POLITECNICO DI MILANO
DIPARTIMENTO DI ELETTRONICA, INFORMAZIONE E BIOINGEGNERIA
DOCTORAL PROGRAMME IN COMPUTER SCIENCE AND ENGINEERING

INTEGRATING GEOMETRIC AND LEARNING-BASED METHODS FOR CAMERA AUTOCALIBRATION AND MULTI-BODY STRUCTURE-FROM-MOTION

Doctoral Dissertation of:
Andrea Porfiri Dal Cin

Supervisor:
Prof. Luca Magri

Tutor:
Prof. Francesco Amigoni

The Chair of the Doctoral Program:
Prof. Cinzia Cappiello

2024 – Cycle XXXVII

Abstract

Three-dimensional (3D) reconstruction from multiple images, known as Structure-from-Motion (SfM), is a fundamental challenge in modern computer vision with broad applications in fields such as robotics, photogrammetry, augmented reality (AR), and virtual reality (VR). SfM aims to achieve two main objectives: estimating the motion of the camera and reconstructing the 3D scene. The first objective involves determining the camera's trajectory relative to the scene, while the second focuses on recovering the scene's three-dimensional geometry from multiple two-dimensional images. The success of SfM heavily relies on the accurate calibration of camera parameters, a process known as camera autocalibration. Accurate camera parameter estimation is crucial because errors can lead to significant inaccuracies in the reconstructed 3D geometry, potentially compromising the entire reconstruction process.

This thesis advances the state-of-the-art in camera autocalibration and Multi-Body Structure-from-Motion (MBSfM), particularly for *in-the-wild* scenarios where images are captured in uncontrolled environments with unknown camera parameters. Traditional SfM methods often assume static scenes and rely on external calibration objects or known scene geometry for camera parameter estimation. However, these assumptions may not hold in uncontrolled environments, necessitating robust autocalibration methods that do not require prior knowledge of camera parameters or known scene geometry.

To address these challenges, this thesis introduces a novel family of solvers for *pinhole* camera autocalibration. These solvers enhance the accuracy and robustness of existing methods, overcoming common failure cases in real-world scenarios. Unlike traditional approaches that use Kruppa's equations and the modulus constraint to perform a metric upgrade of an initial projective reconstruction, the proposed solvers jointly perform camera autocalibration and metric reconstruction using only a minimal set of keypoint correspondences between image pairs or triplets. Extensive experimental evaluations on synthetic and real-world datasets, along with integration into the widely used COLMAP SfM software, demonstrate that our solvers improve the accuracy of both the reconstructed 3D points and the estimated camera parameters compared to existing methods.

To further extend the applicability of these solvers to wide-angle cameras character-

ized by radial distortion, the thesis introduces a novel learning-based zero-shot autocalibration method for radially symmetric cameras, including most fisheye and 360-degree cameras. Our method estimates an implicit camera representation from which distortion parameters can be recovered, allowing for image rectification.

Additionally, the thesis addresses the challenge of reconstructing dynamic scenes with multiple rigidly moving objects—a scenario where traditional SfM algorithms often struggle. This limitation is tackled by developing new algorithms for Multi-Body Structure-from-Motion (MBSfM). The research presents a practical algorithm for multi-body autocalibration, which enhances the accuracy of camera parameter estimation under dynamic conditions. Furthermore, a zero-shot learning-based framework is proposed for depth and camera pose estimation in multi-body scenes, resolving the relative scale ambiguity common in MBSfM. This is achieved using monocular depth estimators to ensure a consistently scaled reconstruction across the entire scene.

Contents

CHAPTER 1

Introduction

Three-dimensional (3D) reconstruction from multiple images, whether from sparse image sets or video sequences, is a fundamental challenge in modern Computer Vision, with applications in robotics, photogrammetry, augmented reality (AR), and virtual reality (VR). Known as Structure-from-Motion (SfM), this problem focuses on two main objectives: camera motion estimation and 3D reconstruction. Camera motion estimation involves computing the camera's trajectory relative to the scene. At the same time, 3D reconstruction aims to recover the scene's three-dimensional geometry, including the shape, structure, and spatial layout of objects from multiple two-dimensional images.

Structure-from-Motion is a complex problem that depends on several critical components functioning correctly, with any failure in the pipeline potentially leading to inaccurate reconstructions. This is especially true when performing SfM "in-the-wild," meaning in uncontrolled, real-world environments where factors such as unknown or varying camera parameters, moving objects, changing lighting conditions, and occlusions can complicate the process. Among the most crucial components for successful SfM is *camera autocalibration*, which involves determining the camera's intrinsic parameters—such as focal length, principal point, skew, and distortion coefficients—without relying on external calibration objects or known scene geometry. These intrinsic parameters are essential for unprojecting 2D image coordinates, enabling the triangulation of three-dimensional points and ultimately leading to an accurate reconstruction of the scene's geometry.

Accurate camera calibration is essential for Structure-from-Motion because inaccuracies in the estimated camera intrinsics can introduce errors in both 3D point locations and camera motion estimations. These errors can propagate throughout the reconstruction process, leading to significant artifacts, inaccuracies, or even the complete

failure of the reconstruction. Although global refinement techniques like Bundle Adjustment [77] can mitigate calibration errors and produce accurate estimates of intrinsic camera parameters, they often struggle to recover from incorrect initial reconstructions. This difficulty can cause the optimization process to converge to a local minimum that does not accurately represent the true 3D geometry and camera motion. Therefore, robust and precise camera autocalibration algorithms are fundamental for achieving reliable SfM results.

Camera autocalibration becomes even more critical when performing *in-the-wild* Structure-from-Motion—reconstructing a scene from images or videos captured in uncontrolled environments and/or from unknown sources, such as those found on the Internet. In many cases, these images lack associated intrinsic and extrinsic camera parameters. Without access to the original camera device, traditional calibration methods [14, 17–19, 25, 33, 49, 50, 57, 60, 71, 76] cannot be employed to obtain the camera intrinsics, making *autocalibration* algorithms indispensable for successful SfM in such contexts.

In this work, we advance the state of the art in camera autocalibration by introducing a family of novel solvers specifically designed for *pinhole* camera autocalibration. The proposed approach improves the accuracy and robustness of existing methodologies while addressing many of the well-known degeneracies and failure cases that limit the applicability of traditional autocalibration methods, such as those based on Kruppa's equations [45] or the modulus constraint [59], in real-world scenarios. Unlike conventional approaches that perform a projective-to-metric upgrade of a given reconstruction using pairwise image constraints—typically encoded in fundamental matrices—our method, for the first time, jointly performs camera autocalibration and metric reconstruction using only a minimal set of pairwise keypoint correspondences between arbitrary image frames. Our autocalibration solvers not only deliver provably more accurate reconstructions and intrinsic camera parameter estimates than existing methods, but are also highly efficient, optimized through a detailed analysis of the most effective autocalibration formulations based on algebraic geometry principles. Furthermore, when integrated into the widely used COLMAP [68, 69] SfM software, these solvers improve both the reconstruction quality and the number of 3D points correctly registered during the reconstruction process.

Additionally, recognizing that our initial family of autocalibration solvers focuses solely on pinhole camera calibration–*i.e.*, non-distorted cameras–we have developed a novel learning-based zero-shot autocalibration method for estimating the distortion parameters of radially symmetric cameras. This method also facilitates the rectification of distorted images. Radially symmetric cameras encompass a wide range of devices, including most fisheye and 360-degree sensors, which are increasingly popular due to their ability to support a larger field-of-view (FOV), particularly in applications like augmented reality (AR) and autonomous driving. This method, akin to other image rectification techniques, can also serve as a preprocessing step for distorted images before applying our pinhole autocalibration solvers for precise calibration and metric reconstruction.

As discussed above, addressing the Structure-from-Motion problem also requires accounting for moving objects in the scene, which is often the case in many real-world, in-the-wild scenarios. Traditionally, SfM has been been designed to reconstruct

static scenes, with algorithms assuming that all motions other than the dominant one–typically defined as the motion covering the majority of the image area–act as outliers and hinder the reconstruction process. Consequently, these algorithms often fail to capture the dynamic parts of the scene, resulting in incomplete reconstructions that are of limited use in many practicaly applications where reconstructing moving objects, such as cars, pedestrians, or animals, is crucial.

To address these limitations, the problem of SfM has been extended to Multi-Body Structure-from-Motion (MBSfM) [2, 13, 35, 55, 67, 75]. MBSfM aims to achieve consistent 3D reconstruction in scenes with multiple independently moving *rigid* bodies. Emphasizing rigid motion is important, as MBSfM is specifically designed for this task and not for reconstructing non-rigid motion. Reconstructing non-rigid motion typically requires temporally consistent input frames, such as those extracted from a continuous video source, or specific assumptions about the camera setup, such as the use of orthographic cameras.

In this work, we advance the state-of-the-art of Multi-Body Structure-from-Motion from two perspectives. First, we tackle the problem of camera autocalibration in MBSfM. Historically, Structure-from-Motion has focused primarily on static scene reconstruction, and as a result, camera autocalibration algorithms have followed suit, assuming the scene is static and treating rigidly moving objects as outliers. However, A. Fitzgibbon and A. Zisserman [16] theorized that since each rigid displacement observed across pairs of images can be used to estimate a fundamental matrix encoding the rigid motion, and since each fundamental matrix can be used to derive two Kruppa equations that constrain the intrinsic camera parameters, segmenting and leveraging the rigid motions in the scene could actually improve calibration accuracy. This is due to the additional constraints provided by these motions, which could also reduce the minimum number of images required to achieve full camera calibration.

In this thesis, we address the challenge of multi-body autocalibration and develop a practical algorithm that improves the accuracy of estimated intrinsic parameters under dynamic scene conditions, demonstrated through both synthetic and real-world examples. Furthermore, we validate the theoretical foundations introduced in [16] by successfully calibrating the full set of camera intrinsics using just two displaced image frames, a reduction from the theoretical lower bound of three displaced frames required under the static scene assumption.

Second, we develop a zero-shot learning-based framework for depth and camera pose estimation, specifically designed for multi-body scenes from calibrated, unstructured image sets. The proposed method addresses a key challenge in Multi-Body Structure-from-Motion (MBSfM): the *relative scale problem*. In traditional MBSfM [2, 13, 75], rigid motions are segmented using motion segmentation techniques, and then each motion is reconstructed independently. However, since each reconstruction is metric, each individual motion is reconstructed in its own scale, making it impossible to achieve a consistently scaled reconstruction across the entire scene.

Our method resolves this scale ambiguity by employing monocular depth estimators and their learned priors to predict depths from single images. These estimators are not affected by the relative scale ambiguity that plagues multi-view approaches, allowing us to achieve a consistently scaled reconstruction across the entire scene. Finally, we introduce a multi-view depth and camera pose refinement network, based on a novel

revisitation of the *plane sweep* [12] algorithm adapted to multi-body scenes. This network refines the initial depth and camera pose estimates generated by the monocular depth estimators, leading to improved accuracy in the final reconstruction.

In summary, this thesis advances Structure-from-Motion by addressing critical challenges in camera autocalibration and multi-body scene reconstruction. By integrating geometric and deep learning approaches, it aims to enhance the robustness and applicability of SfM algorithms, paving the way for more reliable and adaptable applications in real-world, especially *in-the-wild*, environments.

In the following sections and embedded works, we will explore key aspects of camera autocalibration and Structure-from-Motion, and provide an in-depth discussion of the novel algorithms and methods presented in this thesis to address the following long-standing challenges: (i) the process of camera autocalibration within the SfM framework (Sec. 1.1), and (ii) the reconstruction of rigidly moving objects in dynamic scenes, known as Multi-body Structure-from-Motion (Sec. 1.2).

1.1 Autocalibration in Structure-from-Motion

This thesis begins by exploring foundational algorithms in 3D Computer Vision, focusing on minimal problems in camera autocalibration and metric reconstruction. Minimal problems involve determining the smallest number of data points or measurements required to solve a specific task in Vision. Many of these tasks require solving polynomial systems consisting of equations encoding geometric constraints related to the 3D scene. These polynomial systems may be solved using symbolic or numerical methods when symbolic solvers cannot run in reasonable computational time.

In this thesis, we develop practical and efficient numerical solvers for minimal autocalibration by introducing a novel depth formulation that extends the minimal Euclidean reconstruction problem of four points in three calibrated views [27, 63] to the uncalibrated case. Employing tools from algebraic geometry, we develop a general theory of minimal relaxations to address instances where our formulation leads to an over-constrained problem while also identifying which minimal relaxations yield the most efficient autocalibration formulations that can be solved quickly using Homotopy Continuation (HC) methods. Starting from matched 2D points across image pairs or triplets, our solvers simultaneously recover the camera's intrinsic parameters—such as focal lengths along the x- and y-axes, principal point coordinates, camera skew, and distortion parameters—while also determining the 3D structure of the scene through the projective depths of the observed 2D points.

Our primary autocalibration solver can perform *full* camera autocalibration from six-point correspondences across three views, recovering the projective depths of the six points in the three views. Based on prior knowledge about the camera, *i.e.*, which unknown intrinsic camera parameters are unknown, and the number of views available, we have developed a family of specialized numerical solvers, each optimized to run as efficiently as possible through our proposed theory of minimal relaxations. These solvers require as few as four, five, or six matched pixels in the input images.

We demonstrate that integrating our minimal solvers into the widely-used COLMAP Structure-from-Motion software improves the accuracy of estimated intrinsic camera parameters while reducing the overall reprojection error in the 3D reconstruction of

scenes. Besides improving COLMAP, these solvers have the potential to be embedded into future geometric and learning-based depth and camera pose estimation algorithms, advancing the field of Structure-from-Motion.

Nonetheless, our minimal autocalibration solvers are limited to *pinhole* camera calibration and do not support wide field-of-view (FOV) cameras, including fisheye lenses, 360-degree, and spherical cameras. While most of the traditional autocalibration methods generally assume an underlying pinhole camera model, which does not account for radial distortions, the growing use of wide FOV cameras has created a need for autocalibration processes capable of handling the distortions inherent to these larger FOVs. Conventional geometric-based calibration methods, which rely on structured environments and calibration objects, often perform poorly in unstructured or dynamic settings. This challenge has driven the development of learning-based calibration and image rectification methods that leverage learned image priors rather than handcrafted features, improving the generalizability of calibration algorithms in diverse, uncontrolled, real-world scenarios.

Recent advancements in learning-based calibration methods have shown promise in addressing these challenges by using convolutional neural networks (CNNs) to estimate camera parameters directly from images. These approaches offer the advantage of not requiring specific calibration objects and can operate effectively in uncontrolled environments. However, most current learning-based methods are tied to fixed camera models, leading to several significant limitations: a lack of flexibility when switching between different camera models, reduced accuracy due to limited training data diversity, and constraints in selecting suitable camera models due to the need for differentiable loss functions with closed-form undistortion equations.

This thesis introduces a novel two-step calibration framework for radially symmetric cameras to overcome these limitations. The core innovation of the proposed framework is the introduction of the VACR, short for *Viewing-angle Camera Representation*, an implicit camera representation that can be regressed by a specialized CNN that we specifically developed for such task. The VACR maps each image point to the direction of the 3D light ray projecting onto it, independent of the underlying mathematical camera model and parameters. The second step involves a robust non-linear optimization process that uses the VaCR to determine the camera parameters for any radially symmetric model, thus bypassing network retraining when switching between camera models.

Our experimental evaluation shows that our proposed two-step calibration framework generally outperforms state-of-the-art geometric- and learning-based calibration and image rectification methods when using an equivalent camera model. Additionally, we demonstrate that our method can adapt to different camera models without requiring network retraining, depending on the dataset, often yielding improved performance.

In summary, this thesis contributes to the field of camera autocalibration for both undistorted pinhole cameras and wide-angle distorted cameras, enabling the recovery of intrinsic camera parameters for a wide range of cameras, and in-the-wild reconstructions where these parameters may not be available.

1.2 Multi-body Structure-from-Motion

When performing Structure-from-Motion (SfM) using images sourced from real-world videos or those downloaded from the Internet—essentially, in uncontrolled real-world environments—SfM often encounters challenges due to moving objects within the scene. These objects move relative to the static environment, which is a common scenario in various real-world applications of SfM, including autonomous driving, robotics, virtual and augmented reality, in-the-wild photogrammetry, and visual localization.

Most traditional and learning-based SfM methods are primarily designed for *static* scenes, where the constraints of multi-view (epipolar) geometry apply uniformly across the entire scene, assuming a single relative camera pose between view pairs. However, when objects within the scene move relative to the static environment, they generate a distinct relative camera motion in the reference frame of the moving object, causing the epipolar constraints to no longer be valid for the image regions containing these moving objects. Consequently, moving objects are treated as outliers in the reconstruction process and must be excluded from the static scene reconstruction. In many cases, particularly when moving objects occupy a significant portion of the image, or when robustness is not adequately enforced in the SfM pipeline, this can result in suboptimal performance, with significant reconstruction artifacts and inaccuracies in depth and pose estimation.

