

PIAC: Revisiting Absolute Pose From a Single Affine Correspondence

Jonathan Ventura¹ Zuzana Kukelova² Torsten Sattler³ Dániel Baráth⁴

¹ Department of Computer Science & Software Engineering, Cal Poly, San Luis Obispo

² Visual Recognition Group, Faculty of Electrical Engineering, Czech Technical University in Prague

³ Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical University in Prague

⁴ Computer Vision and Geometry Group, ETH Zürich

Abstract

Affine correspondences have traditionally been used to improve feature matching over wide baselines. While recent work has successfully used affine correspondences to solve various relative camera pose estimation problems, less attention has been given to their use in absolute pose estimation. We introduce the first general solution to the problem of estimating the pose of a calibrated camera given a single observation of an oriented point and an affine correspondence. The advantage of our approach (PIAC) is that it requires only a single correspondence, in comparison to the traditional point-based approach (P3P), significantly reducing the combinatorics in robust estimation. PIAC provides a general solution that removes restrictive assumptions made in prior work and is applicable to large-scale image-based localization. We propose a minimal solution to the PIAC problem and evaluate our novel solver on synthetic data, showing its numerical stability and performance under various types of noise. On standard image-based localization benchmarks we show that PIAC achieves more accurate results than the widely used P3P algorithm. Code for our method is available at <https://github.com/jonathanventura/PIAC/>.

1. Introduction

Image-based localization is the process of determining the pose of a query image in a reference coordinate system from pre-registered images in a database. Image-based localization is an important technology for many applications including robotics [41], self-driving cars [20] and augmented reality [51]. It also forms an essential component of large-scale structure-from-motion pipelines [48].

The dominant method for image-based localization is to estimate the camera pose from correspondences between 2D features in the query image and known 3D points in the scene. Since these correspondences are likely to contain mismatches, a best-fit model is found through Random

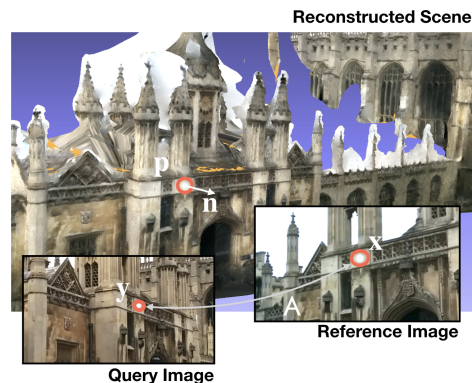


Figure 1: Overview of the perspective-one-affine-correspondence (PIAC) problem. Our goal is to estimate the pose of a query image based on a single 2D observation of a known oriented 3D point and an affine correspondence to a reference image with known pose.

Sample Consensus (RANSAC) [17] or one of its derivatives [12, 32, 5], where camera pose hypotheses are estimated from minimal samples of three point correspondences (PCs) using a perspective-three-point (P3P) algorithm [44].

In cases of very wide-baseline matching or with extreme differences in viewing angle, affine-covariant features increase match reliability in comparison to traditional difference-of-Gaussian detectors such as SIFT [35]. A match between affine-covariant features, which we will call an affine correspondence (AC), estimates a 2×2 affine transformation matrix between the local regions around the corresponding points in the two images.

Recently, several authors have developed a theory of the constraints induced by ACs on two-view geometry [9, 45, 16], and minimal solvers for relative pose estimation from ACs [16, 15]. However, relatively little attention has been given to absolute pose estimation from ACs. We revisit the problem of estimating absolute camera pose from a single affine correspondence, and refer to this

as the perspective-one-affine-correspondence (P1AC) problem. Köser and Koch [29] considered a limited form of the P1AC problem by assuming that the reference view is orthorectified. In contrast, our method removes this limiting assumption and makes more general assumptions: that the reference and query images are perspective images of a locally planar surface in an arbitrary orientation, that we have estimated the normal to the surface at the observed 3D point, and that we have estimated an affine transformation between the reference and the query images. We derive a system of equations for this problem and propose the first general solution for absolute pose from a single AC.

In comparison to P3P solutions, our P1AC solution only requires a single correspondence instead of three in the minimal case, and thus significantly reduces the complexity of robust estimation. We demonstrate these advantages through experiments on synthetic and real data.

Our contributions are as follows:

- We introduce a novel solution to the P1AC problem using reduction to the 3Q3 problem [30]. Our solution is general and does not impose the restrictive assumptions made in previous work.
- Experiments on synthetic data establish the numerical stability of our solver and analyze its performance under various types of noise.
- Experiments on two benchmark datasets for image-based localization show that our P1AC solution produces more accurate pose estimates than P3P.

2. Related work

Affine correspondences are commonly estimated using either frame-to-frame tracking techniques [36, 3] or affine-covariant region detection [37, 39]. In the case of frame-to-frame tracking, the deformation of a locally planar surface observed by a moving camera is well-modeled by an affine transformation. For affine-covariant regions, an affine transformation describing the shape of the local region around the interest point is estimated [8, 39]. While computing ACs may increase feature detection computation time over computing PCs, ACs improve match reliability in wide-baseline image matching [24], introduce new constraints on camera geometry [28, 9] and, therefore, decrease the number of correspondences needed for camera pose estimation [29, 45, 16, 15].

Recent work has examined how ACs might be used to improve and simplify pose estimation and subsequently improve image-based localization, visual odometry, and structure-from-motion. Köser developed a theory of relationship between homographies and ACs [28], and Bento-lila and Francos demonstrated that an AC provides three constraints on the fundamental matrix [9]. Raposo and Barreto re-derived these constraints [45], making them usable for essential matrix estimation using only two ACs in-

stead of five PCs as is normally needed [42]. Eichhardt and Chetverikov extended relative pose estimation using ACs to handle any central camera model [16]. Eichhardt and Barath [15] demonstrated relative pose estimation from a single AC and a depth map estimated by a neural network. Recently, Barath et al. provided recommendations for speeding up RANSAC when using ACs [6], including local feature geometry refinement and uncertainty propagation in RANSAC. The ideas presented in [6] can be used in combination with any pose estimation solver that uses ACs, including the solvers proposed in this paper.

Absolute pose estimation using ACs has received less attention than relative pose. Algorithms for absolute pose estimation from PCs have a long history, dating back to Grunert’s solution [18]. Among many recent P3P solutions [27, 25, 44, 14] the Lambda Twist method [44] and [14] are among the fastest and most numerically stable.

Köser and Koch introduced a solution for absolute pose from a single AC called Differential Pose Resection (DPR), with the restriction that the reference image is a scaled or- thographic image of a planar surface with the normal parallel to the optical axis [29]. In our setting, the reference image is, instead, a perspective image of a locally planar surface in an arbitrary orientation. Their method could be applied in our setting if we rectified the image and then re-computed the affine region in the rectified image. However, this is expensive as it needs to be done per feature. An alternative is to virtually warp the AC so that the reference view is orthorectified. However, this warping cannot be done exactly, and, as we show in our experiments, this approximate warping leads to decreasing accuracy in the DPR solution as the amount of perspective distortion increases.

