Efficient Two-view Estimation using Richer Geometric Correspondences

Astermark, Jonathan

2025

– CENTRUM SCIENTIARUM MATHEMATICARUM –

Efficient Two-view Estimation using Richer Geometric Correspondences

JONATHAN ASTERMARK

Lund University
Faculty of Engineering
Centre for Mathematical Sciences
Mathematics

Efficient Two-view Estimation using Richer Geometric Correspondences

Efficient Two-view Estimation using Richer Geometric Correspondences

by Jonathan Astermark

LUND
UNIVERSITY

LICENTIATE THESIS

which, with due permission of Faculty of Engineering at Lund university, will be publicly defended on Tuesday 18th of March, 2025, at 13:15 in the room MH:309A at the Centre for Mathematical Sciences.

Thesis advisors:
Prof. Anders Heyden
Asst. Prof. Viktor Larsson

Faculty opponent:
Assoc. Prof. Juho Kannala, Aalto University, Finland

Organization
LUND UNIVERSITY
Centre for Mathematical Sciences
Box 118
SE–221 00 LUND
Sweden

Document name
Licentiate thesis

Date of presentation
2025-03-18

Sponsoring organization
ELLIIT

Author(s)
Jonathan Astermark

Title and subtitle
Efficient Two-view Estimation using Richer Geometric Correspondences

Abstract

Two-view estimation is a fundamental problem in 3D computer vision, and an important sub-task of multi-view estimation pipelines such as Structure-from-Motion (SfM) and Simultaneous Localization and Mapping (SLAM). In recent years, the main focus in the field has been on keypoint-based methods, where interest points are first detected and matched across the two images, followed by robust estimation of the geometry based on these keypoints. In this robust estimation step, the traditional approach is to use keypoint coordinates as basis for the estimation, by minimizing a geometric residual such as the reprojection error.

This thesis investigates estimation based on keypoints using richer geometric information, in addition to the keypoint coordinates. The goal is to increase efficiency in the estimation to achieve better runtime with maintained accuracy, which is an important factor for inclusion in multi-view systems for SfM and SLAM.

The thesis is based on three papers; the first two concern sample efficient minimal solvers for relative pose and homography estimation, using keypoints augmented with additional geometric information. In the first paper, we use scale information to constrain relative depths when estimating relative pose. In paper II, we use both scale and orientation information to constrain estimation of a plane-induced Euclidean homography. By combining multiple similar and seemingly redundant constraints, we develop a novel minimal solver allowing us to get noisy but surprisingly good homography estimates from even a single correspondence. The third paper is focused on summarizing semi-dense keypoint matches, to harness recent improvements in dense, detector-free keypoint matching. We introduce a summarization scheme that reduces the redundancy of semi-dense keypoints, which significantly decreases runtime compared to traditional estimation, with negligible reduction in estimation accuracy.

Key words
Two-view Estimation; Relative pose; Essential matrix; Homography; RANSAC

Classification system and/or index terms (if any)

Supplementary bibliographical information

Language
English

ISSN and key title
1404-028X

ISBN
978-91-8104-429-4 (print)
978-91-8104-430-0 (electronic)

Recipient's notes

Number of pages
xiv+78

Price

Security classification

Efficient Two-view Estimation using Richer Geometric Correspondences

by Jonathan Astermark

LUND
UNIVERSITY

Funding information: The thesis work was financially supported by ELLIIT.

Centre for Mathematical Sciences
Lund University
Box 118
SE-221 00 Lund
Sweden
www.maths.lu.se

Licentiate Theses in Mathematical Sciences 2025:01
ISSN: 1404-028X
ISBN: 978-91-8104-429-4 (print)
ISBN: 978-91-8104-430-0 (electronic)
LUTFTM-2001-2025

Printed in Sweden by Media-Tryck, Lund University, Lund 2025

Abstract

Two-view estimation is a fundamental problem in 3D computer vision, and an important sub-task of multi-view estimation pipelines such as Structure-from-Motion (SfM) and Simultaneous Localization and Mapping (SLAM). In recent years, the main focus in the field has been on keypoint-based methods, where interest points are first detected and matched across the two images, followed by robust estimation of the geometry based on these keypoints. In this robust estimation step, the traditional approach is to use keypoint coordinates as basis for the estimation, by minimizing a geometric residual such as the reprojection error.

This thesis investigates estimation based on keypoints using richer geometric information, in addition to the keypoint coordinates. The goal is to increase efficiency in the estimation to achieve better runtime with maintained accuracy, which is an important factor for inclusion in multi-view systems for SfM and SLAM.

The thesis is based on three papers; the first two concern sample efficient minimal solvers for relative pose and homography estimation, using keypoints augmented with additional geometric information. In the first paper, we use scale information to constrain relative depths when estimating relative pose. In paper II, we use both scale and orientation information to constrain estimation of a plane-induced Euclidean homography. By combining multiple similar and seemingly redundant constraints, we develop a novel minimal solver allowing us to get noisy but surprisingly good homography estimates from even a single correspondence. The third paper is focused on summarizing semi-dense keypoint matches, to harness recent improvements in dense, detector-free keypoint matching. We introduce a summarization scheme that reduces the redundancy of semi-dense keypoints, which significantly decreases runtime compared to traditional estimation, with negligible reduction in estimation accuracy.

List of Publications

This thesis is based on the following publications, referred to by uppercase Roman numerals. They are reproduced and included in this thesis with the permission of their respective publishers.

I **Fast Relative Pose Estimation using Relative Depth**

Jonathan Astermark, Yaqing Ding, Viktor Larsson, and Anders Heyden
In *International Conference on 3D Vision (3DV)*, 2024.

Author's contributions: The idea was jointly developed with YD, who implemented the minimal solver. I developed RelScaleNet and did most of the real-data evaluation. I also contributed significantly to the writing.

II **Noisy One-point Homographies are Surprisingly Good**

Yaqing Ding, **Jonathan Astermark**, Magnus Oskarsson, and Viktor Larsson
In *Computer Vision and Pattern Recognition (CVPR)*, 2024.

Author's contributions: The idea and implementation of the minimal solver came from YD. I performed the solver stability evaluation, and the qualitative evaluation. I also contributed to the writing.

III **Dense Match Summarization for Faster Two-view Estimation**

Jonathan Astermark, Anders Heyden, and Viktor Larsson
Accepted and to appear in *Computer Vision and Pattern Recognition (CVPR)*, 2025.

Author's contributions: The idea was jointly developed between all co-authors. I performed most of the experiments and evaluations, while VL implemented the C++ optimizations. I also contributed significantly to the writing.