To address these challenges, the concept of Multi-Body Structure-from-Motion (MBSfM) has been introduced. MBSfM is designed to handle scenes containing objects that move rigidly relative to the background, enabling the reconstruction of 3D geometry for both the static background and the moving objects. It ensures consistent scaling throughout the reconstruction, accurately positioning objects in the 3D world and correctly reflecting their volumetric properties.

Previous approaches to MBSfM have attempted to segment rigid motions to achieve partial reconstructions of each moving object. However, these methods are limited by the inherent scale ambiguity in SfM, where objects are reconstructed at inconsistent scales, preventing a unified reconstruction without additional information about object scales. This issue of scale ambiguity persists across both traditional geometric methods and modern deep learning frameworks, significantly reducing their effectiveness in dynamic scenarios.

This thesis introduces a novel framework for depth and camera pose estimation specifically designed for multi-body scenes to address these challenges. The key innovation of this method is a robust scale estimator that operates without relying on semantic priors or known scene geometry. Instead, it uses monocular depth estimates from each image frame intended for reconstruction and employs a robust voting scheme with a kernel density estimator to jointly determine the most likely scale for each segmented object in the scene. The estimated scale factors, which, when applied to each object, enable a consistently scaled reconstruction throughout the entire scene, are then used to adjust the camera poses. These adjusted poses are subsequently fed into a pose and depth refinement network that is specifically designed for multi-body scenes. This network refines the initial monocular depth estimates and the adjusted pose estimates to produce depth maps and poses that are not only accurate but also geometrically consistent across the entire reconstruction, ensuring that all objects (or bodies) in the scene

are reconstructed within a common reference frame and at a consistent scale.

The core contributions of this work are twofold. First, it introduces a robust scale estimation technique that utilizes monocular depth maps to resolve scale ambiguity, thereby enabling consistent 3D reconstruction across multiple objects. Second, it presents a multi-body plane sweep network that refines both extrinsics and dense depth estimates, ensuring that all rigidly moving objects in the scene are accurately accounted for in the reconstructed environment.

Experimental results on various datasets, including those with static and dynamic scenes, demonstrate the superior performance of the proposed framework. The method achieves higher accuracy in depth estimation and camera pose recovery than existing single-view and multi-view approaches. By effectively leveraging multi-view constraints and resolving scale ambiguities, the proposed framework sets a new benchmark for depth and camera pose estimation in multi-body scenarios, contributing to advancements in autonomous navigation and augmented reality.

When addressing Multi-Body Structure-from-Motion, most autocalibration methodologies, including the minimal solvers presented earlier, excel in static scenes but struggle in environments with moving objects. The reason is that moving objects do not satisfy the epipolar constraints leveraged by multi-view calibration algorithms, leading to their classification as outliers. Treating dynamic objects as outliers reduces the number of samples available for calibration and reconstruction, potentially limiting the accuracy of the final reconstruction or, in particularly challenging cases, making reconstruction unfeasible.

To address the issue of autocalibration in the presence of rigidly moving objects, we introduce the concept of *multi-body autocalibration*, designed for a moving camera capturing dynamic scenes with multiple rigidly moving objects. Our approach leverages motion segmentation to identify and utilize these independent motions, thereby improving the estimation of the camera's intrinsic parameters. A key innovation of our method is a nonlinear optimization scheme that enhances robustness against inaccurate measurements from incorrectly detected moving objects and critical motion sequences. By incorporating information from multiple moving objects, our method uses each rigid motion as an additional constraint, thereby enhancing the accuracy of camera intrinsics estimation.

The proposed multi-body autocalibration method has been tested against state-of-the-art autocalibration techniques in dynamic scenes, where significant improvements were observed, with our method outperforming existing approaches by effectively leveraging multiple motions within the scene. In static scenes, our method compares favorably to the state-of-the-art, underscoring the robustness and versatility of the proposed method and marking a substantial advancement in the field of camera autocalibration.

1.3 Structure of the Thesis

This thesis is structured as a collection of articles, which, as described in the introduction, present the novel technical solutions brought forward in this work.

The following list of articles is included, ordered by publication date:

- Andrea Porfiri Dal Cin, Giacomo Boracchi, Luca Magri.
  "Multi-body Self-Calibration."

In: 33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022. BMVA Press. 2022, pp. 1–14.

- Andrea Porfiri Dal Cin, Giacomo Boracchi, and Luca Magri.
  "Multi-body depth and camera pose estimation from multiple views."
  Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.
- Andrea Porfiri Dal Cin, Timothy Duff, Luca Magri and Tomas Pajdla.
  "Minimal Perspective Autocalibration."
  Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.
- Andrea Porfiri Dal Cin, Giacomo Boracchi, and Luca Magri.
  "Revisiting Calibration of Wide-Angle Radially Symmetric Cameras."
  Proceedings of the European Conference on Computer Vision (ECCV). 2024.

CHAPTER 2

State of the Art

In this chapter we will provide an in-depth overview of the state-of-the-art in the fields of camera autocalibration and Structure-from-Motion, with a particular emphasis on methods aimed at Multi-Body Structure-from-Motion, while also discussing recent developments into non-rigid and monocular depth estimation for moving object, may them rigid or non-rigid. As camera autocalibration for wide-angle and distorted cameras is often addressed with different methodologies and ad-hoc solutions with respect to the calibration of pinhole cameras, we will provide two sections with the goal of providing a more organized taxonomy of the state-of-the-art. This section is aimed at providing a comprehensive overview of the state-of-the-art with the goal of highlighting advantages and limitations of different methodologies to the problems tackled in our work.

2.1 Autocalibration of Pinhole Cameras

In this thesis, we primarily focus on the classical scenario where a single camera captures the scene, or equivalently, where multiple cameras capturing the scene share the same intrinsic parameters K. Historically, autocalibration methods have been categorized into two categories: *direct* and *stratified* methods. In this section, we will explore this taxonomy and describe prominent works in detail.

2.1.1 Direct Autocalibration

Direct methods use the so-called rigidity constraint encoded in fundamental matrices. In theory, K can be recovered from the knowledge of three fundamental matrices resulting from three different camera motions [14, 50]. Direct methods [23, 46, 89] exploit this observation and recover the intrinsic parameters by solving Kruppa's equa-

tions [19, 33]. As methods used to solve the Kruppa equations vary considerably, we will discuss works in the literature that are particularly relevant to our work.

In [46], Luong and Faugeras address the problem of calibrating a moving camera and estimating its three-dimensional motion using point correspondences between multiple images. The main contribution is a method that does not require any calibration object or prior knowledge of the camera's internal parameters. Instead, it uses the fundamental matrices derived from point correspondences to recover the camera's intrinsic parameters, motion, and to reconstruct the 3D structure of the scene. The method hinges on the properties of the fundamental matrix F, which encapsulates the epipolar geometry between pairs of images. The fundamental matrix is computed from point correspondences and is related to the camera matrices through the essential matrix E for calibrated cameras, where $E = K^\top F K$. Here, K is the intrinsic parameter matrix of the camera. The intrinsic parameters include the focal lengths (f, g), the principal point (u, v), and a skew coefficient s, forming the matrix K as follows:

$$K = \begin{pmatrix} f & s & u \\ 0 & g & v \\ 0 & 0 & 1 \end{pmatrix} \tag{2.1}$$

The self-calibration process uses Kruppa's equations, which relate the intrinsic parameters to the fundamental matrix. Given three views of a static scene, the method establishes sufficient constraints to solve for the camera's intrinsic parameters and motion parameters. The Kruppa equations are derived from the dual absolute quadric, a geometric construct that remains invariant under projective transformations. The constraints from multiple views are combined to form a system of equations that can be solved iteratively or through continuation methods. Specifically, instead of considering the complete over-constrained system of six Kruppa equations in the five unknowns that are the five entries of the Dual Image of the Absolute Conic. The authors propose to solve all six square subsystems of five equations in five unknowns and obtain the solution to the parameters of the Dual Image of the Absolute Conic by maximizing consensus. This estimate is refined using non-linear optimization techniques that minimize the geometric error of point correspondences. The intrinsic parameters are then obtained by decomposing the Dual Image of the Absolute Conic via Cholesky decomposition. As discussed later in this work, a common weakness of such direct approaches is that they do not enforce positive-semidefiniteness of the dual image of absolute conic and hence fail when noise in point correspondences and, hence, in the fundamental matrix, is large, as the estimated parameters are also noise and may break the positive-semidefiniteness requirement of the Cholesky decomposition, or, in other words, make the estimated Dual Image of the Absolute Conic indefinite. Overall, the method described in [46] is able to accurately recover the intrinsic parameters and a 3D reconstruction up to a similarity transformation, as evidenced by the simulations and real-world experiments.

In [89], Zeller and Faugeras focus on improving camera self-calibration using video sequences through a detailed revisitation of the Kruppa equations. The authors present an improved approach to utilize these equations effectively for camera calibration, in which the over-constrained system of Kruppa's equations is solved with a nonlinear least squares technique; here, good initialization is needed to obtain an accurate esti-

mate. The primary steps and concepts involved in this method are as follows.

The intrinsic parameters K of the camera, which include the focal length, principal point, and skew coefficient, are defined as follows:

$$K = \begin{pmatrix} f & s & u \\ 0 & \gamma f & v \\ 0 & 0 & 1 \end{pmatrix}$$

where f is the focal length, (u, v) is the principal point, s is the skew coefficient, and $gamma$ is the aspect ratio. The relationship between two views of the same scene is described using the fundamental matrix F, from which the Kruppa equations are derived. The steps in the calibration process are:

(a) *Feature Detection and Matching*, where corresponding points across multiple frames in the video sequence are identified. These correspondences are essential for constructing the epipolar geometry.

(b) *Estimation of Fundamental Matrix*, where using the matched points, the fundamental matrix F is estimated. This estimation can be done using methods like the eight-point algorithm.

(c) *Application of Kruppa Equations*, where the Kruppa equations are applied to the estimated fundamental matrices from different pairs of frames. The Kruppa equations relate the elements of F to the intrinsic parameters K. The Kruppa equations can be expressed as:

$$\begin{aligned} K_2^{-T} F K_1^{-1} &= \begin{bmatrix} e_1' & e_2' & e_3' \end{bmatrix} \\ \text{Let } F &= \begin{bmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \end{bmatrix} \end{aligned} \tag{2.2}$$

$$\text{Let } K_1 = \begin{bmatrix} \alpha_1 & s_1 & u_1 \\ 0 & \beta_1 & v_1 \\ 0 & 0 & 1 \end{bmatrix}, \quad K_2 = \begin{bmatrix} \alpha_2 & s_2 & u_2 \\ 0 & \beta_2 & v_2 \\ 0 & 0 & 1 \end{bmatrix} \tag{2.3}$$

The Kruppa equations can then be written as:

$$\begin{aligned} &\left(\frac{f_{11}}{\alpha_1\alpha_2} - \frac{s_1 f_{21}}{\alpha_1\beta_2} - \frac{s_2 f_{12}}{\beta_1\alpha_2} + \frac{s_1 s_2 f_{22}}{\beta_1\beta_2} \right)^2 + \\ &\left(\frac{f_{13}}{\alpha_1\alpha_2} - \frac{s_1 f_{23}}{\alpha_1\beta_2} - \frac{u_2 f_{12}}{\beta_1\alpha_2} + \frac{s_1 u_2 f_{22}}{\beta_1\beta_2} - \frac{s_2 f_{13}}{\beta_1\alpha_2} + \frac{s_2 u_1 f_{22}}{\beta_1\beta_2} - \frac{u_1 f_{23}}{\alpha_1\beta_2} + \frac{u_1 u_2 f_{33}}{\beta_1\beta_2} \right)^2 + \\ &\left(\frac{v_1 f_{21}}{\alpha_1\beta_2} - \frac{v_1 s_2 f_{22}}{\beta_1\beta_2} + \frac{v_2 f_{12}}{\beta_1\alpha_2} - \frac{v_2 s_1 f_{22}}{\alpha_1\beta_2} + \frac{v_1 v_2 f_{33}}{\beta_1\beta_2} \right)^2 + \\ &2 \left(\frac{f_{11}}{\alpha_1\alpha_2} - \frac{s_1 f_{21}}{\alpha_1\beta_2} \right) \left(\frac{f_{12}}{\beta_1\alpha_2} - \frac{s_2 f_{22}}{\beta_1\beta_2} \right) + \\ &2 \left(\frac{f_{13}}{\alpha_1\alpha_2} - \frac{s_1 f_{23}}{\alpha_1\beta_2} \right) \left(\frac{f_{23}}{\beta_1\alpha_2} - \frac{u_2 f_{22}}{\beta_1\beta_2} \right) = 0 \end{aligned} \tag{2.4}$$

These equations provide a system of polynomial equations in terms of the intrinsic parameters.

(d) *Optimization*, where the system of polynomial equations is solved using numerical optimization techniques to find the best estimates of the intrinsic parameters. This step may involve iterative methods to minimize the error between the predicted and observed correspondences.

(e) *Validation and Refinement*, where the obtained intrinsic parameters are validated by reprojecting the 3D points into the image plane and comparing with the original points. If necessary, the parameters are refined by iteratively adjusting them to minimize reprojection errors.

The revisited Kruppa equations offer a robust framework for camera self-calibration, leveraging the inherent geometric constraints in video sequences. By systematically applying these equations and optimizing the intrinsic parameters, the method achieves accurate camera calibration.

Not all direct methods use Kruppa's equations—in [48], a method analogous to the F4 method for computing Grobner bases is devised for computing the DIAC. Specifically, in [48], a novel non-iterative autocalibration algorithm is presented, which requires only a minimal set of six scene points in three views taken by a camera with fixed but unknown intrinsic parameters. The method is divided into two main stages: projective reconstruction and metric reconstruction. First, the algorithm employs the six-point three-view algorithm to achieve a projective reconstruction of the scene. Given three uncalibrated images of six points from a rigid scene, the method applies the 3-view algorithm to solve for the homogeneous coordinates of the sixth scene point, X_6, while the first five points are set as the standard basis vectors of the projective 3-space. The camera matrices P_i are then recovered by solving twelve linearly independent equations:

$$\mathbf{x}_{ij} \times P_i \mathbf{X}_j = \mathbf{0}_3, \quad j = 1, \ldots, 6,$$

where $\mathbf{x}_{ij}$ is the image of $\mathbf{X}_j$ under the projection P_i. Next, the projective camera matrices are transformed to a specific form using projective ambiguity:

$$P_1' = P_1 H_0 = \begin{bmatrix} I_3 & \mathbf{0}_3 \end{bmatrix}, \quad P_2' = P_2 H_0 = \begin{bmatrix} B_2 & \mathbf{b}_2 \end{bmatrix}, \quad P_3' = P_3 H_0 = \begin{bmatrix} B_3 & \mathbf{b}_3 \end{bmatrix},$$

with

$$H_0 = \begin{bmatrix} A_1^{-1} & -A_1^{-1}\mathbf{a}_1 \\ \mathbf{0}_3^T & 1 \end{bmatrix}.$$

In the metric reconstruction stage, the goal is to upgrade the projective cameras to metric cameras. The metric camera matrices are represented as:

$$P_1^M = K \begin{bmatrix} I_3 & \mathbf{0}_3 \end{bmatrix}, \quad P_2^M = K \begin{bmatrix} R_2 & \mathbf{t}_2 \end{bmatrix}, \quad P_3^M = K \begin{bmatrix} R_3 & \mathbf{t}_3 \end{bmatrix},$$

where R_i and $\mathbf{t}_i$ are the rotation matrix and translation vector, respectively, and K is the upper triangular calibration matrix. The projective matrix H that transforms the projective cameras to metric ones must satisfy:

$$P_i^M = P_i' H, \quad i = 1, 2, 3.$$

This H matrix has the form:

$$H = \begin{bmatrix} K & \mathbf{0}_3 \\ -\mathbf{p}^T K & 1 \end{bmatrix}.$$

The constraints for H give rise to a system of non-linear polynomial equations involving the dual image of the absolute conic $\omega^* = KK^T$ and other parameters. This system is solved using a series of Gauss-Jordan eliminations with partial pivoting to ensure numerical stability. The sequence of matrix transformations reduces the system to a form solvable for the calibration matrix K. The algorithm is validated through experiments on synthetic data, showing high accuracy and robustness even under noise and outliers. Numerical results demonstrate the median error and the algorithm's computational efficiency, making it suitable for real-time applications where camera calibration is critical.