Collins and Bartoli [13] derived novel constraints and a minimal solution for the absolute pose of a camera given four co-planar PCs. They first calculate the homography between the object plane and the camera image. While the homography could be directly decomposed into the camera pose, they propose an alternative solution which addresses the spatially-varying accuracy of the homography. They choose an optimal point on the plane to compute a first-order approximation to the homography (an affine transformation) and then compute the pose of the camera based on the 2D point projection and the first-order approximated local transformation. Their solution, called Infinitesimal Plane-based Pose Estimation (IPPE), is therefore highly related to DPR [29], in that it ultimately computes the pose of the camera based on a single AC between the object plane and the image. To apply IPPE in our setting, we generate three virtual PCs around the AC in the reference image by ray intersection with the object plane. However, as we show in our experiments, the IPPE solution is less accurate than ours because of the inherent inaccuracy in the virtual PCs.

Haug and Jähne developed a solution for computing the

absolute pose of a stereo or RGB-D camera from a single AC [22]. Their solution assumes knowledge of the depth of the point in both the reference and query images and thus is not applicable in our scenario of localizing a monocular camera, where the depth of the point in the query image is unknown.

3. Minimal solvers

Notation: We use a sans-serif capital letter M for a matrix, a bold lower-case letter \mathbf{v} for a vector, and an italic lower-case letter s for a scalar. We use subscripts to indicate matrix and vector indexing. A tilde over a vector indicates a homogeneous version of that vector ($\tilde{\mathbf{x}} = [x_1 \ x_2 \ 1]^T$).

Consider two calibrated pinhole cameras observing a scene: a reference camera, whose pose w.r.t. the world coordinate system is known, and a query camera, whose pose is unknown. Without loss of generality, we assume that the reference camera has extrinsics $[I \mid \mathbf{0}]$, and the query camera has extrinsics $[R \mid \mathbf{t}]$.¹ Here R and \mathbf{t} represent the world-to-camera transformation of the query camera, transforming points from the world coordinate system to the coordinate system of the query camera. This is a standard setup in localization pipelines, where our goal is to estimate the 3D rotation R and translation \mathbf{t} of the query camera.

The inputs to our PIAC method are: corresponding points \mathbf{x} and \mathbf{y} in the reference and query images, respectively; A , the affine transformation from the local region around \mathbf{x} to the local region around \mathbf{y} ; d , the depth of the observed 3D point in the reference image; and \mathbf{n} , the surface normal at the 3D point in the coordinate system of the reference image.

First, we will develop a system of constraints on the absolute pose of the query camera induced by the AC, and then we will describe a minimal solution to this derived system.

3.1. Constraints induced by AC on camera pose

Let us define a function $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ that maps points in the reference image to points in the query image. Let \mathbf{x}, \mathbf{y} be corresponding points in the two images such that $\mathbf{y} = f(\mathbf{x})$, and let \mathbf{u} be a point in the neighborhood of \mathbf{x} . The first-order Taylor approximation of f centered at \mathbf{x} is $\hat{f}(\mathbf{u}) = \mathbf{y} + A(\mathbf{u} - \mathbf{x})$, where $A = \nabla_{\mathbf{u}}f(\mathbf{u})|_{\mathbf{u}=\mathbf{x}}$ is the local affine transformation matrix.

We now aim to analytically derive the form of $\nabla_{\mathbf{u}}f(\mathbf{u})$ under perspective projection. Let $\pi^{-1} : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ be the mapping from a point in the reference image to the local 3D surface. Then we have

$$f(\mathbf{u}) = \pi(R\pi^{-1}(\mathbf{u}) + \mathbf{t}). \quad (1)$$

¹When the reference camera is not at the origin, we first transform the world coordinate system to place the reference camera at the origin and then apply the inverse transform to the resulting solution for R and \mathbf{t} .

The projection function $\pi(\mathbf{p})$ is straightforward: $\pi(\mathbf{p}) = [p_1/p_3, p_2/p_3]^T$. The un-projection function $\pi^{-1}(\mathbf{u})$, however, requires knowledge of the scene geometry. We assume that the cameras observe a locally planar surface defined by 3D point $\mathbf{p} = d\tilde{\mathbf{x}}$ and normal \mathbf{n} . To find $\pi^{-1}(\mathbf{u})$, we intersect ray \mathbf{u} with the plane. The plane is defined by

$$\mathbf{n}^T(\mathbf{p}' - \mathbf{p}) = 0 \quad (2)$$

where \mathbf{p}' is a point on the planar surface. Parameterizing the ray-plane intersection by $\mathbf{p}' = \alpha\tilde{\mathbf{u}}$, we have

$$\mathbf{n}^T(\alpha\tilde{\mathbf{u}} - \mathbf{p}) = 0 \quad (3)$$

and therefore

$$\alpha = \frac{\mathbf{n}^T\mathbf{p}}{\mathbf{n}^T\tilde{\mathbf{u}}}. \quad (4)$$

Thus, under the assumption of a locally planar surface, we have determined $\pi^{-1}(\mathbf{u})$ as follows:

$$\pi^{-1}(\mathbf{u}) = \frac{\mathbf{n}^T\mathbf{p}}{\mathbf{n}^T\tilde{\mathbf{u}}}\tilde{\mathbf{u}}. \quad (5)$$

Let $\mathbf{q} = R\pi^{-1}(\mathbf{u}) + \mathbf{t}$ so that $f(\mathbf{u}) = \pi(\mathbf{q})$. To compute $\nabla_{\mathbf{u}}f(\mathbf{u})$, we first apply the quotient rule:

$$\begin{aligned} \nabla_{\mathbf{u}}f(\mathbf{u}) &= \nabla_{\mathbf{u}}\pi(\mathbf{q}) = \nabla_{\mathbf{u}}\left(\frac{1}{q_3}\begin{bmatrix} q_1 \\ q_2 \end{bmatrix}\right) \\ &= \frac{1}{q_3^2}\left[q_3\left(\nabla_{\mathbf{u}}\begin{bmatrix} q_1 \\ q_2 \end{bmatrix}\right) - \begin{bmatrix} q_1 \\ q_2 \end{bmatrix}(\nabla_{\mathbf{u}}q_3)\right] \\ &= \frac{1}{q_3}\left[\nabla_{\mathbf{u}}\begin{bmatrix} q_1 \\ q_2 \end{bmatrix} - \frac{1}{q_3}\begin{bmatrix} q_1 \\ q_2 \end{bmatrix}(\nabla_{\mathbf{u}}q_3)\right]. \end{aligned} \quad (6)$$

To make the expression linear in R, \mathbf{t} and simplify the later solution of the system of equations, we apply the substitution $\mathbf{v} = \frac{1}{q_3}[q_1 \ q_2]^T$ to obtain

$$\nabla_{\mathbf{u}}f(\mathbf{u}) = \frac{1}{q_3}\left(\nabla_{\mathbf{u}}\begin{bmatrix} q_1 \\ q_2 \end{bmatrix} - \mathbf{v}(\nabla_{\mathbf{u}}q_3)\right). \quad (7)$$

