

Relative Pose from Deep Learned Depth and a Single Affine Correspondence

Ivan Eichhardt^{1,3} and Daniel Barath^{1,2}

¹ Machine Perception Research Laboratory, SZTAKI, Budapest, Hungary

² VRG, Faculty of Electrical Engineering, Czech Technical University in Prague

³ Faculty of Informatics, University of Debrecen, Debrecen, Hungary


{eichhardt.ivan, barath.daniel}@sztaki.mta.hu

Abstract. We propose a new approach for combining deep-learned non-metric monocular depth with affine correspondences (ACs) to estimate the relative pose of two calibrated cameras from a single correspondence. Considering the depth information and affine features, two new constraints on the camera pose are derived. The proposed solver is usable within 1-point RANSAC approaches. Thus, the processing time of the robust estimation is linear in the number of correspondences and, therefore, orders of magnitude faster than by using traditional approaches. The proposed 1AC+D⁴ solver is tested both on synthetic data and on 110 395 publicly available real image pairs where we used an off-the-shelf monocular depth network to provide up-to-scale depth per pixel. The proposed 1AC+D leads to similar accuracy as traditional approaches while being significantly faster. When solving large-scale problems, *e.g.* pose-graph initialization for Structure-from-Motion (SfM) pipelines, the overhead of obtaining ACs and monocular depth is negligible compared to the speed-up gained in the pairwise geometric verification, *i.e.*, relative pose estimation. This is demonstrated on scenes from the 1DSfM dataset using a state-of-the-art global SfM algorithm.

Keywords: pose estimation, minimal solver, depth prediction, affine correspondences, global structure from motion, pose graph initialization

1 Introduction

This paper investigates the challenges and viability of combining deep-learned monocular depth with affine correspondences (ACs) to create minimal pose solvers. Estimating pose is a fundamental problem in computer vision [37, 49, 48, 8, 17, 22, 21, 55, 42, 36, 1, 25–29, 43], enabling reconstruction algorithms such as simultaneous localization and mapping [34, 35] (SLAM) as well as structure-from-motion [57, 45, 51, 44] (SfM). Also, there are several other applications, *e.g.*, in 6D object pose estimation [23, 31] or in robust model fitting [15, 52, 12, 40, 2, 7], where the accuracy and runtime highly depends on the minimal solver applied. The proposed approach uses monocular depth predictions together with ACs to estimate the relative camera pose from a single correspondence.

⁴ Source code:  <https://github.com/eivan/one-ac-pose>

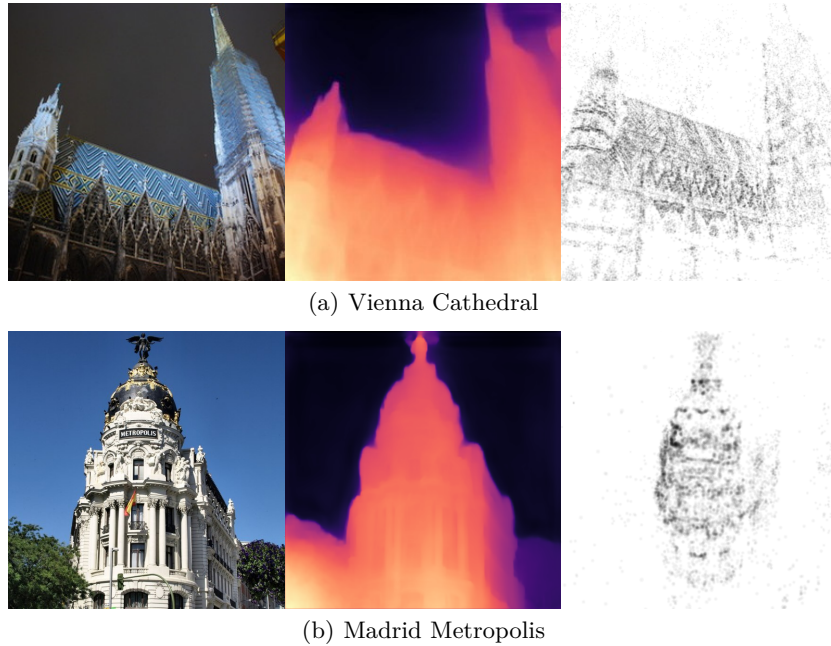


Fig. 1. Scenes from the 1DSfM dataset [56]. From left to right: color image from the dataset, predicted monocular depth [30], and 3D reconstruction using global SfM [50]. Pose graphs were initialized by the proposed solver, 1AC+D, as written in Section 4.3.

Exploiting ACs for geometric model estimation is a nowadays popular approach with a number of solvers proposed in the recent years. For instance, for estimating the epipolar geometry of two views, Bentolila and Francos [10] proposed a method using three correspondences interpreting the problem via conic constraints. Perdoch *et al.* [38] proposed two techniques for approximating the pose using two and three correspondences by converting each affine feature to three points and applying the standard point-based estimation techniques. Raposo *et al.* [41] proposed a solution for essential matrix estimation from two correspondences. Baráth *et al.* [3, 6] showed that the relationship of ACs and epipolar geometry is linear and geometrically interpretable. Eichhardt *et al.* [14] proposed a method that uses two ACs for relative pose estimation based on general central-projective views. Recently, Hajder *et al.* [19] and Guan *et al.* [18] proposed minimal solutions for relative pose from a single AC when the camera is mounted on a moving vehicle. Homographies can also be estimated from two ACs as it was first shown by Köser [24]. In the case of known epipolar geometry, a single correspondence [5] is enough to recover the parameters of the homography, or the underlying surface normal [4, 24]. Pritts *et al.* [39] used affine features for the simultaneous estimation of affine image rectification and lens distortion.

In this paper, we propose to use affine correspondences together with deep-learned priors derived directly into minimal pose solvers – exploring new ways of

utilizing these priors for pose estimation. We use an off-the-shelf deep monocular depth estimator [30] to provide a ‘relative’ (non-metric) depth for each pixel (examples are in Fig. 1) and use the depth together with affine correspondences for estimating the relative pose from a single correspondence. Aside from the fact that predicted depth values are far from being perfect, they provide a sufficiently strong prior about the underlying scene geometry. This helps in reducing the degrees-of-freedom of the relative pose estimation problem. Consequently, fewer correspondences are needed to determine the pose which is highly favorable in robust estimation, *i.e.*, robustly determining the camera motion when the data is contaminated by outliers and noise.

We propose a new minimal solver which estimates the general relative camera pose, *i.e.*, 3D rotation and translation, and the scale of the depth map from a single AC. We show that it is possible to use the predicted imperfect depth signal to robustly estimate the pose parameters. The depth and the AC together constrain the camera geometry so that the relative motion and scale can be determined as the closed-form solution of the implied least-squares optimization problem. The proposed new constraints are derived from general central projection and, therefore, they are valid for arbitrary pairs of central-projective views. It is shown that the proposed solver has significantly lower computational cost, compared to state-of-the-art pose solvers. The imperfections of the depth signal and the AC are alleviated through using the solver with a modern robust estimator [2], providing state-of-the-art accuracy at exceptionally low run-time.