We conclude our discussion of relevant direct autocalibration methods by noting that certain camera motions can cause degenerate autocalibration problems [**?**], [22, Ch.19]. Additionally, specific methods, particularly those based on Kruppa's equations, may exhibit further degeneracies. For example, the method described in [46] is inadequate when the optical centers of all cameras lie on a sphere and the optical axes pass through the center of the sphere [72].

2.1.2 Stratified Autocalibration

Stratified autocalibration methods assume that a projective reconstruction is known and stratify the problem into Affine and Euclidean stages. An affine reconstruction can be obtained by estimating the Plane-at-Infinity (PaI); from this, the assumption of constant K allows its entries to be easily retrieved.

This idea was pioneered in [24], where chirality constraints are used to estimate the location of the PaI. Specifically, Hartley's work presents a practical algorithm for Euclidean reconstruction from multiple uncalibrated views of a scene. The method leverages projective geometry and robust numerical techniques to achieve camera calibration and scene reconstruction using matched image points from multiple views. The camera model uses projective mapping from 3D space ($\mathbb{P}^3$) to 2D space ($\mathbb{P}^2$), represented by a 3×4 matrix M. This matrix can be decomposed into $M = K[R| - Rt]$, where R is a rotation matrix, t is the camera position, and K is the calibration matrix. The calibration matrix K is upper triangular and includes internal camera parameters described in Eq. (2.1). The goal is to determine the camera matrices M_i and 3D points x_j from their image projections u_{ij}. A Euclidean reconstruction ensures that the structure differs from the true scene by at most a similarity transform. For cameras with the same calibration, the reconstruction problem can be solved up to an unknown scaling factor. The Levenberg-Marquardt (LM) algorithm is employed to minimize the reprojection error. The method iteratively optimizes the camera parameters and 3D points to fit the observed image points. A two-step approach is adopted: first, an initial estimation of camera matrices and 3D points without assuming the same camera calibration across views (projective reconstruction); second, using projective invariants and constraints, the projective reconstruction is transformed to a Euclidean one. To achieve Euclidean reconstruction, a transformation matrix H is sought such that the transformed camera matrices $M_i H^{-1}$ have the same calibration matrix K. This involves solving the equa-

tion $A_i(I + t_i v^T)K \approx KR_i$ through an optimization process. The Jacobian matrix in the LM algorithm has a sparse structure due to the block-wise dependency of image points on specific camera parameters and 3D points. This sparsity is exploited to simplify and accelerate the computation, making the algorithm scalable to large datasets. The algorithm was tested on both synthetic and real data, demonstrating robustness to noise and accuracy in camera calibration and 3D reconstruction. It was shown to work well even with significant noise and multiple views, providing reliable Euclidean reconstructions.

The Plane-at-Infinity can also be located via the so-called *modulus constraints*. Specifically, in [61], Pollefeys and Van Gool introduce a method for autocalibration that incrementally achieves metric calibration of a camera setup from an uncalibrated image sequence. The approach is stratified, beginning with projective calibration, followed by affine calibration, and culminating in metric calibration. The camera model is described using projective geometry, where the projection of a 3D point M onto a 2D image point m is given by $m \propto PM$, with P as the 3×4 projection matrix. In the metric case, the projection matrix P can be factored into $P = K[R| - Rt]$, where K is the intrinsic parameter matrix. The method starts with projective calibration, which computes a projective reconstruction of the scene. This involves estimating the fundamental matrix F between pairs of images, robustly computed from point correspondences. The projection matrices for the first two views are obtained using the homography H for an arbitrary reference plane, yielding initial projection matrices P_1 and P_2. Subsequent views are added by ensuring that reconstructed points from previous views are reprojected as closely as possible to their matches in the new view. Affine calibration identifies the Plane-at-Infinity, which is crucial for transitioning from projective to affine space. The Plane-at-Infinity is characterized by the modulus constraint, which states that the homography for this plane must be conjugated to a rotation matrix. This constraint is expressed mathematically as:

$$H_\infty R H_\infty^T = R,$$

where H_∞ is the infinity homography. This results in a system of three quartic polynomials on the coefficient of the Plane-at-Infinity. By solving for the Plane-at-Infinity using vanishing points or other geometric features, the method upgrades the projective reconstruction to an affine reconstruction. Metric calibration then involves identifying the absolute conic, which is invariant under Euclidean transformations. The absolute conic's image, known as the Dual Image of the Absolute Conic (DIAC), is represented as $\omega = K^{-T}K^{-1}$. The calibration process ensures that the DIAC remains consistent across all views, allowing the recovery of the intrinsic parameters. This is achieved through the relationship:

$$\omega_i = P_i \Omega^* P_i^T,$$

where Ω^* is the absolute quadric. To solve the self-calibration equations, the authors employ non-linear least squares optimization, which minimizes the reprojection error across all views. The process iteratively refines the estimates of the intrinsic and extrinsic parameters until convergence.

In general, stratified approaches are more robust to noise than direct ones but require good initialization of the Plane-at-Infinity. Thus, several works [8, 9, 18, 58] focus on optimality guarantees exploiting a branch-and-bound framework. Particularly, [58]

proposes a method that leverages the Branch-and-Bound (BnB) search paradigm to maximize the consensus of polynomials parameterized by entries of either the Dual Image of the Absolute Conic or the Plane-at-Infinity. During the BnB search, the theory of sampling algebraic varieties is exploited to test the positivity of polynomials within a parameter's interval, effectively identifying outliers. The camera model used in the paper assumes constant intrinsic parameters across multiple views, represented by the calibration matrix K of Eq. (2.1). The authors address the challenges of autocalibration due to the non-linear nature of the problem and the presence of Critical Motion Sequences, which can lead to degenerate motions. Existing stratified autocalibration methods fall into two categories: (i) stratified estimation of the Plane-at-Infinity followed by linear retrieval of the Dual Image of the Absolute Conic, and (ii) joint estimation of both Dual Image of the Absolute Conic and Plane-at-Infinity. As opposed to most methods that are locally optimal and sensitive to outliers, the method presented in [58] falls into the category of globally optimal methods, which are computationally expensive and often impractical for long image sequences. [58] addresses the issues of globally optimal methods by integrating BnB with algebraic geometry to maximize polynomial consensus. The polynomials are derived from simplified Kruppa's equations and Modulus constraints. Kruppa's equations relate the DIAC to the fundamental matrix F between two views, while Modulus constraints involve the homography between two images induced by the PaI. The key contribution of the paper is the use of sampling algebraic varieties to detect outlier polynomials with certainty. This involves establishing optimistic and pessimistic sets of inliers, where the positivity of interval polynomials on the given varieties indicates outliers. The authors implement a pre-certificate computation and a poisedness test to ensure the correctness of the outlier detection. The BnB algorithm iteratively refines the search space by branching on parameter intervals, with the goal of maximizing the number of inliers. The local refinement method helps in estimating the pessimistic inlier set, while the outlier test discards measurements that cannot be inliers. The method has been tested on synthetic and real datasets, demonstrating robustness and optimality even with a high number of outliers. Experimental results show that the method in [58] effectively handles up to 90% outliers, consistently detecting the correct number of inliers and providing accurate camera intrinsics. The algorithm's convergence is observed to be efficient, typically within 1000 iterations, and it maintains reasonable memory usage. The method outperforms traditional RANSAC approaches, especially in the presence of many outliers, and yields accurate metric reconstructions from projective reconstructions. For these reasons, [58] will be compared to our proposed autocalibration methodologies as a benchmark for state-of-the-art accuracy and robustness in pinhole camera autocalibration.

2.1.3 Autocalibration with Varying Camera Intrinsics

As previously discussed, this work predominantly assumes that K is constant across views. For a complete presentation of the state-of-the-art, we discuss prominent works in autocalibration of cameras with varying intrinsics K.

In [60], Pollefeys, Koch, and Van Gool investigate the feasibility of autocalibrating cameras with varying intrinsic parameters. The authors propose a versatile method that accommodates different constraints on intrinsic parameters, allowing for metric

reconstruction from uncalibrated image sequences. The camera model used in this study follows the perspective projection described by:

$$m \propto PM$$

where P is a 3×4 projection matrix, M represents the 3D world points in homogeneous coordinates, and m represents the 2D image points in homogeneous coordinates. In the metric case, the projection matrix P factorizes as:

$$P = K[R| - Rt]$$

where K is the intrinsic parameter matrix:

$$K = \begin{pmatrix} f & s & u \\ 0 & g & v \\ 0 & 0 & 1 \end{pmatrix}$$

where R is a rotation matrix and t is a translation vector, representing the camera's extrinsic parameters. A significant aspect of the method involves the use of the absolute conic and the absolute quadric. The absolute conic, embedded in the Plane-at-Infinity, is invariant under Euclidean transformations and is directly related to the intrinsic parameters. The Dual Image of the Absolute Conic (DIAC) is used to impose constraints on these parameters. For each view i, the DIAC is expressed as:

$$\omega_i^* = K_i K_i^T \propto P_i \Omega^* P_i^T$$

where Ω^* represents the absolute quadric. The self-calibration process is broken down into the following steps:

(a) *Projective Reconstruction*: An initial projective reconstruction is obtained using point correspondences to compute the fundamental matrix F. The projection matrices are derived, typically starting with $P_1 = [I|0]$.

(b) *Affine Calibration*: The Plane-at-Infinity (π_∞) is identified using the modulus constraint, which requires the infinity homography to be conjugate to a rotation matrix. The constraint is expressed as:

$$H_\infty R H_\infty^T = R$$

This step upgrades the projective reconstruction to an affine reconstruction.

(c) *Metric Calibration*: The absolute conic is localized using the DIACs, transforming the affine reconstruction to a metric one. The relationship:

$$\omega_i^* = K_i K_i^T \propto P_i \Omega^* P_i^T$$

is used to enforce the calibration constraints. The non-linear optimization approach minimizes the reprojection error across all views:

$$\min \sum_{i=1}^{n} \left\| \frac{K_i K_i^T}{\|K_i K_i^T\|_F} - \frac{P_i \Omega^* P_i^T}{\|P_i \Omega^* P_i^T\|_F} \right\|_F^2$$

where $\|\cdot\|_F$ denotes the Frobenius norm. A simplified linear method is used for initialization when both the principal point and aspect ratio are known:

$$\lambda \begin{pmatrix} f_i^2 & 0 & 0 \\ 0 & f_i^2 & 0 \\ 0 & 0 & 1 \end{pmatrix} = P_i \begin{pmatrix} b_1 & 0 & 0 & b_2 \\ 0 & b_1 & 0 & b_3 \\ 0 & 0 & 1 & b_4 \\ b_2 & b_3 & b_4 & b_5 \end{pmatrix} P_i^T$$

where the parameters b_1, b_2, b_3, b_4, b_5 are solved linearly.

To handle critical motion sequences, a sensitivity analysis is performed to detect sequences where the constraints are nearly degenerate. This involves examining the Jacobian matrix of the system of equations to ensure that small perturbations significantly impact the solution, thereby identifying critical configurations. Experiments on synthetic and real data validate the proposed method. The authors test the algorithm under various conditions, including different noise levels and sequence lengths. The results show that the method can accurately recover intrinsic parameters and achieve metric reconstruction, even with significant variations in camera settings.

On the other hand, the method described in [25] focuses on the minimal conditions required for flexible calibration. The primary theoretical result is that only one intrinsic parameter needs to be constant for a Euclidean reconstruction. The method begins with an initial projective reconstruction, which is then upgraded to a Euclidean reconstruction. The camera model is represented by the equation

$$\lambda \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} \gamma f & sf & u \\ 0 & f & v \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} R & t \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix},$$

where λ is the scale factor accounting for perspective effects, R is a 3×3 rotation matrix, t is a 3×1 translation vector, f is the focal length, γ is the aspect ratio, s is the skew, and (u, v) are the coordinates of the principal point. The intrinsic parameters are contained within the matrix K. Given a sequence of camera matrices

$$P = (P_i = K_i[R_i \mid -R_i t_i])_{i=1,\ldots,m},$$

with different constraints on the intrinsic parameters, the problem is formulated as finding the subset of projective transformations H such that the transformed camera matrices $P_i H$ can be factorized as

$$P_i H \approx K_i[R_i \mid -R_i t_i],$$

where K_i represents an intrinsic calibration matrix with one constant intrinsic parameter. To solve this problem, constraints are derived from the camera matrices. For instance, if $P_i = \begin{bmatrix} u_i^T & v_i^T & w_i^T \end{bmatrix}^\top$ is normalized such that $w_i^T w_i = 1$, then for constant skew, aspect ratio, focal length, and principal point coordinates, certain conditions must be met. For example, the skew is constant if $(u_i - w_i) \cdot (v_i - w_i) / \|v_i - w_i\|^2$ is constant. A bundle adjustment algorithm is employed to refine the estimates of all unknown parameters. This algorithm minimizes the sum of squared differences between the observed

and projected image points. The optimization problem is

$$\min_{K_i,R_i,t_i,X_j} \sum_{i,j} (x_{ij} - \hat{x}_{ij})^2,$$

where $\hat{x}_{ij}$ are the re-projected image coordinates calculated from the parameters. An initial estimate is often obtained using methods that assume reasonable guesses for the intrinsic parameters, followed by iterative refinement through bundle adjustment. The method's efficacy is demonstrated through experiments on both synthetic and real data, showing that accurate calibration can be achieved under the flexible calibration model.

2.1.4 Autocalibration for Dynamic Scenes

In Sec. **??** and Sec. **??**, we discussed direct and stratified autocalibration methods, respectively. These methodologies, however, assume a static 3D scene and consider moving objects in dynamic scenes as outliers. Autocalibration in dynamic scenes has not been extensively explored, primarily because Multi-body Structure-from-Motion [56] has not matured to the level of its single-body counterpart.

Although more challenging, the multi-body scenario offers advantages that have been somewhat overlooked in the literature. A notable work in this area is [16], where Fitzgibbon and Zisserman address the recovery of structure and motion from image sequences involving several independently moving objects. They demonstrate that multi-body analysis allows for Euclidean reconstruction in cases that are under-constrained for a static scene. The key novelty lies in leveraging the fact that, although the motions are independent, they share common camera parameters across different frames. This constraint enables the recovery of Euclidean structure in scenarios where traditional static-scene reconstruction methods fail. The camera model used in this work is the standard pinhole model, where the projection of a 3D point $\mathbf{X}$ into a 2D image point $\mathbf{x}$ is given by:

$$\mathbf{x} = K[R|t]\mathbf{X}.$$

The projection equation for a point p of body β in view v is:

$$\beta x_{vp} = K_v[\beta R|\beta t]\beta X_p$$

The goal is to recover the structure $\{\beta X_p\}$ and motion parameters $\{K_v, \beta R, \beta t\}$ that best fit the observed image data $\{\beta x_{vp}\}$. In the most general case, several high-relief bodies are tracked across multiple views. For each body β, a fundamental matrix βF is computed. The Kruppa equations provide two constraints per body on the camera parameters for two views. For v views and b bodies, there are $b(5v - 8)$ constraints available for calibration, allowing complete determination of the camera parameters if $b(5v - 8) \geq 5v$. The method uses bundle adjustment to refine the estimates of camera calibration and object motion. The objective function minimized during bundle adjustment is:

$$\min_{K_v,R_v,t_v,X_p} \sum_{v} \sum_{p} d^2(x_{vp}, K_v[R_v|t_v]X_p)$$

where $d(x, y)$ is the Euclidean distance between projected and observed points. An important contribution is the capability to handle scenes with few tracked points. The analysis shows that with just four points per object tracked over multiple views, the

camera calibration and Euclidean structure and motion can be recovered. The counting argument for this scenario ensures that the constraints available are sufficient for calibration when $2vbp \geq b(3p+6v-7)$. Furthermore, covariance intersection is employed to combine the estimates from multiple independent motions. The combined estimate of the camera parameters θ is obtained by:

$$\theta = \Lambda_B(\Lambda_A + \Lambda_B)^{-1}\theta_A + \Lambda_A(\Lambda_A + \Lambda_B)^{-1}\theta_B$$

where Λ_A and Λ_B are the covariance matrices of the individual estimates. The method demonstrates significant improvements in reconstruction accuracy over traditional single-body methods, opening new avenues for accurate 3D reconstruction in dynamic environments with multiple moving objects. However, the analysis is mainly theoretical and does not address robustness, as we did in this work.