Noting that $\mathbf{q} = \alpha(R\tilde{\mathbf{u}}) + \mathbf{t}$, to compute $\nabla_{\mathbf{u}}\mathbf{q}$ we apply the product rule:

$$\nabla_{\mathbf{u}}\mathbf{q} = (R\tilde{\mathbf{u}})\nabla_{\mathbf{u}}\alpha + \alpha\nabla_{\mathbf{u}}(R\tilde{\mathbf{u}}). \quad (8)$$

To compute $\nabla_{\mathbf{u}}\alpha$ we apply the quotient rule:

$$\nabla_{\mathbf{u}}\alpha = \nabla_{\mathbf{u}}\left(\frac{\mathbf{n}^T\mathbf{p}}{\mathbf{n}^T\tilde{\mathbf{u}}}\right) = -\frac{(\mathbf{n}^T\mathbf{p})(\mathbf{n}_{1:2}^T)}{(\mathbf{n}^T\tilde{\mathbf{u}})^2} = -\alpha\frac{\mathbf{n}_{1:2}^T}{\mathbf{n}^T\tilde{\mathbf{u}}}. \quad (9)$$

Note also that $\nabla_{\mathbf{u}}(R\tilde{\mathbf{u}}) = R_{:,1:2}$. Now we plug our expressions for $\nabla_{\mathbf{u}}\alpha$ (Eq. 9) and $\nabla_{\mathbf{u}}(R\tilde{\mathbf{u}})$ into Eq. 8 to obtain

$$\nabla_{\mathbf{u}}\mathbf{q} = (R\tilde{\mathbf{u}})\left(-\alpha\frac{\mathbf{n}_{1:2}^T}{\mathbf{n}^T\tilde{\mathbf{u}}}\right) + \alpha R_{:,1:2} \quad (10)$$

$$= \alpha\left(R_{:,1:2} - (R\tilde{\mathbf{u}})\frac{\mathbf{n}_{1:2}^T}{\mathbf{n}^T\tilde{\mathbf{u}}}\right) \quad (11)$$

$$= \frac{\mathbf{n}^T\mathbf{p}}{\mathbf{n}^T\tilde{\mathbf{u}}}\left(R_{:,1:2} - (R\tilde{\mathbf{u}})\frac{\mathbf{n}_{1:2}^T}{\mathbf{n}^T\tilde{\mathbf{u}}}\right). \quad (12)$$

Evaluating this expression at $\mathbf{u} = \mathbf{x}$ we obtain

$$\nabla_{\mathbf{u}} \mathbf{q}|_{\mathbf{u}=\mathbf{x}} = d \left(\mathbf{R}_{:,1:2} - (\mathbf{R}\tilde{\mathbf{x}}) \frac{\mathbf{n}_{1:2}^T}{\mathbf{n}^T \tilde{\mathbf{x}}} \right). \quad (13)$$

Finally, we plug Eq. 13 into Eq. 7 to arrive at a formula for $\mathbf{J} = \nabla_{\mathbf{u}} f(\mathbf{u})|_{\mathbf{u}=\mathbf{x}}$:

$$\mathbf{J} = \frac{d}{m} (\mathbf{R}_{1:2,1:2}(\mathbf{n}^T \tilde{\mathbf{x}}) - (\mathbf{R}_{1:2,:} \tilde{\mathbf{x}}) \mathbf{n}_{1:2}^T - \mathbf{v}(\mathbf{R}_{3,1:2}(\mathbf{n}^T \tilde{\mathbf{x}}) - (\mathbf{R}_{3,:} \tilde{\mathbf{x}}) \mathbf{n}_{1:2}^T)) \quad (14)$$

where $m = \mathbf{n}^T \tilde{\mathbf{x}}(d(\mathbf{R}_{3,:} \tilde{\mathbf{x}}) + t_3)$.

We multiply m on both sides of the constraint $\mathbf{A} = \mathbf{J}$ to cancel out the denominator and make the constraint linear. Our set of six constraints to be solved is thus

$$y_1(\mathbf{R}_{3,:} \mathbf{p} + t_3) - (\mathbf{R}_{1,:} \mathbf{p} + t_1) = 0, \quad (15)$$

$$y_2(\mathbf{R}_{3,:} \mathbf{p} + t_3) - (\mathbf{R}_{2,:} \mathbf{p} + t_2) = 0, \quad (16)$$

$$ma_{11} - mj_{11} = 0, \quad ma_{12} - mj_{12} = 0, \quad (17)$$

$$ma_{21} - mj_{21} = 0, \quad ma_{22} - mj_{22} = 0, \quad (18)$$

where the first two constraints enforce that the 3D point projects to the 2D observation in the query camera, and the remaining constraints enforce that the affine transformation matrix equals the Jacobian matrix derived above. Note that the only unknowns in equations 14 and 15-18 are elements of \mathbf{R} and \mathbf{t} .

3.2. Minimal solution

We have six independent constraints on the entries of \mathbf{R} and \mathbf{t} and six degrees of freedom (three for the rotation and three for the translation). Therefore, we can solve for the absolute pose with a single AC.

We use the Cayley parameterization of the rotation with parameters x, y, z as follows:

$$\mathbf{R} = \frac{1}{s} \begin{bmatrix} 1+x^2-y^2-z^2 & 2(xy-z) & 2(y+xz) \\ 2(xy+z) & 1-x^2+y^2-z^2 & 2(yz-x) \\ 2(xz-y) & 2(x+yz) & 1-x^2-y^2+z^2 \end{bmatrix} \quad (19)$$

where $s = 1 + x^2 + y^2 + z^2$. Note that we cannot represent 180° rotations with this parameterization; however, this degenerate configuration is not a problem in practice since it would mean that the reference and the query camera have no visual overlap. After plugging this parameterization of \mathbf{R} into equations 15–18, we multiply all equations by s to cancel out the denominator in \mathbf{R} .

Let $\mathbf{c} = [st_1, st_2, st_3, x^2, xy, xz, y^2, yz, z^2, x, y, z, 1]^T$. Writing the equations 15-18 as a matrix-vector multiplication $\mathbf{M}\mathbf{c} = 0$, we first eliminate st using Gauss-Jordan elimination. This leaves three quadratic equations in ten monomials on x, y, z – a 3Q3 problem. Kukulova et al. [30] introduced a fast solver for the 3Q3 problem which produces up to eight solutions for x, y, z . Once we obtain solutions for x, y, z , we solve for \mathbf{t} by backsubstitution into equations 15-18.

3.3. Alternative solvers

We also consider two alternative solvers based on previous work, to establish a baseline comparison for the PIAC method proposed in this paper.

3.3.1 Differential pose resection

The Differential Pose Resection (DPR) method [29] solves the PIAC problem under the restriction that the reference image is an ortho-rectified view of the plane. To apply the DPR method in the general case, we need to virtually warp the AC so that it comes from an ortho-rectified view.

