	60.9	61.7	60.8	57.5	60.2	61.8	**61.8**
	AUC@20°	74.5	73.4	74.5	73.3	69.4	72.6	74.6	**74.6**
	t (ms)	29.1	29.8	23.5	**18.8**	24.6	24.8	93.6	55.2
$k, k+10$	AVG	11.7	13.3	11.5	14.6	27.1	16.4	11.1	**11.0**
	MED	3.7	4.0	3.7	4.3	9.1	4.7	3.7	**3.6**
	AUC@5°	33.7	32.1	33.8	31.0	23.6	29.5	34.0	**34.1**
	AUC@10°	49.5	47.7	49.6	45.7	35.3	43.9	49.9	**50.0**
	AUC@20°	63.2	61.1	63.3	58.3	45.7	56.5	63.8	**63.8**
	t (ms)	30.6	140.5	**22.4**	33.7	46.0	60.8	56.7	35.4
$k, k+25$	AVG	30.7	38.4	30.5	40.1	71.9	45.5	26.9	**26.8**
	MED	10.7	24.1	11.1	39.2	75.4	39.5	**8.6**	8.8
	AUC@5°	17.1	11.6	16.7	9.0	4.3	8.2	**18.7**	18.6
	AUC@10°	29.8	20.9	29.2	15.7	7.8	15.3	**32.6**	32.4
	AUC@20°	42.7	31.2	42.2	22.7	11.8	23.3	**46.9**	46.6
	t (ms)	23.9	376.8	14.3	68.3	61.5	121.2	14.2	**11.5**

Table 3: Avg. and med. pose errors (in degrees; max. of the rot. and trans. errors), the AUC score at 5°, 10°, and 20° and average runtime (in milliseconds) on the KITTI dataset [28] with different distances between the consecutive frames I_k and I_{k+d}.

	7PC	6PC	3AC	AC-to-3PC	4SIFT	5ORB	1AC+D	1AC+D$^+$
PhotoT.	23.1	21.9	21.2	20.9	21.7	22.2	21.1	**20.7**
ScanNet	19.8	24.4	**15.9**	17.9	18.3	19.8	20.1	18.9
KITTI	73.6	71.8	70.0	71.1	81.2	76.0	**67.5**	68.5

Table 4: Median relative focal length errors (in %). The best results are in bold, and the second bests are underlined.

Focal length. The median relative focal length errors are in Table 4. Both on the PhotoTourism and KITTI datasets, the proposed solvers lead to the most accurate focal length. On ScanNet, the 3AC method achieves the lowest error. However, the proposed solvers also lead to comparable accuracy.

4 Conclusions

This paper proposes a new approach to semi-calibrated relative pose estimation from a single affine correspondence and predicted monocular depth. The proposed method is the *first* minimal solver for this problem and provides significant improvements in efficiency and accuracy over existing methods. We also propose a second solver that addresses the degeneracies in the data and improves the accuracy of the estimation when the first solver fails. Through extensive experiments on indoor and outdoor datasets, we demonstrate that the proposed method outperforms standard algorithms, achieving more accurate fundamental matrices with fewer correspondences. The proposed method has potential applications in various computer vision tasks, including 3D reconstruction, visual localization, simultaneous localization and mapping, and multi-view stereo. The code is available at `https://github.com/petrhruby97/semicalibrated_1AC_D`.