Acknowledgments

I would like to extend my warmest gratitude to my supervisors, Anders Heyden and Viktor Larsson, for all their advice and guidance. I am grateful for how you always make time in your busy schedules to provide valuable feedback, discuss new ideas, and answer silly questions. I know I still have much to learn from you both, and look forward to having more interesting discussions during the rest of my PhD studies. I would also like to thank my co-authors, Yaqing Ding and Magnus Oskarsson for the collaborations we have had, and I hope there will be more of those in the future. Research, like most things, is the most fun when done together!

I would like to thank my colleagues at the Centre for Mathematical Sciences for creating a joyful work environment, balancing out focused work with regular coffee breaks, journal clubs, and small talk in the corridors. I am in particular grateful to my fellow PhD-students, and friends I have made during these last three years. I hope we will have many more after works, "symposiums", and summer school trips in the coming years! I am also thankful to my friends for always being up for boardgames when I need a break from work!

Finally, I would like to thank my family for always being there for me. I would especially like to thank my mom and dad for their endless support and encouragement, and my grandfather for always sharing wisdom and showing interest in both my life and work.

Funding

This work was funded by the strategic research project ELLIIT under project number B10.

Contents

Introduction

A central problem in 3D computer vision is the estimation of scene geometry from unconstrained image collections, known as Structure-from-Motion (SfM). In the general formulation, SfM involves simultaneous optimization of the 3D structures in a scene and the unknown parameters of the cameras that captured it. SfM, and the related task Simultaneous Localization and Mapping (SLAM), have received a lot of attention in recent decades and emerged as useful tools for visual localization and large-scale photogrammetry, with applications ranging from autonomous navigation [26] and augmented reality [2], to geosciences [46] and cultural heritage preservation [11]. Scene reconstruction can also be used as a prior for novel view synthesis, which has recently gained tremendous attention in the vision research community following the success of neural radiance fields [35] and 3D Gaussian splatting [22].

The main challenge in Structure-from-Motion is to robustly handle image collections taken from different viewpoints, potentially with different cameras, and presented to the system in no particular order. The images may also be taken at widely different points in time, which mean they can contain large variation in lighting conditions, changing seasons, and potentially even changes to the scene itself; see Figure 1 for a few examples.

Figure 1: Wide baseline image pairs. Unconstrained image collections may contain large variations in camera position, illumination, and structural changes in the scene. Even if humans easily can understand the relation between camera positions in these images, automated estimation can be challenging. Image pairs are from the WxBS dataset [36].

Systems with better robustness to these conditions typically come with a trade-off in terms of runtime, *i.e*. more robust and accurate methods tend to have a greater computational burden. This thesis investigates methods for improving this runtime-accuracy trade-off, to achieve shorter runtimes with maintained estimation accuracy. It is based on three papers, with the common theme of estimating geometry from correspondences containing more geometric information compared to the traditional keypoints used in most modern pipelines. The focus is on two-view estimation, which is an important sub-problem in both SfM and SLAM.

The thesis is organized as follows: the next two sections cover the relevant background and context of the presented research. Section 1 gives an overview of the mathematical models used in two-view geometry, while Section 2 covers methods for robustly estimating two-view geometry from real data. Then, in Section 3, a summary of the research contributions in this thesis is given, as well as a discussion on future research directions. The full papers are included at the end of the thesis.

1 Two-view Geometry

This section introduces the necessary mathematical models for two-view geometry: *the fundamental matrix*, *the essential matrix*, and *homographies*, as well as the camera model. In this thesis, we always assume a *pinhole camera model* with no distortion. For a more detailed discourse on camera models and two-view geometry, we refer to [20].

1.1 The Pinhole Camera Model

Consider a camera placed at the origin in some coordinate system, which we will refer to as the *camera's* coordinate system, looking down along the positive z-axis. In the *pinhole camera model*, each *scene point* $\hat{X} \in \mathbb{R}^3$ is projected along a line $\ell = \{\mu\hat{X} \mid \mu \in \mathbb{R}\}$ that passes through the origin, see Figure 2. This line's intersection with the *image plane* $z = 1$ gives the *image point* $(x, y, 1) \in \mathbb{R}^3$, which corresponds to $(x, y) \in \mathbb{R}^2$ in the plane.

However, since all points on the line ℓ projects to the same point, any triplet $(\lambda x, \lambda y, \lambda)$ for $\lambda \neq 0$ is an equivalent representation of (x, y). Any such representation is referred to as the *homogeneous coordinates* of (x, y). Then, we can write

$$\lambda \boldsymbol{x} = \hat{X}, \tag{1}$$

where $\boldsymbol{x}$ is a homogeneous representation of (x, y).

To describe simultaneous projection into multiple cameras, it is necessary to convert between each camera's own coordinate frame and a *world* coordinate system. If the camera is

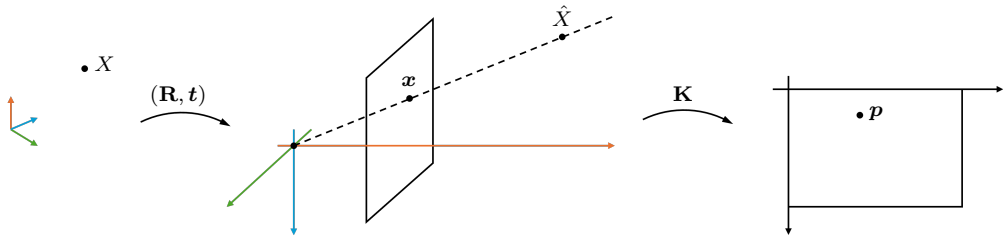

Figure 2: Overview of the pinhole camera model. A scene point $X \in \mathbb{R}^3$ in *world* coordinates is transformed to the *camera's* coordinate system by a rotation $\mathbf{R}$ and a translation $\boldsymbol{t}$. The transformed point $\hat{X}$ is projected along the line $\mu\hat{X}$. The corresponding calibrated image point in homogeneous coordinates is $\boldsymbol{x}$. The calibrated image point $\boldsymbol{x}$ is transformed to homogeneous *image* or *pixel* coordinates $\boldsymbol{p}$ using the calibration matrix $\mathbf{K}$.

rotated by $\mathbf{R} \in SO(3)$, followed by a translation $\boldsymbol{t} \in \mathbb{R}^3$ in the rotated coordinate system, then a scene point $X \in \mathbb{R}^3$ in world coordinates is given in the camera's coordinate system by the transformation

$$\hat{X} = \mathbf{R}X + \boldsymbol{t}. \tag{2}$$

This gives us the projection from world coordinates as

$$\lambda\boldsymbol{x} = \mathbf{R}X + \boldsymbol{t}. \tag{3}$$

To model transformation from the camera's calibrated coordinate system to *image* or *pixel* coordinates, we use the *calibration matrix*

$$\mathbf{K} = \begin{bmatrix} f_x & 0 & x_0 \\ 0 & f_y & y_0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{4}$$

where $(x_0, y_0) \in \mathbb{R}^2$ is the principal point, $f_x, f_y \in \mathbb{R}$ the focal lengths, and we have assumed zero skew. The image point $\boldsymbol{p}$ in homogeneous image coordinates is related to X by the *camera equation*

$$\lambda\boldsymbol{p} = \mathbf{K}(\mathbf{R}X + \boldsymbol{t}). \tag{5}$$

The point $\boldsymbol{p}$ is also referred to as the *uncalibrated* image point.