Let \mathbf{w} represent the 2D coordinates of a point on the plane and let \mathbf{H} be a homography from the plane to the reference image s.t. $\mathbf{u} = \pi(\mathbf{H}\tilde{\mathbf{w}})$ and $\mathbf{x} = \pi(\mathbf{H}[0 \ 0 \ 1]^T)$. The point corresponding to \mathbf{u} is $\mathbf{v} = g(\mathbf{w}) = f(\pi(\mathbf{H}\tilde{\mathbf{w}}))$. g is thus a non-linear transformation from the plane to the query. We replace f with \hat{f} since we only know the AC between the reference and query images. We can approximate g as an affine transformation by computing the matrix $\mathbf{A}' = \nabla_{\mathbf{w}} g|_0$:

$$\mathbf{A}' = \frac{A}{H_{33}} \begin{bmatrix} H_{11} - (H_{13}H_{31})/H_{33} & H_{12} - (H_{13}H_{32})/H_{33} \\ H_{21} - (H_{23}H_{31})/H_{33} & H_{22} - (H_{23}H_{32})/H_{33} \end{bmatrix}. \quad (20)$$

3.3.2 Infinitesimal plane-based pose estimation

The Infinitesimal Plane-based Pose Estimation (IPPE) method [13] solves for the pose of the query camera given four coplanar PCs. In our case, the AC only contains a single PC. To transform a PIAC problem into an IPPE problem, we generate three extra virtual PCs by sampling 2D points around \mathbf{x} in the reference image: the 3D points are found by ray intersection with the plane (Eq. 5), and the 2D points are found by applying the affine matrix \mathbf{A} . We generate three extra virtual PCs in this manner and provide them along with the PC from the original AC as input to IPPE. It is important to note that these virtual correspondences are only approximations generated by the assumption that the AC lies on an actual plane, and that the AC is valid beyond an infinitesimal region around the 2D point.

4. Evaluation

We performed evaluations on synthetic and real data to test the numerical stability of our solver, test robustness to various types of noise, and assess its performance as part of an image-based localization system.

We compared the following algorithms in our evaluation:

- **P3P** Standard P3P absolute pose calculation from three 2D-3D point observations. We used the Lambda Twist P3P solver [44].
- **PIAC**: Our solution to the PIAC problem.

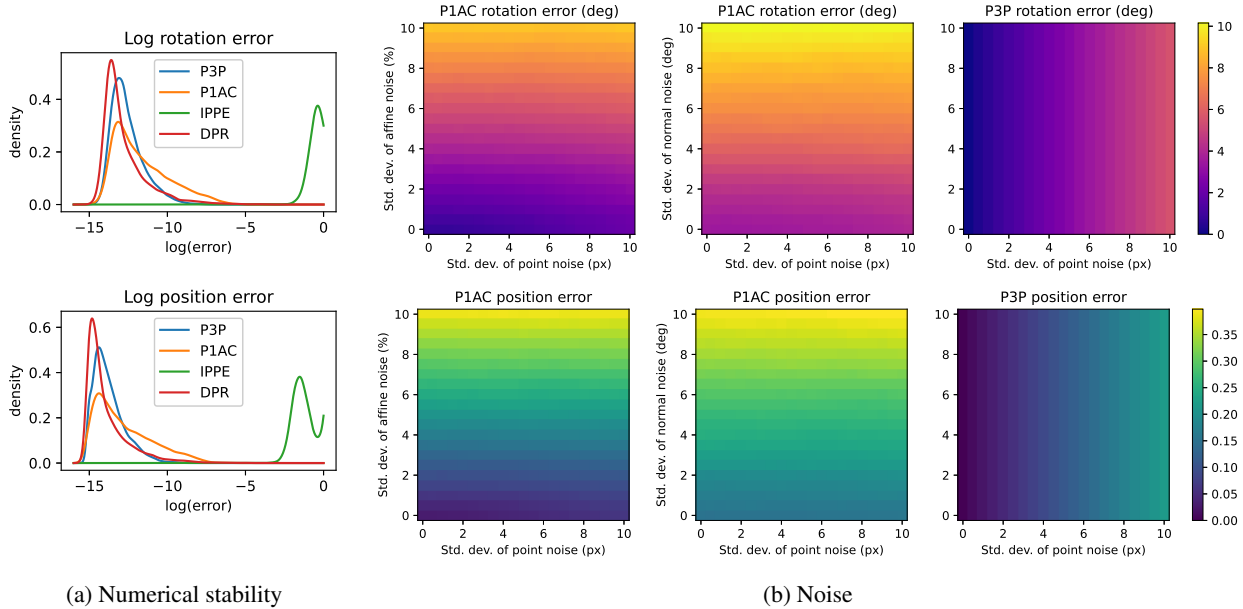


Figure 2: Results of synthetic data experiments. (a) Analysis of numerical stability with zero noise added to observations. The plots are estimates of the distribution produced by Gaussian kernel-density estimation. *Top*: Log rotation error; *Bottom*: Log position error. (b) Analysis of error with respect to various types and levels of noise. *Left*: Error of P1AC solver w.r.t. affine and point noise, with normal noise fixed to 1° . *Center*: Error of P1AC solver w.r.t. normal and point noise, with affine noise fixed to 4%. *Right*: Error of P3P solver w.r.t. affine and point noise. Note that the P3P solver does not use the affine transformation or normal vector and thus is unaffected by affine and normal noise.

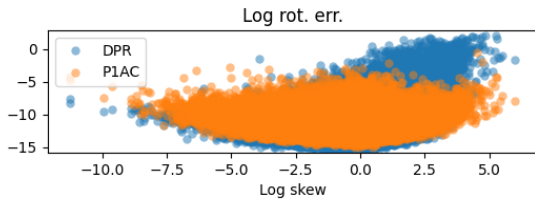


Figure 3: Rotation accuracy versus skew in the affine transformation, over 1M random problem instances. The accuracy of the DPR solver degrades as perspective distortion increases, whereas our P1AC solver is stable.

- **DPR**: Differential Pose Resection [29].
- **IPPE**: Infinitesimal Plane Pose Estimation [13].

In the experiments, we computed two pose accuracy metrics. Given the true camera rotation R and the estimated camera rotation \hat{R} we computed the *rotation error* as $\|\ln(\hat{R}R^T)\|$, where $\ln(\cdot)$ denotes the $SO(3)$ logarithm [23, 21]. While $\arccos((\text{trace}(R) - 1)/2)$ is often used as a rotation error metric, we instead use a robust implementation of $\|\ln(\hat{R}R^T)\|$, which we found to be more stable for small rotations. Given the true camera center \mathbf{c} and the estimated camera center $\hat{\mathbf{c}}$, we computed the *position error* as $\|\mathbf{c} - \hat{\mathbf{c}}\|$. In the synthetic experiments, when considering

multiple solutions produced by a method, we choose the solution that minimizes the max. of these two error metrics.

4.1. Synthetic data experiments

We generated synthetic problem instances with varying levels of noise added to the 2D point observations, affine transformations, and normal vectors.