References

1. Agarwal, S., Furukawa, Y., Snavely, N., Simon, I., Curless, B., Seitz, S.M., Szeliski, R.: Building rome in a day. Commun. ACM **54**(10) (2011)
2. Arun, K.S., Huang, T.S., Blostein, S.D.: Least-squares fitting of two 3-d point sets. IEEE Transactions on Pattern Analysis and Machine Intelligence (1987)
3. Barath, D., Chin, T.J., Chum, O., Mishkin, D., Ranftl, R., Matas, J.: RANSAC in 2020 tutorial. In: Conference on Computer Vision and Pattern Recognition (2020), `http://cmp.felk.cvut.cz/cvpr2020-ransac-tutorial/`
4. Barath, D.: Five-point fundamental matrix estimation for uncalibrated cameras. In: CVPR 2018 (2018)
5. Barath, D., Hajder, L.: A theory of point-wise homography estimation. Pattern Recognit. Lett. (2017)
6. Barath, D., Hajder, L.: Efficient recovery of essential matrix from two affine correspondences. IEEE Trans. Image Process. (2018)
7. Barath, D., Kukelova, Z.: Relative pose from SIFT features. In: ECCV 2022 (2022)
8. Barath, D., Matas, J.: Graph-cut RANSAC. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018 (2018)
9. Barath, D., Mishkin, D., Eichhardt, I., Shipachev, I., Matas, J.: Efficient initial pose-graph generation for global sfm. In: CVPR (2021)
10. Barath, D., Molnár, J., Hajder, L.: Novel methods for estimating surface normals from affine transformations. In: Computer Vision, Imaging and Computer Graphics Theory and Applications - 10th International Joint Conference, VISIGRAPP 2015, Berlin, Germany, March 11-14, 2015, Revised Selected Papers. Communications in Computer and Information Science (2015)
11. Barath, D., Polic, M., Förstner, W., Sattler, T., Pajdla, T., Kukelova, Z.: Making affine correspondences work in camera geometry computation. In: ECCV 2020 (2020)
12. Bartoli, A., Sturm, P.: Nonlinear estimation of the fundamental matrix with minimal parameters. IEEE Transactions on Pattern Analysis and Machine Intelligence **26**(3), 426–432 (2004)
13. Bentolila, J., Francos, J.M.: Conic epipolar constraints from affine correspondences. Comput. Vis. Image Underst. (2014)
14. Bujnak, M., Kukelova, Z., Pajdla, T.: A general solution to the P4P problem for camera with unknown focal length. In: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA. IEEE Computer Society (2008). `https://doi.org/10.1109/CVPR.2008.4587793`, `https://doi.org/10.1109/CVPR.2008.4587793`
15. Bujnak, M., Kukelova, Z., Pajdla, T.: Robust focal length estimation by voting in multi-view scene reconstruction. In: Asian Conference on Computer Vision. pp. 13–24. Springer (2009)
16. Chen, R., Han, S., Xu, J., Su, H.: Point-based multi-view stereo network. In: ICCV (2019)
17. Chum, O., Matas, J., Obdrzalek, S.: Epipolar geometry from three correspondences. In: Computer Vision Winter Workshop. pp. 1057–7149 (2003)
18. Chum, O., Werner, T., Matas, J.: Two-view geometry estimation unaffected by a dominant plane. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). vol. 1, pp. 772–779. IEEE (2005)

19. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: CVPR (2017)
20. DeTone, D., Malisiewicz, T., Rabinovich, A.: Toward geometric deep SLAM. CoRR (2017)
21. DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperPoint: Self-supervised interest point detection and description. In: CVPR (2018)
22. Eichhardt, I., Barath, D.: Relative pose from deep learned depth and a single affine correspondence. In: Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XII (2020)
23. Eichhardt, I., Chetverikov, D.: Affine correspondences between central cameras for rapid relative pose estimation. In: Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VI. Lecture Notes in Computer Science (2018)
24. Engel, J., Schöps, T., Cremers, D.: LSD-SLAM: large-scale direct monocular SLAM. In: ECCV (2014)
25. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6), 381–395 (1981)
26. Furukawa, Y., Curless, B., Seitz, S.M., Szeliski, R.: Towards internet-scale multi-view stereo. In: CVPR (2010)
27. Furukawa, Y., Hernández, C.: Multi-view stereo: A tutorial. FTCGV **9**(1-2) (2015)
28. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
29. Guan, B., Zhao, J., Li, Z., Sun, F., Fraundorfer, F.: Minimal solutions for relative pose with a single affine correspondence. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020 (2020)
30. Hajder, L., Barath, D.: Relative planar motion for vehicle-mounted cameras from a single affine correspondence. In: 2020 IEEE International Conference on Robotics and Automation, ICRA 2020, Paris, France, May 31 - August 31, 2020 (2020)
31. Hartley, R.: In defense of the eight-point algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence (1997)
32. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press (2004)
33. Heinly, J., Schönberger, J.L., Dunn, E., Frahm, J.: Reconstructing the world* in six days. In: CVPR (2015)
34. Jin, Y., Mishkin, D., Mishchuk, A., Matas, J., Fua, P., Yi, K.M., Trulls, E.: Image Matching across Wide Baselines: From Paper to Practice. IJCV (2020)
35. 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)
36. Kar, A., Häne, C., Malik, J.: Learning a multi-view stereo machine. In: NeurIPS (2017)
37. Köser, K.: Geometric estimation with local affine frames and free-form surfaces. Ph.D. thesis, University of Kiel (2009)
38. Kukelova, Z., Bujnak, M., Pajdla, T.: Polynomial eigenvalue solutions to the 5-pt and 6-pt relative pose problems. In: BMVC (2008)
39. Kukelova, Z., Heller, J., Fitzgibbon, A.: Efficient intersection of three quadrics and applications in computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)