The reduced number of data points required for the model estimation leads to linear time complexity when combined with robust estimators, *e.g.* RANSAC [15]. The resulting 1-point RANSAC has to check only n model hypotheses (where n is the point number) instead of, *e.g.*, $\binom{n}{5}$ which the five-point solver implies. This improvement in efficiency has a significant positive impact on solving large-scale problems, *e.g.*, on view-graph construction (*i.e.*, a number of 2-view matching and geometric verification) [45, 46, 51] which operates over thousands of image pairs. We provide an evaluation of the proposed solver both on synthetic and real-world data. It is demonstrated, on 110 395 image pairs from the 1DSFM dataset [56], that the proposed methods are similarly robust to image noise while being up to 2 orders of magnitude faster, when applied within Graph-Cut RANSAC [2], than the traditional five-point algorithm. Also, it is demonstrated that when using the resulting pose-graph in the global SfM pipeline [11, 56] as implemented in the Theia library [50], the accuracy of the obtained reconstruction is similar to when the five-point algorithm is used.

2 Constraints from Relative Depths and Affine Frames

Let us denote a local affine frame (LAF) as (\mathbf{x}, \mathbf{M}) , where $\mathbf{x} \in \mathbb{R}^2$ is an image point and $\mathbf{M} \in \mathbb{R}^{2 \times 2}$ is a linear transformation describing the local coordinate system of the associated image region. The following expression maps points \mathbf{y} from a local coordinate system onto the image plane, in the vicinity of \mathbf{x} .

$$\mathbf{x}(\mathbf{y}) \approx \mathbf{x} + \mathbf{M}\mathbf{y}. \quad (1)$$

LAFs typically are acquired from images by affine-covariant feature extractors [33]. In practice, they can easily be obtained by, *e.g.*, the VLFeat library [54].

Two LAFs $(\mathbf{x}_1, \mathbf{M}_1)$ and $(\mathbf{x}_2, \mathbf{M}_2)$ form an affine correspondence, namely, \mathbf{x}_1 and \mathbf{x}_2 are corresponding points and $\mathbf{M}_2\mathbf{M}_1^{-1}$ is the linear transformation mapping points between the infinitesimal vicinities of \mathbf{x}_1 and \mathbf{x}_2 [13].

2.1 Local Affine Frames and Depth through Central Projection

Let $q : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ be a function that maps image coordinates to bearing vectors. $\mathbf{R} \in \text{SO}(3)$ and $\mathbf{t} \in \mathbb{R}^3$ are the relative rotation and translation of the camera coordinate system, such that $\mathbf{x} \in \mathbb{R}^2$ is the projection of $\mathbf{X} \in \mathbb{R}^3$, as follows:

$$\exists! \lambda : \mathbf{X} = \lambda \mathbf{R} q(\mathbf{x}) + \mathbf{t}. \quad (2)$$

That is, a unique *depth* λ corresponds to \mathbf{X} , along the bearing vector $q(\mathbf{x})$.

When image coordinates \mathbf{x} are locally perturbed, the corresponding \mathbf{X} and λ are also expected to change as per Eq. (2). The first order approximation of the projection expresses the connection between local changes of \mathbf{x} and \mathbf{X} , *i.e.*, it is an approximate linear description how local perturbations would relate the two. Differentiating Eq. (2) along image coordinates $k = u, v$ results in an expression for the first order approximation of projection, as follows:

$$\exists! \lambda, \partial_k \lambda : \partial_k \mathbf{X} = \mathbf{R} (\partial_k \lambda q(\mathbf{x}) + \lambda \partial_k q(\mathbf{x}) \partial_k \mathbf{x}). \quad (3)$$

Observe that \mathbf{t} is eliminated from the expression above. Note that, considering the k -th image coordinate, $\partial_k \mathbf{X}$ is the the Jacobian of the 3D point \mathbf{X} , and $\partial_k \lambda$ is the Jacobian of the depth λ . Using the differential operator $\nabla = [\partial_u, \partial_v]$ it is more convenient to use a single compact expression, that includes the local affine frame \mathbf{M} , as $\nabla \mathbf{x} = [\partial_u \mathbf{x}, \partial_v \mathbf{x}] = \mathbf{M}$.

$$\exists! \lambda, \partial_u \lambda, \partial_v \lambda : \nabla \mathbf{X} = \mathbf{R} (q(\mathbf{x}) \nabla \lambda + \lambda \nabla q(\mathbf{x}) \mathbf{M}). \quad (4)$$

All in all, Eqs. (2) and (4) together describe all constraints on the 3D point \mathbf{X} and on its vicinity $\nabla \mathbf{X}$, imposed by a LAF (\mathbf{x}, \mathbf{M}) .

2.2 Affine Correspondence and Depth Constraining Camera Pose

Assume two views observing \mathbf{X} under corresponding LAFs $(\mathbf{x}_1, \mathbf{M}_1)$ and $(\mathbf{x}_2, \mathbf{M}_2)$. Without the loss of generality, we define the coordinate system of the first view as the identity – thus putting it in the origin – while that of the second one is described by the relative rotation and translation, \mathbf{R} and \mathbf{t} , respectively. In this two-view system, Eqs. (2) and (4) lead to two new constraints on the camera pose, depth and local affine frames, as follows:

$$\underbrace{\lambda_1 q_1(\mathbf{x}_1)}_{\mathbf{a}} = \mathbf{R} \underbrace{\lambda_2 q_2(\mathbf{x}_2)}_{\mathbf{b}} + \mathbf{t}, \quad (5)$$

$$\underbrace{q_1(\mathbf{x}_1) \nabla \lambda_1 + \lambda_1 \nabla q_1(\mathbf{x}_1) \mathbf{M}_1}_{\mathbf{A}} = \mathbf{R} \underbrace{(q_2(\mathbf{x}_2) \nabla \lambda_2 + \lambda_2 \nabla q_2(\mathbf{x}_2) \mathbf{M}_2)}_{\mathbf{B}},$$

where the projection functions $q_1(\mathbf{x}_1)$ and $q_2(\mathbf{x}_2)$ assign bearing vectors to image points \mathbf{x}_1 and \mathbf{x}_2 of the first and second view, respectively. The first equation comes from Eq. (2) through the point correspondence, and the second one from Eq. (4) through the AC, by equating \mathbf{X} and $\nabla\mathbf{X}$, respectively. Note that depths λ_1, λ_2 and their Jacobians $\nabla\lambda_1, \nabla\lambda_2$ are intrinsic to each view. Also observe that 3D point \mathbf{X} and its Jacobian $\nabla\mathbf{X}$ are now eliminated from these constraints.

In the rest of the paper, we are resorting to the more compact notations using $\mathbf{a}, \mathbf{b}, \mathbf{A}$ and \mathbf{B} , as highlighted above, *i.e.*, the two lines of Eq. (5) take the simplified forms of $\mathbf{a} = \mathbf{R}\mathbf{b} + \mathbf{t}$ and $\mathbf{A} = \mathbf{R}\mathbf{B}$, respectively.

Relative depth. In this paper, as monocular views are used to provide relative depth predictions, λ_1 and λ_2 are only known up to a common scale Λ , so that Eq. (5), with the simplified notation, is modified as follows:

$$\mathbf{a} = \Lambda\mathbf{R}\mathbf{b} + \mathbf{t}, \quad \mathbf{A} = \Lambda\mathbf{R}\mathbf{B}. \quad (6)$$

These constraints describe the relationship of relative camera pose and ACs in the case of known relative depth.