1.2 The Fundamental Matrix

Image points that represent projections of the same 3D scene point into different cameras is called a *keypoint correspondence*. For a keypoint correspondence, the camera equation (5) must be satisfied for both keypoints simultaneously, for the same scene point. Formally, let $(\boldsymbol{p}, \boldsymbol{p}')$ be an uncalibrated keypoint correspondence associated with the scene point X. If the relative pose of the two cameras is $(\mathbf{R}, \boldsymbol{t}) \in SE(3)$, then the camera equations are

$$\begin{cases} \lambda\boldsymbol{p} = \mathbf{K}_1 X, \\ \lambda'\boldsymbol{p}' = \mathbf{K}_2(\mathbf{R}X + \boldsymbol{t}), \end{cases} \tag{6}$$

where $\mathbf{K}_1$, $\mathbf{K}_2$ are the calibration matrices of the two cameras.

Substituting $X = \lambda \mathbf{K}_1^{-1} \boldsymbol{p}$ from the first row into the second and multiplying from the left with $(\mathbf{K}_2^{-1}\boldsymbol{p}')^\mathsf{T}[\boldsymbol{t}]_\times$, where $[\cdot]_\times$ is the skew-symmetric matrix representation of the cross-product, gives the well-known epipolar constraint

$$(\boldsymbol{p}')^\mathsf{T}\mathbf{F}\boldsymbol{p} = 0, \tag{7}$$

were $\mathbf{F} = \mathbf{K}_2^{-\mathsf{T}}[\boldsymbol{t}]_\times \mathbf{R} \mathbf{K}_1^{-1}$ is known as the *Fundamental Matrix* [32]. From (7), the fundamental matrix can be recovered from eight keypoint correspondences using the Direct Linear Transform (DLT) algorithm. By also using the non-linear constraint $\det \mathbf{F} = 0$, the problem can even be solved from seven correspondences [20]. Since this is the least number of keypoint correspondences required to constrain the fundamental matrix, it is called the *minimal sample*, and the 7-point solver is called the *minimal solver* for the fundamental matrix. However, if additional geometric constraints are known, the minimal sample can be less than seven points. For example, Bentolila and Francos [10] introduced a solver requiring only three *affine correspondences*, *i.e*. point correspondences with associated local affine transformations.

1.3 The Essential Matrix

If the camera parameters are known, the keypoint correspondences can be replaced by their calibrated counterparts $\boldsymbol{x} = \mathbf{K}_1^{-1}\boldsymbol{p}$ and $\boldsymbol{x}' = \mathbf{K}_2^{-1}\boldsymbol{p}'$. The corresponding camera equations are

$$\begin{cases} \lambda \boldsymbol{x} = X, \\ \lambda' \boldsymbol{x}' = \mathbf{R}X + \boldsymbol{t}. \end{cases} \tag{8}$$

The epipolar constraint for a calibrated keypoint correspondence is

$$(\boldsymbol{x}')^\mathsf{T}\mathbf{E}\boldsymbol{x} = 0, \tag{9}$$

where $\mathbf{E} = [t]_\times \mathbf{R}$ is known as the *Essential Matrix* [30]. The Essential matrix thus encodes the relative rotation and translation between cameras with known calibration.

Similarly to the fundamental matrix, the DLT algorithm can be used to recover the essential matrix from eight keypoint correspondences. The minimal problem, however, only requires five correspondences, and was efficiently solved by Nistér in 2004 [38]. If more geometric constraints are known, in addition to the keypoint coordinates, the problem can be solved from even fewer than five correspondences. For example, in [19] a minimal solver was introduced for the case of known vertical direction, obtained from an inertial measurement unit. In this case, three correspondences were enough to solve the minimal problem. Other works have used constraints from monocular depth [8] or affine correspondences [3].

In Paper I, we introduce a minimal solver leveraging the inter-image relative scale of keypoints, reducing the minimal number of correspondences to three. A similar idea was presented in [29]; however, we present an alternative parametrization that we show is more stable to noisy estimates of the relative depth. If we also assume known vertical direction, our solver is able to recover the pose using only two correspondences.

1.4 Homographies

If a collection of scene points $\{X_i\}_{i\in\mathbb{N}}$ lie on a plane $\boldsymbol{\pi} = (\boldsymbol{n}, d)$ in 3D, parameterized by the normal $\boldsymbol{n} \in \mathbb{R}^3$ and depth $d \in \mathbb{R}$, then in addition to the camera equations (6) or (8) the plane equation

$$\boldsymbol{n}^\mathsf{T} X_i + d = 0 \tag{10}$$

must be satisfied. We shall only treat the calibrated case here; inserting (10) into the camera equations (8) gives

$$\lambda' \boldsymbol{x}' = \lambda \left(\mathbf{R} + \boldsymbol{t}\boldsymbol{n}^\mathsf{T} \right) \boldsymbol{x}, \tag{11}$$

where

$$\mathbf{H} = \mathbf{R} + \boldsymbol{t}\boldsymbol{n}^\mathsf{T} \tag{12}$$

is the plane-induced Euclidean homography. Using the DLT algorithm, the homography can be recovered from four keypoint correspondences.

Just like for fundamental and essential matrix, previous works have introduced additional constraints to estimate the homography from fewer correspondences. For example, Barath and Hajder [4] used just two affine correspondences to constrain the homography. Similarly, Barath and Kukelova [5] used scale and orientation estimates to approximate the affine correspondences, to also solve from two correspondences. In paper II, we introduce an alternative minimal solver from two correspondences, also by leveraging both scale and orientation constraints. This is then extended by including the affine constraint from [5], along with a heuristic line-normal constraint, to give noisy but surprisingly good homography estimates from just a single correspondence.

2 Robust Two-view Estimation

The dominant approach for robust two-view estimation is based on first establishing keypoint correspondences in the image pair, and then trying to find a model that best fit to these keypoints. Since the matching step may introduce large errors caused by incorrect matches, keypoint-based geometry estimation must be robust to such errors. An alternative approach is to learn direct regression of a geometry model from the images. An overview of methods for keypoint detection, robust estimation, and direct regression is given below.