40. Larsson, V., Åström, K., Oskarsson, M.: Efficient solvers for minimal problems by syzygy-based reduction. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017 (2017)
41. Larsson, V., Zobernig, N., Taskin, K., Pollefeys, M.: Calibration-free structure-from-motion with calibrated radial trifocal tensors. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part V. Lecture Notes in Computer Science, vol. 12350, pp. 382–399. Springer (2020). https://doi.org/10.1007/978-3-030-58558-7_23, https://doi.org/10.1007/978-3-030-58558-7_23
42. Li, H.: A simple solution to the six-point two-view focal-length problem. In: European Conference on Computer Vision. pp. 200–213. Springer (2006)
43. Li, H., Hartley, R.I.: Five-point motion estimation made easy. In: ICPR (2006)
44. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision (IJCV) **60**(2), 91–110 (2004)
45. Lynen, S., Zeisl, B., Aiger, D., Bosse, M., Hesch, J.A., Pollefeys, M., Siegwart, R., Sattler, T.: Large-scale, real-time visual-inertial localization revisited. IJRR **39**(9) (2020)
46. Malis, E., Vargas, M.: Deeper understanding of the homography decomposition for vision-based control. INRIA Research Report (01 2007)
47. Megyesi, Z., Kós, G., Chetverikov, D.: Dense 3d reconstruction from images by normal aided matching. Machine Graphics & Vision International Journal archive (2006)
48. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Gool, L.V.: A comparison of affine region detectors. Int. J. Comput. Vis. (2005)
49. Mishchuk, A., Mishkin, D., Radenovic, F., Matas, J.: Working Hard to Know Your Neighbor's Margins: Local Descriptor Learning Loss. In: NeurIPS (2017)
50. Mishkin, D., Radenovic, F., Matas, J.: Repeatability is Not Enough: Learning Affine Regions via Discriminability. In: European Conference on Computer Vision (2018)
51. Mishkin, D., Matas, J., Perdoch, M.: MODS: fast and robust method for two-view matching. Comput. Vis. Image Underst. (2015)
52. Morel, J., Yu, G.: ASIFT: A new framework for fully affine invariant image comparison. SIAM J. Imaging Sci. (2009)
53. Mur-Artal, R., Montiel, J.M.M., Tardós, J.D.: ORB-SLAM: A versatile and accurate monocular SLAM system. IEEE Trans. Robotics **31**(5) (2015)
54. Nistér, D.: An efficient solution to the five-point relative pose problem. In: CVPR (2003)
55. Nistér, D., Naroditsky, O., Bergen, J.R.: Visual odometry. In: CVPR (2004)
56. Nistér, D., Naroditsky, O., Bergen, J.R.: Visual odometry for ground vehicle applications. J. Field Robotics **23**(1) (2006)
57. Panek, V., Kukelova, Z., Sattler, T.: Meshloc: Mesh-based visual localization. In: ECCV (2022)
58. Pautrat, R., Liu, S., Hruby, P., Pollefeys, M., Barath, D.: Vanishing point estimation in uncalibrated images with prior gravity direction. In: International Conference on Computer Vision (ICCV) (2023)
59. Perdoch, M., Matas, J., Chum, O.: Epipolar geometry from two correspondences. In: 18th International Conference on Pattern Recognition (ICPR 2006), 20-24 August 2006, Hong Kong, China. pp. 215–219. IEEE Computer Society (2006). https://doi.org/10.1109/ICPR.2006.497, https://doi.org/10.1109/ICPR.2006.497

60. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. ICCV (2021)
61. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(3) (2022)
62. Raposo, C., Barreto, J.P.: Theory and practice of structure-from-motion using affine correspondences. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016 (2016)
63. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: Computer Vision and Pattern Recognition. pp. 4938–4947 (2020)
64. Sattler, T., Leibe, B., Kobbelt, L.: Improving image-based localization by active correspondence search. In: ECCV (2012)
65. Sattler, T., Maddern, W., Toft, C., Torii, A., Hammarstrand, L., Stenborg, E., Safari, D., Okutomi, M., Pollefeys, M., Sivic, J., Kahl, F., Pajdla, T.: Benchmarking 6dof outdoor visual localization in changing conditions. In: CVPR (2018)
66. Schönberger, J.L., Frahm, J.: Structure-from-motion revisited. In: CVPR (2016)
67. Snavely, N., Seitz, S.M., Szeliski, R.: Modeling the world from internet photo collections. IJCV **80**(2) (2008)
68. Stewénius, H., Nistér, D., Kahl, F., Schaffalitzky, F.: A minimal solution for relative pose with unknown focal length. Image Vis. Comput. (2008)
69. Stewénius, H., Engels, C., Nistér, D.: Recent developments on direct relative orientation. ISPRS Journal of Photogrammetry and Remote Sensing **60**(4), 284–294 (2006). https://doi.org/https://doi.org/10.1016/j.isprsjprs.2006.03.005, https://www.sciencedirect.com/science/article/pii/S092427160600030X
70. Ventura, J., Kukelova, Z., Sattler, T., Baráth, D.: P1ac: Revisiting absolute pose from a single affine correspondence. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 19751–19761 (October 2023)
71. Zhu, S., Zhang, R., Zhou, L., Shen, T., Fang, T., Tan, P., Quan, L.: Very large-scale global sfm by distributed motion averaging. In: CVPR (2018)