3 Relative Pose and Scale from a Single Correspondence

Given a single affine correspondence, the optimal estimate for the relative pose and scale is given as the solution of the following optimization problem.

$$\min_{\mathbf{R}, \mathbf{t}, \Lambda} \underbrace{\frac{1}{2} \|\mathbf{a} - (\Lambda\mathbf{R}\mathbf{b} + \mathbf{t})\|_2^2 + \frac{1}{2} \|\mathbf{A} - \Lambda\mathbf{R}\mathbf{B}\|_F^2}_{f(\Lambda, \mathbf{R}, \mathbf{t})} \quad (7)$$

To solve the problem, the first order optimality conditions have to be investigated. At optimality, differentiating $f(\Lambda, \mathbf{R}, \mathbf{t})$ by \mathbf{t} gives:

$$\nabla_{\mathbf{t}} f(\Lambda, \mathbf{R}, \mathbf{t}) = \mathbf{a} - \Lambda\mathbf{R}\mathbf{b} - \mathbf{t} \stackrel{!}{=} \mathbf{0}, \quad (8)$$

that is, \mathbf{t} can only be determined once the rest of the unknowns, \mathbf{R} and Λ , are determined. This also means that above optimization problem can be set free of the translation, by performing the following substitution $\mathbf{t} \leftarrow \Lambda\mathbf{R}\mathbf{b} - \mathbf{a}$.

The optimization problem, where the translation is replaced as previously described, only contains the rotation and scale as unknowns, as follows:

$$\min_{\mathbf{R}, \Lambda} \underbrace{\frac{1}{2} \|\mathbf{A} - \Lambda\mathbf{R}\mathbf{B}\|_F^2}_{g(\mathbf{R}, \Lambda)}. \quad (9)$$

Below, different approaches for determining the unknown scale and rotation are described, solving Eq. (9) exactly or, in some manner, approximately.

1AC+D (Umeyama): An SVD solution. In the least-squares sense, the optimal rotation and scale satisfying Eq. (9) can be acquired by the singular value decomposition of the covariance matrix of \mathbf{A} and \mathbf{B} , as follows:

$$\mathbf{USV}^\top = \mathbf{AB}^\top. \quad (10)$$

Using these matrices, the optimal rotation and scale are expressed as

$$\mathbf{R} = \mathbf{UDV}^\top, \quad \Lambda = \frac{\text{tr}(\mathbf{SD})}{\text{tr}(\mathbf{B}^\top \mathbf{B})}, \quad (11)$$

where \mathbf{D} is a diagonal matrix with its diagonal being $[1, 1, \det(\mathbf{UV}^\top)]$, to constrain $\det \mathbf{R} = 1$. Given \mathbf{R} and Λ , the translation is expressed as $\mathbf{t} = \mathbf{a} - \Lambda \mathbf{R} \mathbf{b}$.

This approach was greatly motivated by Umeyama’s method [53], where a similar principle is used for solving point cloud alignment.

1AC+D (Proposed): Approximate relative orientation solution. Matrices \mathbf{A} and \mathbf{B} , together with their 1D left-nullspaces, define two non-orthogonal coordinate systems. By orthogonalizing the respective basis vectors, one can construct matrices $\mathbf{R}_\mathbf{A}$ and $\mathbf{R}_\mathbf{B}$, corresponding to \mathbf{A} and \mathbf{B} . The relative rotation between these two coordinate systems can be written as follows:

$$\mathbf{R} = \mathbf{R}_\mathbf{A}^\top \mathbf{R}_\mathbf{B}. \quad (12)$$

$\mathbf{R}_\mathbf{A}$ and $\mathbf{R}_\mathbf{B}$ can be determined, *e.g.*, using Gram-Schmidt orthonormalization [47]. For our special case, a faster approach is shown in Alg. 1.

Using this approach, the solution is biased towards the first columns of \mathbf{A} and \mathbf{B} , providing perfect alignment of those two axes. In our experiments, this was rather a favorable property than an issue. The first axis of a LAF represents the orientation of the feature while the second one specifies the affine shape, as long as the magnitude of $\nabla \lambda_i$ is negligible compared to λ_i , which is usually the case. The orientation of a LAF is usually more reliable than its shape.

As \mathbf{R} is now known and \mathbf{t} has been ruled out, Λ is to be determined. Differentiating $g(\mathbf{R}, \Lambda)$ by Λ gives:

$$\nabla_\Lambda g(\mathbf{R}, \Lambda) = \text{tr}(\mathbf{A}^\top \mathbf{R} \mathbf{B}) + \Lambda \text{tr}(\mathbf{B}^\top \mathbf{B}), \quad (13)$$

that is, once \mathbf{R} is known, Λ can be determined as follows:

$$\Lambda = -\frac{\text{tr}(\mathbf{A}^\top \mathbf{R} \mathbf{B})}{\text{tr}(\mathbf{B}^\top \mathbf{B})}. \quad (14)$$

Finally, the translation parameters are expressed as $\mathbf{t} = \mathbf{a} - \Lambda \mathbf{R} \mathbf{b}$.

The complete algorithm is shown in Alg. 1. Note that, although, we have tested various approaches to computing the relative rotation between the frames \mathbf{A} and \mathbf{B} , such as the above described SVD approach or the Gram-Schmidt process [47], the one introduced in Alg. 1 proved to be the fastest with no noticeable deterioration in accuracy.

Table 1. The theoretical computational complexity of the solvers used in RANSAC. The reported properties are: the number of operations of each solver (steps; 1st row); the computational complexity of one estimation (1 iter; 2nd); the number of correspondences required for the estimation (m ; 3rd); possible outlier ratios ($1 - \mu$; 4th); the number of iterations needed for RANSAC with the required confidence set to 0.99 (# iters; 5th); and computational complexity of the full procedure (# comps; 6th). The major operations are: singular value decomposition (SVD), eigenvalue decomposition (EIG), LU factorization (LU), and QR decomposition (QR).

	point-based	AC-based	AC+depth	
	5PT [48]	2AC [6]	1AC+D (Umeyama)	1AC+D (Proposed)
steps	5×9 SVD + 10×20 LU + 10×10 EIG	6×9 SVD + 10×9 QR	3×3 SVD + cov. + trans.	$4 \times$ (cross + norm) + trans. + etc.
1 iter	$5^2 * 9 + 10^2 * 20 + 10^2$ = 2325	$6^2 * 9 + 10^3 + 10^2$ = 1424	$11 * 3^3 + 27 + 21$ = 345	$72 + 21 + 40$ = 133
m	5	2	1	1
$1 - \mu$	0.50 0.75 0.90	0.50 0.75 0.90	0.50 0.75 0.90	0.50 0.75 0.90
# iters	145 4713 $\sim 10^6$	16 71 458	7 16 44	7 16 44
# comps	$\sim 10^5 \sim 10^7 \sim 10^9$	$\sim 10^4 \sim 10^5 \sim 10^5$	$\sim 10^3 \sim 10^3 \sim 10^4$	$\sim 10^2 \sim 10^3 \sim 10^3$

