POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Two-View Motion Segmentation with Semi-Calibrated Cameras via Hypothesis Filtering

Tesi di Laurea Magistrale in
Computer Science and Engineering - Ingegneria Informatica

Marco Cerino, 11011193

Advisor:
Prof. Luca Magri

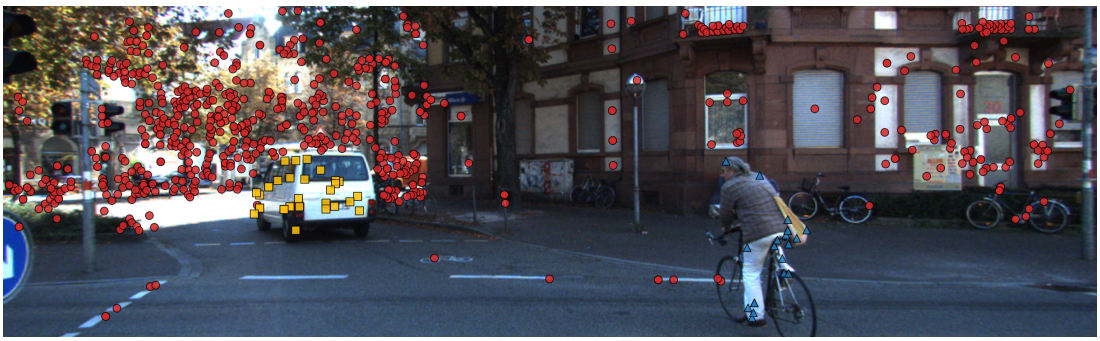
Academic year:
2024-2025

Abstract: Two-view motion segmentation, the task of partitioning sparse correspondences into multiple rigid motions, is inherently challenging in a purely uncalibrated setting. Standard approaches often treat motion models as independent geometric entities, which neglects a strong geometric prior available in many practical scenarios: images are often acquired by a single sensor with constant, albeit unknown, intrinsic parameters. In this paper, we address this problem by operating in a *semi-calibrated regime*. We propose a robust adaptation of the T-Linkage algorithm, a preference-based clustering method, by introducing a motion hypothesis filtering step based on focal length estimation. Instead of relying on all generated candidate motions, we exploit the constant-intrinsics assumption to identify and discard geometrically inconsistent models before the clustering phase. Extensive experiments on the HOPE-F and AdelaideRMF benchmarks demonstrate that leveraging this semi-calibrated constraint significantly improves segmentation accuracy compared to the uncalibrated baseline.

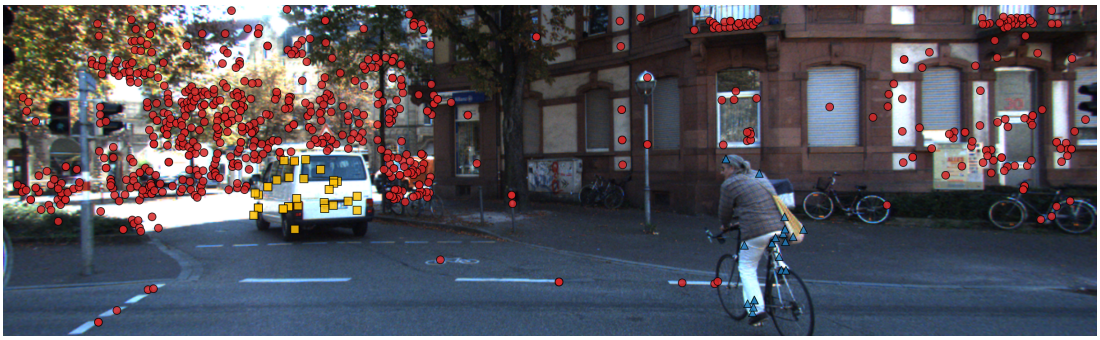
Key-words: Motion Segmentation, Multi-Model Fitting, Semi-Calibrated Geometry

1. Introduction

Motion segmentation, the task of grouping point correspondences according to their underlying rigid 3D motions, is a fundamental step for reconstructing and understanding dynamic scenes [26]. Existing techniques typically rely on different input assumptions, ranging from *trajectory clustering* [7] (which requires reliable long-term tracking) to the case of *unknown correspondences* (which is often computationally intractable). A strategic trade-off is represented by the usage of *two-frame correspondences*, where the task is restricted to clustering sparse matches between pairs of images without assuming temporal continuity [18, 20].
Adopting this setting, in this document we specifically address the *two-view motion segmentation* problem: given two images of a dynamic scene, the goal is to segment point correspondences into disjoint subsets, each associated with a distinct rigid motion.



(a) View 1



(b) View 2

Figure 1: **Robust Two-View Motion Segmentation in the Wild.** An example result from the KITTI benchmark using our Semi-Calibrated T-Linkage.

An example of this task in a real-world driving scenario is illustrated in Figure 1. In this complex urban scene, our method successfully identifies $k = 3$ distinct motion groups. Specifically, the algorithm separates the dominant motion of the static background (red circles), induced by the camera's ego-motion, from the independent motions of two dynamic actors: a cyclist in the foreground (blue triangles) and a commercial vehicle transiting on the left (yellow squares). The ability to distinguish these multiple rigid structures, despite the limited sparse correspondences and the geometric similarity between certain motion vectors, highlights the effectiveness of our proposed approach, which leverages novel geometric constraints as detailed below.

1.1. Motivation and Applications

Resolving multiple motions from two views is a critical capability for Visual SLAM and dynamic scene understanding. This task is particularly relevant in the context of *semi-calibrated devices*, which constitute the standard setup in many real-world applications.
In *Mobile Vision* and *Augmented Reality*, where smartphones provide approximate focal lengths via API or EXIF metadata, motion grouping is essential for robust mapping and realistic occlusion handling. In *autonomous driving* [9] and robotics, monocular cameras often operate with intrinsics that are constant over short time windows, even if factory calibration has drifted. Similarly, in *video analysis* "in the wild" (e.g., body-cams or dash-cams), the focal length is often estimable or assumed constant between consecutive frames, providing a valuable constraint that is typically ignored by purely uncalibrated solvers.

1.2. The Challenge

The problem constitutes a challenging instance of multi-model fitting, as it requires the simultaneous estimation of multiple geometric structures from data contaminated by noise and outliers, without prior knowledge of the number of motions [11]. A primary difficulty in two-view motion segmentation arises from the numerical instability of epipolar geometry estimation within an uncalibrated framework. In the absence of intrinsic parameters, the Fundamental matrix remains weakly constrained by available correspondences, particularly when affected by measurement noise or the limited spatial support typical of cluttered environments.
This susceptibility is further exacerbated when correspondences lie on a dominant plane, a configuration where the Fundamental matrix cannot be uniquely or accurately recovered, resulting in ambiguous solutions [5, 31]. Conversely, the Essential matrix does not suffer from this specific ambiguity, as the internal metric properties

of the camera provide the necessary constraints to ensure a more robust estimation.

1.3. The Gap in the Literature: The Semi-Calibrated Regime

In the literature, when dealing with sparse matches from two views, research has predominantly focused on two extremes: the purely *uncalibrated* case, where Fundamental matrices are estimated without any prior on the sensors [2, 18, 25], or the *fully calibrated* case, where Essential matrices are recovered assuming known intrinsic parameters [19]. Both scenarios can be cast as a multi-model fitting framework where multiple instances of epipolar geometry are estimated simultaneously.
In this work, we explore a less addressed but practical scenario: the *semi-calibrated* regime. In this context, we assume that images are acquired by a sensor with constant, albeit unknown, intrinsic parameters, typically simplified to a single unknown focal length. While semi-calibrated two-view geometry has been extensively studied for static scene reconstruction (single motion) [14, 23], most motion segmentation approaches overlook this constraint: they either simplify the problem by assuming affine cameras or rely on purely uncalibrated models that treat each motion as an independent geometric entity.

1.4. Proposed Approach

Exploiting the fact that if multiple motions are captured by the same camera within the same scene, their respective fundamental matrices must necessarily share the same focal length value, we propose a robust customization of the T-Linkage method [18]. We introduce a motion hypothesis pre-filtering step that enforces *focal length consistency* as a requirement for model admissibility. This allows us to identify and discard hypotheses that, while algebraically valid in an uncalibrated sense, are geometrically inconsistent with the underlying camera sensor.
We exploit focal length estimates derived from each Fundamental matrix to identify and discard geometrically inconsistent models. Specifically, we integrate a robust focal length solver [14] into the sampling process of T-Linkage and enforce a scene-level consensus via Kernel Density Estimation (KDE) [21]. This acts as a pre-filtering mechanism designed to prune the pool of generated hypotheses; by discarding models that violate the constant intrinsics assumption, we ensure that the subsequent clustering operates on a cleaner set of hypotheses that are more likely to represent valid camera geometries.
Our main contributions are:

- **Geometric Hypothesis Pruning:** We introduce a rejection-based sampling strategy that adapts T-Linkage to the semi-calibrated regime. Instead of seeking optimal models, we focus on identifying and discarding geometrically inconsistent hypotheses prior to the clustering phase.
- **Scene-Level Focal Consensus:** We propose a consensus mechanism based on KDE to identify the dominant focal length, effectively pruning algebraic artifacts that do not correspond to a valid calibration matrix.

We demonstrate the effectiveness of this approach through extensive experiments on the HOPE-F [13] and AdelaideRMF [30] benchmarks. Our results suggest that even minimal calibration cues, such as the constancy of focal length, can significantly improve two-view motion segmentation when properly integrated into the model fitting process.

2. Background

This section establishes the formal mathematical and geometric foundations required to comprehend the problem of two-view motion segmentation and the proposed semi-calibrated framework. We systematically introduce the algebraic representation of camera projection, the strict constraints governing two-view epipolar geometry, and the sequential pipeline adopted for 3D reconstruction.