2.1 Keypoint Detection and Matching

Some of the earliest work on detection and matching of salient keypoints in images was done by Moravec [37] in 1981. In a seminal work from 2004, Lowe later introduced the Scale Invariant Feature Transform (SIFT) [31], detecting interest points at multiple scales and estimating an associated orientation. This enabled keypoint matching that is covariant to scale and rotation changes. As a further consequence, SIFT-keypoints provide a rough estimate of the associated scale and orientation of matched features, in addition to establishing keypoint correspondences. In papers I and II, we leverage this additional information to present novel minimal solvers for relative pose and homography.

In contrast to the hand-crafted features in SIFT, more recent developments in keypoint detection, such as SuperPoint [15], have focused on deep neural networks that learn both the detection and description of interest points through self-supervised learning. The matching of learned feature descriptors was later improved with SuperGlue [41] and LightGlue [28], by formulating matching as an optimal assignment problem on a Graph Neural Network, where features are aggregated using attention [44].

Recently, an emerging trend has also been seen in *detector-free matching*, sparked by the introduction of the Local Feature TRansformer (LoFTR) in 2021 [42]. Instead of first detecting salient and repeatable interest points and then matching them, the detector-free paradigm works by first matching images pixel-wise and then extracting refined keypoints. In LoFTR, the matching is done using self- and cross-attention on feature maps encoded by a Convolutional Neural Network (CNN). In addition to generally giving more matches per image pair, detector-free matchers are able to find correspondences even where salient keypoints are not available, such as on homogeneous surfaces or in repetitive patterns; see Figure 3 for an example.

The feature matching in LoFTR is however limited by the coarse level at which features are extracted, due to down-sampling in the CNN. In ASpanFormer [12], attention at different scales is used to get a more fine-grained matching. Dense Kernalized Matching (DKM) [16] instead refines coarse initial matches through depthwise convolutions on a pyramid of feature maps, resulting in the prediction of pixel-dense warps between images. In RoMa [17], coarse features are encoded using the visual foundation model DINOv2 [39], which has shown a remarkable ability to match semantically similar features under extreme pose changes, different image styles, and even between different objects. These coarse features are then decoded into coarse matches using a transformer architecture, and refined with a separate CNN-based feature encoding.

For the purpose of geometry estimation using traditional keypoint-based methods, pixel-dense warps such as the ones established by DKM and RoMa need to be sampled into semi-dense keypoint correspondences. In both DKM and RoMa, this sampling is done by

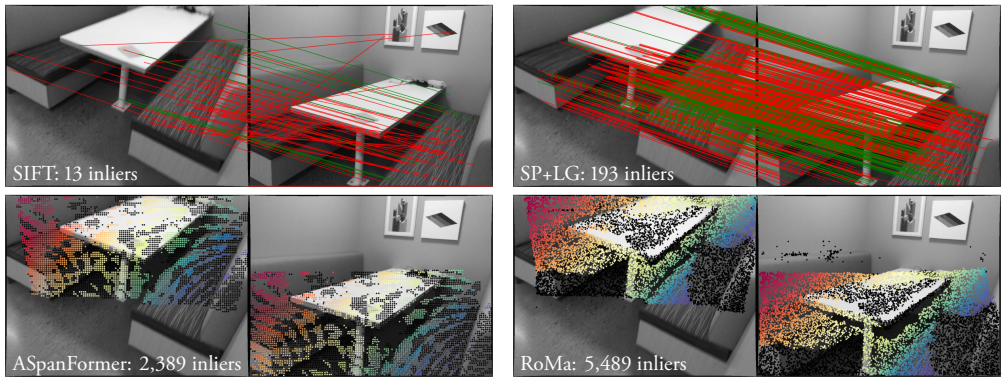

Figure 3: Qualitative comparison of keypoint correspondences using different matchers on an image pair from ScanNet [14]. The two top images show sparse keypoint matches from SIFT (left) and SuperPoint+LightGlue (right). Here, green lines denote matches with a Sampson error lower than 2.5 pixels (*i.e.* inliers) based on the provided ground-truth pose. Conversely, red lines denote matches with an error larger than 2.5 pixels (*i.e.* outliers). The two bottom images show semi-dense keypoints from ASpanFormer (left), and dense matching from RoMa sub-sampled to 10,000 semi-dense matches (right). Here, black points denote outliers while a color gradient is applied to inliers such that corresponding keypoints have the same color in both images.

balancing a predicted warp certainty per pixel with a reciprocal match density to enforce diversity; see [16] for details.

2.2 Robust Estimation

When estimating geometry on real data, it can be expected that the measurements, in the form of keypoint correspondences, contain some noise. It is well known that if all noise is normally distributed, the maximum likelihood estimator is found by minimizing squared residuals. However, correspondence sets may also contain mismatched keypoints, called *outliers*, which do not follow a normal distribution. For difficult image pairs, this outlier ratio can be significant, even for state-of-the-art image matchers like SuperPoint+LightGlue (SP+LG) or RoMa, as observed in Figure 3. In order to find an optimal model by least-squares optimization, it is necessary to filter away outliers. This needs to be done simultaneously with estimation of the model, since the two problems are interconnected. The problem of simultaneously estimating a model and its support is colloquially known as *robust estimation*.

The most common paradigm for robust estimation in computer vision is RANdom SAmple Consensus (RANSAC), introduced by Fishler and Bolles in 1981 [18]. RANSAC uses a hypothesize-and-verify approach, alternating between generating new candidate models and verifying them. The candidate model generation is done by randomly drawing a minimal sample, and then estimating the model using a minimal solver. Verification is then done by calculating residuals using the candidate model (or models, as minimal solvers may generate multiple possible solutions). From these residuals, a *consensus set* is found

that supports the candidate model according to an *inlier threshold* hyperparameter. This process is repeated until the probability of finding a model with a new largest consensus set falls below a specific threshold, typically less than 1 %. The final model is then refined by local optimization on its consensus set.

Much attention has been given to improving the initial RANSAC in terms of convergence speed, robustness, and accuracy. Notably, in Locally Optimized RANSAC (LO-RANSAC) [13, 23] local optimization is done each time a new best model is found. This not only leads to faster convergence, but also better final model estimates. Following [43], it is common practice to use truncated residuals for the local optimization. In Graph-Cut RANSAC (GC-RANSAC) [7], the consensus set is updated based on spatial coherence by alternating local optimization with solving a graph-cut problem.

2.3 Regression-based Methods

An alternative approach to keypoint-based estimation is direct regression of the geometry with a neural network. Early works on relative pose regression required highly overlapping camera views, while later works were able to handle wider baselines [34]. Instead of regressing just a single pose, RelPose [47] estimates probabilistic relative rotations, enabling joint reasoning from multiple views despite pairwise initial estimation. RelPose++ [27] extends this to 6D poses, while also processing multiple views jointly.

Recently, Wang *et al*. [45] introduced DUSt3R, instead regressing 2D-3D *pointmaps* for both images with a single pass through a Siamese vision transformer, followed by separate transformer-based decoders utilizing cross-image attention. The pointmaps can then be used for a variety of downstream estimation tasks; for example can relative pose be estimated through either robust perspective-n-point, or global alignment of the 3D coordinates.