We generated synthetic random problems in the following manner adapted from [15]. We randomly place two cameras around the origin at a distance sampled from $[1, 2]$, both oriented towards a random target point sampled from $[-0.5, 0.5]^3$. To generate a correspondence between the two cameras, we select a random 3D point from $\mathcal{N}(\mathbf{0}, I_{3 \times 3})$ with a random normal vector. We project the 3D point to the two cameras using a focal length of 400. We then calculate the affine transformation from the local homography between the first and second cameras [4]. We transform the cameras and 3D points so that the first camera (the reference camera) is at the origin with identity rotation.

We add Gaussian noise to the 2D point observations. Following prior studies [16, 15], we add Gaussian noise to the elements of the 2×2 affine transformation. We select the level of affine noise based on the chosen percentage error; *i.e.*, if the chosen percentage error is e , the std. dev. of the noise added to a_{ij} is $e \cdot a_{ij}$. Finally, we add noise to

	P3P	P1AC	DPR	IPPE
Time (μ s)	0.54	2.73	3.23	25.27

Table 1: Average timing in μ s over 10,000 trials.

the normal vectors by rotating them by a random rotation with angles randomly chosen from a Gaussian distribution. For each problem instance, we generate three correspondences between the two cameras, so that we can test the 1AC solvers (using the first correspondence) and P3P (using all three correspondences) on the same problem instance.

Numerical stability. To evaluate the numerical stability of each solver, we computed the rotation and position errors of each solver over 10,000 random problem instances with no added observational noise. Figure 2a shows the results. In the context of the minimal solver literature, the primary objective of such stability tests is *not* to facilitate comparisons between different solvers but to verify the individual stability of each solver, which is indicated by an absence of peaks above 10^{-2} to 10^{-1} . It is important to note that $> 99.9\%$ of the errors for P3P and P1AC are located below 10^{-5} . Practically speaking, this suggests these solvers would yield usable results. IPPE has substantially worse performance, as expected due to the unavoidable error in the virtual correspondences used to calculate IPPE.

To further investigate the difference between P1AC and DPR, we analyzed the performance of each solver with respect to the amount of perspective distortion between the reference and query images. To quantify the distortion, we applied a QR decomposition of the affine matrix $QR = A$ such that Q is a 2D rotation matrix and R is upper-triangular. The value at R_{12} represents the amount of skew in the transformation, and indicates the level of perspective distortion. Figure 3 shows a scatter plot of rotation accuracy versus skew for both solvers over 1M random problem instances. DPR has a noticeable increase in error as skew increases, whereas P1AC is unaffected. This increase in error is caused by the approximate warping of the AC required for the DPR method.

Timings. We computed the average processing times of the solvers over 10,000 random problem instances. The solvers were all implemented in C++ and timings were made on an Apple M1 Pro CPU. We used the implementations of Lambda Twist and 3Q3 provided in PoseLib [31] and the implementation of IPPE provided in OpenCV’s solvePnP function. The results are given in Table 1.

Among the 1AC variants, P1AC is the fastest with an average timing of 2.73 μ s. The Lambda Twist P3P solver is about $6\times$ faster than P1AC. Note that this implementation of the P3P solver is based on a careful elimination of the rotation and translation parameters to reveal the underlying elliptic equations in the P3P problem which are solved

through diagonalization.

Noise. We tested the P1AC solver and the P3P solver over a range of noise levels in the 2D point observations, affine transformations, and surface normals. We tested 1,000 random problem instances for each noise level. Figure 2b plots the median errors for each method at each setting. The chosen noise ranges are the same as used in [15]. The left column presents the results for increasing levels of point and affine noise, with the normal noise set to 1 degree. The center column presents the results for increasing levels of point and normal vector noise, with the affine noise set to 4% [15]. The right column presents the results for the P3P solver, for comparison. The P3P solver does not use the affine transformation or normal vector and thus is unaffected by affine and normal noise.

The synthetic experiments show that the P1AC solver is affected by affine and normal noise much more than by 2D point observation noise. In the presence of affine and normal noise, it returns larger errors than the P3P solver, which is not affected by these noises. Nevertheless, for noise levels that correspond to realistic values, *i.e.*, 2-4% for affine noise and 1° for normal noise, the P1AC solver still returns reasonably small errors that are sufficient for the initialization of local optimization [32] inside RANSAC.

Robust estimation. We investigated the performance of P1AC and P3P when used in robust estimation with an increasing outlier ratio, and confirmed that within a robust estimation loop, P1AC outperforms P3P in terms of speed.

We tested a range of outlier ratios from 0 to 0.9, with 10,000 trials at each setting. In each trial we randomly generate 1,000 correspondences with a point noise set at 1 pixel, affine noise at 4%, and normal noise at 1° . We replace a portion of the correspondences with random values according to the desired outlier ratio. We then run vanilla RANSAC [17] or LO-RANSAC [32] using either P1AC or P3P as the minimal solver and EPnP [33] as the non-minimal solver in local optimization. We tested vanilla RANSAC to analyze the trade-off in our approach between having a smaller minimal sample size but greater noise sensitivity. LO-RANSAC mitigates noise sensitivity in the minimal solver by incorporating local optimization (LO) iterations to grow the inlier set from the initial estimate.

The average error was stable for each method across the range of outlier ratios. Table 2 shows the average error for each method across all outlier ratios. When employing vanilla RANSAC, P1AC is less accurate than P3P at this level of noise. When employing LO-RANSAC, while the error for P1AC is marginally higher than that for P3P, the error rate for both methods remains extremely low, attesting to their effectiveness. Note that on real data, P1AC substantially outperforms P3P (see Section 4.2).

Figure 4 plots the average timing for each method. When

Robust estimator	P1AC		P3P	
	Rot. ($^{\circ}$)	Pos.	Rot. ($^{\circ}$)	Pos.
RANSAC	0.450	0.019	0.056	0.002
LO-RANSAC	0.085	0.003	0.019	0.001

Table 2: Average rotation and pose error for each robust estimation method on synthetic data over a range of outlier ratios.

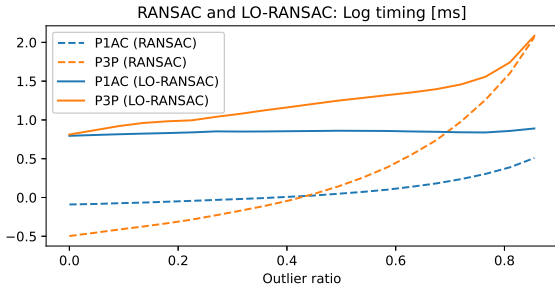


Figure 4: Average timing for each robust estimation method across a range of outlier ratios.

using vanilla RANSAC, P1AC is faster than P3P past an outlier ratio of about 0.45; with LO-RANSAC, P1AC is faster across all outlier ratios. This result indicates that our P1AC solver is accurate enough to provide a good initialization for LO and reach convergence more quickly than P3P, even with a low outlier ratio. In our real data experiments (Section 4.2), to be consistent with the state-of-the-art, we use Graph Cut RANSAC (GC-RANSAC) [5], which improves upon LO-RANSAC by incorporating graph cut optimization into the LO iterations. We did not use GC-RANSAC in our synthetic data experiments because the observations, being randomly generated, did not exhibit spatial coherence.