2.1. Geometric and Algebraic Foundations

Computer vision fundamentally deals with the recovery of 3D structure from 2D image projections. In a standard Euclidean space $\mathbb{R}^3$, the perspective projection is a non-linear operation involving division by depth. To treat this and other geometric transformations, such as rigid body motions, within a unified linear algebraic framework, we map points to the *Projective Space* $\mathbb{P}^n$ using homogeneous coordinates.

2.1.1. Projective Geometry and Homogeneous Coordinates

The projective space $\mathbb{P}^n$ is defined as the set of lines passing through the origin of $\mathbb{R}^{n+1}$. A point in the Euclidean space $\mathbf{X}_{euc} = [X, Y, Z]^T \in \mathbb{R}^3$ is represented in homogeneous coordinates as a 4-vector $\mathbf{X} \in \mathbb{P}^3$ by appending a non-zero scale factor w:

$$\mathbf{X} = \begin{bmatrix} wX \\ wY \\ wZ \\ w \end{bmatrix} = \begin{bmatrix} x \\ y \\ z \\ w \end{bmatrix}, \quad w \neq 0 \tag{1}$$

Crucially, homogeneous vectors are defined only up to a scale factor. Two vectors $\mathbf{X}$ and $\mathbf{X}'$ represent the same physical point if $\mathbf{X} = \lambda \mathbf{X}'$ for any scalar $\lambda \neq 0$. This equivalence relation implies that the mapping from $\mathbb{P}^3$ back to $\mathbb{R}^3$ involves the division by the last component:

$$\mathbf{X}_{euc} = \left[\frac{x}{w}, \frac{y}{w}, \frac{z}{w}\right]^T \tag{2}$$

Points with $w = 0$ have no counterpart in the finite Euclidean space; they represent *points at infinity* (or ideal points), which allow modeling the intersection of parallel lines and play a key role in camera calibration theory. Similarly, 2D image points $\mathbf{x}_{euc} = [u, v]^T \in \mathbb{R}^2$ are represented in $\mathbb{P}^2$ as homogeneous 3-vectors $\mathbf{x} = [u, v, 1]^T$.

2.1.2. Rigid Body Transformations

Motion in a dynamic scene is modeled as a rigid body transformation mapping a point from a source coordinate system to a target coordinate system (e.g., from World to Camera frame). The set of all valid rigid transformations forms the *Special Euclidean group* $SE(3)$, which preserves both distances and orientation.
A transformation $T \in SE(3)$ is parameterized by a rotation R and a translation $\mathbf{t}$. The rotation belongs to the *Special Orthogonal group* $SO(3)$, defined as:

$$SO(3) = \{R \in \mathbb{R}^{3\times 3} \mid R^T R = I, \det(R) = +1\} \tag{3}$$

The translation is a vector $\mathbf{t} \in \mathbb{R}^3$. In Euclidean coordinates, the transformation is affine:

$$\mathbf{X}'_{euc} = R\mathbf{X}_{euc} + \mathbf{t} \tag{4}$$

By leveraging homogeneous coordinates, this affine mapping becomes a linear operation represented by a 4×4 matrix multiplication:

$$\mathbf{X}' = \begin{bmatrix} R & \mathbf{t} \\ \mathbf{0}^T & 1 \end{bmatrix} \mathbf{X} \tag{5}$$

This matrix formulation is essential for chaining multiple transformations (e.g., camera ego-motion followed by object motion) via simple matrix multiplication.

2.2. Image Formation and Camera Models

The physical process of capturing a 3D scene onto a 2D sensor is mathematically abstracted by the *pinhole camera model*. While ideal, this model provides an accurate geometric approximation for most modern optical systems once lens distortion has been corrected. Geometrically, the model assumes that light rays from 3D objects pass through a single point in space, the *camera center* (or optical center) $\mathbf{C}$, and intersect a specific plane, the *image plane* or *retinal plane*, located at a distance f (the focal length) from $\mathbf{C}$ along the *principal axis*.

2.2.1. The Projective Mapping

Let the camera center $\mathbf{C}$ be the origin of the Euclidean Camera Coordinate System. Consider a 3D point in space $\mathbf{X}_{cam} = [X_c, Y_c, Z_c]^T$. Under the pinhole projection, this point is mapped to the intersection of the line joining $\mathbf{X}_{cam}$ and $\mathbf{C}$ with the image plane $Z = f$. By similar triangles, the 2D coordinates on the image plane (u, v) are given by:

$$u = f\frac{X_c}{Z_c}, \quad v = f\frac{Y_c}{Z_c} \tag{6}$$

This non-linear division by depth Z_c is linearized in the projective space $\mathbb{P}^2$ using homogeneous coordinates. The projection can be expressed as a linear mapping:

$$Z_c \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} \tag{7}$$

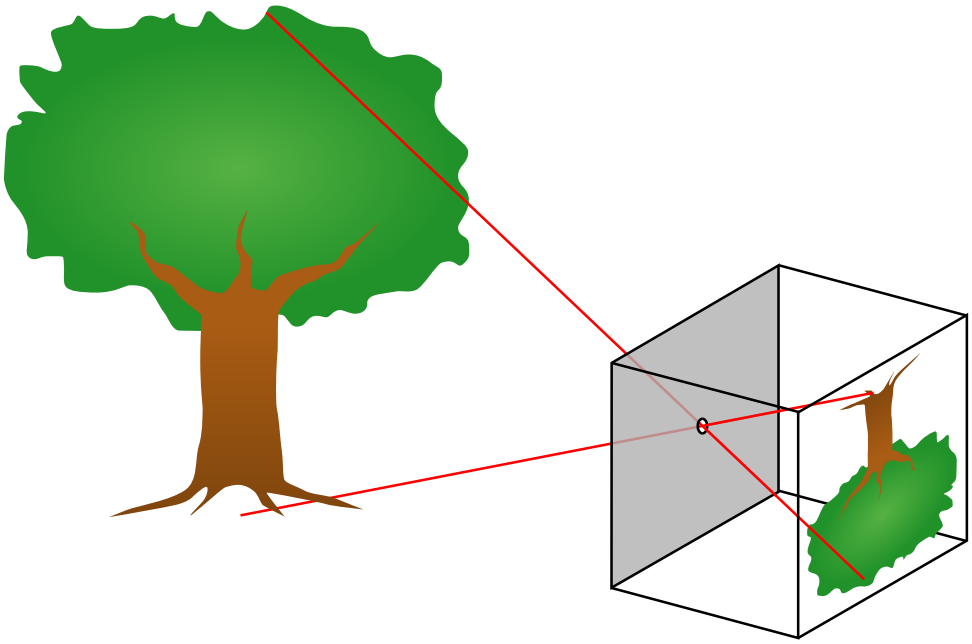


Figure 2: **The Pinhole Camera Model.** A 3D point $\mathbf{X}$ is projected through the camera center C onto the image plane, resulting in the 2D point $\mathbf{x}$. The distance between the camera center and the image plane corresponds to the focal length f. Reproduced from Wikipedia [28].

2.2.2. Intrinsic Parameters

The mapping in Equation (7) assumes the image coordinate system is centered at the principal point with isotropic scaling. In real digital sensors, the origin is typically at the top-left corner, and pixels may not be perfectly square. To account for this, we introduce the *Intrinsic Calibration Matrix* K.
In its most general form, K is an upper-triangular matrix containing 5 degrees of freedom:

$$K = \begin{bmatrix} \alpha_u & s & c_u \\ 0 & \alpha_v & c_v \\ 0 & 0 & 1 \end{bmatrix} \tag{8}$$

where the diagonal elements α_u and α_v represent the focal length expressed in horizontal and vertical pixel units, respectively, allowing for the modeling of non-square pixels. The vector (c_u, c_v) defines the coordinates of the *principal point*, corresponding to the intersection of the optical axis with the image plane, which is typically located near the geometric center of the image $(W/2, H/2)$. Finally, the off-diagonal term s constitutes the *skew parameter*, which accounts for any potential non-orthogonality of the sensor axes.
For modern digital cameras, it is standard practice to assume square pixels ($\alpha_u = \alpha_v = f$) and zero skew ($s = 0$). Under these assumptions, widely adopted in motion segmentation and SLAM literature, the calibration matrix simplifies to the form utilized in this work:

$$K = \begin{bmatrix} f & 0 & c_u \\ 0 & f & c_v \\ 0 & 0 & 1 \end{bmatrix} \tag{9}$$

2.2.3. Extrinsic Parameters and the Full Projection Matrix

In general, 3D points are defined in an arbitrary World Coordinate System $\mathbf{X}_{world}$. To project them, they must first be transformed into the Camera Coordinate System via a rigid body transformation $[R \mid \mathbf{t}] \in SE(3)$, where

R is the orientation of the camera frame w.r.t. the world frame, and $\mathbf{t} = -R\tilde{\mathbf{C}}$ is the translation (with $\tilde{\mathbf{C}}$ being the coordinates of the camera center in the world frame).
Combining the intrinsic and extrinsic parameters, the full projection of a world point $\mathbf{X}_{world}$ to a pixel $\mathbf{x}$ is given by the 3×4 projection matrix P:

$$\lambda \mathbf{x} = K[R \mid \mathbf{t}]\mathbf{X}_{world} = P\mathbf{X}_{world} \tag{10}$$

where λ is the projective depth of the point. The matrix P encapsulates the complete geometric description of the camera, decomposing the imaging process into metric calibration (K) and pose ($R, \mathbf{t}$).