As an extension of [45], MASt3R [24] introduces feature-based matching alongside the pointmap regression, by adding a second regression head per image to predict feature maps. Semi-dense keypoint correspondences are then established by sub-sampled mutual nearest-neighbor matching. Robust estimation from these keypoints further improves performance of downstream tasks, such as multi-view relative pose estimation.

In general, since direct regression involves no intermediate match assignment, this approach avoids propagation of errors from strict match decisions performed early in the pipeline. This also means that expensive robust estimation with RANSAC is typically not required in regression-based methods, although DUSt3R, for example, does run RANSAC for estimation from the regressed pointmaps. Another consequence of not running intermediate match assignment is that direct regression typically does not rely on the detection of salient and repeatable image features, which is essential for sparse keypoint-based methods. How-

Figure 4: Relative depth from scales. Paper I builds on the idea that relative scale observed from the images is inversely proportional to the relative depth. By estimating the relative scale, either using SIFT-features or a proposed CNN, we can calculate the relative depth. We introduce a novel three-point solver for relative pose using this additional constraint, as well as a two-point solver for known vertical direction.

ever, estimation from keypoints is usually more interpretable compared to direct regression. In addition, MASt3R shows benefits of adding keypoint matching back on top of a regression method. For these reasons, it is not obvious whether one of these paradigms is superior to the other for all applications.

3 Summary of Research Contributions

The research behind this thesis is focused on keypoint-based two-view estimation. However, unlike purely point-based methods, we focus on extending keypoints with additional geometric constraints, creating richer geometric correspondences, to get more efficient estimation. In papers I and II, we focus on minimal solvers leveraging scale and/or orientation estimates to reduce the number of correspondences required for minimal solvers of relative pose and homography, respectively. In Paper III, we instead summarize semi-dense keypoints into sparse clusters, with a small matrix approximating the total residual contributions of all keypoints within a cluster. In all papers, the increased efficiency leads to faster convergence in RANSAC, with at worst a relatively small loss in accuracy. In paper II, the sample efficiency also leads to increased accuracy for homography estimation.

3.1 Paper I

In paper I, we revisit the minimal problem for two-view relative pose estimation. Similarly to [29], we observe that the relative depth between keypoints gives one additional constraint on relative pose and can be obtained from the relative scale in the images, see Figure 4. With this additional constraint, only three correspondences are needed to solve the minimal problem. Using fewer samples in the minimal solver increases the chance of finding all-

Table 1: Results from paper I. Our solver is compared to the traditional, purely coordinate-based five-point solver from [38], and the solver from [6] approximating affine correspondences from SIFT-keypoints. Evaluation is done on ScanNet-1500 [14, 41], using both SIFT and SuperPoint+SuperGlue keypoints. Note that our method only requires scale estimates, while [6] requires scale and orientation. Thus when using RelScaleNet to estimate the scales, our method is applicable to any type of keypoints. All compared solvers are using in LO-RANSAC for robust estimation.

		SIFT keypoints			SP+SG keypoints		
Method		AUC@5°	AUC@10°	RT(ms)	AUC@5°	AUC@10°	RT(ms)
5 pt.	[38]	**11.06**	**21.99**	8.2	17.55	34.21	59.4
3 pt. + SIFT	[6]	4.94	10.33	3.7	-	-	-
3 pt. + SIFT	*Ours*	9.90	20.59	**2.9**	-	-	-
3 pt. + RelScaleNet	*Ours*	10.43	21.21	**2.9**	**18.39**	**35.46**	**2.9**

inlier samples, and this can significantly reduce the number of RANSAC iterations required for convergence, especially in scenarios with a high outlier ratio. In the paper, we present an alternative parametrization to that of [29], which we show is more stable to noisy estimates of the relative depth. Since it is typically harder to get good relative depth estimates than keypoint coordinates, this stability has a significant impact on the solver performance.

Our parametrization of the relative pose estimation is based on first solving for the unknown depths λ, λ' in the camera equations (8). Given three keypoint correspondences $(\boldsymbol{x}_i, \boldsymbol{x}'_i)$, $i \in \{1, 2, 3\}$ with associated depths λ_i, λ'_i, we find that if two of the correspondences have known *relative depths* $\sigma_i = \lambda'_i/\lambda_i$, we can formulate the constraints

$$\|\sigma_1 \boldsymbol{x}'_1 - \sigma_2 \lambda_2 \boldsymbol{x}'_2\|^2 = \|\boldsymbol{x}_1 - \lambda_2 \boldsymbol{x}_2\|^2, \tag{13}$$

$$\|\sigma_1 \boldsymbol{x}'_1 - \lambda'_3 \boldsymbol{x}'_3\|^2 = \|\boldsymbol{x}_1 - \lambda_3 \boldsymbol{x}_3\|^2, \tag{14}$$

$$\|\sigma_2 \lambda_2 \boldsymbol{x}'_2 - \lambda'_3 \boldsymbol{x}'_3\|^2 = \|\lambda_2 \boldsymbol{x}_2 - \lambda_3 \boldsymbol{x}_3\|^2. \tag{15}$$

The three unknowns $\lambda_2, \lambda_3, \lambda'_3$ can then be found by solving two quadratics. The key to the stability of our solver is that we only use two of the relative depths; so if all correspondences have known estimates of σ_i, we can use all possible permutations to generate candidate solutions.

We further demonstrate two ways of obtaining the relative scales. First, we show that scale estimates from SIFT-keypoints are sufficiently accurate for our solver. Then, to make our method work with any type of keypoints we introduce a CNN, *RelScaleNet*, that can estimate the relative scale for any keypoint correspondences based on image patches. Through experiments, we show that our solver performs on par with the five-point solver for both types of scale estimates in high-inlier scenarios. In low-inlier scenarios, however, our solver has significantly lower runtime compared to the five-point solver, while maintaining most of the estimation accuracy, see Table 1. Furthermore, we extend the three-point solver to a two-point solver for the special case of known vertical direction.

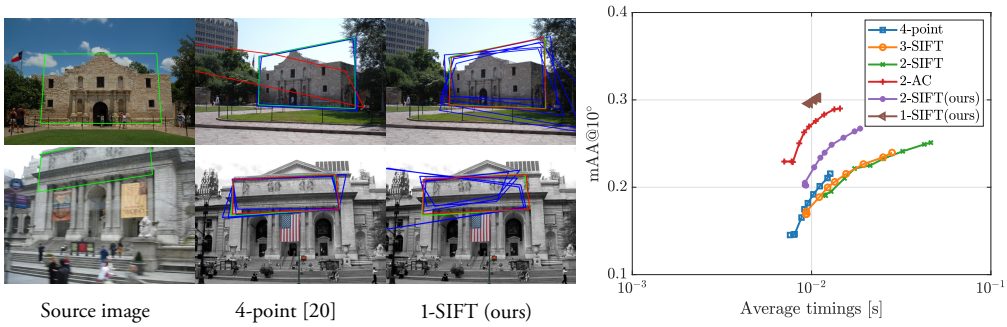

Figure 5: Results from paper II. *Left:* Qualitative examples showing top-5 homographies obtained from minimal samples (**blue**) and output from GC-RANSAC (**red**), compared to ground-truth (**green**). We note that our solver converges in GC-RANSAC even in some cases where all minimal estimates are noisy.) *Right:* mAA vs. runtime trade-off on HEB [9].