2.2 Autocalibration of Wide-Angle Cameras

Camera autocalibration of wide-angle cameras is a critical component in numerous Computer Vision applications, especially with the proliferation of fisheye lenses, 360-degree cameras, and spherical cameras. These types of cameras have become indispensable in fields such as autonomous driving and robotics, and are increasingly integrated into popular consumer devices like smartphones and tablets. The primary challenge posed by wide-angle cameras is their significant radial distortions, which are intentionally introduced to achieve a larger field-of-view. Addressing these distortions is essential because the traditional pinhole camera model, represented by the intrinsic matrix K as introduced in Eq. (2.1), is insufficient to accurately model the projection (3D to 2D) and back-projection (2D to 3D) functions. These functions are critical for tasks such as structure-from-motion, visual localization, and Simultaneous Localization and Mapping (SLAM).

Wide-angle cameras introduce complexities in calibration due to their unique optical properties. The strong radial distortions necessitate the use of more sophisticated mathematical models beyond the conventional pinhole model. These advanced models incorporate additional distortion parameters, which must be accurately estimated for effective calibration. The inadequacy of traditional pinhole camera autocalibration methods, as discussed in Sec. 2.1, underscores the need for specialized techniques tailored for wide-angle optics.

This review of the state-of-the-art in wide-angle camera autocalibration will begin with an introduction and definition of the popular camera models used to represent wide-angle lenses. These models are essential for capturing the unique characteristics of fisheye lenses, 360-degree cameras, and spherical cameras. Following this foundational overview, we will delve into prominent works in the field of wide-angle camera autocalibration. We will provide a thorough analysis of the common taxonomy of these methods, which are typically categorized into two main approaches: *geometric*-based methods and *learning*-based methods.

2.2.1 Wide-angle Camera Models

In [80], the authors provide a comprehensive review of wide-angle camera models, starting with the Unified Camera Model (UCM). Although the UCM is typically used

for modelling systems with catadioptric cameras, it can also apply to fisheye lenses. It has five parameters $i = [f, g, u, v, \xi]$. The projection function is defined as:

$$\pi(x, i) = \left[f\frac{x}{\xi d + z}, g\frac{y}{\xi d + z} \right] + [u, v]$$

where

$$d = \sqrt{x^2 + y^2 + z^2}$$

The point is first projected onto the unit sphere and then onto the image plane of the pinhole camera, shifted by ξ from the sphere's center.

The Extended Unified Camera Model (EUCM) extends the UCM by adding a sixth parameter β, making it suitable for a wider range of lenses. The projection function is:

$$\pi(x, i) = \left[f\frac{x}{\alpha d + (1-\alpha)z}, g\frac{y}{\alpha d + (1-\alpha)z} \right] + [u, v]$$

where

$$d = \sqrt{\beta(x^2 + y^2) + z^2}$$

This model transforms the sphere into an ellipsoid, allowing it to represent more complex lens geometries.

The Kannala-Brandt (KB) model is a generic model that fits regular, wide-angle, and fisheye lenses. It has two versions with six or eight parameters. The projection function is:

$$\pi(x, i) = \left[fd(\theta)\frac{x}{r}, gd(\theta)\frac{y}{r} \right] + [u, v]$$

where

$$r = \sqrt{x^2 + y^2}$$

$$\theta = \text{atan2}(r, z)$$

$$d(\theta) = \theta + k_1\theta^3 + k_2\theta^5 + k_3\theta^7 + k_4\theta^9$$

The displacement from the optical center is proportional to a polynomial of the angle θ between the point and the optical axis.

The Field-of-View Camera Model (FOVCM) assumes that the distance between an image point and the principal point is proportional to the angle between the corresponding 3D point and the optical axis. The projection function is:

$$\pi(x, i) = \left[f\frac{r_d x}{r_u}, g\frac{r_d y}{r_u} \right] + [u, v]$$

where

$$r_u = \sqrt{x^2 + y^2}$$

$$r_d = \frac{\text{atan2}(2r_u \tan(\frac{w}{2}), z)}{w}$$

Here, w approximately corresponds to the field-of-view of the lens.

Finally, The Double Sphere model (DSCM) is proposed to better fit fisheye cameras, offering a closed-form inverse without requiring computationally expensive trigonometric operations. It has six parameters $i = [f, g, u, v, \xi, \alpha]$. The projection function is:

$$\pi(x, i) = \left[f \frac{x}{\alpha d_2 + (1-\alpha)(\xi d_1 + z)}, g \frac{y}{\alpha d_2 + (1-\alpha)(\xi d_1 + z)} \right] + [u, v]$$

where

$$d_1 = \sqrt{x^2 + y^2 + z^2}$$
$$d_2 = \sqrt{x^2 + y^2 + (\xi d_1 + z)^2}$$

This model projects a point onto two unit spheres, then onto the image plane using a pinhole model.

In this work, the authors also evaluate the effectiveness of the aforementioned camera models in modelling several real-world cameras. For this task, the authors estimate the camera parameters for each model using a grid of AprilTag markers detected in images. The optimization function for calibration depends on the state $s = [i, T_{ca1}, ..., T_{caN}]$, minimizing the reprojection error:

$$E(s) = \sum_{n=1}^{N} \sum_{k \in K} \rho((\pi(T_{can} x_k, i) - u_{nk})^2)$$

where ρ is the Huber norm, π is the projection function, and T_{can} transforms the coordinates from the calibration pattern to the camera coordinate system. The camera models are evaluated using various metrics, including reprojection error and computation time. The Double Sphere model demonstrates comparable accuracy to the Kannala-Brandt model with eight parameters but with significantly reduced computation time. The evaluation shows that the DSCM and EUCM models are efficient and accurate, making them suitable for real-time applications in vision-based motion estimation. For these reasons, the DSCM and EUCM models will be used extensively in the context of our work to develop accurate and generalizable autocalibration methods for wide-angle cameras.

2.2.2 Geometric-based methods

Geometric-based methods utilize calibration objects [78, 91], line detection [1, 4, 7, 20, 66, 90], or vanishing points [42, 62] to reproject 2D points into 3D, enabling distortion and intrinsic camera parameter estimation. Although these methods are effective, they encounter difficulties in unstructured environments without manual input. Although our work primarily focuses on learning-based methodologies, we also discuss two widely used geometric-based methods, namely Auto-DE [7] and Auto-DC [1], which we compare against our method in this study.

In [7], the authors propose a robust method for radial distortion estimation using circle fitting techniques and employing Fitzgibbon's division model for distortion correction. The authors extract edge contours from the distorted image using a modified Canny edge detector with adaptive thresholds. Each detected contour with a length exceeding 10 pixels is considered for circular arc fitting through a random sampling approach inspired by RANSAC, preserving non-overlapping models that had significant

support. Subsequent to circle identification and fitting, the authors employ a RANSAC-based approach to estimate the distortion parameters, considering the support from longer arcs to ensure accuracy. The objective is to determine parameters (λ, x_0, y_0) that maximally undistorted the image, verified by the alignment of undistorted arcs to straight lines using orthogonal regression. This is done by iteratively sampling arcs and refining the model based on pixel support, terminating when a satisfactory probability threshold is met.

Finally, in [1], the authors propose an unsupervised method to correct the radial distortion caused by wide-angle lenses using a single image. The approach relies on the one-parameter division model, which simplifies the problem by using a single distortion parameter k_1. The method consists of four main stages: edge detection, initial parameter estimation using an improved Hough transform, parameter optimization, and image correction. First, edge detection is performed using the Canny method. This involves applying a Gaussian convolution, computing the image gradient using 3x3 convolution masks, and then performing non-maximum suppression and hysteresis to detect edges. The detected edges, along with their positions and orientations, are stored for further processing. Next, the initial distortion parameter is estimated using an improved Hough transform. The classical Hough transform is revisited by incorporating the radial distortion parameter into the Hough space, which now becomes a three-dimensional space defined by the distance d, orientation α, and normalized distortion parameter p. The normalization is given by $p = (r_{\text{max}} - r'_{\text{max}})/r_{\text{max}}$, where r_{max} is the distance from the center of distortion to the farthest point in the original image, and r'_{max} is the same distance after applying the distortion model. The relationship between p and k_1 is $k_1 = -p/((1+p)r^2_{\text{max}})$. In the voting step of the Hough transform, for each value of p in a discretized interval, the edge point coordinates and orientation are corrected using the lens distortion model associated with p. Each edge point then votes for a line if the edge point orientation is coherent with the line orientation, within a threshold $\delta\alpha$. The votes are weighted by the distance from the point to the line. The best distortion parameter p_0 is the one that maximizes the sum of votes for the most voted lines. Once p_0 is determined, it is optimized using a numerical scheme that minimizes the average squared distance from the corrected points to their associated lines. This is done using a modified Newton-Raphson method with a damping parameter γ to ensure convergence. The error function to minimize is $E(p) = \sum_{j=1}^{N_l} \sum_{i=1}^{N(j)} (\cos(\alpha_j^p) x_{ji}^p + \sin(\alpha_j^p) y_{ji}^p + d_j^p)^2 / \sum_{j=1}^{N_l} N(j)$. Finally, the optimized distortion parameter is applied to correct the image. The inverse transformation for each pixel is computed by solving a second-degree polynomial, yielding the corrected coordinates. The corrected image is generated by interpolating the input image at these coordinates.

2.2.3 Learning-based methods

Learning-based methods, which rely on Convolutional Neural Networks (CNNs), can calibrate wide-angle cameras end-to-end from single images using priors learned from large-scale datasets of radially distorted images, whether these images come from real-world cameras or are synthetically distorted. These methods are particularly well-regarded for their ability to accurately calibrate cameras in uncontrolled environments, where geometric-based methods often struggle. However, geometric methods are still

preferred in controlled environments due to their reliability and lower computational costs. Learning-based autocalibration methods primarily target intrinsic and distortion parameter estimation, though recent works [11, 21, 64] have also addressed extrinsic parameter calibration. In this work, we predominantly focus on intrinsic parameter estimation and will present methods addressing this scenario.

We further categorize learning-based methods into regression-based and reconstruction-based autocalibration methods in our taxonomy.

Regression models

Regression models [5,26,29,37,38,43,65,81,86] estimate intrinsic and distortion camera parameters based on a predefined camera model specific to each method. Specifically, depending on which cameras are targeted by the method, these methods can rely on: (i) *fisheye* camera models, characterized by their equidistant projection, where the angle of incidence is directly proportional to the distance from the image center, (ii) *spherical* camera models, mapping 3D points onto a spherical surface and supporting 360-degree imaging, and (iii) *unified* camera models, combining elements of perspective and fisheye models to offer a versatile calibration approach suitable for a range of wide-angle lenses.

Despite their general effectiveness in uncontrolled environments, these methods lack accuracy and generalization due to their reliance on a single model. This dependency hinders the calibration network's adaptability to various camera types. The single-model approach also limits the range of distortion parameters in the training datasets, resulting in poor calibration accuracy when parameters fall outside predefined ranges. While general camera models have been proposed [81], they lack the representational capacity to handle cameras with complex projection functions, leading to low accuracy. Although several prior works have attempted to eliminate this single-model dependency, their efficacy has been limited for reasons that we will discuss next.

Liao et al. [40] model-free approach first uses adversarial learning for image rectification, estimating a parameter-free distortion distribution map (DDM) without relying on a predefined camera model. The methods begins by constructing the DDM, which serves as a representation of the global distortion features in an image. Formally, the DDM for a distorted image I_d is defined as a pixel-wise map where each pixel value represents the distortion level. Mathematically, if (x, y) is a pixel in the distorted image and (x', y') is the corresponding pixel in the rectified image, the DDM value $D(x, y)$ can be expressed as:

$$D(x, y) = \frac{x'}{x} = \frac{y'}{y} = k_1 + k_2 r_d + k_3 r_d^2 + \dots$$

where r_d is the radial distance from the distortion center and k_i are the distortion parameters. The full framework consists of three main modules: SSD (semantics, structure, and distortion) learning, multimodal attention fusion, and distortion rectification.

(a) *SSD Learning Module*: This module extracts semantic, structural, and distortion features using three specialized neural networks. The distortion features are captured using a fully convolutional network similar to U-Net, which outputs the DDM. The semantic features are extracted using a network trained from scratch with guidance from the DDM. The structural features are captured using a network

inspired by PointNet, which processes the sparse and grayscale attributes of handcrafted features like Canny edges.

(b) *Multimodal Attention Fusion Module*: This module fuses the features extracted by the SSD learning module. The fusion is performed using an attention mechanism that selectively integrates the local structural and global semantic features. Formally, if V_{sem} and V_{str} are the semantic and structural feature vectors, the fused feature vector V_{hyb} is given by:

$$V_{\text{hyb}} = V_{\text{sem}} \oplus (\sigma(\text{fc}(V_{\text{sem}})) \otimes V_{\text{str}})$$

where σ is the sigmoid function, fc is a fully connected layer, $\otimes$ denotes element-wise multiplication, and $\oplus$ denotes concatenation.

(c) *Distortion Rectification Module*: This module takes the hybrid features and reconstructs the rectified image using a decoder network. The architecture of the decoder includes several hierarchies with upsampling layers followed by convolutional layers.

The training of the framework involves a combination of losses. The distortion distribution loss L_d ensures the accuracy of the DDM, the rectification loss L_r measures the difference between the rectified and ground truth images, the perceptual loss L_p ensures high-level feature consistency, and the adversarial loss L_a improves the realism of the rectified images. The overall loss function is:

$$L = \lambda_d L_d + \lambda_r L_r + \lambda_p L_p + \lambda_a L_a$$

Although experimental results demonstrate the framework's generalizability, which allows it to handle various types of distortions without being limited to specific camera models, the formulation of the DDM does not exploit radial symmetries and the overall method underperforms in accuracy and efficiency compared to the state-of-the-art. The adversarial training process further introduces challenges in loss balancing between the generator and discriminator, leading to the frequent need for manual adjustments and generally suboptimal results.

Another limitation of successful, recent calibration methods, *e.g.*, Wakai et al. [81], lies in their reliance on loss functions formulated from the camera model's undistortion equations, which, to be differentiable, require the model's 2D-to-3D back-projection function to have a unique, closed-form solution for computing bearing vectors. However, sophisticated camera models like EUCM [32] and DSCM [80] do not fulfill this condition, restricting their use in these methods. In scenarios lacking a differentiable back-projection function, other approaches, such as direct loss computation from camera parameters [5] or iterative solutions for the back-projection function [43] have been proposed, yet these often fall short in accuracy. Moreover, computing the loss directly from camera parameters, *e.g.*, using the root-mean-square error, is problematic for models with ambiguities, where multiple parameter sets can represent the same camera.

Finally, we discuss [29], a recent and successful end-to-end method that predicts fixed camera parameters from per-pixel perspective fields, in a similar fashion to the contributions presented in this thesis. In [29], the authors propose *perspective fields*, a novel representation that models the local perspective properties of an image. Perspective Fields contain per-pixel information about the camera view, parameterized as an

Up-vector $\mathbf{u_x}$ and a Latitude value ϕ_x. The Up-vector $\mathbf{u_x}$ gives the world-coordinate up direction at each pixel, equivalent to the inverse gravity direction of the 3D scene projected onto the image. The Latitude ϕ_x is defined as the angle between the incoming light ray $\mathbf{R}$ and the horizontal plane. Mathematically, the Up-vector $\mathbf{u_x}$ and the Latitude ϕ_x can be expressed as:

$$\mathbf{u_x} = \lim_{c \to 0} \frac{P(\mathbf{X} - c\mathbf{g}) - P(\mathbf{X})}{\|P(\mathbf{X} - c\mathbf{g}) - P(\mathbf{X})\|_2}$$

$$\phi_x = \arcsin\left(\frac{\mathbf{R} \cdot \mathbf{g}}{\|\mathbf{R}\|_2}\right)$$

Here, $P(\mathbf{X})$ denotes an arbitrary projection function mapping a point in the world $\mathbf{X}$ to the image plane, and $\mathbf{g}$ is the unit vector representing the gravity direction. To estimate Perspective Fields from a single image, the authors introduce PerspectiveNet, an end-to-end neural network that predicts the Up-vector and Latitude for each pixel. PerspectiveNet utilizes a pixel-to-pixel architecture suitable for predicting per-pixel values. In sum, the core contribution of this work consists in the introduction of Perspective Fields, which provide a robust, interpretable, and generalizable representation of image perspective that can be leveraged for camera calibration and image compositing applications.

Reconstruction models

Reconstruction models [10,15,39,87,93] employ adversarial learning, leveraging multi-scale information to train generators and discriminators for image undistortion without explicitly estimating camera parameters.