2.2.4. Lens Distortion

While the linear pinhole model is convenient, real optical systems introduce non-linear distortions due to lens curvature and assembly imperfections. The most significant is *radial distortion*, which displaces points radially from the principal point. According to the standard Brown-Conrady model, the distorted coordinates (x_d, y_d) are related to the ideal normalized coordinates (x_n, y_n) by:

$$\begin{bmatrix} x_d \\ y_d \end{bmatrix} = (1 + k_1 r^2 + k_2 r^4 + k_3 r^6) \begin{bmatrix} x_n \\ y_n \end{bmatrix} \tag{11}$$

where $r^2 = x_n^2 + y_n^2$ is the squared radial distance from the optical center, and k_1, k_2, k_3 are the radial distortion coefficients. Tangential distortion terms (p_1, p_2) are also often included to model lens misalignment.
In the context of epipolar geometry estimation and motion segmentation, it is standard practice to assume that images have been pre-processed (undistorted) or that distortion is negligible. Consequently, throughout the remainder of this work, we will rely on the linear projective model described by Equation (10).

2.3. Two-View Epipolar Geometry

When a rigid scene is observed from two distinct viewpoints $\mathcal{V}_1$ and $\mathcal{V}_2$, the geometric relationship between a 3D point $\mathbf{X}$ and its 2D projections $\mathbf{x}_1, \mathbf{x}_2$ is strictly constrained by the epipolar geometry. This geometry relies solely on the internal parameters of the cameras and their relative pose, effectively decoupling the structure of the scene from the viewing configuration.

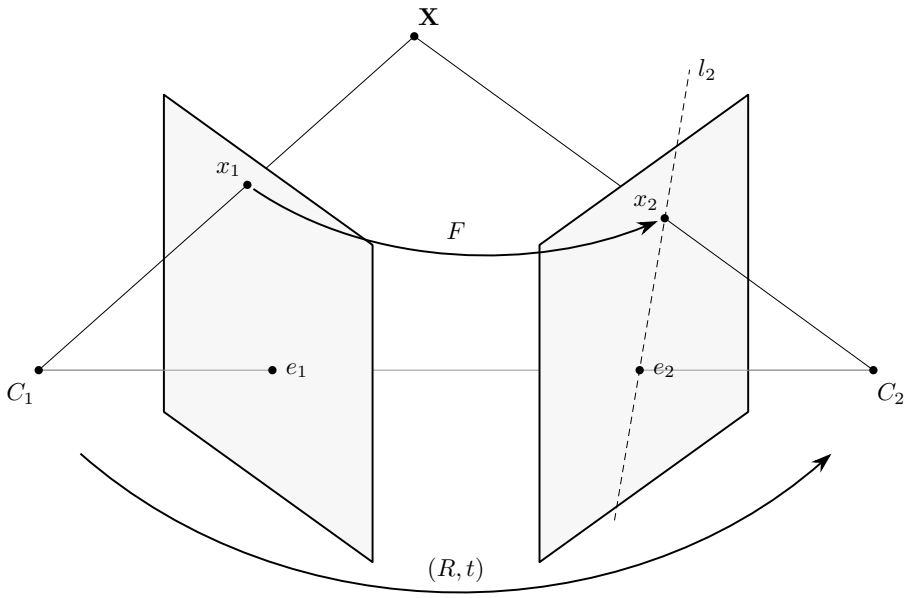


Figure 3: **Two-View Epipolar Geometry.** The camera centers C_1, C_2 and the 3D point P (denoted as $\mathbf{X}$ in the text) define the epipolar plane π. The projections $\mathbf{x}_1, \mathbf{x}_2$ must lie on the conjugate epipolar lines $\mathbf{l}_1, \mathbf{l}_2$. The relative motion is defined by rotation R and translation $\mathbf{t}$.

Let the relative motion from the first to the second camera be described by the rigid transformation $(R, \mathbf{t}) \in SE(3)$, such that a point $\mathbf{X}$ expressed in the first camera frame is transformed into the second as:

$$\mathbf{X}_2 = R\mathbf{X}_1 + \mathbf{t} \tag{12}$$

Geometrically, the two camera centers C_1, C_2 and the 3D point $\mathbf{X}$ form a triangle in space, defining the *epipolar plane* π. This implies that the vector $\mathbf{X}_2$, the translated vector $\mathbf{t}$, and the rotated vector $R\mathbf{X}_1$ must be coplanar. In algebraic terms, their scalar triple product must be zero:

$$\mathbf{X}_2^T(\mathbf{t} \times R\mathbf{X}_1) = 0 \tag{13}$$

This coplanarity constraint is the foundation for deriving the algebraic relations governing two-view geometry.

2.3.1. The Essential Matrix (Calibrated Case)

In the calibrated setting, we assume the intrinsic calibration matrices K_1, K_2 are known. This allows us to work with *normalized image coordinates*, which represent the projection on a virtual image plane with unit focal length:

$$\hat{\mathbf{x}}_1 = K_1^{-1}\mathbf{x}_1, \quad \hat{\mathbf{x}}_2 = K_2^{-1}\mathbf{x}_2 \tag{14}$$

Substituting these normalized coordinates into the coplanarity constraint (Eq. 13) and expressing the cross product with $\mathbf{t}$ as a multiplication by the skew-symmetric matrix $[\mathbf{t}]_\times$, we derive the defining equation of the *Essential Matrix* E:

$$\hat{\mathbf{x}}_2^T E \hat{\mathbf{x}}_1 = 0, \quad \text{where } E = [\mathbf{t}]_\times R \tag{15}$$

The matrix $E \in \mathbb{R}^{3\times3}$ encodes purely the relative pose between the views. Although it has 9 entries, it possesses only 5 degrees of freedom (3 for rotation, 2 for translation direction up to scale). Crucially, a valid Essential Matrix is characterized by strict internal constraints on its singular values. Its Singular Value Decomposition (SVD), $E = U\Sigma V^T$, must satisfy:

$$\Sigma = \text{diag}(\sigma, \sigma, 0) \tag{16}$$

This implies that the two non-zero singular values must be identical. This strong metric property is what makes calibrated estimation generally more robust than the uncalibrated counterpart.

2.3.2. The Fundamental Matrix (Uncalibrated Case)

In the general uncalibrated scenario, the intrinsics K_1, K_2 are unknown. The relationship is derived by substituting the normalized coordinates with raw pixel coordinates $\mathbf{x} = K\hat{\mathbf{x}}$ into Eq. (15):

$$(K_2^{-1}\mathbf{x}_2)^T E (K_1^{-1}\mathbf{x}_1) = 0 \implies \mathbf{x}_2^T (K_2^{-T} E K_1^{-1}) \mathbf{x}_1 = 0 \tag{17}$$

This defines the *Fundamental Matrix* F:

$$F = K_2^{-T} E K_1^{-1} = K_2^{-T} [\mathbf{t}]_\times R K_1^{-1} \tag{18}$$

The algebraic error $\mathbf{x}_2^T F \mathbf{x}_1 = 0$ is the generalized epipolar constraint. Like E, F is a rank-2 matrix ($\det(F) = 0$). However, because it absorbs the intrinsic parameters, it loses the strict singular value constraint of E. Consequently, F has 7 degrees of freedom (9 parameters minus 1 global scale minus 1 rank constraint). Geometrically, F maps a point in one image to a line in the other: $\mathbf{l}_2 = F\mathbf{x}_1$ is the epipolar line in the second view corresponding to $\mathbf{x}_1$.

2.3.3. The Semi-Calibrated Regime

Standard motion segmentation approaches typically treat F as a generic rank-2 projective entity. However, in many practical applications (e.g., autonomous driving, robotics), images are acquired by a single camera with constant internal parameters. This defines the *semi-calibrated* manifold, where:

$$K_1 = K_2 = K = K = \begin{bmatrix} f & 0 & c_u \\ 0 & f & c_v \\ 0 & 0 & 1 \end{bmatrix} \tag{19}$$

In this regime, the Fundamental matrix acquires additional structure compared to the general uncalibrated case. Specifically, it must satisfy the *Kruppa equations* or, equivalently, the trace constraint derived from the

Essential matrix properties. A matrix F is valid in this semi-calibrated setting if and only if there exists a focal length f and a pose $(R, \mathbf{t})$ such that F can be decomposed according to Eq. (18) with equal intrinsics. This imposes a strong prior: while a generic F allows for varying focal lengths (zooming) or skew, a semi-calibrated F resides in a lower-dimensional manifold. Exploiting this restriction allows for the rejection of mathematically valid but physically impossible motion hypotheses, which is the core principle of the proposed method.

2.4. Error Metrics for Epipolar Geometry

Quantifying the agreement between a set of point correspondences $\{(\mathbf{x}_{1,i}, \mathbf{x}_{2,i})\}$ and a hypothesized epipolar geometry represented by a Fundamental matrix F is critical for robust estimation. In the presence of measurement noise, the exact epipolar constraint $\mathbf{x}_{2,i}^T F \mathbf{x}_{1,i} = 0$ is rarely satisfied. Therefore, we require a scalar cost function to evaluate the residual error of each match.
In this work, we distinguish between algebraic measures, which are computationally efficient but lack physical interpretation, and geometric measures, which represent Euclidean distances in the image plane.

2.4.1. Algebraic Error

The simplest residual is derived directly from the definition of the Fundamental matrix. The *algebraic error* for a correspondence i is defined as:

$$\epsilon_{alg}(\mathbf{x}_{1,i}, \mathbf{x}_{2,i}, F) = \mathbf{x}_{2,i}^T F \mathbf{x}_{1,i} \tag{20}$$

While linear and easy to minimize (e.g., via SVD in the 8-point algorithm [16]), this metric is geometrically meaningless. Since F is defined only up to a scale factor, the algebraic error can be arbitrarily minimized by scaling F towards zero unless a norm constraint (e.g., $\|F\|_F = 1$) is enforced. Furthermore, ϵ_{alg} is not invariant to transformations of the image coordinates, often necessitating data normalization (Hartley's preconditioning) to ensure numerical stability.