Algorithm 1 The 1AC+D (Proposed) algorithm for relative pose computation.

```

1: procedure 1AC+D (PROPOSED)(a, b, A, B)      ▷ Computes the relative camera
   pose.
2:   RA ← ORTHONORM(A)
3:   RB ← ORTHONORM(B)
4:   R ← RATRB                                     ▷ Applying Eq. (12).
5:    $\Lambda \leftarrow -\text{tr}(\mathbf{A}^T \mathbf{R} \mathbf{B}) / \text{tr}(\mathbf{B}^T \mathbf{B})$        ▷ Applying Eq. (14).
6:   t ← a −  $\Lambda \mathbf{R} \mathbf{b}$ 
7:   return R, t,  $\Lambda$ 
8: function ORTHONORM(Y)                               ▷ Quick orthonormalization of Y.
9:   rx ← normalize (Y(:,1) × Y(:,2))      ▷ rx is a normal to the underlying plane.
10:  rz ← normalize (Y(:,2))
11:  return [rx, rz × rx, rz]                ▷ Return a 3 × 3 rotation matrix.

```

4 Experimental results

In this section, experiments and complexity analyses compare the performance of the two versions of the proposed 1AC+D solver, *i.e.*, 1AC+D (Umeyama) and 1AC+D (Proposed), 2AC [6] and 5PT [48] methods for relative pose estimation.

4.1 Processing time

In Table 1, we compare the computational complexity of state-of-the-art pose solvers used in our evaluation. The first row consists of the major steps of each solver. For instance, 5×9 SVD + 10×20 EIG means that the major steps are: the SVD decomposition of a 5×9 matrix and eigendecomposition of a 10×20 matrix. In the second row, the implied computational complexities are summed. In the third one, the number of correspondences required for the solvers are written.

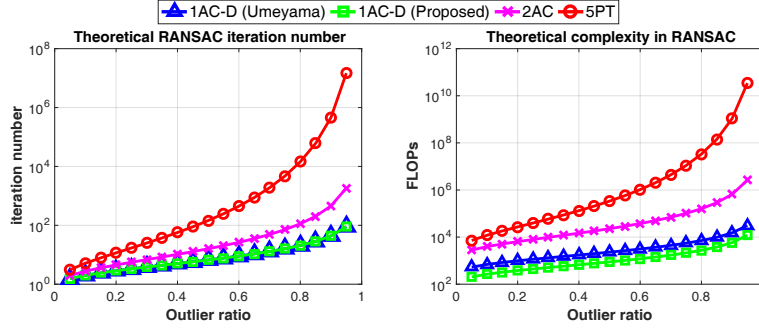


Fig. 2. *Left:* the theoretical number of RANSAC iterations – calculated as in [20]. *Right:* the number of floating point operations (horizontal axes) plotted as the function of the outlier ratio (vertical), displayed on logarithmic scales.

The fourth row lists example outlier ratios. In the fifth one, the theoretical numbers of iterations of RANSAC [20] are written with confidence set to 0.99. The last row reports the total complexity, *i.e.*, the complexity of one iteration multiplied by the number of RANSAC iterations.

The proposed methods have significantly smaller computational requirements than the five-point solver, 5PT, or the affine correspondence-based 2AC method [6] when included in RANSAC. Note that while the reported values for the 1AC+D methods are the actual FLOPs, for 5PT and 2AC, it is in practice about an order of magnitude higher due to the iterative manner of various linear algebra operations, *e.g.*, SVD. In Fig. 2, the theoretical number of RANSAC iterations – calculated as in [20] – and the number of floating point operations are plotted as the function of the outlier ratio (horizontal axes), displayed on logarithmic scales (vertical axes). The proposed solvers lead to fewer iterations compared to the solvers solely based on point or affine correspondences.

In summary of the above analysis of the computational complexity, the proposed approach speeds up the robust estimation procedure in three ways. First, it requires at least an order of magnitude fewer operations to complete a single iteration of RANSAC than by using the five-point algorithm. Second, it leads to significantly fewer iterations due to the fact that 1AC+D takes only a single correspondence to propose a solution to pose estimation. Third, 1AC+D returns a single solution from a minimal sample, in contrast to the five-point algorithm. Therefore, each RANSAC iteration requires the validation of a single model.

4.2 Synthetic evaluation

The synthetic evaluation was carried out in a setup consisting of two cameras with randomized poses represented by rotation $\mathbf{R}_i \in \text{SO}(3)$ and translation $\mathbf{t}_i \in \mathbb{R}^3$ ($i = 1, 2$). The cameras were placed around the origin at a distance sampled from $[1.0, 2.0]$, oriented towards a random point sampled from $[-0.5, 0.5]^3$. Both cameras had a common intrinsic camera matrix \mathbf{K} with focal length $f = 600$

and principal point $[300, 300]$, representing pinhole projection $\pi : \mathbb{R}^3 \rightarrow \mathbb{R}^2$. For generating LAFs $(\mathbf{x}_i, \mathbf{M}_i)$, depths λ_i and their derivatives $\nabla \lambda_i$, randomized 3D points $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{3 \times 3})$ with corresponding normals $\mathbf{n} \in \mathbb{R}^3$, $\|\mathbf{n}\|_2 = 1$ were projected into the two image planes using Eqs. (15) and (16), for $i = 1, 2$.

$$\mathbf{x}_i = \pi(\mathbf{R}_i \mathbf{X} + \mathbf{t}_i), \quad \mathbf{M}_i = \nabla \pi(\mathbf{R}_i \mathbf{X} + \mathbf{t}_i) \mathbf{R}_i \nabla \mathbf{X}, \quad (15)$$

$$\lambda_i = \mathbf{R}_i|_{(3,:)} \mathbf{X} + \mathbf{t}_i|_{(3)}, \quad \nabla \lambda_i = \mathbf{R}_i|_{(3,:)} \nabla \mathbf{X}, \quad (16)$$

where $\mathbf{R}_i|_{(3,:)}$ is the 3rd row of \mathbf{R}_i , $\nabla \mathbf{X} = \text{nulls}(\mathbf{n})$ simulates the local frame of the surface as the nullspace of the surface normal \mathbf{n} . Note that [14] proposed the approach for computing local frames \mathbf{M}_i , *i.e.*, Eq. (15). Their synthetic experiment are also similar in concept, but contrast to them, we also had to account for the depth in Eq. (16), and we used the nullspace of \mathbf{n} to express the Jacobian $\nabla \mathbf{X}$. The 2nd part of Eq. (16), the expression for the Jacobian of the depth, $\nabla \lambda_i$, was derived from the 1st part by differentiation. In this setup λ_i is the projective depth, a key element of perspective projection. Note that the depth and its derivatives are different for various other camera models. Finally, zero-mean Gaussian-noise was added to the coordinates/components of \mathbf{x}_i , \mathbf{M}_i , λ_i and $\nabla \lambda_i$, with σ , $\sigma_{\mathbf{M}}$, σ_{λ} and $\sigma_{\nabla \lambda}$ standard deviation (STD), respectively.