[93] is particularly relevant in the context of this thesis, as the authors propose a novel method for correcting radial distortion in fisheye images by leveraging the radial symmetry inherent in such distortions, employing polar coordinates within to estimate a 1D distorted-to-rectified flow instead of a 2D flow. In this thesis, we will also employ polar coordinates to reduce the dimensionality of our novel implicit camera model. In [93], the authors explain how the conventional approach of predicting displacement fields in Cartesian coordinates is challenged due to its complexity and the inefficiency of convolutional kernels that do not account for the radial nature of distortion. Instead, the proposed method transforms the problem into polar coordinates, where the distortion correction is simplified into a one-dimensional flow prediction task, taking advantage of the radial symmetry. The proposed method is implemented through the Polar Coordinates Distortion Rectification Network (PCDRN), which is designed to predict the one-dimensional flow in the polar coordinate system. The transformation from Cartesian to polar coordinates aligns the convolutional kernel's sampling strategy with the distortion characteristics, improving the accuracy of the flow prediction. Specifically, each row in the polar coordinates corresponds to pixels at the same radial distance from the optical center, thus having the same distortion. This enables the network to predict a single distortion value per row, simplifying the learning process and enhancing prediction accuracy. To address artifacts and blurs induced by the coordinates transformation, the authors introduce a Polar-To-Cartesian Appearance Enhancement Network (PAEN). This network refines the appearance of the rectified images by eliminating ring

artifacts and enhancing local details. The PAEN uses an encoder-decoder architecture, similar to U-Net, to effectively interpolate and restore the image details lost during the polar-to-Cartesian conversion.

2.3 Multi-Body Structure-from-Motion

Traditional geometric-based Structure-from-Motion (SfM) purely focuses on reconstructing *static* scenes, where the number of motions, denoted by μ, is equal to 1. In such cases, moving objects are typically treated as outliers and excluded from the reconstruction process because they do not satisfy the epipolar constraints imposed by the relative camera poses, which are estimated with reference to the static components of the observed 3D scene. This exclusion not only prevents the reconstruction of moving objects but can also negatively impact the overall reconstruction quality if the SfM algorithm lacks inherent robustness against such outliers.

To address these limitations, recent studies have explored the problem of Multi-Body Structure-from-Motion (MBSfM), which extends SfM to handle scenarios involving multiple motions, i.e., where $\mu > 0$. In this section, we first review the few existing geometric MBSfM methods that are capable of reconstructing 3D scenes with rigidly moving bodies from unstructured image sets captured using perspective cameras. Following this, we examine recent advancements in MBSfM that employ learning-based approaches. These approaches utilize monocular depth estimators and implicit scene representations, such as Neural Radiance Fields (NeRF) and 3D Gaussian Splatting, to model scene dynamics in a globally consistent scale across the entire 3D scene, providing a more comprehensive solution to the challenges posed by multi-body movements.

2.3.1 Geometric-based Multi-body Structure-from-Motion

A few studies have extended Multi-Body Structure-from-Motion to unstructured images $\mathbf{I}ii = 1^n$, treating it as a generalization of traditional SfM to multi-body scenes. We identify two primary research directions in this area.

The first direction involves early factorization methods [13, 75], which simultaneously segment rigidly moving objects $\mathcal{B}$ and perform sparse 3D reconstruction in a single step. However, these methods are limited by restrictive assumptions on camera models and the requirement for complete feature tracks, which are rarely available in real-world scenarios. Due to their lack of robustness, factorization methods have not been widely adopted for practical applications and are mainly confined to short, unrealistic sequences.

Specifically [75] presents a novel method for separating and recovering the motion and shape of multiple independently moving objects in an image sequence, without prior knowledge of the number of objects or the need for feature grouping at the image level. The key concept introduced is the *shape interaction matrix*, which is invariant to object motions and can be computed solely from observed trajectories of image features. This invariant structure allows for the segmentation of features into distinct objects and facilitates the recovery of their shapes and motions.

The method builds upon the factorization approach initially developed for a single moving object, extending it to handle multiple objects. Given a set of features tracked in an image sequence, the goal is to factorize the measurement matrix W into motion

M and shape S matrices. The measurement matrix $W \in \mathbb{R}^{2F\times N}$ contains the image coordinates of N tracked features across F frames. The motion matrix $M \in \mathbb{R}^{2F\times 4}$ encapsulates the motion parameters, and the shape matrix $S \in \mathbb{R}^{4\times N}$ represents the 3D shape of the objects. This factorization is achieved using Singular Value Decomposition (SVD), expressed as:

$$W = U\Sigma V^T,$$

where $U \in \mathbb{R}^{2F\times 4}$, $\Sigma \in \mathbb{R}^{4\times 4}$, and $V \in \mathbb{R}^{N\times 4}$. The matrices M and S are then defined as:

$$M = U\Sigma^{1/2}A, \quad S = A^{-1}\Sigma^{1/2}V^T,$$

with A being an arbitrary invertible 4×4 matrix. Constraints on M, derived from the orthonormality of the camera axes and the translational motion, ensure a proper decomposition into orthogonal vectors corresponding to the image plane and motion vectors.

In scenarios involving multiple moving objects, the measurement matrix W will contain features from various objects, leading to a higher rank—up to 8 for the case of two full 3D objects. The challenge lies in segmenting W into submatrices that correspond to each object's motion and shape, without prior knowledge of which features belong to which object. To address this, the paper introduces the *shape interaction matrix* Q, defined as:

$$Q = VV^T,$$

where V is obtained from the SVD of W. The matrix Q has a block-diagonal structure in its canonical form, reflecting the distinct objects in the scene. Each block corresponds to the interaction between features of the same object, with non-zero values indicating features from the same object and zero values indicating features from different objects. This block-diagonal structure is invariant to the specific motions of the objects, making Q a powerful tool for segmentation.

The segmentation and recovery process begins by sorting the shape interaction matrix Q into a block-diagonal form through permutations of its rows and columns, which align with grouping features in W by their respective objects. This reorganization transforms W into its canonical form, where features from one object are grouped into contiguous columns. The corresponding canonical measurement matrix can then be factorized as:

$$W' = \begin{bmatrix} W_1 & W_2 \end{bmatrix},$$

where W_1 and W_2 correspond to different objects. Each submatrix W_i can be independently factorized to recover the motion and shape of each object as:

$$W_i = M_i S_i, \quad i = 1, 2, \ldots, k,$$

where k is the number of objects. The invariant nature of the shape interaction matrix Q ensures that its values remain consistent across different object motions, providing a robust basis for segmentation without the need for prior knowledge of the objects' identities or motions.

The algorithm follows a structured sequence: first, features are extracted and tracked to form the measurement matrix W. The rank of W is determined, followed by its decomposition using SVD to obtain the motion and shape components. The shape interaction matrix Q is then computed and sorted into a block-diagonal form. The resulting permutations of the columns of V correspond to the correct grouping of features, allowing the factorization of each submatrix to recover the motion and shape of each object.

The second, more recent, research direction in geometric-based Multi-Body Structure-from-Motion, more recent, research direction separates motion segmentation of objects $\mathcal{B}$ and their 3D reconstruction. [2, 67] cluster sparse correspondences based on their rigid motion and then recover the 3D structure of each object independently using a SfM pipeline [68].

Specifically, in [2], the authors propose a motion segmentation and trajectory clustering algorithm that they suggest can be applied to achieve Multi-Body Structure-from-Motion. The idea is to segment motions using two-frame correspondences for each view pair independently and, then, upgrading the pairwise motion segmentation from local, two-view, consistency to multi-view consistency using the robust permutation synchronization algorithm based on the spectral decomposition of a binary correspondence matrix described in [2]. The segmented motions can then be provided independently to an off-the-shelf Structure-from-Motion algorithm to reconstruct each individually moving object in the 3D scene.

As discussed in [35,54,55], the relative scale problem is a common challenge affecting all the aforementioned methods, as each object tends to be reconstructed at its own independent scale. In SLAM, this issue has been tackled under certain assumptions: for instance, [55] assumes that objects move within a one-parameter family of motions, while [35] requires video input and continuous object motion to resolve relative scales.

2.3.2 Learning-based Depth and Camera Pose Estimation

Learning-based methods for Structure-from-Motion from unstructured images regress camera poses and dense depth maps using multi-view stereo matching.

DeMoN [79] is recognized as the first widely adopted zero-shot Structure-from-Motion method that operates on image pairs. It functions by taking image pairs as input and iteratively predicting and refining depth and camera pose estimates. The initial estimates are derived from monocular depth estimation priors and are further refined using pairwise correspondences obtained from an optical flow regression head, which is shared with the monocular depth estimator, structured as a stacked encoder-decoder network.

DeMoN's reliance on monocular depth priors allows it to predict consistent depths for both moving objects and static regions of the scene. As a result, it can sometimes recover the depths of dynamic objects when their motion between frames is minimal. However, when there is significant motion between the two frames, the accuracy of the predicted optical flow correspondences deteriorates, adversely affecting the quality of the zero-shot depth estimation produced by the network.

Recent approaches, inspired by DeMoN, have been developed to enable two-view Structure-from-Motion without relying on constraints from epipolar geometry, instead leveraging learning-based priors to infer geometric relationships between image pairs.

Methods such as DUSt3R [84] and MASt3R [36] generalize across a broad range of intrinsic and extrinsic camera parameters, eliminating the need for camera calibration or prior knowledge of camera positions when performing Structure-from-Motion. These methods differ by treating the reconstruction task as a regression problem of pointmaps, which allows them to bypass the rigid constraints of conventional camera models. This novel formulation of the Structure-from-Motion problem, which does not rely on epipolar constraints for regressing multi-view geometry, enables generalization to scenes containing both static and dynamic elements, with moving objects being reasonably well reconstructed in practice.

However, this line of work shares similar limitations with monocular depth estimation methods regarding depth and camera pose estimation. Relying exclusively on learning-based priors means that while these methods can perform well in many real-world scenarios that do not demand high accuracy and robustness in depth and camera pose estimates, they are generally unreliable and benefit significantly from multi-view refinement. This refinement is typically achieved through multi-view correspondences obtained from either optical flow [74] or 2D point tracks [30].

To address these limitations, recent methodologies [6, 70, 82] have introduced differentiable, learning-based Structure-from-Motion pipelines that reparameterize depth, extrinsic, and intrinsic parameters, allowing for end-to-end learning from unstructured image sets or video sequences. These methods begin with initialization from monocular depth estimators, and the predicted depth maps are iteratively refined by overfitting Convolutional Neural Networks (CNNs) or Multi-Layer Perceptrons (MLPs) to the specific scene, in a manner similar to NeRFs [52] and Gaussian Splatting [31]. While these approaches are not specifically designed for reconstructing moving objects within scenes, the initialization from monocular depth estimators can result in fewer reconstruction artifacts in the depth maps for dynamic regions of the 3D scene.

Specifically, in [44, 92], monocular depth maps are estimated from video frames, followed by a *sceneflow*—a motion prediction—network that is fine-tuned on the entire video sequence. Although these methods mitigate the static scene assumption common in both geometric- and learning-based Structure-from-Motion approaches, they are limited to video inputs and do not function with unstructured image sets. Moreover, they require multiple frames for accurate depth prediction, pose tracking, intrinsic parameter refinement, and fine-tuning. In [92], depth estimation for moving objects still relies primarily on depth priors, which can lead to inaccuracies similar to those found in standard monocular depth estimation. Additionally, the scale of moving objects may drift if the monocular depth estimates include outliers.

Another research direction draws inspiration from traditional Bundle Adjustment for multi-view depth and camera pose refinement using multi-view correspondences. DeepSfM [85] introduces dedicated cost volumes for iterative refinement of depth and camera pose, alternating predictions in a bundle-adjustment-like strategy that allows for reciprocal improvement. Similarly, [73] uses a BA-layer for regression from basis depth maps. Wang et al. [83] propose a scale-invariant network for depth and pose estimation from view pairs. The primary drawbacks of this line of research are the long computational times and high memory requirements, even on the latest GPU architectures, making large-scale Structure-from-Motion infeasible. Additionally, these methods often require knowledge of intrinsic camera parameters, which complicates

their application in practice when camera specifications are unknown, particularly in in-the-wild settings.

CHAPTER 3

Contributions

In this chapter we will review, for each of the works composing this thesis, the contributions to the state-of-the-art.

3.1 Minimal Perspective Autocalibration

Camera autocalibration is the long-standing problem in 3D Computer Vision of recovering the intrinsic parameters of a camera from point correspondences without requiring calibration objects or a specifically designed scene layout or geometry.

This paper presents a comprehensive characterization of two- and three-view minimal autocalibration problems in the case of a perspective camera with constant intrinsics $K \in \mathbb{R}^{3\times3}$, defined assuming a traditional pinhole camera model:

$$K = \begin{pmatrix} f & s & u \\ 0 & g & v \\ 0 & 0 & 1 \end{pmatrix} . \tag{3.1}$$

In our work, the goal is to recover the entries of K, *i.e.*, focal lengths f and g, principal point (u, v) and camera skew s. To this end, we design practical and efficient solvers to recover K from a set of image points $x_{ip} \in \mathbb{R}^2$, indexed by the image $i \in [M]$ and point $p \in [N]$.

Our approach rethinks the traditional paradigm of autocalibration, where a projective reconstruction is obtained and is later upgraded to a metric reconstruction using the Kruppa equations or the Modulus constraint. Instead, we introduce a novel formulation that extends the minimal Euclidean reconstruction problem of four points in three calibrated views to the uncalibrated case where K is unknown. Formally, we jointly estimate camera intrinsics K and the unknown *projective depths* λ_{ip}, *i.e.*, the depth of

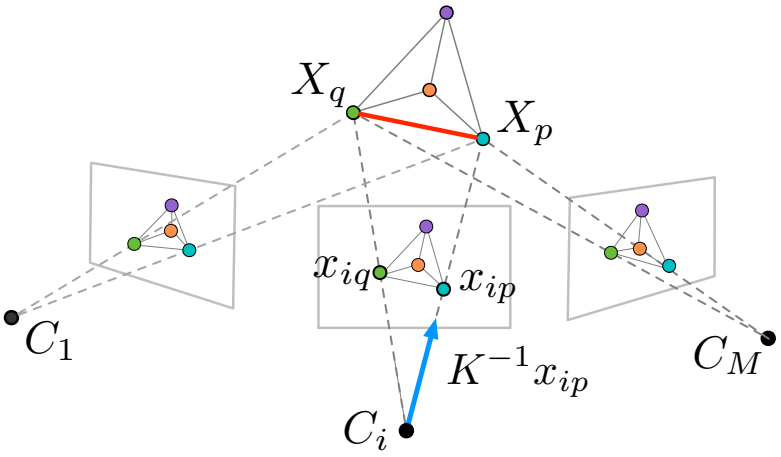

Figure 3.1: *Illustrating the setup in the paper* Minimal Perspective Autocalibration *and of equations* (3.2)*. In our work, we recover the projective depths λ_{ip} and the camera intrinsics K from a set of matched points x_{ip}, where indexes $i \in [M]$ and $p \in [N]$ refer to images and points respectively. In this illustration, the geometric constraint used throughout the paper is highlighted in red, i.e., the fact that the Euclidean distance between a given pair X_p and X_q of 3D points is constant regardless of the images i and j from which these points are observed.*

the p-th point as seen in the i-th image. This effectively amounts to solving the Euclidean reconstruction problem for the uncalibrated camera, as not only the position of 3D points X_p is recovered by back-projection of the 2D point x_{ip} using its projective depth λ_{ip}:

$$X_p = K^{-1}\lambda_{ip}x_{ip}, \tag{3.2}$$

but also the camera poses $\{P_i\}_{i=1}^{M}$ for each of the i-th images can be obtained by SVD decomposition of the recovered positions of the 3D points. As usual, we express a camera pose $P_i = R_i\,(I \mid -C_i)$ as a combination of a rotation matrix $R_i \in \mathbb{R}^{3\times3}$ and a camera center $C_i \in \mathbb{R}^3$.

This proposed formulation used throughout this work is based on the geometric constraint that the Euclidean distance between a pair X_p and X_q of 3D points $\|X_p - X_q\|$ is the same regardless of the images i and j from which they are reconstructed, for any $i, j \in [M]$. Starting from Eq. 3.2, this amounts to the vanishing of the function:

$$\begin{aligned} d_{i,j,pq}(\lambda, K; x) :=& (\lambda_{ip}x_{ip} - \lambda_{iq}x_{iq})^T\omega(\lambda_{ip}x_{ip} - \lambda_{iq}x_{iq}) \\ &-(\lambda_{jp}x_{jp} - \lambda_{jq}x_{jq})^T\omega(\lambda_{jp}x_{jp} - \lambda_{jq}x_{jq}), \end{aligned} \tag{3.3}$$

where $\omega = K^{-T}K^{-1}$ is the image of the absolute conic. From the above formulation, it is clear that our formulation consists in eliminating the camera poses P_i, which do not appear in Eq. 3.3.