2.4.2. Geometric Reprojection Error

The "Gold Standard" metric is the geometric reprojection error. It seeks to find the corrected points $\hat{\mathbf{x}}_{1,i}, \hat{\mathbf{x}}_{2,i}$ that perfectly satisfy the epipolar constraint ($\hat{\mathbf{x}}_{2,i}^T F \hat{\mathbf{x}}_{1,i} = 0$) while minimizing the Euclidean distance to the measured points. Formally, we seek to minimize the cost function:

$$\mathcal{C}_{geom} = \sum_i d(\mathbf{x}_{1,i}, \hat{\mathbf{x}}_{1,i})^2 + d(\mathbf{x}_{2,i}, \hat{\mathbf{x}}_{2,i})^2 \tag{21}$$

subject to $\hat{\mathbf{x}}_{2,i}^T F \hat{\mathbf{x}}_{1,i} = 0$, where $d(\cdot, \cdot)$ denotes the Euclidean distance in pixel coordinates. Minimizing this error requires estimating the optimal triangulation of the 3D points $\mathbf{X}_i$ alongside the camera motion, typically via Bundle Adjustment. While statistically optimal (Maximum Likelihood Estimate under Gaussian noise), it is computationally expensive and requires iterative non-linear optimization over high-dimensional spaces.

2.4.3. The Sampson Distance

To achieve a trade-off between the simplicity of the algebraic error and the accuracy of the geometric error, we employ the *Sampson distance* (or first-order geometric error). As demonstrated by Hartley and Zisserman [10], the Sampson distance approximates the geometric reprojection error by a first-order Taylor expansion of the algebraic cost function.
Geometrically, the term $F\mathbf{x}_{1,i}$ represents the epipolar line $\mathbf{l}_{2,i} = [l_u, l_v, l_w]^T$ in the second image. The Euclidean distance from point $\mathbf{x}_{2,i}$ to this line is:

$$d(\mathbf{x}_{2,i}, \mathbf{l}_{2,i}) = \frac{|\mathbf{x}_{2,i}^T F \mathbf{x}_{1,i}|}{\sqrt{(F\mathbf{x}_{1,i})_1^2 + (F\mathbf{x}_{1,i})_2^2}} \tag{22}$$

Symmetrically, the distance in the first image is related to $F^T \mathbf{x}_{2,i}$. The squared Sampson distance combines these to approximate the joint reprojection error:

$$d_S(\mathbf{x}_{1,i}, \mathbf{x}_{2,i}, F)^2 = \frac{(\mathbf{x}_{2,i}^T F \mathbf{x}_{1,i})^2}{(F\mathbf{x}_{1,i})_1^2 + (F\mathbf{x}_{1,i})_2^2 + (F^T\mathbf{x}_{2,i})_1^2 + (F^T\mathbf{x}_{2,i})_2^2} \tag{23}$$

where $(\mathbf{v})_j$ denotes the j-th component of vector $\mathbf{v}$. This metric effectively normalizes the algebraic error by the magnitude of the gradient of the epipolar constraint. Unlike the algebraic error, the Sampson distance

represents a valid geometric cost in pixel units and is invariant to the scaling of F. Consequently, it serves as the standard objective function for the robust IRLS refinement steps in modern multi-model fitting frameworks, including the one proposed in this work.

2.4.4. Symmetric Epipolar Distance

Another common geometric approximation is the Symmetric Epipolar Distance, which sums the squared distances of each point to the epipolar line corresponding to the other point:

$$d_{sym}^2 = d(\mathbf{x}_{2,i}, F\mathbf{x}_{1,i})^2 + d(\mathbf{x}_{1,i}, F^T\mathbf{x}_{2,i})^2 \tag{24}$$

While intuitive, the Sampson distance is generally preferred for optimization as it is a closer approximation to the true reprojection error minima [10].

2.5. The Standard 3D Reconstruction Pipeline

The mathematical constraints formalized in the previous sections form the backbone of the standard Structure-from-Motion (SfM) pipeline. Given a pair of images capturing a scene, the goal is to recover the relative camera motion $(R, \mathbf{t})$ and the sparse 3D structure $\mathcal{X}$. This process is traditionally formalized into a sequential algorithmic pipeline, as illustrated in Figure 4, comprising four distinct stages: feature extraction, matching, robust geometric fitting and triangulation.

Figure 4: **3D Reconstruction Pipeline.**

2.5.1. Feature Extraction: Scale-Invariant Detection

The first step involves detecting salient, repeatable interest points (keypoints) $\mathbf{x}$ in each image $I(u, v)$. To ensure robustness against scale changes and rotation, the *Scale-Invariant Feature Transform* (SIFT) [17] is the standard choice. SIFT identifies keypoints as extrema in the *Scale Space*, constructed by convolving the image with variable-scale Gaussian kernels $G(u, v, \sigma)$:

$$L(u, v, \sigma) = G(u, v, \sigma) * I(u, v) \tag{25}$$

To efficiently approximate the Laplacian of Gaussian (LoG), SIFT computes the *Difference of Gaussians* (DoG):

$$D(u, v, \sigma) = L(u, v, k\sigma) - L(u, v, \sigma) \tag{26}$$

Keypoints are localized at maxima/minima of $D(u, v, \sigma)$ across scales. For each keypoint, a 128-dimensional descriptor vector $\mathbf{d} \in \mathbb{R}^{128}$ is generated based on local gradient histograms, ensuring invariance to illumination and viewpoint changes. Let $\mathcal{P}_1 = \{\mathbf{x}_{1,i}, \mathbf{d}_{1,i}\}_{i=1}^{N_1}$ and $\mathcal{P}_2 = \{\mathbf{x}_{2,j}, \mathbf{d}_{2,j}\}_{j=1}^{N_2}$ be the feature sets for the two views.

2.5.2. Matching

Correspondences are established by searching for the nearest neighbor in the descriptor space. The similarity between two features is measured via the Euclidean distance $\|\mathbf{d}_{1,i} - \mathbf{d}_{2,j}\|_2$. However, the nearest neighbor is not always correct due to repetitive textures (perceptual aliasing). To reject ambiguous matches, *Lowe's Ratio Test* is applied. A match is accepted only if the ratio of the distance to the nearest neighbor (d_1) to the distance to the second-nearest neighbor (d_2) is below a strict threshold τ (typically 0.8):

$$\frac{d_1}{d_2} < \tau \tag{27}$$

This yields a set of *putative correspondences* $\mathcal{M} = \{(\mathbf{x}_{1,k}, \mathbf{x}_{2,k})\}_{k=1}^{M}$, which inevitably contains false positives (outliers).

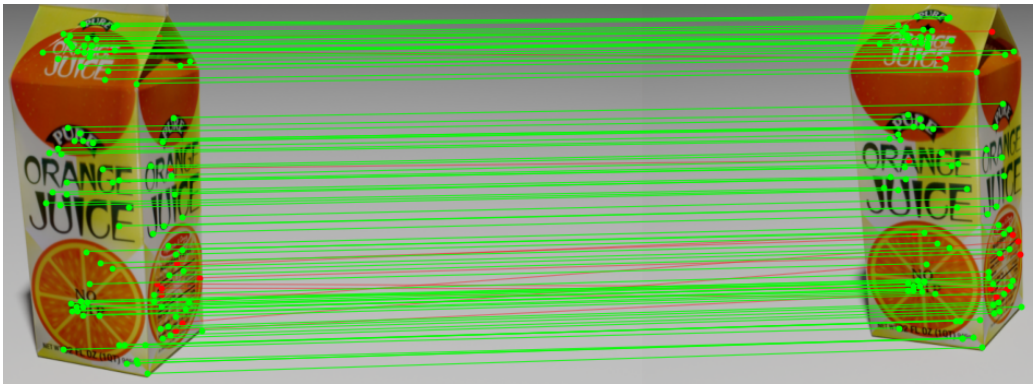


Figure 5: **Feature Matching.** Correspondences established via descriptor similarity, still contaminated by outliers(red lines), shown on an example from the HOPE-F dataset[13] .

2.5.3. Robust Geometric Model Fitting

Standard least-squares estimation is extremely sensitive to outliers. Therefore, the geometric relation is estimated using a robust estimator, most commonly **RANSAC** (RANdom SAmple Consensus) [8]. RANSAC operates by repeatedly drawing minimal subsets of data to hypothesize a model and verifying it against the entire dataset. The number of iterations k required to ensure with probability p (usually 0.99) that at least one sample is outlier-free is given by:

$$k = \frac{\log(1-p)}{\log(1-w^s)} \tag{28}$$

where w is the inlier ratio and s is the minimal sample size. At this stage, the pipeline strictly bifurcates based on the available calibration priors:

A. The Calibrated Case (E Matrix). If intrinsic matrices K_1, K_2 are known, points are normalized to $\hat{\mathbf{x}}$. The minimal sample size is $s = 5$. The *5-point algorithm* (Nistér [19]) is employed to solve for the Essential matrix E. It involves finding the roots of a tenth-degree polynomial derived from the epipolar constraint and the trace constraint $2EE^TE - \text{tr}(EE^T)E = 0$. This method is highly efficient and enforces the metric validity of the reconstruction.

B. The Uncalibrated Case (F Matrix). If intrinsics are unknown, the solver operates on raw pixels. The minimal sample size is $s = 7$ (using the singularity constraint $\det(F) = 0$ to solve a cubic equation) or $s = 8$ (linear solution). The *normalized 8-point algorithm* [10] is the standard reference. It solves the linear system $A\mathbf{f} = 0$ via SVD, where $\mathbf{f}$ is the flattened Fundamental matrix. A final enforcement of rank-2 is applied by setting the smallest singular value to zero.

C. The Semi-Calibrated Case (Focus of this Work). In many scenarios, such as the one addressed in this thesis, we can assume constant but unknown focal lengths ($f_1 = f_2 = f$). This "semi-calibrated" regime lies between the two extremes. While standard uncalibrated solvers (8-point) can still be used, they ignore the constraint $f_1 = f_2$, often resulting in solutions that violate the metric rigidity of the camera (e.g., yielding complex focal lengths upon decomposition). Specialized solvers, such as the *6-point algorithm* [23] or the recent robust trace-constraint solver [14], exploit this prior to reject degenerate configurations that plague the purely uncalibrated 7-point or 8-point methods.