The reported errors for rotation and translation both were measured in degrees ($^\circ$) in this evaluation. The rotation error was calculated as follows: $\epsilon_{\mathbf{R}} = \cos^{-1} \left(\frac{1}{2} \left(\text{tr}(\hat{\mathbf{R}} \mathbf{R}^T) - 1 \right) \right)$, where $\hat{\mathbf{R}}$ is the measured and \mathbf{R} is the ground truth rotation matrix. Translation error $\epsilon_{\mathbf{t}}$ is the angular difference between the estimated and ground truth translations. For some of the tests, we also show the avg. Sampson error [20] of the implied essential matrix.

Solver stability. We randomly generated 30 000 instances of the above described synthetic setup, except for adding any noise to the LAFs and depths, in order to evaluate the stability of the various solver in a noise-free environment. Figs. 3(a), 3(b) and 3(c) show the distribution of the Sampson, rotation and translation errors on logarithmic scales. The figures show that the proposed method is one of the best performers given that its distribution of errors is shifted lower compared to the other methods. However, all methods are quite stable having no peaks on the right side of the figures.

Noise study. We added noise to the 30 000 instances of the synthetic setup. The estimated poses were evaluated to determine how sensitive the solvers are to the noise in the data, *i.e.*, image coordinates, depth, and LAFs. As expected, the noise in the depth or parameters of LAFs has no effect on the 5PT [48] solver. This can be seen in Fig. 5(b), where 5PT [48] is only affected by image noise. Increasing noise on either axes has a negative effect on the output of the proposed method as seen on Fig. 5(a). The effect of noise on rotation and translation estimated by 1AC+D and 5PT [48] are visualized on diagrams Fig. 4(a)-(f). The curves show that the effect of image noise – on useful scales, such as 2.5 px – in itself has a less significant effect on the degradation of rotation and translation estimates of the proposed method, compared to 5PT [48]. However, adding realistic scales of depth or affine noise, *e.g.* Fig. 4(a), has an observable negative

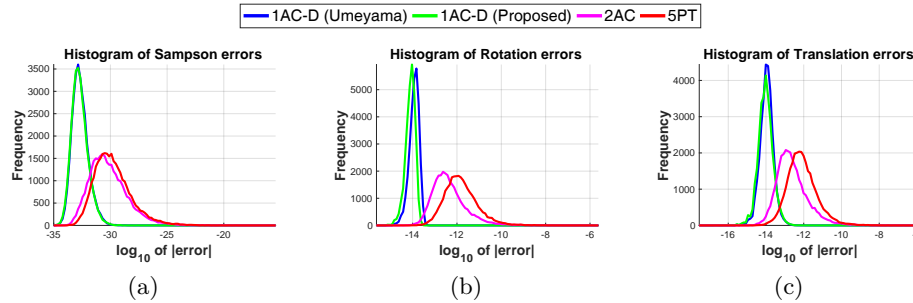


Fig. 3. Stability study of four different relative pose solvers: two versions of the proposed 1AC+D method, 1AC+D (Umeyama) and 1AC+D (Proposed), 2AC [6] solver using two affine correspondences and the five point solver 5PT [48]. The three plots show the distribution of various errors on logarithmic scales as histograms, namely (a) Sampson error, (b) rotation and (c) translation errors in degrees.

effect on the accuracy of the proposed ones. Although 2AC [6] is unaffected by the noise on depth, it is moderately affected by noise on the affinities of LAFs. Its rotation estimates are worse than that of any other methods in the comparison. The provided translation vectors are of acceptable quality.

Summarizing the synthetic evaluation, we can state that, on realistic scales of noise, the accuracy of the rotation and translation estimates of 5PT [48] and both versions of 1AC+D are comparable. 2AC is the worst performer among all.

4.3 Real-world evaluation

We tested the proposed solver on the 1DSfM dataset [56]⁵. It consists of 13 scenes of landmarks with photos of varying sizes collected from the internet. 1DSfM provides 2-view matches with epipolar geometries and a reference reconstruction from incremental SfM (computed with Bundler [45, 46]) for measuring error. We iterated through the provided 2-view matches, detected ACs [32] using the VLFeat library [54], applying the Difference-of-Gaussians algorithm combined with the affine shape adaptation procedure as proposed in [9]. In our experiments, affine shape adaptation only had a small $\sim 10\%$ extra time demand, *i.e.*, (0.31 ± 0.25) s per view, over regular feature extraction. This extra overhead is negligible, compared to the more time-consuming feature matching. Matches were filtered by the standard ratio test [32]. We did not consider image pairs in the evaluation with fewer than 20 corresponding features between them. For the evaluation, we chose scenes Piccadilly, NYC Library, Vienna Cathedral, Madrid Metropolis, and Ellis Island. To get monocular depth for each image, we applied the detector of Li *et al.* [30]. In total, the compared relative pose estimators were tested on a total of 110 395 image pairs.

⁵ <http://www.cs.cornell.edu/projects/1dsfm/>

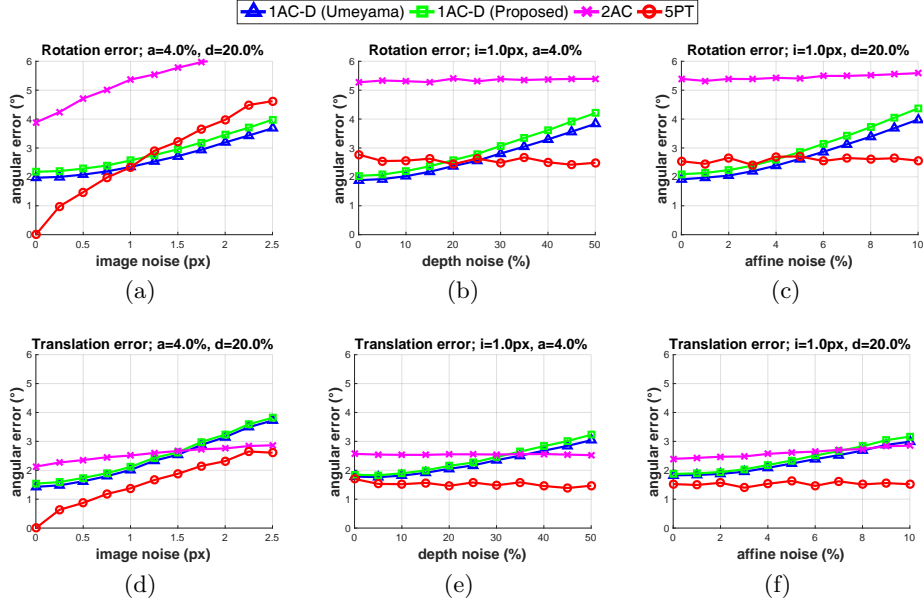


Fig. 4. Synthetic evaluation – the effect of different levels of noise on various relative pose solvers. Plots (a–f) compare the two proposed 1AC+D solvers, 2AC [6] and 5PT [48]. In the 1st row (a–c), the rotation error is shown. In the 2nd one (d–f) translation errors are plotted. All errors are angular errors in degrees. In each column, different setups are shown, where we fixed two of the three sources of noise – image, affine or depth noise – to analyse the negative effect of the third as its level increases.