3.2 Paper II

In Paper II, we continue on the idea of using the scale-derived relative depth constraint from paper I, and apply it to Euclidean homography estimation. For a point correspondence $(\boldsymbol{x}, \boldsymbol{x}')$ with relative depth σ, the Euclidean homography $\mathbf{H}\colon \boldsymbol{x} \mapsto \boldsymbol{x}'$ fulfills

$$\sigma \boldsymbol{x}' = \mathbf{H}\boldsymbol{x}. \tag{16}$$

Additionally, we use the orientation estimates from SIFT to form a line correspondence. Taking $\boldsymbol{l}$ to be the line through $\boldsymbol{x}$ with orientation equal to the SIFT orientation (and analogously for $\boldsymbol{l}'$), we get a line correspondence $(\boldsymbol{l}, \boldsymbol{l}')$ that fulfills [20]

$$[\boldsymbol{l}]_\times \mathbf{H}^\top \boldsymbol{l}' = \mathbf{0}. \tag{17}$$

Using these constraints, together with the constraint from [33, 48] that the second singular value should be one, we developed a novel homography solver from two SIFT correspondences.

Similarly, a 2-point homography solver using SIFT correspondences was also introduced in [5], but this differed from ours by using the scale and orientation to approximate an affine correspondence. By also including this approximation in our solver, we can constrain our solver further. Although this uses seemingly redundant constraints, we argue that they are algebraically independent due to relying on different approximations.

Combining these constraints means a single SIFT correspondence gives eight linear constraints on the homography, out of which seven are linearly independent. In order to make the system full rank, we finally add a heuristic constraint that, although formally incorrect, turns out to have little impact on the result. This constraint is achieved by treating the line normal as an image point that is mapped by the homography.

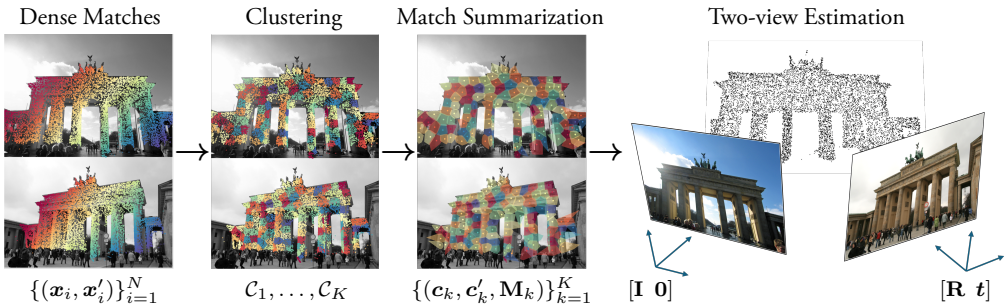

Figure 6: Overview of the method in paper III. Dense matches are first clustered, and then replaced with a representative match and a 9×9 matrix summarizing the residuals. The summarized matches are then used for two-view relative pose estimation.

Through experiments, we show that the resulting one-point solver, although noisy, has similar stability as the two-point solver in [5] but with superior sample efficiency. When integrated into GC-RANSAC, this leads to surprisingly good results as shown in our evaluation, see Figure 5. We draw the conclusion that sample efficiency might be more important for homography estimation than model accuracy, when used in a robust framework with strong local optimization such as GC-RANSAC.

3.3 Paper III

In paper III, our focus is to take advantage of recent developments in dense and semi-dense matching, described in Section 2.1. The superior ability of these methods to find accurate correspondences, even for very wide baselines and homogeneous image regions, enables downstream estimation even in very challenging scenarios, as demonstrated by the experiments in [17]. However, traditional keypoint-based estimation pipelines were originally designed for sparse keypoint matches. The computational complexity of robust estimation in RANSAC scales poorly with number of keypoints, making it prohibitively expensive to run the same pipelines for pixel-dense sampling of correspondences. Thus, both DKM and RoMa employ a balanced sub-sampling scheme to extract 5,000-10,000 semi-dense correspondences for downstream estimation.

In paper III, we show through experiments that semi-dense keypoints provide heavily redundant geometric constraints, which makes traditional robust estimation inefficient, and suggests that a different approach is needed for semi-dense matches. Our experiments further show that too heavy subsampling, on the other hand, leads to a loss in estimation accuracy.

Instead, we suggest a scheme for clustering correspondences with similar contributions of geometric constraints, and summarizing their residuals. Our approach is summarized in Figure 6. In the first step of our approach, we cluster correspondences using K-means in 4D

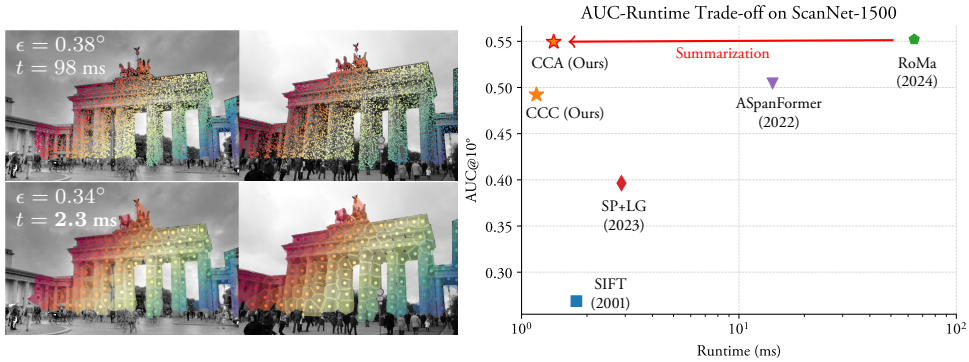

Figure 7: Results from paper III. *Left:* Example of using our clustering and summarization approach on an image pair from MegaDepth [25]. In the top images, 10,000 DKM-matches are used to calculate the relative pose. This gives a pose error is $\epsilon = 0.38°$. Our approach (bottom images) has comparable accuracy but is 43x faster on this image pair. *Right:* AUC vs runtime trade-off ScanNet-1500 using different keypoint correspondences, and our summarized correspondences (CCC and CCA).

match-space and select the match closest to the cluster centroid as a *representative match*. In our experiments, we show that using the representative matches for model verification leads to a runtime improvement of more than 50x compared to fully dense robust estimation, at only a small cost in pose accuracy (see method CCC in Figure 7).