2.5.4. Triangulation

Once a robust geometric model is estimated and outliers are removed, the relative pose $(R, \mathbf{t})$ (and intrinsics K if uncalibrated) is extracted. The final step is to recover the 3D coordinates $\mathbf{X}$ for each inlier correspondence $(\mathbf{x}_1, \mathbf{x}_2)$. This is the *Triangulation* problem. Given projection matrices $P_1 = K_1[I|\mathbf{0}]$ and $P_2 = K_2[R|\mathbf{t}]$, we seek $\mathbf{X}$ such that $\mathbf{x}_1 \times (P_1\mathbf{X}) = \mathbf{0}$ and $\mathbf{x}_2 \times (P_2\mathbf{X}) = \mathbf{0}$. This leads to a linear system of the form $A\mathbf{X} = \mathbf{0}$, where

$A \in \mathbb{R}^{4\times4}$ is composed of the rows of P_1 and P_2:

$$A = \begin{bmatrix} x_1\mathbf{p}_1^{3T} - \mathbf{p}_1^{1T} \\ y_1\mathbf{p}_1^{3T} - \mathbf{p}_1^{2T} \\ x_2\mathbf{p}_2^{3T} - \mathbf{p}_2^{1T} \\ y_2\mathbf{p}_2^{3T} - \mathbf{p}_2^{2T} \end{bmatrix} \tag{29}$$

where $\mathbf{p}_i^{rT}$ denotes the r-th row of matrix P_i. The solution is the right singular vector of A corresponding to the smallest singular value (Direct Linear Transformation - DLT). In practice, this algebraic solution is refined by minimizing the geometric reprojection error via non-linear optimization.

2.5.5. Limitation: The Single Rigid Motion Assumption

The standard pipeline described above relies on a fundamental assumption: *the scene is static*, or equivalently, it contains a single dominant rigid motion. RANSAC treats any deviation from the dominant model as a random outlier. However, in dynamic scenes containing multiple moving objects, "outliers" are not random noise but structured data belonging to other valid motions. This failure of the single-model assumption necessitates the shift to *Multi-Model Fitting* techniques, which is the core subject of the following chapter.

2.6. Robust Estimation and Multi-Model Fitting

The geometric pipeline described in Section 2.5 relies on robust estimators like RANSAC to handle measurement noise and mismatches. While highly effective for static scenes, the standard consensus maximization paradigm exhibits a fundamental failure mode when applied to dynamic environments containing multiple moving objects.

2.6.1. The Failure of Single-Model Estimators

When dealing with a multi-body scenario, we are looking at an inherently multimodal data distribution. If RANSAC tries to fit a model to one specific motion $\mathcal{M}_k$, the points tied to other independent motions $\{\mathcal{M}_j\}_{j\neq k}$ do not just act as unstructured random noise. They actually form *pseudo-outliers*: spatially coherent groups of data that possess a valid geometric structure of their own. Because of this structural contamination, the consensus landscape becomes filled with local maxima that easily trap greedy estimators. For instance, a single Fundamental matrix found by RANSAC will often bridge across two completely distinct motions just to grab a higher inlier count, resulting in a physically meaningless "merged" model. To properly resolve dynamic scenes, we therefore have to move away from simple outlier rejection (single-model) and treat the problem strictly as *data clustering* (multi-model fitting).

2.6.2. Taxonomy of Motion Segmentation

To contextualize the operational setting of this work, it is useful to categorize motion segmentation approaches based on their input assumptions, as formalized by Arrigoni [1]. The landscape can be visualized as a spectrum of complexity versus applicability (Figure 6).

Figure 6: **Taxonomy of Motion Segmentation Approaches.** Existing methods are categorized based on their input assumptions along a spectrum ranging from realistic but computationally unfeasible tasks (left) to heavily constrained but easily solvable scenarios (right). The two-frame correspondence paradigm offers an optimal structural trade-off.

On the right end of the spectrum lies *Trajectory Clustering*. Historically, methods like Subspace Clustering (SSC) or Generalized PCA rely on the assumption that feature points are tracked reliably over long video sequences (typically > 10 frames). This allows modeling motion as linear subspaces under affine projection assumptions. While computationally elegant, these methods are fragile in "in-the-wild" scenarios where long-term tracking is broken by occlusions or rapid camera motion.

At the opposite extreme lies the case of *Unknown Correspondences*, involving the simultaneous estimation of matching and segmentation. This avoids reliance on pre-computed matches but leads to a combinatorial explosion in the search space, making it computationally intractable for real-time applications without heavy priors.

Positioned optimally in the middle of this spectrum is the *Two-View Correspondence Clustering* paradigm, which constitutes the operational setting of this work. Here, the input is a set of sparse putative matches between just two views. This formulation does not require temporal continuity (unlike trajectory clustering) and leverages the geometric rigidity of epipolar constraints (unlike affine methods). It is robust enough to handle wide baselines and unstructured image pairs, yet computationally tractable. However, as noted in Section 2.3, working with only two views makes the uncalibrated estimation (F-matrix) particularly unstable, necessitating the semi-calibrated constraints proposed in our approach.

3. Problem Formulation

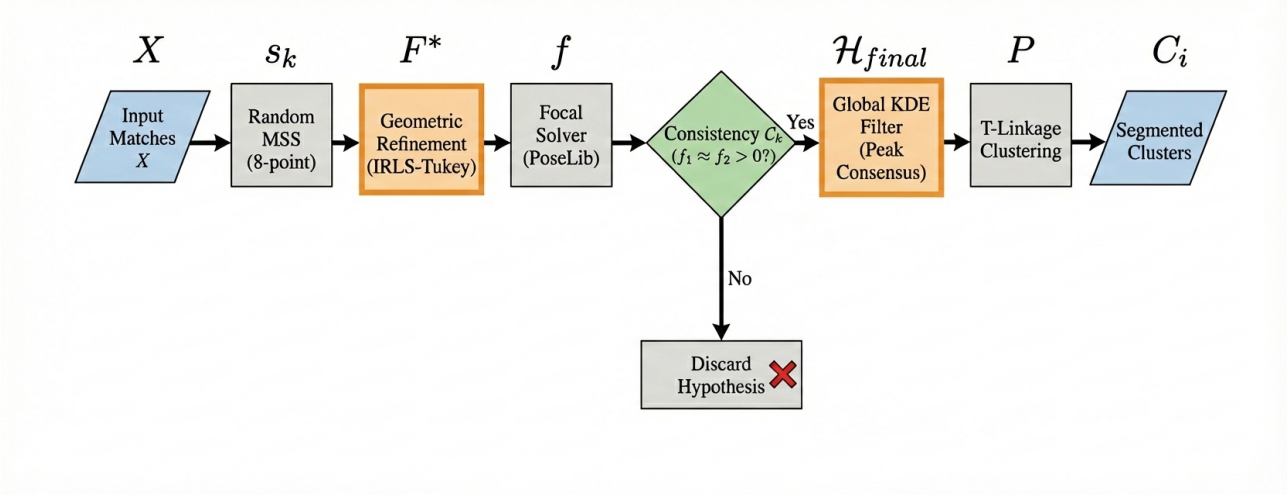


Figure 7: **Pipeline Overview of the Semi-Calibrated T-Linkage.** From input matches X, random minimal sample sets s_k are drawn to generate Fundamental matrices F, which are then geometrically refined via IRLS. A robust solver extracts the corresponding focal length pair (f_1, f_2). Hypotheses satisfying the local metric consistency ($f_1 \approx f_2 > 0$) are collected as valid candidates. Subsequently, a Global KDE filter enforces scene-level consensus by retaining only models consistent with the dominant focal peak f_{peak}, yielding the final hypothesis space $\mathcal{H}_{final}$. Finally, T-Linkage groups the data based on geometric preferences P, outputting the segmented clusters C_i.

Let $X = \{(\mathbf{x}_i, \mathbf{x}'_i)\}_{i=1}^N$ be a set of point correspondences between two views, potentially contaminated by a significant fraction of outliers. Our goal is to partition the set of inliers into disjoint clusters $\{C_k\}$, each associated with a distinct rigid motion represented by a Fundamental matrix F_k.

Unlike standard approaches that operate in a purely uncalibrated setting, where F is treated as a generic projective entity, we restrict the search space to the manifold of *geometrically consistent* matrices $\mathcal{F}_{semicalib}$. We define this manifold such that all valid motions in the scene share the same constant intrinsic calibration matrix K:

$$\mathcal{F}_{semicalib} = \{F_k \in \mathbb{R}^{3\times 3} \mid \det(F_k) = 0, \exists K \text{ s.t. } F_k = K^{-T}[\mathbf{t}_k]_\times R_k K^{-1}\} \tag{30}$$

where $R_k \in \mathrm{SO}(3)$ and $\mathbf{t}_k \in \mathbb{R}^3$ represent the relative rotation and translation of each distinct motion. Under the semi-calibrated assumption, we enforce that the camera intrinsics remain constant across the two views ($K = K'$). As introduced in Section 2.2, we assume a simplified pinhole model with zero skew. By pre-processing the point correspondences to shift the image coordinates' origin to the principal point $(w/2, h/2)$, the calibration matrix simplifies to a purely diagonal form $K = \mathrm{diag}(f, f, 1)$. Specifically, we enforce the physical constraint that the focal length must be positive ($f > 0$).