Mirrored solutions and degeneracies. The solutions returned by our P1AC solver include mirrored solutions, meaning pairs of solutions where the cameras are rotated exactly 180 degrees from each other. Previous work on plane-based pose estimation has also mentioned the existence of mirrored solutions [13]; they are indeed common for minimal problems and are an artifact of not including inequalities in the constraints. In practice they are not a problem since such solutions are filtered out inside RANSAC.

In addition to the 180° degeneracy inherent to the Cayley rotation representation, we noticed a degeneracy for rotations very close to identity when using the PoseLib 3Q3 solver implementation [31]. However, when using a 3Q3 solver produced by the GAPS automatic generator [34], this degeneracy for near-identity rotations does not appear. Another observation we made is that the translation accuracy of both P1AC and P3P degrades in weak-perspective situa-

tions, where the plane is very far from the camera. These aspects are discussed further in the supplemental material.

4.2. Real data experiments

To evaluate the performance of our method for large-scale image-based localization, we used the Cambridge Landmarks [26] and Aachen Day-Night v1.1 [47, 46, 55] benchmark datasets. Both datasets are commonly used in the image-based localization literature.

The Cambridge Landmarks dataset consists of six scenes, each recorded through multiple video sequences taken by a mobile phone, depicting parts of Cambridge, UK. Out of the sequences recorded for each dataset, some are used to obtain database images that represent the scene while the others are used to obtain query images. Ground truth poses and intrinsic camera calibrations for all images were obtained using the VisualSfM [52, 53] Structure-from-Motion (SfM) software. Image-based localization performance is typically measured by reporting the median position and orientation error. In addition, we measure the percentage of images localized within $10\text{cm}/1^{\circ}$ and $20\text{cm}/1^{\circ}$ of their ground truth poses.

Cambridge Landmarks is a relatively easy dataset since each scene is rather small and the query images were taken around the same time as the database images. A much more challenging case is the Aachen Day-Night dataset, which depicts the historical inner city of Aachen, Germany. The database images were all taken during daytime conditions and the dataset provides query images taken at nighttime conditions over a longer period of time. Ground truth for the database images were obtained via COLMAP [48]. The nighttime queries were later registered by refining initial pose estimates [55]. We follow the common evaluation protocol and report the percentage of images localized within three error thresholds ($0.25\text{m}/2^{\circ}$, $0.5\text{m}/5^{\circ}$, $5.0\text{m}/10^{\circ}$).

There are multiple ways to obtain affine features from real images. The most standard is to use a local feature detector, like DoG [35] or Key.Net [7], estimate keypoint locations and scales, and use the patch-based AffNet [39] to get affine shapes. Finally, a patch-based descriptor, like HardNet [38] or SOSNet [49], is applied. In our experiments, we combine DoG for feature detection, AffNet for affine shape recovery, and HardNet for patch description, an approach that is among the leaders in the IMC 2020 benchmark [24]. Given a SfM point cloud, we apply the normal estimator implemented in MeshLab to obtain an oriented point cloud. The normals are fitted to the 200 nearest neighbors.

Image localization procedure. For both datasets, we use image retrieval to identify a small set of potentially relevant database images for each query photo. We use DenseVLAD [50] descriptors for Cambridge Landmarks and NetVLAD [2] descriptors for Aachen Day-Night. We then establish 2D-2D correspondences between each query

Scene	Position (cm) ↓				Rotation (°) ↓				Recall (0.1m/1°) ↑				Recall (0.2m/1°) ↑			
	P3P	DPR	IPPE	P1AC	P3P	DPR	IPPE	P1AC	P3P	DPR	IPPE	P1AC	P3P	DPR	IPPE	P1AC
Great Court	8	6	7	6	0.1	0.08	0.1	0.08	61.8	74	65.3	73.6	86.4	90.1	86.2	91.1
King’s College	10	8	9	7	0.16	0.14	0.2	0.14	49.3	65.3	57.7	64.7	79.3	84.9	85.4	86.9
Old Hospital	9	7	8	7	0.15	0.12	0.14	0.11	58.8	68.9	62.6	72.5	79.7	89.6	85.2	93.4
Shop Façade	4	3	4	3	0.13	0.15	0.14	0.13	85.4	86.4	84.5	87.4	94.2	93.2	92.2	95.1
St Mary’s Church	7	5	6	5	0.21	0.17	0.2	0.17	61.7	70.1	67.6	72.5	80.9	85.2	84.3	87.7
Street	32	50	248	20	0.57	0.8	0.25	0.35	10.7	16.2	19.1	19.8	18	24	26.7	29.7
Avg.	12	13	47	8	0.22	0.24	0.16	0.16	54.62	63.48	59.5	65.08	73.08	77.83	76.7	80.65
Weighted avg.	22	33	153	14	0.40	0.53	0.21	0.26	30.44	38.13	37.4	40.62	43.91	49.33	50.1	53.53

Table 3: **Cambridge Landmarks** [26] median position (centimeters) and rotation (degrees) errors, and recalls (percentages), at 0.1m/1° and 0.2m/1°, of GC-RANSAC [5] combined with various solvers. The average over all scenes and average weighted by number of images in each scene are in the last row.

	Recall (0.25m/2°) ↑			
	P3P	DPR	IPPE	P1AC
Day	62.0	62.4	62.0	62.0
Night	47.1	47.6	46.1	51.3
	Recall (0.5m/5°) ↑			
	Day	83.4	84.3	83.4
Night	60.2	64.9	63.9	66.0
	Recall (5m/10°) ↑			
	Day	96.0	95.6	96.0
Night	74.3	75.9	76.4	82.2

Table 4: **Aachen Day-Night** pose error recalls [47], in percentages, at 0.25m/2°, 0.5m/5°, and 5.0m/10° of GC-RANSAC [5] combined with various solvers.

and its retrieved database images. These 2D-2D matches are then lifted to 2D-3D correspondences. For Cambridge Landmarks, we use the 3D points in SfM point clouds corresponding to the features extracted from the database images. For Aachen, we use a dense depth maps obtained by rendering a mesh model of the scene to obtain the 3D points (and their corresponding normals) for the database features [43].

In these experiments we aim to analyze the usefulness of the P1AC solver in practice. To this end we use advanced RANSAC techniques and tune them per solver, to achieve the best performance possible for each method.

To estimate the pose of each query image, we iterate through the retrieved database images. Between each pair, we estimate the pose by GC-RANSAC [5] using the implementation provided by the authors. The final pose for a query image comes from the database image with which the highest inlier number is achieved. GC-RANSAC requires two types of solvers, one for minimal and one for non-minimal pose estimation. We test the proposed P1AC solver as the minimal solver and compare it to the P3P solver [44]. The non-minimal solver, running both in the local optimization and in the final refitting on all inliers, is always

EPnP [33] followed by a few iterations of the Levenberg-Marquardt optimization [40] minimizing the reprojection error. The proposed P1AC solver enables running GC-RANSAC exhaustively, without severely affecting the runtime. Therefore, we perform an exhaustive search over all correspondences without using a random sampler. For P3P solver, we use the PROSAC sampler [11] on matches ordered by SNN ratio.