As a robust estimator, we chose the Graph-Cut RANSAC [2] algorithm (GC-RANSAC) since it is state-of-the-art and has publicly available implementation⁶. In GC-RANSAC, and other locally optimized RANSACs, two different solvers are applied: (i) one for fitting to a minimal sample and (ii) one for fitting to a larger-than-minimal sample when improving the model parameters by fitting to all found inliers or in the local optimization step. For (i), the main objective is to solve the problem using as few points as possible since the overall wall-clock time is a function of the point number required for the estimation. For (ii), the goal is to estimate an accurate model from the provided set of points. In practice, step (ii) is applied rarely and, therefore, its processing time is not so crucial for achieving efficient robust estimation. We used the normalized eight-point algorithm followed by a rank-2 projection to estimate the essential matrix from a larger-than-minimal sample. We applied GC-RANSAC with a confidence set to 0.99 and inlier-outlier threshold set to be the 0.05 % of the image diagonal size. For the other parameters, the default values were used.

Fig. 6 reports the cumulative distribution functions – being accurate or fast is interpreted as a curve closer to the top-left corner – calculated on all image

⁶ <https://github.com/danini/graph-cut-ransac>

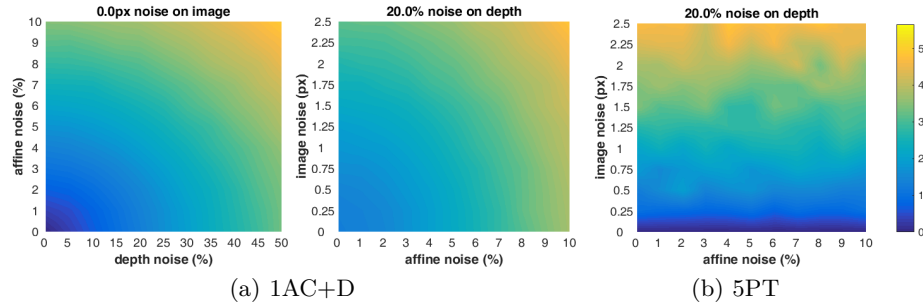


Fig. 5. Synthetic evaluation – the effect of image, affine and depth noise on the proposed 1AC+D and the point-based 5PT [48] – rotation errors, in degrees, displayed as heatmaps. As expected, 1AC+D is affected by all three noise types, as seen on (a). Being a point-based relative pose solver, 5PT [48] is affected only by the image noise (b).

pairs from each scene which was connected by an edge in the provided pose graph made by incremental SfM. The left plot shows the processing time, in seconds, required for the full robust estimation. The pose estimation is more than an order of magnitude faster when using the 1AC-D solver than the 5PT algorithm. The right two plots show the rotation and translation errors calculated using the reference reconstructions of the 1DSfM dataset [56]. 1AC-D leads to accuracy similar to that of the 5PT algorithm, both in terms of rotation and translation.

Note that, in practice, multiple solvers are used for creating the initial pose graph, considering homography, fundamental and essential matrix estimation simultaneously, *e.g.*, by QDEGSAC [16], to improve the accuracy and avoid degenerate configurations. It is nevertheless out of the scope of this paper, to include 1AC+D in state-of-the-art SfM pipelines. However, in the next section we show as a proof-of-concept that the proposed solver can be used within global SfM pipelines and leads to similar accuracy as the widely-used 5PT method [48].

Applying global SfM algorithm. Once relative poses are estimated for camera pairs of a given dataset, along with the inlier correspondences, they are fed to the Theia library [50] that performs global SfM [11, 56] using its internal implementation. That is, feature extraction, image matching and relative pose estimation were performed by our implementation either using the 1AC+D or the 5PT [48] solvers, as described above. The key steps of global SfM are robust orientation estimation, proposed by Chatterjee *et al.* [11], followed by robust nonlinear position optimization by Wilson *et al.* [56]. The estimation of global rotations and positions enables triangulating 3D points, and the reconstruction is finalized by the bundle adjustment of camera parameters and points.

Table 2 reports the results of Theia initialized by different solvers. There is no clear winner in terms of accuracy or run-time (of the reconstruction). Both solvers perform similarly when used for initializing global structure-from-motion.

Table 2. The results of a global SfM [50] algorithm, on scenes from the 1DSfM dataset [56], initialized with pose-graphs generated by the 5PT [48] and 1AC+D solvers. The reported properties are: the scene from the 1DSfM dataset [56] (1st column), relative pose solver (2nd), total runtime of the global SfM procedure given an initial pose-graph (3rd), rotation error of the reconstructed global poses in degrees (4th), position error in meters (5th) and focal length errors (6th).

scene	solver	runtime (s)	orientation (°)			position (m)			focal len. ($\times 10^{-2}$)		
			AVG	MED	STD	AVG	MED	STD	AVG	MED	STD
Piccadilly	1AC+D	51.2	6.1	4.3	8.0	6.5	3.4	7.7	2.3	1.6	3.2
	5PT	48.4	6.8	2.5	10.1	4.8	3.5	7.8	2.4	1.7	3.3
NYC Library	1AC+D	10.4	6.0	2.0	6.3	3.4	3.6	3.3	2.7	1.5	4.1
	5PT	5.9	5.9	1.9	6.7	4.2	3.7	5.5	2.9	1.5	4.6
Vienna Cathedral	1AC+D	56.5	4.5	1.7	11.6	6.1	1.4	10.3	2.3	1.3	3.7
	5PT	81.4	4.6	2.4	12.1	8.4	1.6	12.5	2.3	1.3	4.0
Madrid Metropolis	1AC-D	14.6	5.2	2.4	6.1	13.6	1.3	15.5	1.2	0.5	2.4
	5PT	20.5	6.9	4.7	7.2	18.3	3.6	21.1	1.2	0.5	2.3
Ellis Island	1AC-D	9.0	4.4	5.6	7.2	11.2	2.6	11.2	1.6	1.1	1.7
	5PT	5.7	3.1	6.0	3.9	11.8	1.9	12.0	1.6	1.0	1.7

5 Conclusions

In this paper, we propose a new approach for combining deep-learned non-metric monocular depth with affine correspondences (ACs) to estimate the relative pose of two calibrated cameras from a single correspondence. To the best of our knowledge, this is the first solution to the general relative camera pose estimation problem, from a single correspondence. Two new general constraints are derived interpreting the relationship of camera pose, affine correspondences and relative depth. Since the proposed solver requires a single correspondence, robust estimation becomes significantly faster, compared to traditional techniques, with speed depending linearly on the number of correspondences.

The proposed 1AC+D solver is tested both on synthetic data and on 110 395 publicly available real image pairs from the 1DSfM dataset. It leads to an accuracy similar to traditional approaches while being significantly faster. When solving large-scale problems, *e.g.*, pose-graph initialization for Structure-from-Motion (SfM) pipelines, the overhead of obtaining affine correspondences and monocular depth is negligible compared to the speed-up gain in the pairwise geometric verification. As a proof-of-concept, it is demonstrated on scenes from the 1DSfM dataset, via using a state-of-the-art global SfM algorithm, that acquiring the initial pose-graph by the proposed method leads to reconstruction of similar accuracy to the commonly used five-point solver.