3.1. Instability in Uncalibrated Estimation

A critical challenge in two-view motion segmentation is the numerical instability of epipolar geometry estimation in a purely uncalibrated setting. As demonstrated by Xu et al. [31], this instability arises because F is a flexible, over-parameterized model that tends to overfit data in degenerate or near-degenerate configurations.
This phenomenon is particularly evident when correspondences lie on a dominant plane or undergo specific motions (e.g., pure rotation). In these scenarios, the Fundamental matrix becomes *ill-defined* as the epipolar geometry cannot be uniquely recovered from the available matches; any unconstrained algebraic solver (e.g., the standard 8-point algorithm) will minimize the algebraic error by converging toward arbitrary solutions that satisfy the epipolar constraint $\mathbf{x}_i'^T F_k \mathbf{x}_i \approx 0$ but lack any valid geometric interpretation. Specifically, Xu et al. highlight that while F is theoretically the most general model, its performance degrades significantly compared to more constrained models (e.g., Homographies or Affine transformations) when the data do not provide sufficient independent constraints to resolve its 7 degrees of freedom [31].
Conversely, the Essential matrix, leveraged in calibrated or semi-calibrated regimes, remains robustly constrained by its internal metric properties. By filtering out hypotheses that fall outside the $\mathcal{F}_{semicalib}$ manifold, we effectively discard spurious solutions that, while algebraically plausible, are inconsistent with a physical camera sensor.
Crucially, while these unconstrained solutions are mathematically valid in an uncalibrated sense, they typically violate the metric constraints required for a valid Euclidean reconstruction. As shown in the decomposition in Eq. (30), an arbitrary Fundamental matrix implies specific camera configurations. In an uncalibrated framework, since K and K' are unknown, the algebraic decomposition often implies impossible physical setups characterized by:

- **Complex Intrinsics:** The decomposition of the Kruppa equations may yield imaginary focal lengths ($f^2 < 0$).
- **Inconsistent Calibration:** The estimated focal lengths may differ drastically between views ($f_1 \ll f_2$ or $f_1 \gg f_2$), directly violating the assumption of a single physical sensor with constant intrinsics.
- **Cheirality Violations:** The triangulated 3D points may lie behind the camera, resulting in negative depths.

3.2. The Consistency Criterion

Given the manifold definition in Eq. (30), the segmentation problem relies on a robust measure of consistency between data and models. Leveraging the geometric metrics introduced in Section 2.4, we formally define the *Consensus Set* (CS) of a semi-calibrated hypothesis $F \in \mathcal{F}_{semicalib}$ as the subset of correspondences whose residual error falls below a statistically derived inlier threshold ϵ:

$$\mathrm{CS}(F) = \{i \in \{1, \dots, N\} \mid d_S(\mathbf{x}_i, \mathbf{x}_i', F) < \epsilon\} \tag{31}$$

where $d_S(\cdot)$ is the Sampson distance defined in Eq. (23). Consequently, the motion segmentation task is equivalent to finding a set of models $\{F_k\}_{k=1}^K \subset \mathcal{F}_{semicalib}$ that maximizes the joint cardinality of their consensus sets while minimizing model complexity. Unlike standard approaches that maximize this consensus over the generic projective manifold $\mathcal{F}_{proj}$, our formulation strictly constrains the optimization to $\mathcal{F}_{semicalib}$, ensuring that the recovered inliers support a physically plausible 3D motion.

4. Related Work

4.1. From Affine Subspaces to Perspective Multi-Model Fitting

Motion segmentation is a long-standing problem in computer vision. Early approaches, such as Subspace Clustering (SSC) [7] or Generalized PCA [27], typically rely on the assumption of affine cameras. Under this approximation, the complex non-linearities of perspective projection are ignored, and feature trajectories are constrained to lie in low-dimensional linear subspaces, allowing for elegant algebraic solutions based on factorization. However, the affine assumption is highly restrictive: it requires dense tracking over long sequences with minimal depth variations and inherently fails to model the significant perspective effects and wide baselines that are common in unconstrained, real-world scenarios.
To address these limitations, modern motion segmentation is widely formulated as a robust multi-model fitting problem operating on perspective two-view geometry. In this context, consensus maximization methods like RANSAC [8] are the gold standard for single models, but they struggle fundamentally with multiple structures due to the severe interference of structured pseudo-outliers. Advanced sampling and clustering techniques have

thus been proposed to navigate this complex landscape. J-Linkage [25] revolutionized the field by clustering points based on their binary preference for a large set of randomly generated hypotheses. T-Linkage [18] successfully refines this concept with a continuous relaxation of the preference function, significantly improving robustness to noise and intersecting structures.
More recent state-of-the-art methods explore alternative optimization strategies: Progressive-X [2] combines sequential hypothesis sampling with Graph-Cut based spatial coherence constraints, while methods like Fast-CP [20] leverage spectral clustering and Christoffel polynomials for rapid algebraic grouping. Despite their diverse algorithmic foundations, these state-of-the-art methods have been historically and successfully used in purely uncalibrated motion segmentation problems, treating each motion as an independent projective entity.

4.2. The Role of Calibration Priors

Furthermore, there is a distinct dichotomy in the literature regarding the use of camera calibration priors. In static Structure-from-Motion (SfM) and single-model pose estimation, shifting from purely uncalibrated to semi-calibrated or fully calibrated solvers is a well-established standard practice. Algorithms like the 5-point solver for calibrated cameras [19] or the 6-point solver for unknown but constant focal length [14, 23] explicitly solve for the intrinsic parameters alongside the camera pose. By doing so, they leverage the internal metric rigidity of the physical scene to reject degenerate configurations—such as dominant planes—which otherwise cause the unconstrained geometric estimation to fail catastrophically.
Conversely, in *multi-body motion segmentation*, this critical physical prior is largely overlooked. Most existing approaches persist in estimating unconstrained Fundamental matrices. Because individual moving objects often occupy a limited spatial support within the image, estimating 7 degrees of freedom per motion invariably leads to algebraic overfitting, yielding solutions that minimize the residual error but lack a valid Euclidean interpretation.
Our work bridges this significant gap by transferring the benefits of semi-calibrated estimation, standard in static SfM, directly to the multi-body domain. By enforcing a scene-level consensus on the focal length, we provide a mathematically rigorous geometric regularizer that mitigates the inherent instability of uncalibrated multi-model fitting.

4.3. Deep Learning for Motion Segmentation

In recent years, the success of deep learning in dense prediction tasks has naturally extended to motion segmentation. Early learning-based approaches formulated the problem as a dense foreground-background segmentation task, relying heavily on pre-computed optical flow fields to train Convolutional Neural Networks (CNNs) [24]. More recent architectures leverage Graph Neural Networks (GNNs) or Transformer-based attention mechanisms to perform end-to-end clustering of sparse feature trajectories or two-view correspondences, effectively learning the consensus function directly from data [12]. Recently, this neural-guided sampling paradigm has been extended to achieve real-time performance through parallel hypothesis generation, as demonstrated by PARSAC [13].
While deep learning models achieve impressive performance on standardized benchmarks where training and testing domains are closely aligned, they suffer from critical limitations in unconstrained environments. Deep multi-model fitting networks require massive amounts of annotated dynamic scenes, a type of data that is notoriously difficult and expensive to acquire accurately. Indeed, to train supervised multi-model estimators like PARSAC, researchers have been forced to generate large-scale synthetic datasets (such as HOPE-F) to compensate for the severe scarcity of real-world ground truth. Furthermore, these models often overfit to the motion statistics and camera intrinsics present in the training set, leading to severe performance degradation when deployed "in the wild" on novel sensors or unusual scene geometries.
Conversely, geometric, optimization-based approaches like our proposed Semi-Calibrated T-Linkage remain fundamentally domain-agnostic. By strictly relying on the physical laws of projective geometry and leveraging the internal metric constraints of the camera, our method does not require any prior training phase. This ensures robust generalization across highly diverse scenarios, from indoor planar scenes to outdoor autonomous driving, making geometric methods indispensable where data-driven priors fail or are unavailable.

5. Method

Our method builds upon the T-Linkage framework [18], a well-established approach for multi-model fitting. T-Linkage operates on a “Hypothesize and Cluster” paradigm: it first generates a large pool of random hypotheses from minimal subsets of data, as in RANSAC [8], and then clusters data points based on their preference for

these models.
Standard motion segmentation pipelines typically operate in a purely uncalibrated setting, estimating Fundamental matrices as unconstrained algebraic entities. Transitioning from an uncalibrated to a semi-calibrated framework is non-trivial, as it requires extracting reliable intrinsic information from noisy, minimal data.
To address this, our proposed method introduces an additional layer of robustness to the multi-model fitting process. Instead of attempting to regularize the fitting of all generated models, we intervene directly in the hypothesis generation phase by introducing a geometric consistency filter. By integrating a robust focal length solver, we analyze the compatibility of each sampled Fundamental matrix with the semi-calibrated assumption. This allows us to identify and discard spurious hypotheses that, although algebraically satisfying the epipolar geometry on a small subset of points, yield inconsistent or non-Euclidean camera parameters. Ultimately, this filtering approach effectively distinguishes valid motions from unstable algebraic fits prior to the clustering phase, without relying on scene-specific assumptions.