Our formulation seamlessly enables the integration of any partial knowledge of the camera intrinsics in the form of linear constraints on the camera calibration matrix $f_l(K) = 0$. For instance, if the camera has zero-skew, the constraint $s = 0$ can be enforced on K. Other common cases for autocalibration, *e.g.*, square pixel aspect ratio $f = g$, can also be seamlessly considered. In such cases, we may write $f_1(K) = s$, $f_2(K) = f - g$ and impose f_1 and f_2 to zero. This gives rise to a variety of two- and three-view minimal autocalibration problems, which, for the first time, are organized

in a comprehensive and easy-to-reference taxonomy, providing indications on the feasibility of the problem as well as an indication of its complexity in terms of solutions in complex space to the polynomial system describing the autocalibration problem.

The definition and taxonomy of autocalibration problems are followed by the development of a general theory of minimal relaxations to address cases where our formulation leads to an over-constrained problem. These minimal relaxations of our depth formulation can be completely enumerated, and each instance of a specific autocalibration problem can be solved offline by applying numerical homotopy continuation (HC) methods to one such relaxation. Crucially, the offline analysis with HC methods also enables us to identify the most efficiently solvable minimal relaxations.

The contributions in this work are also from a practical perspective, as we make use of our newly proposed depth formulation and taxonomy of autocalibration problems to design and implement three numerical solvers that are particularly useful in practice. Our main solver is designed to solve the *full* camera calibration problem, *i.e.*, calibration of all five unknown parameters of a perspective camera, referenced in the work as `fguvs`. The other solvers we design are specialized for autocalibration problems with a partially calibrated camera. For this, we design a solver to address the autocalibration problem when the camera has zero skew, which is a common assumption in practice as many camera sensors today have zero or negligible skew. This solver is referred to as `fguv0` in our work. Furthermore, we design a solver, `ffuv0`, designed specifically when the camera has squared pixel aspect ratio and zero skew, another popular setup.

These solvers are designed by leveraging the depth formulation and our taxonomy to identify which minimal relaxations of the problems lead to the most efficient solution to the autocalibration problem. As a result, for the first time, our solvers can be fast enough for many online calibration applications, meaning that they can be embedded in robust frameworks, such as RANSAC, to bootstrap solutions for Structure-from-Motion, camera localization, and other 3D vision tasks with high accuracy in both online and offline calibration settings.

Among the strengths of our approach, we avoid well-known degeneracies of Kruppa's equations. Specifically, we empirically tested that our solvers do not suffer from the degeneracies arising from a singularity of the Kruppa equations when the optical centers of cameras lie on a sphere, and their optical axes intersect at the sphere's center. The reason behind this increased range of applicability is that our solvers recover the intrinsic parameters K directly and are not estimated indirectly from the dual image of the absolute conic (DIAC). The DIAC can be decomposed into an upper-triangular matrix K by Cholesky decomposition, which can be attained if the DIAC is positive-semidefinite. As a result, Kruppa-based methods fail when the estimated DIAC does not satisfy the positive-semidefiniteness condition, leading, as demonstrated in our experimental validation, to failures under certain camera setups or in the case of noisy 2D point correspondences.

Our final contribution is the experimental evaluation of these solvers compared to the state-of-the-art autocalibration methods. The experimental evaluation of the proposed minimal perspective autocalibration methods demonstrates their efficacy and robustness in various settings. Synthetic experiments were conducted to evaluate the performance of the new solvers under different noise levels, where results showed that the proposed solvers achieved superior accuracy compared to traditional methods, even

in the presence of significant noise. In particular, the solvers maintained high precision in estimating both the camera intrinsic parameters and the 3D point depths, outperforming existing approaches such as those based on Kruppa's equations. Real-world experiments further validated the effectiveness of the proposed methods. The solvers were integrated into the COLMAP [68] Structure-from-Motion pipeline and tested on several real image sequences. These tests highlighted the practical applicability of the new solvers, demonstrating their ability to provide accurate camera calibration in real-world scenarios. The solvers exhibited robustness against common degenerate configurations, such as cameras revolving around an object, a situation where traditional methods often fail.

Moreover, the implementation of the numerical solver for full camera calibration, addressing all five unknown parameters of a perspective camera, and the design of fast solvers for specialized problems with partially calibrated cameras, were key practical contributions. These solvers were shown to be efficient enough for online calibration applications and provided high accuracy when used in offline settings with RANSAC-based frameworks. The experiments conclusively demonstrated that the proposed minimal perspective autocalibration methods offer a significant improvement over existing techniques in terms of both accuracy and robustness, making them highly suitable for practical use in various computer vision applications.

3.2 Multi-Body Self-Calibration

This paper addresses a fundamental challenge in Structure-from-Motion (SfM): the assumption of a rigid scene. Traditional SfM techniques typically rely on a static or single moving object to maintain the rigidity constraint, which is integral to autocalibration and Euclidean upgrading from uncalibrated images. Given these premises, our work extends autocalibration to dynamic scenes comprising multiple moving rigid objects, proposing a method that leverages these independent motions to improve camera intrinsic estimation.

The problem is formulated as follows. Given M images $\mathcal{I} = \{I_1, \ldots, I_M\}$ of a 3D scene captured by a projective camera from different poses $\{P_i\}_{i=1}^{M}$, where $P_i = R_i \left(I \mid -C_i\right)$ is a combination of a rotation matrix $R_i \in \mathbb{R}^{3\times3}$ and a camera center $C_i \in \mathbb{R}^3$, and the camera intrinsics K are defined as in Eq. 3.1. Unlike the *Minimal Perspective Autocalibration* discussed in Sec. 3.1, this work assumes the skew of the camera is zero ($s = 0$) and radial distortion is not considered. The scene contains $B \geq 1$ independently moving rigid bodies $\mathcal{B} = \{\beta_1, \ldots, \beta_B\}$. The goal is to recover the intrinsic parameters K given only the collection of images $\mathcal{I}$.

In practice, autocalibration is a challenging problem due to the need to solve a system of non-linear polynomial equations. This is further complicated by noise and outliers in image correspondences and degenerate motions that can lead to indeterminate solutions. As a result, we propose to leverage multiple rigid motions in the scene to better constrain the autocalibration process, as more information can be inferred from dynamic scenes compared to static ones.

The major contribution of this work is represented by our novel approach of leveraging available rigid motions in a dynamic scene to better constrain the autocalibration problem. Unlike classical approaches that treat non-dominant motions as outliers, in

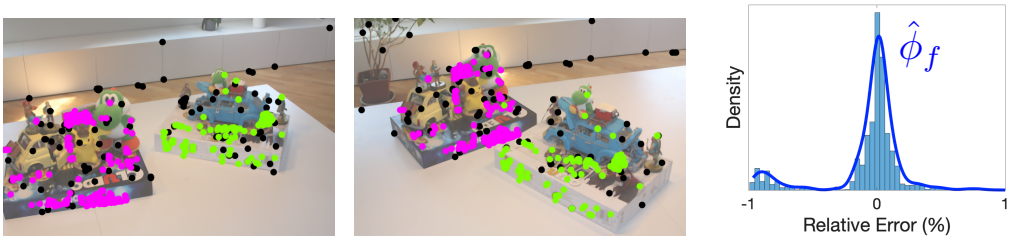

Figure 3.2: ***Left & Middle.*** *Motion segmentation for a sample image pair from the proposed* `Amiibo` *dataset. Matched 2D points are color-coded depending to the motion they are assigned to, whereas outliers are marked as black points.* ***Right.*** *Kernel Density Estimation (KDE) distribution $\hat{\phi}_f$ of focal lengths obtained through Kernel Voting.*

this work, we demonstrate that using the additional motions in the scene can improve the accuracy and robustness of camera parameter calibration. Furthermore, we demonstrate that autocalibration can be achieved with fewer images than otherwise would have been possible without leveraging multiple motions, as suggested in prior theoretical work [16]. Specifically, for each image pair, one fundamental matrix can be estimated from each rigid motion identified in the scene. Using Kruppa equations, for each fundamental matrix, two constraints on the intrinsics K can be obtained. As a result, the calibration of four unknown parameters in K requires two image pairs, *i.e.*, three images. By identifying and segmenting the multiple motions in the scene in a single view-pair, multiple fundamental matrices are derived from a single image pair and *full* calibration of K can be attained from just a single image pair.

Our second contribution is demonstrating that, by considering multiple rigid motions, our method can provide a more robust initialization of the camera parameters within the context of an optimization routine that considers all available images in the input sequence to refine and obtain accurate estimates of the parameters K. In the proposed optimization scheme, we introduce a motion segmentation pre-processing step that is specifically tailored for the problem. This approach uses rigidity constraints to recover fundamental matrices that describe rigid motions and simultaneously estimate the focal length. The segmentation step involves robust multi-model fitting using T-linkage [47], which clusters correspondences according to their rigid motion, applying a 6-point algorithm to sample fundamental matrices and corresponding focal lengths. The optimization routine is initialized using a subsequent initialization step that derives epipolar constraints from the fundamental matrices to compute a distribution of focal lengths $\hat{\phi}_f$, from which the most likely focal length for initialization is derived using a hypothesize-and-verify framework. Specifically, this framework consists in using a Kernel Density Estimation (KDE) and Kernel Voting to identify the most likely focal length from which the final optimization step is started.

Whereas in the initialization only the focal length is our optimization target, in the final non-linear optimization step we refine all camera parameters in K using a non-linear optimization routine augmented with multiple robustness layers. The optimization step samples initial guesses from the focal length distribution $\hat{\phi}_f$ and uses subsets of fundamental matrices to define the cost function. This step is crucial for achieving faster convergence and numerical stability. The optimization is structured as a multi-

start procedure, where multiple initial guesses are sampled from the distribution $\hat{\phi}_f$. Each guess is used to initialize a Levenberg-Marquardt optimization that minimizes the Mendonça-Cipolla cost function [51]. The cost function is designed to encourage the essential matrix $E = K^\top F K$ derived from each fundamental matrix F to have identical non-zero singular values σ_1 and σ_2. This optimization process ensures that the estimated camera parameters are robust to noise and outliers.

The effectiveness of the proposed method is demonstrated through extensive experiments on both synthetic and real-world datasets. The experiments show that the proposed method outperforms state-of-the-art autocalibration techniques, particularly in dynamic scenes. The method's robustness to noise and outliers is highlighted by its ability to accurately estimate camera parameters in challenging scenarios with multiple moving objects. The datasets used for evaluation include both static scenes, such as the popular SfM benchmarks, and dynamic scenes from the `Hopkins155` dataset. For a thorough testing procedure, we also introduce a new dataset, the `Amiibo` dataset, which includes sequences with multiple independent motions.

3.3 Revisiting Calibration of Wide-Angle Radially Symmetric Cameras

The primary contribution of this work lies in its novel approach to calibrating radially symmetric cameras, which are increasingly popular due to their wide field-of-view capabilities in autonomous driving, augmented and virtual reality, and robotics. Traditional camera calibration methods, broadly categorized into geometric-based and learning-based approaches, have significant limitations when applied to such cameras. Geometric-based methods [1, 4, 7, 20, 42, 62, 66, 78, 90, 91] typically rely on structured environments and calibration objects, which restrict their applicability in unstructured settings. Learning-based methods [5, 26, 29, 37, 38, 40, 43, 65, 81, 86], while more adaptable to uncontrolled environments, often depend on specific camera models, limiting their generalizability and accuracy when different types of cameras are involved.

The main contribution of this work is a novel *two-step* learning-based framework that addresses the limitations of existing learning-based end-to-end autocalibration methods. This is achieved by introducing a model-agnostic camera representation we term VaCR, short for *Viewing-angle Camera Representation*.

The VaCR is a model-independent representation of the intrinsic properties of radially symmetric cameras. Specifically, the VaCR maps each image pixel to the direction of the 3D light ray that projects onto it. In the context of learning-based approaches, regressing a one-to-one mapping between each image pixel and its 3D direction is an expensive task. For this reason, the VaCR leverages the inherent symmetries in radially symmetric camera models to simplify the representation and reduce the number of parameters required to characterize any radially symmetric camera fully. The VaCR is also independent of specific camera models, which represents a significant advancement, as it allows the proposed autocalibration framework to generalize across various camera types without requiring retraining or architectural modifications to the network when the model of the camera changes, enabling zero-shot autocalibration for many radially symmetric cameras.

The proposed autocalibration framework, illustrated in Fig. 3.3, contributes to the state-of-the-art by articulating camera calibration in two steps for the first time.

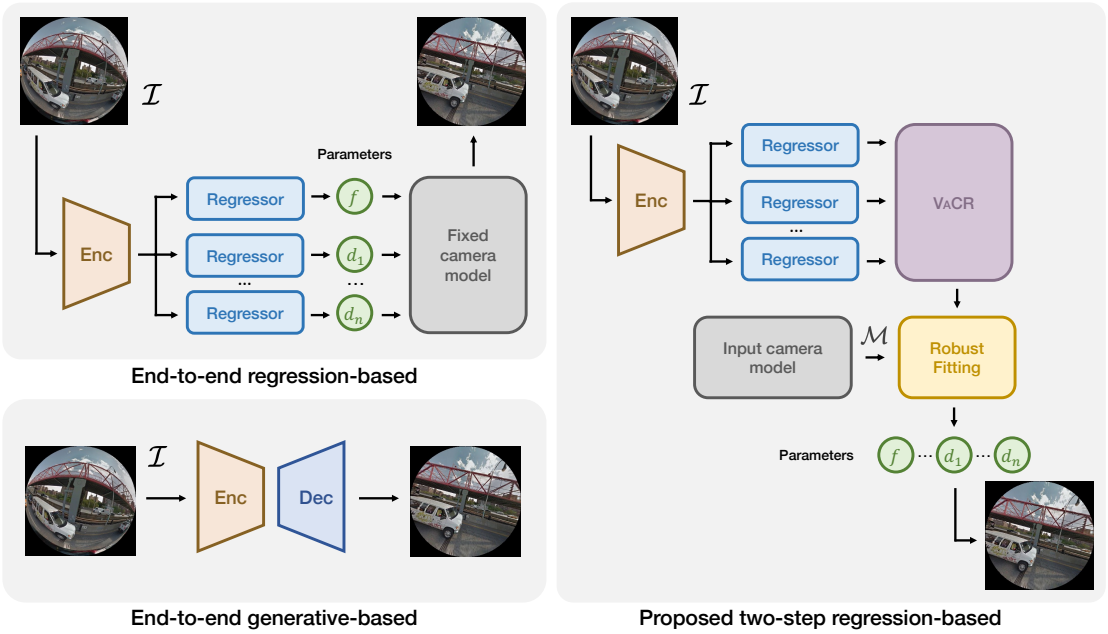

Figure 3.3: ***Two-step Wide-angle Autocalibration vs. End-to-end Methods.*** *Our two-step autocalibration method first regresses an implicit camera representation, the VaCR, that fully characterizes the radially symmetric camera that captured the input RGB image. During the second step, our method takes as input a radially symmetric camera model $\mathcal{M}$ and the VaCR. Together, these inputs drive a robust fitting procedure that outputs the camera parameters specific to the radially symmetric model $\mathcal{M}$. This contrasts with end-to-end regression-based and generative-based approaches, which directly estimate parameters for a fixed camera model or produce undistorted images without intermediary steps, respectively.*

(i) The first step involves the estimation of the VaCR from input images. This is achieved using a Convolutional Neural Network (CNN) specifically designed to handle both polar and Cartesian representations of the image. This work's major contribution is transforming the input image into polar coordinates before feature extraction. As a result of the polar transformation, the method effectively captures the radial symmetries inherent in wide field-of-view cameras. The extracted features are then used to estimate the VaCR, which maps each image point to a bearing vector on the unit sphere. This approach ensures that the VaCR is unambiguous and model-independent, allowing for accurate parameter estimation across a wide range of camera models.

(ii) The second step involves fitting the estimated VaCR to a specific camera model to recover the camera parameters. This is done using a robust fitting procedure that minimizes the differences between the predicted and actual viewing angles. Using a Cauchy loss function in this optimization process helps reduce the impact of outliers, further enhancing the accuracy of the parameter estimation. By separating the model-specific parameter estimation from the VaCR prediction, the method avoids retraining when switching between camera models, making it highly flexible and efficient.

In addition to its novel approach to camera representation and parameter estimation, the proposed method introduces a synthetic dataset generation strategy that ensures a uniform distribution of the angular field-of-view (AFOV) across different camera types. This is achieved by first sampling the AFOV uniformly and then generating cor-

responding synthetic images using sophisticated camera models like the Double Sphere Camera Model (DSCM) [80] and the Extended Unified Camera Model (EUCM) [32]. This approach mitigates biases in the training data and enhances the network's ability to generalize to various camera types.