Acknowledgement Supported by Exploring the Mathematical Foundations of Artificial Intelligence (2018-1.2.1-NKP-00008), ‘Intensification of the activities of HU-MATHS-IN—Hungarian Service Network of Mathematics for Industry and Innovation’ under grant number EFOP-3.6.2-16-2017-00015, the Ministry of Education OP VVV project CZ.02.1.01/0.0/0.0/16 019/0000765 Research Center for Informatics, and the Czech Science Foundation grant GA18-05360S.

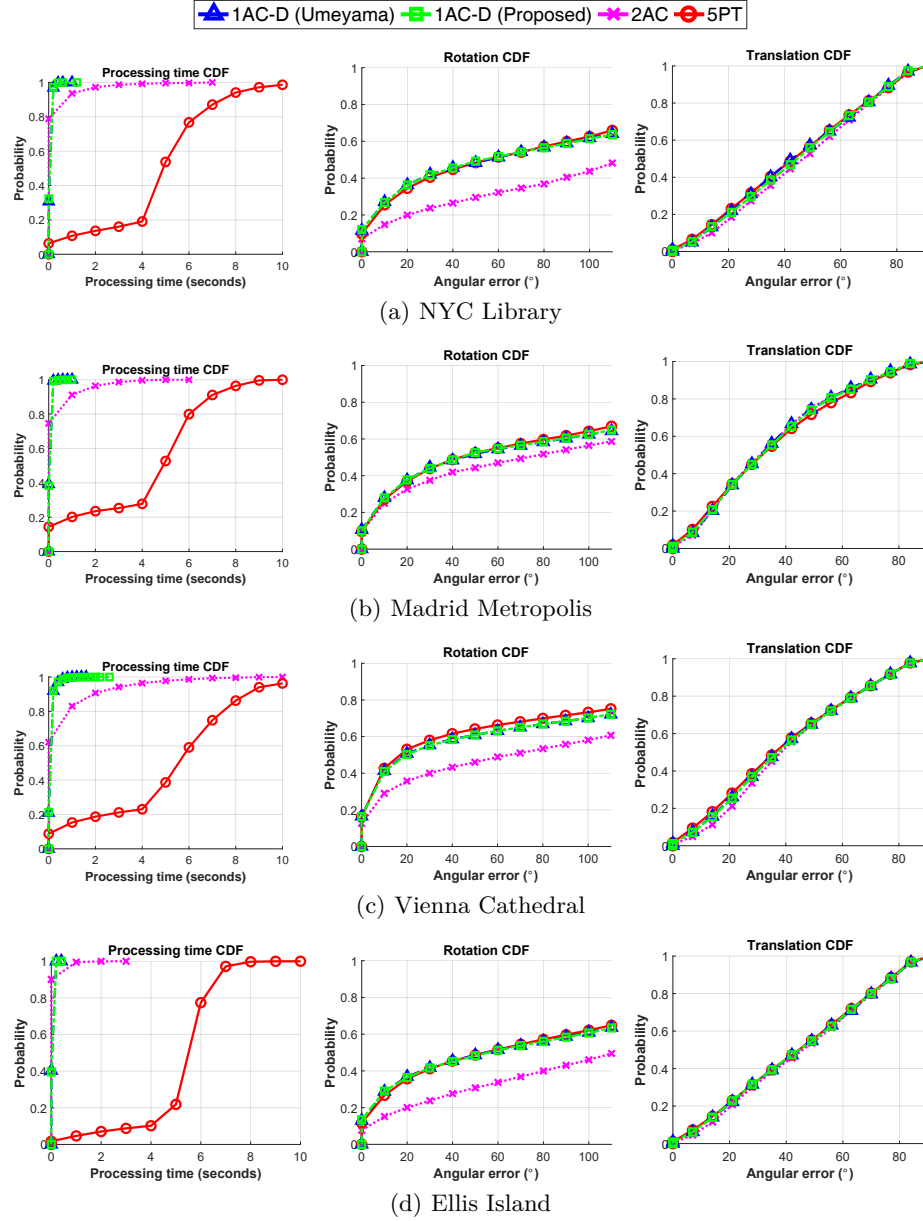


Fig. 6. Relative pose estimation on a total of 110395 image pairs from the 1DSfM dataset. The cumulative distribution functions are shown for the processing time (in seconds), angular error of the estimated rotations and translations (in degrees). Being accurate or fast is interpreted as a curve close to the top-left corner.

References

1. Albl, C., Kukulova, Z., Fitzgibbon, A., Heller, J., Smid, M., Pajdla, T.: On the two-view geometry of unsynchronized cameras. In: *Computer Vision and Pattern Recognition* (July 2017)
2. Barath, D., Matas, J.: Graph-cut RANSAC. In: *Computer Vision and Pattern Recognition*. pp. 6733–6741 (2018)
3. Baráth, D., Tóth, T., Hajder, L.: A minimal solution for two-view focal-length estimation using two affine correspondences. In: *Computer Vision and Pattern Recognition* (2017)
4. Barath, D., Eichhardt, I., Hajder, L.: Optimal multi-view surface normal estimation using affine correspondences. *IEEE Trans. Image Processing* **28**(7), 3301–3311 (2019)
5. Barath, D., Hajder, L.: A theory of point-wise homography estimation. *Pattern Recognition Letters* **94**, 7–14 (2017)
6. Barath, D., Hajder, L.: Efficient recovery of essential matrix from two affine correspondences. *IEEE Trans. Image Processing* **27**(11), 5328–5337 (2018)
7. Barath, D., Matas, J., Noskova, J.: MAGSAC: marginalizing sample consensus. In: *Computer Vision and Pattern Recognition*. pp. 10197–10205 (2019)
8. Batra, D., Nabbe, B., Hebert, M.: An alternative formulation for five point relative pose problem. In: *Workshop on Motion and Video Computing*. pp. 21–21. IEEE (2007)
9. Baumberg, A.: Reliable feature matching across widely separated views. In: *Computer Vision and Pattern Recognition*. vol. 1, pp. 774–781. IEEE (2000)
10. Bentolila, J., Francos, J.M.: Conic epipolar constraints from affine correspondences. *Computer Vision and Image Understanding* **122**, 105–114 (2014)
11. Chatterjee, A., Madhav Govindu, V.: Efficient and robust large-scale rotation averaging. In: *Proc. International Conf. on Computer Vision*. pp. 521–528 (2013)
12. Chum, O., Matas, J.: Matching with PROSAC-progressive sample consensus. In: *Computer Vision and Pattern Recognition*. IEEE (2005)
13. Eichhardt, I., Barath, D.: Optimal multi-view correction of local affine frames. In: *British Machine Vision Conf.* (September 2019)
14. Eichhardt, I., Chetverikov, D.: Affine correspondences between central cameras for rapid relative pose estimation. In: *Proc. European Conf. on Computer Vision*. pp. 482–497 (2018)
15. 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* (1981)
16. Frahm, J.M., Pollefeys, M.: RANSAC for (quasi-) degenerate data (QDEGSAC). In: *Computer Vision and Pattern Recognition*. pp. 453–460. IEEE (2006)
17. Fraundorfer, F., Tanskanen, P., Pollefeys, M.: A minimal case solution to the calibrated relative pose problem for the case of two known orientation angles. In: *Proc. European Conf. on Computer Vision*. pp. 269–282. Springer (2010)
18. Guan, B., Zhao, J., Li, Z., Sun, F., Fraundorfer, F.: Minimal solutions for relative pose with a single affine correspondence. In: *Computer Vision and Pattern Recognition* (2020)
19. Hajder, L., Baráth, D.: Relative planar motion for vehicle-mounted cameras from a single affine correspondence. In: *Proc. International Conf. of Robotics and Automation* (2020)

20. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge university press (2003)
21. Hartley, R., Li, H.: An efficient hidden variable approach to minimal-case camera motion estimation. *IEEE Trans. Pattern Analysis and Machine Intelligence* **34**(12), 2303–2314 (2012)
22. Hesch, J.A., Roumeliotis, S.I.: A direct least-squares (DLS) method for PnP. In: *Proc. International Conf. on Computer Vision*. pp. 383–390. IEEE (2011)
23. Hodan, T., Michel, F., Brachmann, E., Kehl, W., GlentBuch, A., Kraft, D., Drost, B., Vidal, J., Ihrke, S., Zabulis, X., et al.: BOP: Benchmark for 6D object pose estimation. In: *Proc. European Conf. on Computer Vision*. pp. 19–34 (2018)
24. Köser, K.: Geometric Estimation with Local Affine Frames and Free-form Surfaces. Shaker (2009)
25. Kukeleva, Z., Bujnak, M., Pajdla, T.: Polynomial eigenvalue solutions to minimal problems in computer vision. *IEEE Trans. Pattern Analysis and Machine Intelligence* **34**(7), 1381–1393 (2011)
26. Larsson, V., Kukeleva, Z., Zheng, Y.: Camera pose estimation with unknown principal point. In: *Computer Vision and Pattern Recognition* (June 2018)
27. Larsson, V., Sattler, T., Kukeleva, Z., Pollefeys, M.: Revisiting radial distortion absolute pose. In: *Proc. International Conf. on Computer Vision* (October 2019)
28. Li, B., Heng, L., Lee, G.H., Pollefeys, M.: A 4-point algorithm for relative pose estimation of a calibrated camera with a known relative rotation angle. In: *International Conf. on Intelligent Robots and Systems*. pp. 1595–1601. IEEE (2013)
29. Li, H., Hartley, R.: Five-point motion estimation made easy. In: *Proc. International Conf. on Pattern Recognition*. vol. 1, pp. 630–633. IEEE (2006)
30. Li, Z., Snavely, N.: MegaDepth: Learning single-view depth prediction from internet photos. In: *Computer Vision and Pattern Recognition* (June 2018)
31. Li, Z., Wang, G., Ji, X.: CDPN: Coordinates-based disentangled pose network for real-time RGB-based 6-DoF object pose estimation. In: *Proc. International Conf. on Computer Vision*. pp. 7678–7687 (2019)
32. Lowe, D.G.: Object recognition from local scale-invariant features. In: *Proc. International Conf. on Computer Vision*. IEEE (1999)
33. Mikolajczyk, K., Schmid, C.: Comparison of affine-invariant local detectors and descriptors. In: *European Signal Processing Conf.* pp. 1729–1732. IEEE (2004)
34. Mur-Artal, R., Montiel, J.M.M., Tardos, J.D.: ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Trans. on Robotics* **31**(5), 1147–1163 (2015)
35. Mur-Artal, R., Tardós, J.D.: ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras. *IEEE Trans. on Robotics* **33**(5), 1255–1262 (2017)
36. Nakano, G.: A versatile approach for solving PnP, PnPf, and PnPfr problems. In: *Proc. European Conf. on Computer Vision*. pp. 338–352. Springer (2016)
37. Nistér, D.: An efficient solution to the five-point relative pose problem. *IEEE Trans. Pattern Analysis and Machine Intelligence* pp. 756–770 (2004)
38. Perdoch, M., Matas, J., Chum, O.: Epipolar geometry from two correspondences. In: *Proc. International Conf. on Computer Vision* (2006)
39. Pritts, J., Kukeleva, Z., Larsson, V., Lochman, Y., Chum, O.: Minimal solvers for rectifying from radially-distorted conjugate translations. *arXiv preprint arXiv:1911.01507* (2019)
40. Raguram, R., Chum, O., Pollefeys, M., Matas, J., Frahm, J.M.: USAC: a universal framework for random sample consensus. *IEEE Trans. Pattern Analysis and Machine Intelligence* (2013)

41. Raposo, C., Barreto, J.P.: Theory and practice of structure-from-motion using affine correspondences. In: *Computer Vision and Pattern Recognition*. pp. 5470–5478 (2016)
42. Saurer, O., Pollefeys, M., Lee, G.H.: A minimal solution to the rolling shutter pose estimation problem. In: *International Conf. on Intelligent Robots and Systems*. pp. 1328–1334. IEEE (2015)
43. Scaramuzza, D.: 1-point-RANSAC structure from motion for vehicle-mounted cameras by exploiting non-holonomic constraints. *International Journal of Computer Vision* **95**(1), 74–85 (2011)
44. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: *Computer Vision and Pattern Recognition*. pp. 4104–4113 (2016)
45. Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: exploring photo collections in 3d. In: *ACM Trans. Graphics*. vol. 25, pp. 835–846. ACM (2006)
46. Snavely, S., Seitz, S.M., Szeliski, R.: Modeling the world from internet photo collections. *International Journal of Computer Vision* **80**(2), 189–210 (2008)
47. Solomon, J.: *Numerical algorithms: methods for computer vision, machine learning, and graphics*. AK Peters/CRC Press (2015)
48. Stewenius, H., Engels, C., Nistér, D.: Recent developments on direct relative orientation. *Journal of Photogrammetry and Remote Sensing* **60**(4), 284–294 (2006)
49. Stewenius, H., Nistér, D., Kahl, F., Schaffalitzky, F.: A minimal solution for relative pose with unknown focal length. In: *Computer Vision and Pattern Recognition*. vol. 2, pp. 789–794. IEEE (2005)
50. Sweeney, C.: Theia multiview geometry library. <http://theia-sfm.org>
51. Sweeney, C., Sattler, T., Hollerer, T., Turk, M., Pollefeys, M.: Optimizing the viewing graph for structure-from-motion. In: *Proc. International Conf. on Computer Vision*. pp. 801–809 (2015)
52. Torr, P.H.S., Zisserman, A.: MLESAC: A new robust estimator with application to estimating image geometry. *Computer Vision and Image Understanding* (2000)
53. Umeyama, S.: Least-squares estimation of transformation parameters between two point patterns. *IEEE Trans. Pattern Analysis and Machine Intelligence* (4), 376–380 (1991)
54. Vedaldi, A., Fulkerson, B.: VLFeat: An open and portable library of computer vision algorithms. <http://www.vlfeat.org/> (2008)
55. Ventura, J., Arth, C., Reitmayr, G., Schmalstieg, D.: A minimal solution to the generalized pose-and-scale problem. In: *Computer Vision and Pattern Recognition*. pp. 422–429 (2014)
56. Wilson, K., Snavely, N.: Robust Global Translations with 1DSfM. In: *Proc. European Conf. on Computer Vision*. pp. 61–75 (2014)
57. Wu, C.: Towards linear-time incremental structure from motion. In: *International Conf. on 3D Vision*. pp. 127–134. IEEE (2013)