While the representative matches can also be used for model refinement, we found that more of the accuracy could be recovered by approximating the Sampson error for all correspondences $(\boldsymbol{x}, \boldsymbol{x}')$, in a cluster with representative match $(\boldsymbol{c}, \boldsymbol{c}')$, as

$$\mathcal{E}(\mathbf{E}, \boldsymbol{x}, \boldsymbol{x}') = \frac{((\boldsymbol{x}')^\mathsf{T}\mathbf{E}\boldsymbol{x})^2}{\|\mathbf{E}_{12}\boldsymbol{x}\|^2 + \|(\mathbf{E}^\mathsf{T})_{12}\boldsymbol{x}'\|^2} \approx \frac{((\boldsymbol{x}')^\mathsf{T}\mathbf{E}\boldsymbol{x})^2}{\|\mathbf{E}_{12}\boldsymbol{c}\|^2 + \|(\mathbf{E}^\mathsf{T})_{12}\boldsymbol{c}'\|^2}. \tag{18}$$

This lets us write the sum of residuals for a cluster as a matrix expression

$$f_{\text{cl}}(\mathbf{E}; \mathcal{C}) \approx \frac{1}{\|\mathbf{E}_{12}\boldsymbol{c}\|^2 + \|(\mathbf{E}^\mathsf{T})_{12}\boldsymbol{c}'\|^2} \|\mathbf{A}\boldsymbol{e}\|^2, \tag{19}$$

where $\mathbf{A} \in \mathbb{R}^{n\times 9}$ has rows $\boldsymbol{A}_i = (\boldsymbol{x}_i \otimes \boldsymbol{x}_i')^\mathsf{T}$, using $\otimes$ to denote the Kronecker product. Through Cholesky factorization, the matrix A can equivalently and efficiently be replaced with a *reduced measurement matrix* [40] $M \in \mathbb{R}^{9\times 9}$ which, together with the representative match, summarizes the geometric constraints of the cluster. By further assuming that each cluster is either all-inlier or all-outlier, the robust approximate cost function for all matches, using truncated residuals, is approximated with

$$f(\mathbf{E}) \approx \sum_{k=1}^{K} \min \left\{ \frac{1}{\|\mathbf{E}_{12}\boldsymbol{c}\|^2 + \|(\mathbf{E}^\mathsf{T})_{12}\boldsymbol{c}'\|^2} \|\mathbf{M}_k\boldsymbol{e}\|^2, |\mathcal{C}_k|\tau^2 \right\}, \tag{20}$$

where τ is a threshold hyperparameter, K is the number of clusters, and $|\mathcal{C}_k|$ is the size of cluster k. This is fast to compute for a reasonably sized K, which can be made at least

two orders of magnitude smaller than the number of dense matches. By using (20) as cost function in the model refinement, we recover more of the estimation accuracy compared to only using representative matches, while still achieving over 40x speedup compared to dense estimation (see method CCA in Figure 7).

3.4 Conclusions and Future Work

The research presented in this thesis was focused on efficient keypoint-based two-view estimation, using keypoint correspondences with additional geometric information along with 2D coordinates. In papers I and II, the extra information was relative depth, inferred from relative scale, and relative orientation; both obtainable from SIFT keypoints. We used this to solve minimal problems with fewer correspondences than purely point-based methods. Paper III instead focused on summarizing the residuals from semi-dense matches, allowing the use of dense geometric information in sparse estimation. The main improvements in the presented work have been in terms of runtime, although we have also shown that increased sample efficiency can lead to better estimates, at least for homography estimation in GC-RANSAC.

In future work, the summarization scheme in paper III can be extended to other estimation problems, for example homographies where the transfer errors in a cluster can be summarized similarly to the Sampson error as

$$\sum_{(\boldsymbol{x},\boldsymbol{x}')\in\mathcal{C}} \frac{\|\mathbf{H}\boldsymbol{x} - (\mathbf{H}_3\boldsymbol{x})\boldsymbol{x}'\|^2}{(\mathbf{H}_3\boldsymbol{x})^2} \approx \frac{\|A\boldsymbol{h}\|^2}{(\mathbf{H}_3\boldsymbol{c})^2}, \tag{21}$$

where the approximation $\mathbf{H}_3\boldsymbol{x} \approx \mathbf{H}_3\boldsymbol{c}$ corresponds to fronto-parallelism. If the clusters are approximately planar, we can further use this to reformulate relative pose estimation as the optimization of K homographies

$$\mathbf{H}_i = \mathbf{R} + \boldsymbol{t}\boldsymbol{n}_i^\mathsf{T}, \quad i \in \{1, \dots, K\}, \tag{22}$$

with shared $\mathbf{R}$ and $\boldsymbol{t}$. While this introduces additional parameters to optimize for the two-view case, for multi-view optimization we may benefit from the fact that the normals are shared across different views. To ensure validity of the co-planarity approximation, it could be helpful to include the use of a plane detector, similar to [21, 1]. Furthermore, the approximation of fronto-parallelism could potentially be validated though monocular depth estimation.

Another possible multi-view extension of paper III could be the formation of tracks of summarized clusters. However, this would require the clustering to be multi-view consistent, and it is not immediately clear how this should be achieved. Perhaps a more sophisticated

clustering approach is needed for this case, using learned priors such as semantic-aware segmentation. However, the clusters would simultaneously need to be constrained to ensure tightness of the used approximations.

References

[1] Samir Agarwala et al. “PlaneFormers: From Sparse View Planes to 3D Reconstruction”. In: *European Conference on Computer Vision (ECCV)*. 2022.

[2] Hyojoon Bae et al. “Fast and scalable structure-from-motion based localization for high-precision mobile augmented reality systems”. In: *mUX: The Journal of Mobile User Experience* 5 (2016), pp. 1–21.

[3] Daniel Barath and Levente Hajder. “Efficient recovery of essential matrix from two affine correspondences”. In: *IEEE Trans. on Image Processing (TIP)* 27.11 (2018), pp. 5328–5337.

[4] Daniel Barath and Levente Hajder. “Novel ways to estimate homography from local affine transformations”. In: *International Conference on Computer Vision Theory and Applications (VISAPP)*. 2016.

[5] Daniel Barath and Zuzana Kukelova. “Homography from two orientation-and scale-covariant features”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2019.

[6] Daniel Barath and Zuzana Kukelova. “Relative Pose from SIFT Features”. In: *European Conference on Computer Vision (ECCV)*. 2022.

[7] Daniel Barath and Jiří Matas. “Graph-Cut RANSAC”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2018.