The method's effectiveness is demonstrated through extensive experiments on several public datasets, including KITTI-360 [41], StreetLearn [53], SILDa [3], and WoodScape [88]. The results show that the proposed method outperforms state-of-the-art learning-based calibration techniques regarding both parameter prediction accuracy and image rectification quality. Furthermore, the ablation studies conducted as part of this work reveal the critical role of the proposed polar coordinate transformation and the dual-branch feature extraction network in achieving high calibration accuracy. These studies demonstrate that the combination of polar and Cartesian features significantly enhances the network's performance, validating the design choices made in developing the VaCR estimation network. The dual-branch architecture allows the network to leverage the strengths of both coordinate systems, leading to more accurate and robust VaCR estimations.

Another notable contribution of this work is the comprehensive evaluation of the proposed method's performance in image rectification tasks. Image rectification, which involves correcting distortions in images captured by wide-angle cameras, is a challenging problem that requires precise calibration. The proposed method's ability to produce high-quality rectified images, as evidenced by the high PSNR and SSIM scores, underscores its effectiveness in practical applications. This capability is beneficial in fields such as autonomous driving, virtual reality, and photogrammetry, where accurate image rectification is crucial.

The method also addresses the challenge of calibrating cameras with inherent ambiguities in their models, such as the EUCM [32] and DSCM [80]. These models can have multiple parameter sets that describe the same physical camera, making traditional calibration methods less effective. Using the VaCR as an intermediate representation, the proposed method can handle these ambiguities and still produce accurate parameter estimates. This is a significant advancement, as it expands the range of camera models that can be effectively calibrated using learning-based methods.

3.4 Multi-body depth and camera pose estimation from multiple views

Traditional and deep Structure-from-Motion methods operate under the assumption that the scene is rigid, meaning the environment is static or consists of a single moving object. This assumption is challenged by real-world scenarios where multiple independently moving objects are present. This work addresses the reconstruction of multiple rigid bodies in a scene and resolves the inherent scale ambiguity of Structure-from-Motion, ensuring that objects are reconstructed at consistent scales.

The problem tackled in this work is the estimation of depth and camera poses in dynamic scenes with multiple moving objects, also known as Multi-Body Structure-from-Motion (MBSfM). Traditional MBSfM methods segment rigid motions to obtain partial reconstructions of individually moving objects, but they suffer from the relative scale problem, where the 3D structure of each object is estimated up to a similarity transformation, resulting in inconsistent scales across objects. This issue prevents a

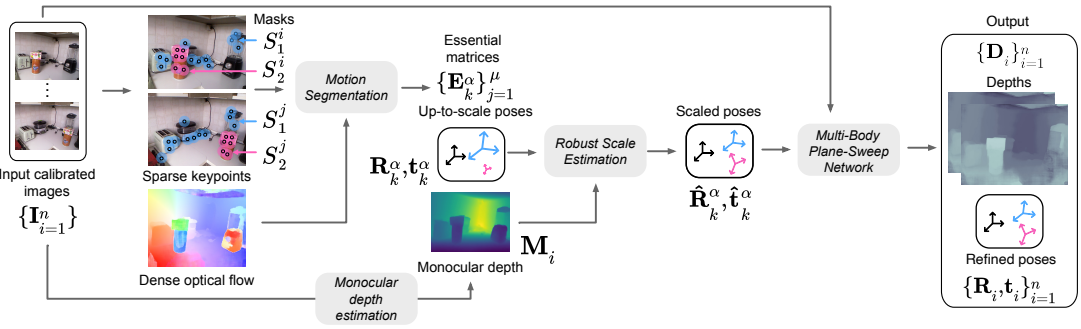

Figure 3.4: ***Multi-Body Structure-from-Motion:*** *This diagram outlines the key components of the proposed MBSfM framework. (i)* ***Motion Segmentation*** *is performed on sparse keypoint correspondences, which are enhanced with dense optical flow matches derived from the sparse set of input images. This process yields essential matrices that represent each motion in the scene as viewed from different image pairs. (ii) The* ***Robust Scale Estimator*** *uses monocular depth maps generated from each input image along with the essential matrices obtained from motion segmentation. These matrices are decomposed into scale-ambiguous rotations and translations, with a robust voting mechanism based on Gaussian kernel density estimation applied to rescale each camera pose, ensuring consistency in both reference frame and scale. (iii) The* ***Multi-Body Plane-Sweep Network*** *refines the camera poses and depth estimates from step (ii), achieving precise and improved depth maps and camera poses.*

unified reconstruction of all objects under a common global scale factor without additional information or learning priors.

The primary contribution of this paper is the introduction of a general dee learning-based Multi-Body Structure-from-Motion framework that aims to overcome the limitations of existing methodologies listed above, that is, not only the independent reconstruction of rigidly moving objects in the scene, but also the goal is to address the relative scale problem that prevents a unified reconstruction where all objects in the scene are reconstructed up to a single scale factor that is consistent throughout the scene.

Formally, we frame the Multi-Body Structure-from-Motion problem as follows. Our method receives as input a set of M *unstructured* images $\mathcal{I} = \{I_1, \dots, I_M\}$. The captured scene can be either static or dynamic, comprising of $B \geq 1$ independently moving rigid bodies $\mathcal{B} = \{\beta_1, \dots, \beta_B\}$. It is important to note that the Multi-Body Structure-from-Motion problem is different from non-rigid Structure-from-Motion [28, 34], which addresses a wider class of object deformations, but makes additional assumptions about the input images $\mathcal{I}$, which often are required to be sourced from a video sequence or captured using cameras with an orthographic projection. The goal of our method is to estimate dense depth maps $\{D_i\}_{i=1}^{M}$ for each input image $\mathcal{I}$ and the absolute camera poses $\{R_i, C_i\}_{i=1}^{M}$ of the cameras that captured the images in $\mathcal{I}$ in the reference frame of object β_1 that, without any loss of generality, is assumed to be static in the image sequence.

The core contributions of the proposed method align with the three main steps in the proposed MBSfM pipeline:

(i) Motion segmentation is performed using a combination of traditional sparse feature matching and dense optical flow matches, followed by robust multi-model fitting to estimate essential matrices that encode rigid motions. The motion segmentation step is critical for identifying the different rigid motions present in the scene. This involves

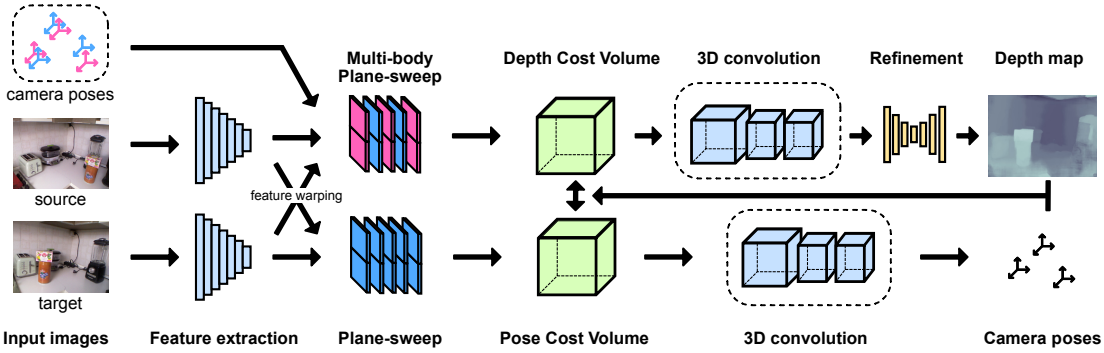

Figure 3.5: ***Multi-Body Plane-Sweep Network:*** *The figure illustrates the workflow of the multi-body plane-sweep network for joint estimation of camera poses and depth maps. The process begins with input images (source and target) and camera poses from the Robust Estimator step. Features are extracted from the images, followed by a multi-body plane-sweep step that generates multi-body feature maps through feature warping. These feature maps are then used to construct depth cost and pose cost volumes. 3D convolutions are applied to both volumes to estimate refined depth maps and camera poses. The refinement stage further enhances the depth map for more accurate results.*

extracting SIFT keypoints from each image and matching them across pairs of images. The matched keypoints are then used to estimate essential matrices using the RPA algorithm, which clusters the correspondences based on their rigid motion. This process results in a set of essential matrices, each representing the motion of a distinct rigid body in the scene.

(ii) Robust scale estimator uses monocular depth maps to estimate the scale factors for each object, applying a Kernel Density Estimator to derive a probability density function for the scale factors and resolve the scale ambiguity. Specifically, each essential matrix is factorized to obtain the relative camera poses, which are initially up to an unknown scale factor. To resolve these scale ambiguities, monocular depth maps are regressed for each image using an off-the-shelft monocular depth estimation network. These monocular depths provide a way to estimate the scale factors by comparing them to the triangulated depths obtained from the essential matrices. The scale factors are aggregated using a Kernel Density Estimator to derive the most likely scale for each object, ensuring consistency across the scene.

(iii) Multi-body plane sweep network The proposed network architecture, illustrated and commented in Fig. 3.5, represents the main contribution of this work. For the first time, we present a deep neural network that is capable of refines the depth maps and camera poses using the scale-consistent poses obtained from the *robust scale estimator*, which provides scale-consistent relative camera poses but not depth maps that are regressed from a network that is conditioned by multiple input images. In practice, this network operates by assigning virtual planes at different depths to each of the rigid motions idenitfied in the scene during the motion segmentation step. For each image pair, deep image features are then warped from one image to the other according to the estimated extrinsic camera parameters and the known calibration matrix K between image pairs and then are concatenated to construct a multi-body feature volume. The resulting feature volume is then processed by a series of 3D convolutional layers to produce the final depth maps and refined camera poses.

The contributions of this work are substantial, addressing several key challenges in multi-body SfM. By introducing a robust scale estimation technique and a novel multi-body plane sweep network, the authors have provided a comprehensive solution for depth and camera pose estimation in dynamic scenes. The method's ability to handle multiple rigid motions and resolve scale ambiguities sets it apart from traditional SfM approaches, making it a valuable tool for applications such as robot navigation and augmented reality.

The experimental results demonstrate the effectiveness of the proposed method in both static and dynamic scenes. On static scenes, the method performs comparably to state-of-the-art SfM methods, while on dynamic scenes, it significantly outperforms existing techniques in depth estimation. The experiments show that the method can accurately reconstruct multiple moving objects and provide consistent camera poses, even in challenging scenarios with complex motions.

In conclusion, this work advances the field of SfM by effectively addressing the scale ambiguity in multi-body scenes and leveraging deep learning to estimate depth and camera poses. The authors suggest future research directions, including extending the method to handle non-rigid deformations and integrating it into complete SfM pipelines. This innovative approach sets the stage for more robust and accurate reconstructions in complex dynamic environments, benefiting applications such as robot navigation and augmented reality.

CHAPTER 4

Conclusions & Future Work

In the works embedded in this thesis, we have presented advancements in the field of Structure-from-Motion by developing novel geometric- and learning-based autocalibration algorithms and progressing towards a comprehensive Multi-Body Structure-from-Motion framework. Our approach leverages learning priors to unify inconsistent object scales in 3D reconstruction and utilizes a geometric plane-sweep-inspired algorithm embedded within a neural network to refine depths and camera poses using multi-view epipolar constraints.

In the area of autocalibration, we introduce a family of minimal autocalibration solvers capable of performing joint Euclidean reconstruction and intrinsic parameter estimation from keypoint correspondences between image pairs, and notably, triplets. We optimized these solvers for efficiency by developing a general theory of minimal relaxations grounded in algebraic geometry principles. By employing numerical methods, such as homotopy continuation and monodromy, to solve these minimal problems, we harness the parallel computing capabilities of modern computer architectures, making these solvers fast enough for practical use in Structure-from-Motion software like COLMAP. Integration of our minimal solvers within COLMAP results in enhanced 3D reconstructions, particularly in challenging datasets.

With the recent advances in GPU-accelerated computing, driven by the rapid growth of Artificial Intelligence and the expansion of parallel computing solutions on edge devices, our solvers are particularly well-positioned relative to other autocalibration and 3D reconstruction algorithms. Our algorithms exhibit runtime scaling linearly with the number of available computing cores, suggesting that these parallel computing architectures can make our solvers significantly faster, potentially enabling real-time applications in visual localization, and navigation for robots and AR devices. Several studies [] have demonstrated the effectiveness of running numerical solvers on GPUs, achieving

substantial runtime reductions for various tasks in 3D Computer Vision. Advancing in this direction could further improve the performance of our solvers, whose autocalibration formulations, while optimized through our proposed minimal relaxation theory, remain inherently complex. For instance, our formulations involve solving problems with approximately 2000 complex solutions, whereas the previously most challenging problem, known as the "Scranton" problem (reconstructing four points in three calibrated views), involves only 272 solutions. Overall, as more computing cores become available, our solvers stand to benefit significantly, since each of the 2000 solutions can be processed in parallel on individual GPU cores, thereby enhancing the speed of the algorithms proportionally to the number of cores available.

Recognizing that our family of minimal solvers is currently limited to calibrating pinhole cameras and does not account for radial and tangential distortions present in wide-angle, fisheye, and 360-degree cameras commonly used in automotive applications, robot navigation, and AR devices. While geometric methods generally perform well for pinhole camera autocalibration, even in real-world scenarios, they face additional challenges when calibrating distortion parameters of wide-angle cameras. Our study demonstrates that learning-based approaches have outperformed purely geometric methods in terms of accuracy and robustness for image rectification. Therefore, we directed our efforts towards advancing autocalibration using a learning-based, zero-shot autocalibration network for radially symmetric cameras. This approach leverages the inherent symmetries of these camera sensors to efficiently and robustly regress distortion parameters through a camera model-free implicit representation.

Although this approach already competes with and often surpasses the state-of-the-art performance in wide-angle camera calibration, there remains significant potential for future improvements. Currently, the method combines learned image features in both polar and Cartesian coordinate systems by concatenating feature volumes, which are then fed into the regression head to produce the implicit camera representation. However, recent developments in cross-attention architectures, inspired by the success of transformer-based models like Vision Transformers (ViT) [], have shown improved performance over the traditional deep learning practice of concatenating feature volumes, which is employed in our current method. By incorporating cross-attention mechanisms and a token-based approach to feature extraction and processing, we anticipate enhancements in the performance of our autocalibration method compared to the originally published results.

Further improvements could include extending the method to support tangential distortion and even more complex camera models, such as the FisheyeRadTanThinPrism model commonly used for advanced fisheye cameras found in popular AR devices and consumer action cameras. Additionally, the proposed camera representation is both general and compact, making it highly suitable for integration into other learning-based algorithms within 3D reconstruction and camera calibration domains, such as differentiable Structure-from-Motion pipelines [], which are gaining popularity in the 3D Computer Vision community.

In developing a general Multi-Body Structure-from-Motion (MBSfM) framework, we have successfully integrated learning-based monocular depth estimation priors with a multi-body adapted geometric plane-sweep algorithm. Additionally, we developed a practical autocalibration algorithm that utilizes the multi-view constraints induced by

rigidly moving objects in the scene to refine camera intrinsic parameters. This approach improves calibration accuracy and/or reduces the number of views required for full camera calibration. The resulting framework is versatile and performs effectively in practice, yielding consistently scaled reconstructions throughout the entire scene.

In retrospect, given the rapid advancements in monocular depth estimation solutions for metric depth estimation, and the impressive depth estimation capabilities of recent diffusion methods from single images, our framework is well-positioned to remain relevant both now and in the future. As learning-based methods continue to provide increasingly accurate initial depth estimates, they will offer improved starting points for our geometric, multi-body approach to depth and camera pose refinement.

However, a significant limitation of our method is its inability to support non-rigid deformations of moving objects within the scene, which restricts its applicability to specific scenarios, such as industrial robotics or autonomous navigation in controlled or semi-controlled environments. Despite this limitation, recent studies [**?**] have shown that deformable motion can be effectively modeled as a linear combination of a limited number of rigid motion transformations. These transformations can be either learned by a multi-layer perceptron tailored to each specific image sequence or identified through geometric-based motion segmentation algorithms, as demonstrated in our original work.

Furthermore, another area for potential improvement lies in our reliance on purely geometric object motion segmentation algorithms. It is important to note that recent advancements in foundational learning-based models have exhibited impressive performance in general object segmentation, often surpassing the capabilities of geometric methods while also providing dense object masks. Integrating these advanced segmentation techniques could further enhance the robustness and versatility of our MBSfM framework.

Bibliography

[1] Miguel Alemán-Flores, Luis Alvarez, Luis Gomez, and Daniel Santana-Cedrés. Automatic lens distortion correction using one-parameter division models. *Image Processing On Line*, 4:327–343, 2014.

[2] Federica Arrigoni, Elisa Ricci, and Tomas Pajdla. Multi-frame motion segmentation by combining two-frame results. *International Journal of Computer Vision*, 130(3):696–728, 2022.