For Cambridge Landmarks, we set the inlier-outlier threshold to 4 pixels. For Aachen Day-Night, we set it to 12 pixels. We tuned the SNN ratio threshold for both P3P and P1AC. For P3P, 0.9 works the best. Surprisingly, P1AC and other 1AC solvers are the most accurate on Cambridge Landmarks without any SNN ratio filtering. Even though this leads to slightly increased run-times due to running on much more points, it increases both the accuracy and inlier number significantly.

Pose accuracy. A summary of errors on the Cambridge Landmarks dataset is given in Table 3. We report the median position (in meters) and rotation errors (in degrees), and the recall (in percentages) at 0.1m/1° and 0.2m/1°. The proposed P1AC method has lower errors and thus, higher recall on *all* scenes. In many cases, the recall increase is significant compared to the P3P solver. For example, on scene Street (the most challenging scene in the dataset) both recall values increase by around 10 percentage points (pp). The average 0.1m/1° over all scenes is increased by 10.5pp and the average 0.2m/1° by 7.6pp. Compared to DPR and IPPE, the proposed P1AC solver increases the recall by 1-3pp on all scenes, while significantly reducing the median position error.

The recalls at 0.25m/2°, 0.5m/5°, and 5.0m/10° on the Aachen Day-Night [47] dataset are in Table 4. In this case, all methods use an SNN ratio threshold set to 0.9 since increasing it led to a deterioration in the accuracy. This is likely caused by the dataset being more challenging than the Cambridge Landmarks. While the accuracy difference is smaller than in Table 3, the proposed P1AC consistently

Scene	Time (secs) ↓			
	P3P	DPR	IPPE	P1AC
Great Court	0.14	0.17	0.21	0.26
King’s College	0.19	0.58	0.37	0.48
Old Hospital	0.12	0.59	0.39	0.42
Shop Facade	0.13	0.52	0.36	0.37
St Mary’s Church	0.12	0.45	0.30	0.38
Street	0.10	0.37	0.28	0.24
Day	0.24	0.18	0.17	0.09
Night	0.12	0.05	0.07	0.09

Table 5: Avg. times of GC-RANSAC [5] on the Cambridge Landmarks [26] and Aachen Day-Night [47] datasets.

improves all metrics on the Night sequence and leads to similar accuracy to all other methods in the Day sequence.

Timing. Table 5 reports the average processing times (in seconds) of all compared methods. On Cambridge Landmarks, as expected from the excluded SNN ratio filtering and the exhaustive search used for 1AC methods, their runtime is larger than that of P3P. On Aachen Day-Night, all methods use the same SNN ratio threshold and the proposed P1AC is faster on both scenes than P3P. P1AC is the fastest on average on the Day sequence, and runs at a comparable speed to other methods in the Night sequence.

In our experiments, we used the same set of matches (DoG+HardNet+AffNet) for all methods, with the P3P method disregarding the affine components. As such, no method had a timing advantage related to feature extraction. However, for completeness, we measured feature extraction time and found that DoG+HardNet needs 0.18 ms per feature, and DoG+HardNet+AffNet 0.53 ms. Note that AffNet plays a crucial role in maintaining high accuracy and is not solely used for generating affine features [24].

4.3. Discussion

In terms of the timing of the minimal solver by itself, our P1AC solver is slower than the P3P solver. However, our P1AC solver is still quite fast ($< 3\mu\text{s}$) and moreover the minimal solver computation time is usually negligible when compared to other steps in the random sampling process such as local optimization and inlier testing.

While our P1AC solver itself is slightly slower than P3P, and is susceptible to more types of noise, it has a number of beneficial properties due to estimating the pose from a single AC, such as deterministic and fast robust estimation. On real data, we show that, when employed inside a robust estimator, our P1AC solver produces more accurate results and also is faster in some cases (e.g., Aachen Day-Night).

In comparison to alternative baseline methods, DPR and IPPE, our P1AC method successfully localizes more frames

and is more accurate. This indicates that our proposed approach, based on differential analysis of the AC, more accurately models the constraint imposed by an AC on the absolute pose of the query camera.

Improving the AC quality [6] might lead to a speed up in sample consensus because the initial pose estimates by the P1AC solver would be more accurate, so the system would spend less time on LO steps. Other minimal solvers that use a single correspondence often use histogram or kernel density voting in place of RANSAC, although they typically estimate a single parameter such as the focal length [10, 19]. Future work lies in analyzing how ACs and the P1AC constraints could be applied to 6DOF camera pose voting [54, 1]. While we only use point reprojection error for inlier testing due to the inherent noise in the affine components [6], future work could include testing first-order compatibility.

5. Conclusions

In this work, we derived novel constraints imposed by an affine correspondence on the absolute pose of a camera, given knowledge of a 3D point in the scene, the surface normal at that point, and the pose of the reference image. Using these novel constraints we developed a minimal solution for absolute pose from a single AC. Through experiments on synthetic data we established the numerical stability of our solver and its behavior under increasing levels of noise. We demonstrated the usefulness of our solver in an image-based localization system. In our experiments, replacing P3P with our P1AC solver improved localization accuracy and recall.

Acknowledgements. This work was supported by NSF Award No. 2144822, EU Horizon 2020 project RICAIP (No. 857306), the European Regional Development Fund under project IMPACT (No. CZ.02.1.01/0.0/0.0/15.003/0000468), and the Czech Science Foundation (GAČR) JUNIOR STAR Grant No. 22-23183M. D. Barath was supported by the ETH Postdoc Fellowship.

References

- [1] Dror Aiger, Simon Lymen, Jan Hosang, and Bernhard Zeisl. Efficient large scale inlier voting for geometric vision problems. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3243–3251, October 2021.
- [2] Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pa-jdla, and Josef Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5297–5307, 2016.
- [3] Simon Baker and Iain Matthews. Lucas-Kanade 20 years on: A unifying framework. *International Journal of Computer Vision*, 56(3):221–255, 2004.
- [4] Daniel Barath and Levente Hajder. Novel ways to estimate homography from local affine transformations. In *Proceed-*