[8] Daniel Barath and Chris Sweeney. “Relative Pose Solvers using Monocular Depth”. In: *International Conference on Pattern Recognition (ICPR)*. 2022.

[9] Daniel Barath et al. “A Large-Scale Homography Benchmark”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2023.

[10] Jacob Bentolila and Joseph M Francos. “Conic epipolar constraints from affine correspondences”. In: *Computer Vision and Image Understanding (CVIU)* 122 (2014), pp. 105–114.

[11] Grazia Caradonna et al. “Multi-image 3D Reconstruction: A Photogrammetric and Structure from Motion Comparative Analysis”. In: *Computational Science and Its Applications (ICCSA)*. 2018.

[12] Hongkai Chen et al. “ASpanFormer: Detector-Free Image Matching with Adaptive Span Transformer”. In: 2022.

[13] Ondřej Chum, Jiří Matas, and Josef Kittler. “Locally optimized RANSAC”. In: *Joint pattern recognition symposium*. 2003.

[14] Angela Dai et al. “ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2017.

[15] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. "Superpoint: Self-supervised interest point detection and description". In: *Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2018.

[16] Johan Edstedt et al. "DKM: Dense Kernelized Feature Matching for Geometry Estimation". In: *Computer Vision and Pattern Recognition (CVPR)*. 2023.

[17] Johan Edstedt et al. "RoMa: Robust Dense Feature Matching". In: *Computer Vision and Pattern Recognition (CVPR)*. 2024.

[18] Martin A Fischler and Robert C Bolles. "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography". In: *Communications of the ACM* 24.6 (1981), pp. 381–395.

[19] Friedrich Fraundorfer, Petri Tanskanen, and Marc Pollefeys. "A minimal case solution to the calibrated relative pose problem for the case of two known orientation angles". In: *European Conference on Computer Vision (ECCV)*. 2010.

[20] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[21] Linyi Jin et al. "Planar Surface Reconstruction from Sparse Views". In: *International Conference on Computer Vision (ICCV)*. 2021.

[22] Bernhard Kerbl et al. "3D Gaussian Splatting for Real-Time Radiance Field Rendering". In: *ACM Transactions on Graphics* 42.4 (2023).

[23] Karel Lebeda, Jirı Matas, and Ondrej Chum. "Fixing the locally optimized ransac–full experimental evaluation". In: *British Machine Vision Conference (BMVC)*. 2012.

[24] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. "Grounding Image Matching in 3D with MASt3R". In: *European Conference on Computer Vision (ECCV)*. 2024.

[25] Zhengqi Li and Noah Snavely. "MegaDepth: Learning Single-View Depth Prediction from Internet Photos". In: *Computer Vision and Pattern Recognition (CVPR)*. 2018.

[26] Hyon Lim et al. "Real-time image-based 6-DOF localization in large-scale environments". In: *Computer Vision and Pattern Recognition (CVPR)*. 2012.

[27] Amy Lin et al. "RelPose++: Recovering 6D Poses from Sparse-view Observations". In: *International Conference on 3D Vision (3DV)*. 2024.

[28] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. "Lightglue: Local feature matching at light speed". In: *International Conference on Computer Vision (ICCV)*. 2023.

[29] Stephan Liwicki and Christopher Zach. "Scale Exploiting Minimal Solvers for Relative Pose with Calibrated Cameras." In: *British Machine Vision Conference (BMVC)*. 2017.

[30] H Christopher Longuet-Higgins. “A computer algorithm for reconstructing a scene from two projections”. In: *Nature* 293.5828 (1981), pp. 133–135.

[31] David G Lowe. “Distinctive image features from scale-invariant keypoints”. In: *International Journal of Computer Vision (IJCV)* 60 (2004), pp. 91–110.

[32] Quan-Tuan Luong and Olivier Faugeras. “The Fundamental Matrix: Theory, Algorithms, and Stability Analysis”. In: *International Journal of Computer Vision (IJCV)* 17 (1996), pp. 43–75.

[33] Ezio Malis and Manuel Vargas Villanueva. “Deeper understanding of the homography decomposition for vision-based control”. In: *INRIA, Tech. Rep.* (2007).

[34] Iaroslav Melekhov et al. “Relative camera pose estimation using convolutional neural networks”. In: *Advanced Concepts for Intelligent Vision Systems (ACIVS)*. 2017.

[35] Ben Mildenhall et al. “NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis”. In: *European Conference on Computer Vision (ECCV)*. 2020.

[36] Dmytro Mishkin et al. “WxBS: Wide Baseline Stereo Generalizations”. In: *British Machine Vision Conference (BMVC)*. 2015.

[37] Hans P. Moravec. “Rover Visual Obstacle Avoidance”. In: *International Joint Conference on Artificial Intelligence*. 1981.

[38] David Nistér. “An efficient solution to the five-point relative pose problem”. In: *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)* 26.6 (2004), pp. 756–770.

[39] Maxime Oquab et al. “DINOv2: Learning Robust Visual Features without Supervision”. In: *arXiv preprint arXiv:2304.07193* (2023).

[40] Antonio L Rodríguez, Pedro E López-de-Teruel, and Alberto Ruiz. “Reduced epipolar cost for accelerated incremental SfM”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2011.

[41] Paul-Edouard Sarlin et al. “Superglue: Learning feature matching with graph neural networks”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2020.

[42] Jiaming Sun et al. “LoFTR: Detector-Free Local Feature Matching with Transformers”. In: *Computer Vision and Pattern Recognition (CVPR)* (2021).

[43] Philip HS Torr and Andrew Zisserman. “MLESAC: A New Robust Estimator with Application to Estimating Image Geometry”. In: *Computer Vision and Image Understanding (CVIU)* 78.1 (2000), pp. 138–156.

[44] Ashish Vaswani et al. “Attention is All you Need”. In: *Neural Information Processing Systems (NeurIPS)*. 2017.

[45] Shuzhe Wang et al. “DUSt3R: Geometric 3D Vision Made Easy”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2024.

[46] M.J. Westoby et al. "'Structure-from-Motion' photogrammetry: A low-cost, effective tool for geoscience applications". In: *Geomorphology* 179 (2012), pp. 300–314.

[47] Jason Y. Zhang, Deva Ramanan, and Shubham Tulsiani. "RelPose: Predicting Probabilistic Relative Rotation for Single Objects in the Wild". In: *European Conference on Computer Vision (ECCV)*. 2022.

[48] Zhongfei Zhang and Allen R Hanson. "Scaled Euclidean 3D reconstruction based on externally uncalibrated cameras". In: *Proceedings of International Symposium on Computer Vision (ISCV)*. 1995.

Scientific Publications

Printed by Media-Tryck, Lund 2025 NORDIC SWAN ECOLABEL 3041 0903

Licentiate Theses in Mathematical Sciences 2025:01

ISBN 978-91-8104-429-4
ISSN 1404-028X