5.1. The Algorithm

The complete procedure is detailed in Algorithm 1.
Our pipeline encompasses several key stages: robust hypothesis generation with geometric refinement, hierarchical filtering via focal consensus, and preference-based clustering

Algorithm 1 Semi-Calibrated T-Linkage

Require: Matches X, Number of samples M, Thresholds $\tau_{diff}, \tau_{range}, \tau$
Ensure: Cluster Labels L
1: $\mathcal{H}_{valid} \leftarrow \emptyset$ {Initialize valid hypothesis set}
2: **for** $k \leftarrow 1$ to M **do**
3: $s_k \leftarrow \text{RandomMSS}(X, 8)$
4: $F \leftarrow \text{EightPoint}(s_k)$
5: $F \leftarrow \text{Refine}(F, \text{IRLS-Sampson}, \tau)$ {Geometric refinement}
6: $f_1, f_2 \leftarrow \text{EstimateFocal}(F)$ {Using Solver [14]}
7: $f_{avg} \leftarrow (f_1 + f_2)/2$
8: **if** $f_1 > 0 \wedge f_2 > 0$ **and** $\frac{|f_1 - f_2|}{f_{avg}} < \tau_{diff}$ **then**
9: $\mathcal{H}_{valid} \leftarrow \mathcal{H}_{valid} \cup \{(F, f_{avg})\}$
10: **end if**
11: **end for**
12: **Scene-Level Consensus Filtering:**
13: $f_{vals} \leftarrow \{f \mid (F, f) \in \mathcal{H}_{valid}\}$
14: **if** $f_{vals} \neq \emptyset$ **then**
15: $\hat{p}(f) \leftarrow \text{KDE}(f_{vals})$ {Density Estimation}
16: $f_{peak} \leftarrow \arg\max \hat{p}(f)$
17: $\mathcal{H}_{final} \leftarrow \{F \in \mathcal{H}_{valid} \mid |f - f_{peak}| < \tau_{range} \cdot f_{peak}\}$
18: **else**
19: $\mathcal{H}_{final} \leftarrow \mathcal{H}_{valid}$
20: **end if**
21: **Clustering:**
22: $P \leftarrow \text{BuildPreferenceSet}(\mathcal{H}_{final}, X)$
23: $L \leftarrow \text{T-LinkageClustering}(P)$ {Using Tanimoto Dist}
24: **return** L

5.2. Pipeline Stages

5.2.1. Step 1: Robust Sampling and Geometric Refinement

Standard multi-model fitting approaches typically generate hypotheses by applying the normalized 8-point algorithm directly to Minimal Sample Sets (MSS) (Algorithm 1, lines 3–4). However, exact algebraic solutions derived from noisy minimal subsets are notoriously unstable. Before attempting to extract the sensitive intrinsic parameters, we must stabilize the Fundamental matrix F by refining it against its local consensus set (line 5).

Instead of merely minimizing the algebraic error, which lacks physical meaning, we formulate the refinement as an M-estimation problem minimizing the robust Sampson distance. Formally, given an initial estimate $F^{(0)}$ derived from the MSS, we seek the optimal matrix F^* that minimizes the robustified geometric cost:

$$F^* = \arg\min_F \sum_{i \in \mathcal{I}} \rho\left(d_S(\mathbf{x}_i, \mathbf{x}'_i, F)^2\right) \tag{32}$$

where $\mathcal{I}$ is the active set of inlier correspondences, $d_S(\cdot)$ is the Sampson distance defined in Section 2.4, and $\rho(\cdot)$ is a robust loss estimator.

Because the objective function relies on a non-convex distance metric and a non-linear robust loss, a closed-form solution does not exist. We solve this via Iteratively Reweighted Least Squares (IRLS). The core principle of IRLS is to convert the complex M-estimation problem into a sequence of tractable Weighted Linear Least Squares (WLLS) problems. By taking the derivative of the objective function with respect to the model parameters and equating it to zero, the influence of each data point is encapsulated into a dynamic weight w_i.

For our framework, we specifically adopt **Tukey's biweight (bisquare) function** [22]. Unlike standard Huber estimators that merely dampen the influence of large residuals linearly, Tukey's function is a *redescending* estimator. This means the weight of severe outliers smoothly drops to exactly zero, completely nullifying the impact of gross mismatches from the refinement process. The continuous weight function $w(r_i)$ assigned to a residual $r_i = d_S(\mathbf{x}_i, \mathbf{x}'_i, F)$ is defined as:

$$w(r_i) = \begin{cases} \left(1 - \left(\frac{r_i}{c\sigma}\right)^2\right)^2 & \text{if } |r_i| \leq c\sigma \\ 0 & \text{otherwise} \end{cases} \tag{33}$$

where $c = 4.685$ is the standard tuning constant ensuring 95% asymptotic efficiency on Gaussian noise, and σ represents the robust scale (standard deviation) of the noise distribution.

Dynamic Scale Estimation. A critical challenge in uncalibrated dynamic scenes is that the noise scale σ varies unpredictably depending on the motion's distance from the camera and the sensor quality. Fixing a static σ inevitably leads to poor convergence. To address this, we dynamically estimate the noise scale at each iteration using the Median Absolute Deviation (MAD), a highly robust statistic with a breakdown point of 50%:

$$\sigma^{(t)} = 1.4826 \cdot \text{median}_i \left(\left|r_i^{(t-1)} - \text{median}_j(r_j^{(t-1)})\right|\right) \tag{34}$$

The constant 1.4826 scales the MAD to be a consistent estimator for the standard deviation under the assumption of normally distributed inlier noise.

The IRLS Iteration. The refinement proceeds iteratively. At each step t:

1. We compute the Sampson residuals $r_i^{(t-1)}$ for all points using the previous model $F^{(t-1)}$.
2. We estimate the dynamic scale $\sigma^{(t)}$ via the MAD formulation to establish the current statistical boundary of the inliers.
3. We compute the Tukey weights $w_i^{(t)}$ for each correspondence.
4. We assemble the weighted design matrix $X^T W X \mathbf{f} = \mathbf{0}$ (where $W = \text{diag}(w_1, \ldots, w_N)$ and X contains the Kronecker products of the matched coordinates) and solve for the new $F^{(t)}$ via Singular Value Decomposition, subsequently enforcing the rank-2 constraint ($\det(F) = 0$).

This iterative sequence strictly confines the algebraic fit to the true underlying geometry. This profound geometric stabilization is a mandatory prerequisite: as we will demonstrate in the ablation studies, initializing the subsequent semi-calibrated solver with raw, unrefined minimal samples yields mathematically unstable intrinsic estimates that frequently violate physical camera constraints.

5.2.2. Step 2: Semi-Calibrated Hypothesis Generation

A key component of our method is the extraction of the focal length from the Fundamental matrix (Algorithm 1, line 6). Assuming a semi-calibrated setting with zero skew and constant intrinsics, by shifting the image coordinates such that the principal point is moved to the origin, the calibration matrix simplifies to a diagonal form $K = \text{diag}(f, f, 1)$. Under this parameterization, the Fundamental matrix F is related to the Essential matrix E by $F = K^{-T} E K^{-1}$.

The metric constraints on F can be derived from the properties of the Essential matrix ($2EE^T E - \text{trace}(EE^T)E = 0$) [10]. In terms of the Fundamental matrix and unknown focal lengths, this leads to the Kruppa equations. For our idealized case of equal focal lengths ($f_1 = f_2 = f$), the relation can be expressed as:

$$\det\left(F \text{diag}(1, 1, 0) F^T\right) = 0 \tag{35}$$

However, solving this polynomial directly from a noisy F is notoriously unstable [4]. Singularities arise when the optical axes are parallel or when the epipole is close to the principal point.
To overcome this, we employ the robust solver proposed by Kocur et al. [14]. This method reformulates the problem as an optimization that incorporates soft priors on the focal length range. Instead of seeking an exact algebraic root for a single f, it relaxes the problem and extracts two focal length estimates (f_1, f_2) that minimize the violation of the trace constraint while staying within a plausible range. This allows us to recover valid focal estimates even from noisy minimal samples, provided they correspond to a non-critical geometry.
For each refined hypothesis F^*, we evaluate the extracted pair (f_1, f_2). We retain the hypothesis only if it satisfies the semi-calibrated assumption:

$$f_1 > 0, \quad f_2 > 0, \quad \frac{|f_1 - f_2|}{(f_1 + f_2)/2} < \tau_{diff} \tag{36}$$

This step acts as a local filter, rejecting models that require complex or drastically inconsistent focal lengths (e.g., $f_1 \ll f_2$) that violate the constant-intrinsics premise.

5.2.3. Step 3: KDE Filtering

To further resolve ambiguities, we assume that all valid motions in the scene share a single dominant focal length. We estimate the probability density function (PDF) of the extracted focals using Kernel Density Estimation (KDE) (Algorithm 1, line 15):

$$\hat{p}(f) = \frac{1}{mh} \sum_{i=1}^{m} K\left(\frac{f - f_i}{h}\right) \tag{37}$$

where $K(\cdot)$ is a Gaussian kernel and h is the bandwidth parameter, automatically selected via Scott's Rule [21]. We define the valid consensus range centered on the dynamically estimated local peak f_{peak}. To construct this range, we apply a globally fixed tolerance threshold τ_{range} (empirically derived from the whole dataset, see Subsection 7.5) and filter out hypotheses falling outside this band (line 17). This acts as a global consensus check, pruning algebraic artifacts that passed the local filter but are inconsistent with the scene's global geometry.

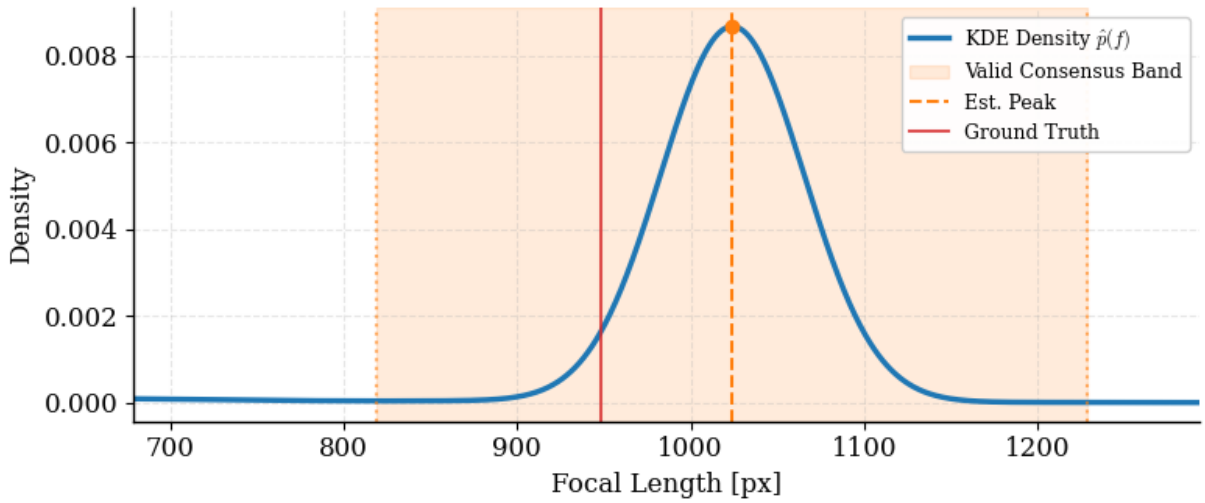


Figure 8: **Focal Consensus on a Single Image Pair.** KDE density estimation of focal length hypotheses dynamically computed for one specific scene from the HOPE-F dataset. While the peak f_{peak} (dashed orange line) is estimated *locally* for each image pair, the width of the valid consensus band (shaded area) is defined by a *global* threshold ($\tau_{range} = 0.20$). This globally fixed tolerance is derived empirically from the entire dataset to prioritize gross-outlier rejection, as detailed in Subsection 7.5 and Figure 9.