- ings of the Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP), pages 434–445, 2016.
- [5] Daniel Barath and Jiri Matas. Graph-Cut RANSAC: local optimization on spatially coherent structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4961–4974, 2021.
 - [6] Daniel Barath, Michal Polic, Wolfgang Förstner, Torsten Sattler, Tomas Pajdla, and Zuzana Kukelova. Making affine correspondences work in camera geometry computation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 723–740, 2020.
 - [7] Axel Barroso-Laguna, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Key.Net: Keypoint Detection by Handcrafted and Learned CNN Filters. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5836–5844, 2019.
 - [8] Adam Baumberg. Reliable feature matching across widely separated views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 774–781, 2000.
 - [9] Jacob Bentolila and Joseph M Francos. Conic epipolar constraints from affine correspondences. *Computer Vision and Image Understanding*, 122:105–114, 2014.
 - [10] Martin Bujnak, Zuzana Kukelova, and Tomas Pajdla. Robust focal length estimation by voting in multi-view scene reconstruction. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 13–24, 2009.
 - [11] Ondrej Chum and Jiri Matas. Matching with PROSAC-progressive sample consensus. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 220–226, 2005.
 - [12] Ondřej Chum, Jiří Matas, and Josef Kittler. Locally optimized ransac. In *Proceedings of the DAGM Symposium on Pattern Recognition*, pages 236–243, 2003.
 - [13] Toby Collins and Adrien Bartoli. Infinitesimal plane-based pose estimation. *International Journal of Computer Vision*, 109(3):252–286, 2014.
 - [14] Yaqing Ding, Jian Yang, Viktor Larsson, Carl Olsson, and Kalle Åström. Revisiting the P3P problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4872–4880, June 2023.
 - [15] Ivan Eichhardt and Daniel Barath. Relative pose from deep learned depth and a single affine correspondence. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 627–644, 2020.
 - [16] Iván Eichhardt and Dmitry Chetverikov. Affine correspondences between central cameras for rapid relative pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 482–497, 2018.
 - [17] Martin A Fischler and Robert C Bolles. 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.
 - [18] J Grunert. Das Pothenotische Problem in erweiterter Gestalt nebst Über seine Anwendungen in der Geodäsie. *Grunerts Archiv für Mathematik und Physik*, 1:238–248, 1841.
 - [19] Levente Hajder and Daniel Barath. Relative planar motion for vehicle-mounted cameras from a single affine correspondence. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 8651–8657, 2020.
 - [20] Christian Häne, Lionel Heng, Gim Hee Lee, Friedrich Fraundorfer, Paul Furgale, Torsten Sattler, and Marc Pollefeys. 3D visual perception for self-driving cars using a multi-camera system: Calibration, mapping, localization, and obstacle detection. *Image and Vision Computing*, 68:14–27, 2017.
 - [21] Richard Hartley, Jochen Trumpf, Yuchao Dai, and Hongdong Li. Rotation averaging. *International Journal of Computer Vision*, 103(3):267–305, 2013.
 - [22] Florian Haug and Bernd Jähne. 6 DoF appearance-based object localization with local covariant features. *Forum Bildverarbeitung*, pages 13–24, 2016.
 - [23] Du Q Huynh. Metrics for 3D rotations: Comparison and analysis. *Journal of Mathematical Imaging and Vision*, 35(2):155–164, 2009.
 - [24] Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiri Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Image matching across wide baselines: From paper to practice. *International Journal of Computer Vision*, 129(2):517–547, 2021.
 - [25] Tong Ke and Stergios I Roumeliotis. An efficient algebraic solution to the perspective-three-point problem. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7225–7233, 2017.
 - [26] Alex Kendall, Matthew Grimes, and Roberto Cipolla. PoseNet: A convolutional network for real-time 6-DOF camera relocalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2938–2946, 2015.
 - [27] Laurent Kneip, Davide Scaramuzza, and Roland Siegwart. A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2969–2976, 2011.
 - [28] Kevin Köser. *Geometric estimation with local affine frames and free-form surfaces*. PhD thesis, University of Kiel, 2009.
 - [29] Kevin Köser and Reinhard Koch. Differential spatial resection-pose estimation using a single local image feature. In *Proceedings of the European Conference on Computer Vision*, pages 312–325, 2008.
 - [30] Zuzana Kukelova, Jan Heller, and Andrew Fitzgibbon. Efficient intersection of three quadrics and applications in computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1799–1808, 2016.
 - [31] Viktor Larsson. PoseLib - Minimal Solvers for Camera Pose Estimation. <https://github.com/vlarsson/PoseLib>, 2020.
 - [32] Karel Lebeda, Jiri Matas, and Ondrej Chum. Fixing the locally optimized RANSAC. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2012.

- [33] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. EPnP: An accurate $O(n)$ solution to the PnP problem. *International Journal of Computer Vision*, 81(2):155, 2009.
- [34] Bo Li and Viktor Larsson. GAPS: Generator for automatic polynomial solvers. *arXiv preprint arXiv:2004.11765*, 2020.
- [35] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
- [36] Bruce D Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. *Proceedings of the DARPA Image Understanding Workshop*, 1981.
- [37] Krystian Mikolajczyk and Cordelia Schmid. Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.
- [38] A. Mishchuk, D. Mishkin, F. Radenovic, and J. Matas. Working hard to know your neighbor’s margins: Local descriptor learning loss. In *Advances in Neural Information Processing Systems*, volume 30, 2017.
- [39] Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Repeatability is not enough: Learning affine regions via discriminability. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 284–300, 2018.
- [40] Jorge J Moré. The Levenberg-Marquardt algorithm: implementation and theory. In *Numerical analysis*, pages 105–116. Springer, 1978.
- [41] Raul Mur-Artal and Juan D Tardós. ORB-SLAM2: An open-source slam system for monocular, stereo, and RGB-D cameras. *IEEE Transactions on Robotics*, 33(5):1255–1262, 2017.
- [42] David Nistér. An efficient solution to the five-point relative pose problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(6):756–770, 2004.
- [43] Vojtech Panek, Zuzana Kukelova, and Torsten Sattler. Meshloc: Mesh-based visual localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 589–609, 2022.
- [44] Mikael Persson and Klas Nordberg. Lambda Twist: An accurate fast robust perspective three point (P3P) solver. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 318–332, 2018.
- [45] Carolina Raposo and Joao P Barreto. Theory and practice of structure-from-motion using affine correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5470–5478, 2016.
- [46] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, Fredrik Kahl, and Tomas Pajdla. Benchmarking 6DOF urban visual localization in changing conditions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8601–8610, 2018.
- [47] Torsten Sattler, Tobias Weyand, Bastian Leibe, and Leif Kobbelt. Image retrieval for image-based localization revisited. In *Proceedings of the British Machine Vision Conference (BMVC)*, volume 1, page 4, 2012.
- [48] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [49] Yurun Tian, Xin Yu, Bin Fan, Fuchao Wu, Huub Heijnen, and Vassileios Balntas. SOSNet: Second order similarity regularization for local descriptor learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [50] Akihiko Torii, Relja Arandjelović, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 24/7 Place Recognition by View Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [51] Jonathan Ventura, Clemens Arth, Gerhard Reitmayr, and Dieter Schmalstieg. Global localization from monocular slam on a mobile phone. *IEEE Transactions on Visualization and Computer Graphics*, 20(4):531–539, 2014.
- [52] Changchang Wu. Towards linear-time incremental structure from motion. In *Proceedings of the IEEE International Conference on 3D Vision (3DV)*, pages 127–134, 2013.
- [53] Changchang Wu, Sameer Agarwal, Brian Curless, and Steven M Seitz. Multicore bundle adjustment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3057–3064, 2011.
- [54] Bernhard Zeisl, Torsten Sattler, and Marc Pollefeys. Camera pose voting for large-scale image-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.
- [55] Zichao Zhang, Torsten Sattler, and Davide Scaramuzza. Reference pose generation for long-term visual localization via learned features and view synthesis. *International Journal of Computer Vision*, 129:821–844, 2021.