[3] V. Balntas. SILDa: A Multi-Task Dataset for Evaluating Visual Localization. *Medium*, Apr 2019.

[4] Burak Benligiray and Cihan Topal. Blind rectification of radial distortion by line straightness. In *2016 24th European Signal Processing Conference (EUSIPCO)*, pages 938–942. IEEE, 2016.

[5] Oleksandr Bogdan, Viktor Eckstein, Francois Rameau, and Jean-Charles Bazin. Deepcalib: A deep learning approach for automatic intrinsic calibration of wide field-of-view cameras. In *Proceedings of the 15th ACM SIGGRAPH European Conference on Visual Media Production*, pages 1–10, 2018.

[6] Eric Brachmann, Jamie Wynn, Shuai Chen, Tommaso Cavallari, Áron Monszpart, Daniyar Turmukhambetov, and Victor Adrian Prisacariu. Scene coordinate reconstruction: Posing of image collections via incremental learning of a relocalizer. *arXiv preprint arXiv:2404.14351*, 2024.

[7] Faisal Bukhari and Matthew N Dailey. Automatic radial distortion estimation from a single image. *Journal of mathematical imaging and vision*, 45:31–45, 2013.

[8] Manmohan Krishna Chandraker, Sameer Agarwal, David J. Kriegman, and Serge J. Belongie. Globally optimal affine and metric upgrades in stratified autocalibration. In *IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14-20, 2007*, pages 1–8. IEEE Computer Society, 2007.

[9] Manmohan Krishna Chandraker, Sameer Agarwal, David J. Kriegman, and Serge J. Belongie. Globally optimal algorithms for stratified autocalibration. *Int. J. Comput. Vis.*, 90(2):236–254, 2010.

[10] Chun-Hao Chao, Pin-Lun Hsu, Hung-Yi Lee, and Yu-Chiang Frank Wang. Self-supervised deep learning for fisheye image rectification. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2248–2252. IEEE, 2020.

[11] Yuhua Chen, Cordelia Schmid, and Cristian Sminchisescu. Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7063–7072, 2019.

[12] Robert T Collins. A space-sweep approach to true multi-image matching. In *Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 358–363. Ieee, 1996.

[13] Joao Costeira and Takeo Kanade. A multi-body factorization method for motion analysis. In *Proceedings of IEEE International Conference on Computer Vision*, pages 1071–1076. IEEE, 1995.

[14] Olivier D. Faugeras, Quang-Tuan Luong, and Stephen J. Maybank. Camera self-calibration: Theory and experiments. In Giulio Sandini, editor, *Computer Vision - ECCV'92, Second European Conference on Computer Vision, Santa Margherita Ligure, Italy, May 19-22, 1992, Proceedings*, volume 588 of *Lecture Notes in Computer Science*, pages 321–334. Springer, 1992.

Bibliography

[15] Hao Feng, Wendi Wang, Jiajun Deng, Wengang Zhou, Li Li, and Houqiang Li. Simfir: A simple framework for fisheye image rectification with self-supervised representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12418–12427, 2023.

[16] Andrew W Fitzgibbon and Andrew Zisserman. Multibody structure and motion: 3-d reconstruction of independently moving objects. pages 891–906. Springer, 2000.

[17] Andrea Fusiello. Uncalibrated Euclidean reconstruction: a review. *Image and Vision Computing*, 18(6-7):555–563, 2000.

[18] Andrea Fusiello, Arrigo Benedetti, Michela Farenzena, and Alessandro Busti. Globally convergent autocalibration using interval analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(12):1633–1638, 2004.

[19] Guillermo Gallego, Elias Mueggler, and Peter F. Sturm. Translation of "Zur Ermittlung eines Objektes aus zwei Perspektiven mit innerer Orientierung" by Erwin Kruppa (1913). *CoRR*, abs/1801.01454, 2018.

[20] Diego Gonzalez-Aguilera, Javier Gomez-Lahoz, and Pablo Rodríguez-Gonzálvez. An automatic approach for radial lens distortion correction from a single image. *IEEE Sensors journal*, 11(4):956–965, 2010.

[21] Ariel Gordon, Hanhan Li, Rico Jonschkowski, and Anelia Angelova. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8977–8986, 2019.

[22] R. Hartley and A. Zisserman. *Multiple view geometry in Computer Vision*. Cambridge University Press, Cambridge, second edition, 2003. With a foreword by Olivier Faugeras.

[23] R.I. Hartley. Kruppa's equations derived from the fundamental matrix. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2):133–135, 1997.

[24] Richard I Hartley. Euclidean reconstruction from uncalibrated views. In *Joint European-US workshop on applications of invariance in computer vision*, pages 235–256. Springer, 1993.

[25] Anders Heyden and Kalle Åström. Flexible calibration: Minimal cases for auto-calibration. In *Proceedings of the International Conference on Computer Vision, Kerkyra, Corfu, Greece, September 20-25, 1999*, pages 350–355. IEEE Computer Society, 1999.

[26] Masaki Hosono, Edgar Simo-Serra, and Tomonari Sonoda. Self-supervised deep fisheye image rectification approach using coordinate relations. In *2021 17th International Conference on Machine Vision and Applications (MVA)*, pages 1–5. IEEE, 2021.

[27] Petr Hruby, Timothy Duff, Anton Leykin, and Tomás Pajdla. Learning to solve hard minimal problems. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 5522–5532. IEEE, 2022.

[28] Sebastian Hoppe Nesgaard Jensen, Mads Emil Brix Doest, Henrik Aanæs, and Alessio Del Bue. A benchmark and evaluation of non-rigid structure from motion. *International Journal of Computer Vision*, 129(4):882–899, 2021.

[29] Linyi Jin, Jianming Zhang, Yannick Hold-Geoffroy, Oliver Wang, Kevin Blackburn-Matzen, Matthew Sticha, and David F Fouhey. Perspective fields for single image camera calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17307–17316, 2023.

[30] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. *arXiv preprint arXiv:2307.07635*, 2023.

[31] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

[32] Bogdan Khomutenko, Gaëtan Garcia, and Philippe Martinet. An enhanced unified camera model. *IEEE Robotics and Automation Letters*, 1(1):137–144, 2015.

[33] Erwin Kruppa. *Zur Ermittlung eines Objektes aus zwei Perspektiven mit innerer Orientierung*. Hölder, 1913.

[34] Suryansh Kumar. Non-rigid structure from motion: Prior-free factorization method revisited. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 51–60, 2020.

[35] Abhijit Kundu, K Madhava Krishna, and CV Jawahar. Realtime multibody visual slam with a smoothly moving monocular camera. In *2011 International Conference on Computer Vision*, pages 2080–2087. IEEE, 2011.

[36] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. *arXiv preprint arXiv:2406.09756*, 2024.

[37] Kang Liao, Chunyu Lin, Lixin Liao, Yao Zhao, and Weiyao Lin. Multi-level curriculum for training a distortion-aware barrel distortion rectification model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4389–4398, 2021.

[38] Kang Liao, Chunyu Lin, Yunchao Wei, Feng Li, Shangrong Yang, and Yao Zhao. Towards complete scene and regular shape for distortion rectification by curve-aware extrapolation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14569–14578, 2021.

[39] Kang Liao, Chunyu Lin, Yao Zhao, and Moncef Gabbouj. Dr-gan: Automatic radial distortion rectification using conditional gan in real-time. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(3):725–733, 2019.

[40] Kang Liao, Chunyu Lin, Yao Zhao, and Mai Xu. Model-free distortion rectification framework bridged by distortion distribution map. *IEEE Transactions on Image Processing*, 29:3707–3718, 2020.

[41] Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3292–3310, 2022.

[42] Yaroslava Lochman, Oles Dobosevych, Rostyslav Hryniv, and James Pritts. Minimal solvers for single-view lens-distorted camera auto-calibration. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2887–2896, 2021.

[43] Manuel Lopez, Roger Mari, Pau Gargallo, Yubin Kuang, Javier Gonzalez-Jimenez, and Gloria Haro. Deep single image camera calibration with radial distortion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11817–11825, 2019.

[44] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent video depth estimation. *ACM Transactions on Graphics (ToG)*, 39(4):71–1, 2020.

[45] Quang-Tuan Luong, Rachid Deriche, Olivier Faugeras, and Theodore Papadopoulo. *On determining the fundamental matrix: Analysis of different methods and experimental results*. PhD thesis, Inria, 1993.

[46] Quang-Tuan Luong and Olivier D. Faugeras. Self-calibration of a moving camera from point correspondences and fundamental matrices. *Int. J. Comput. Vis.*, 22(3):261–289, 1997.

[47] Luca Magri and Andrea Fusiello. T-linkage: A continuous relaxation of j-linkage for multi-model fitting. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3954–3961. IEEE.

[48] Evgeniy Martyushev, Jana Vráblíková, and Tomas Pajdla. Optimizing elimination templates by greedy parameter search. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15733–15743. IEEE, 2022.

[49] Evgeniy V. Martyushev. A minimal six-point auto-calibration algorithm. *http://arxiv.org/abs/1307.3759*, 2013.

[50] Stephen J. Maybank and Olivier D. Faugeras. A theory of self-calibration of a moving camera. *Int. J. Comput. Vis.*, 8(2):123–151, 1992.

[51] P.R.S. Mendonça and R. Cipolla. A simple technique for self-calibration. pages I:500–505, 1999.

[52] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[53] Piotr Mirowski, Andras Banki-Horvath, Keith Anderson, Denis Teplyashin, Karl Moritz Hermann, Mateusz Malinowski, Matthew Koichi Grimes, Karen Simonyan, Koray Kavukcuoglu, Andrew Zisserman, et al. The streetlearn environment and dataset. *arXiv preprint arXiv:1903.01292*, 2019.

[54] Kemal E Ozden, Konrad Schindler, and Luc Van Gool. Multibody structure-from-motion in practice. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(6):1134–1141, 2010.

[55] Kemal Egemen Ozden, Kurt Cornelis, Luc Van Eycken, and Luc Van Gool. Reconstructing 3d trajectories of independently moving objects using generic constraints. *Computer Vision and Image Understanding*, 96(3):453–471, 2004.

[56] Kemal Egemen Ozden, Konrad Schindler, and Luc Van Gool. Multibody structure-from-motion in practice. 32(6):1134–1141, 2010.

[57] Danda Pani Paudel and Luc Van Gool. Sampling algebraic varieties for robust camera autocalibration. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 265–281, 2018.

[58] Danda Pani Paudel and Luc Van Gool. Sampling algebraic varieties for robust camera autocalibration. pages 265–281, 2018.

[59] Marc Pollefeys, Luc Van Gool, and André Oosterlinck. The modulus constraint: a new constraint self-calibration. In *13th International Conference on Pattern Recognition, ICPR 1996, Vienna, Austria, 25-19 August, 1996*, pages 349–353. IEEE Computer Society, 1996.

[60] Marc Pollefeys, Reinhard Koch, and Luc Van Gool. Self-calibration and metric reconstruction inspite of varying and unknown intrinsic camera parameters. *Int. J. Comput. Vis.*, 32(1):7–25, 1999.

[61] Marc Pollefeys and Luc Van Gool. A stratified approach to metric self-calibration. In *Proceedings of IEEE computer society conference on computer vision and pattern recognition*, pages 407–412. IEEE, 1997.

[62] James Pritts, Zuzana Kukelova, Viktor Larsson, and Ondřej Chum. Radially-distorted conjugate translations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1993–2001, 2018.

[63] Long Quan, Bill Triggs, and Bernard Mourrain. Some results on minimal Euclidean reconstruction from four points. *Journal of Mathematical Imaging and Vision*, 24:341–348, 2006.

[64] Liangliang Ren, Yangyang Song, Jiwen Lu, and Jie Zhou. Spatial geometric reasoning for room layout estimation via deep reinforcement learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pages 550–565. Springer, 2020.

[65] Jiangpeng Rong, Shiyao Huang, Zeyu Shang, and Xianghua Ying. Radial lens distortion correction using convolutional neural networks trained with synthesized images. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part III 13*, pages 35–49. Springer, 2017.

[66] Daniel Santana-Cedrés, Luis Gomez, Miguel Alemán-Flores, Agustín Salgado, Julio Esclarín, Luis Mazorra, and Luis Alvarez. An iterative optimization algorithm for lens distortion correction using two-parameter models. *Image Processing On Line*, 6:326–364, 2016.

[67] Muhamad Risqi U Saputra, Andrew Markham, and Niki Trigoni. Visual slam and structure from motion in dynamic environments: A survey. *ACM Computing Surveys (CSUR)*, 51(2):1–36, 2018.

[68] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[69] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016.

[70] Cameron Smith, David Charatan, Ayush Tewari, and Vincent Sitzmann. Flowmap: High-quality camera poses, intrinsics, and depth via gradient descent. *arXiv preprint arXiv:2404.15259*, 2024.

[71] Christoph Strecha, Wolfgang Von Hansen, Luc Van Gool, Pascal Fua, and Ulrich Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. Ieee, 2008.

[72] P. Sturm. A case against Kruppa's equations for camera self-calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1199–1204, 2000.

[73] Chengzhou Tang and Ping Tan. Ba-net: Dense bundle adjustment network. *arXiv preprint arXiv:1806.04807*, 2018.

[74] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020.

[75] Carlo Tomasi and Takeo Kanade. Shape and motion from image streams under orthography: a factorization method. *International journal of computer vision*, 9(2):137–154, 1992.

[76] Bill Triggs. Autocalibration and the absolute quadric. In *1997 Conference on Computer Vision and Pattern Recognition (CVPR '97), June 17-19, 1997, San Juan, Puerto Rico*, pages 609–614. IEEE Computer Society, 1997.

[77] Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment—a modern synthesis. In *Vision Algorithms: Theory and Practice: International Workshop on Vision Algorithms Corfu, Greece, September 21–22, 1999 Proceedings*, pages 298–372. Springer, 2000.

[78] Roger Tsai. A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses. *IEEE Journal on Robotics and Automation*, 3(4):323–344, 1987.

[79] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. Demon: Depth and motion network for learning monocular stereo. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5038–5047, 2017.

[80] Vladyslav Usenko, Nikolaus Demmel, and Daniel Cremers. The double sphere camera model. In *2018 International Conference on 3D Vision (3DV)*, pages 552–560. IEEE, 2018.

[81] Nobuhiko Wakai, Satoshi Sato, Yasunori Ishii, and Takayoshi Yamashita. Rethinking generic camera models for deep single image camera calibration to recover rotation and fisheye distortion. In *European Conference on Computer Vision*, pages 679–698. Springer, 2022.

[82] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Vggsfm: Visual geometry grounded deep structure from motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21686–21697, 2024.

[83] Jianyuan Wang, Yiran Zhong, Yuchao Dai, Stan Birchfield, Kaihao Zhang, Nikolai Smolyanskiy, and Hongdong Li. Deep two-view structure-from-motion revisited. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8953–8962, 2021.

[84] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024.

[85] Xingkui Wei, Yinda Zhang, Zhuwen Li, Yanwei Fu, and Xiangyang Xue. Deepsfm: Structure from motion via deep bundle adjustment. In *European conference on computer vision*, pages 230–247. Springer, 2020.

[86] Zhucun Xue, Nan Xue, Gui-Song Xia, and Weiming Shen. Learning to calibrate straight lines for fisheye image rectification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1643–1651, 2019.

[87] Shangrong Yang, Chunyu Lin, Kang Liao, Chunjie Zhang, and Yao Zhao. Progressively complementary network for fisheye image rectification using appearance flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6348–6357, 2021.

[88] Senthil Yogamani, Ciarán Hughes, Jonathan Horgan, Ganesh Sistu, Padraig Varley, Derek O'Dea, Michal Uricár, Stefan Milz, Martin Simon, Karl Amende, et al. Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9308–9318, 2019.

[89] Cyril Zeller and Olivier Faugeras. *Camera self-calibration from video sequences: the Kruppa equations revisited*. PhD thesis, INRIA, 1996.

[90] Mi Zhang, Jian Yao, Menghan Xia, Kai Li, Yi Zhang, and Yaping Liu. Line-based multi-label energy optimization for fisheye image rectification and calibration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4137–4145, 2015.

[91] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence*, 22(11):1330–1334, 2000.

[92] Zhoutong Zhang, Forrester Cole, Zhengqi Li, Michael Rubinstein, Noah Snavely, and William T Freeman. Structure and motion from casual videos. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pages 20–37. Springer, 2022.

[93] Keyao Zhao, Chunyu Lin, Kang Liao, Shangrong Yang, and Yao Zhao. Revisiting radial distortion rectification in polar-coordinates: A new and efficient learning perspective. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(6):3552–3560, 2021.

CHAPTER 5

Published Works