5.2.4. Step 4: Preference Representation and Clustering

The surviving hypotheses $\mathcal{H}_{final} = \{h_1, \ldots, h_M\}$ form the basis of the conceptual space. We represent each point $\mathbf{x}_i$ by a preference vector $\mathbf{p}_i$, whose components represent the compatibility with each hypothesis h_j (Algorithm 1, line 22).

Using the geometric residual r_{ij} (Sampson distance), the preference function is defined using the characteristic function of T-Linkage [18]:

$$\phi(\mathbf{x}_i, h_j) = \begin{cases} \exp\left(-\frac{r_{ij}}{\tau}\right) & \text{if } r_{ij} < 5\tau \\ 0 & \text{otherwise} \end{cases} \tag{38}$$

where τ is the inlier threshold. Consequently, the preference vector is $\mathbf{p}_i = [\phi(\mathbf{x}_i, h_1), \ldots, \phi(\mathbf{x}_i, h_M)]$.
The clustering is performed using an agglomerative linkage scheme based on the Tanimoto distance between these vectors (line 23):

$$d_T(\mathbf{p}_i, \mathbf{p}_k) = 1 - \frac{\langle \mathbf{p}_i, \mathbf{p}_k \rangle}{\|\mathbf{p}_i\|^2 + \|\mathbf{p}_k\|^2 - \langle \mathbf{p}_i, \mathbf{p}_k \rangle} \tag{39}$$

At each step of the agglomerative process, the two clusters with the minimum Tanimoto distance are merged. Crucially, when two clusters A and B are merged into a new cluster C, the preference vector of the newly formed cluster is defined as the element-wise minimum of its parents' preference vectors:

$$\mathbf{p}_C = \min(\mathbf{p}_A, \mathbf{p}_B) \tag{40}$$

This operation corresponds to the continuous logic intersection of the two preference sets, meaning that the merged cluster only retains strong preferences for geometric hypotheses that effectively model *both* original sets of points.
Consequently, the Tanimoto distance between the new cluster C and all other existing clusters changes and must be re-evaluated. This continuous updating of the preference vectors dynamically invalidates previously computed pairwise distances. As we will detail in Section 6, this specific mathematical property of T-Linkage destroys the validity of standard distance caching and strictly motivates the engineering of the custom *lazy deletion* Min-Heap data structure used to accelerate our pipeline.
The number of distinct motions is determined dynamically during this phase. Agglomeration proceeds as long as there exist clusters with a Tanimoto distance $d_T < 1$, indicating a shared preference for at least one geometric hypothesis. The process terminates when the preference sets of the remaining clusters become mutually exclusive (i.e., $d_T = 1$). To ensure that the recovered motions are statistically significant, we finally retain only those clusters that contain a sufficient number of correspondences to define a stable model, effectively treating smaller groups as outliers.
Crucially, by filtering $\mathcal{H}_{final}$ via focal consensus before this step, we ensure that the preference space is not contaminated by geometrically inconsistent models, leading to sharper separation between motions and avoiding the erroneous agglomeration of disparate structures.

6. Implementation Details

Moving from the theoretical framework of the Semi-Calibrated T-Linkage to a practically scalable algorithm comes with its fair share of computational hurdles. The inherent complexity of multi-model fitting, especially during the continuous clustering phase with its dense preference matrices, can quickly become unmanageable when dealing with high-resolution images and thousands of feature points. To bridge this gap, this section outlines the engineering choices we made to guarantee both numerical stability and execution speed. Specifically, we detail the hybrid software architecture and the custom data structures we designed to accelerate the clustering process.

6.1. Software Architecture and Stack

The proposed framework is implemented in Python 3.10 to leverage its rich ecosystem for data handling and rapid prototyping. However, the overhead of the Python interpreter and dynamic type checking is prohibitive for tight algorithmic loops required in geometric estimation. To bridge this performance gap, we adopted a hybrid architectural approach. The high-level orchestration, including data loading and evaluation metrics, relies on `NumPy` for vectorized operations. Conversely, the computationally intensive geometric kernels, specifically the robust residual computation and Tanimoto distance evaluation, are optimized using Numba. Numba is a Just-In-Time (JIT) compiler that translates Python functions into optimized machine code at runtime using the LLVM compiler infrastructure. This allows us to bypass the interpreter overhead for critical sections, achieving performance comparable to compiled C++ implementations while maintaining code flexibility. Furthermore, for the delicate task of focal length estimation, we interface with PoseLib [14], a highly optimized C++ library with Python bindings that provides state-of-the-art implementations for minimal solvers, ensuring that the roots of the polynomial systems are found with maximum numerical precision.

6.2. Accelerating Geometric Kernels via JIT

A critical computational bottleneck in T-Linkage is the generation of the continuous preference matrix. Given N data points and M geometric hypotheses, the algorithm must compute $N \times M$ Sampson errors. For a typical scene with $N = 2000$ points and $M = 3000$ hypotheses, this results in millions of floating-point evaluations per image pair.
To handle this, we leverage Numba's `@njit` mode with the `fastmath=True` directive. This configuration instructs the compiler to relax strict IEEE 754 compliance, enabling aggressive algebraic simplifications and, crucially, SIMD (Single Instruction, Multiple Data) vectorization. By manually unrolling loop iterations and keeping the Fundamental matrix coefficients in CPU registers, we minimize memory access latency, which is the primary constraint in large-scale geometric verification.

6.3. Min-Heap for Efficient Continuous Clustering

The standard implementation of agglomerative clustering requires identifying the pair of clusters with the minimum distance at each iteration. In a naive implementation using an adjacency matrix, finding this minimum requires an exhaustive search over the active clusters, leading to a complexity of $\mathcal{O}(C^2)$ per step and an overall cubic complexity $\mathcal{O}(N^3)$, which is prohibitive for datasets with thousands of points.
To strictly bound this complexity, we engineered a custom Min-Heap (Priority Queue) data structure specifically tailored for the continuous Tanimoto distance. Unlike standard libraries, our implementation supports the dynamic nature of the T-Linkage process via a *lazy deletion* strategy. When two clusters u and v are merged into a new cluster w, the pairwise distances involving u and v currently stored in the heap become invalid. Instead of performing a costly linear scan to remove these outdated entries, we simply flag the old clusters as "inactive". When the heap's `pop` operation extracts an entry involving an inactive cluster, it is discarded in $\mathcal{O}(1)$ time. This optimization ensures that the maintenance of the priority queue remains logarithmic $\mathcal{O}(\log K)$, drastically reducing the runtime of the clustering phase compared to the baseline implementation.

6.4. Hyperparameters and Reproducibility

For the sake of complete algorithmic reproducibility, the primary hyperparameters utilized in our testing framework are detailed in Table 1. These values were determined empirically on a validation subset of the HOPE-F dataset and were kept strictly constant throughout all automated evaluations on the test split (sequences 900-999 for each category).

Table 1: **Algorithmic Hyperparameters.** Values used for the evaluation on the HOPE-F test split.

Parameter	Value	Description
`NUM_MSS_SAMPLES`	3000	Number of random minimal sample sets generated
`MSS_SIZE`	8	Points used for the initial 8-point algorithm
`BASE_THR_PX` (τ)	2.5	Baseline inlier threshold (pixels) for IRLS and preferences
`MAX_REL_F12_DIFF`	0.05	Max allowed relative difference between f_1 and f_2
`FOCAL_PEAK_RANGE`	0.20	KDE acceptance range around the dominant peak f_{peak}
`IRLS_ITER`	5	Maximum outer iterations for non-linear refinement

The configuration of the `FOCAL_PEAK_RANGE` at 0.20 (a 20% relative tolerance) is of particular geometric importance. The distribution of estimated focal lengths derived from noisy, real-world correspondences, especially in near-degenerate planar scenes, does not form a perfect Dirac delta function but exhibits a measurable variance. Setting this threshold too strictly would inadvertently reject valid physical models, whereas a looser threshold would fail to adequately prune algebraic hallucinations.

7. Experiments

This section details the experimental evaluation of the proposed Semi-Calibrated T-Linkage framework. We present the datasets utilized, the selected state-of-the-art competitors, the evaluation metrics, and a comprehensive analysis of the results, encompassing both quantitative performance and computational complexity.

7.1. Datasets

To comprehensively evaluate the quantitative accuracy of the proposed framework and demonstrate its broad applicability across diverse real-world scenarios, we selected three challenging benchmarks.

- **HOPE-F [13]:** This dataset comprises 4000 image pairs derived from ScanNet [6], evenly distributed across four distinct categories representing different scene configurations. It features realistic indoor man-made environments populated by furniture, clutter, and regular structures. These configurations represent a significant challenge for fundamental matrix estimation due to the complex arrangement of rigid bodies, varying object scales, and the measurement noise inherent to real-world sensor data.
- **AdelaideRMF [30]:** A well-known benchmark in the multi-model fitting literature that includes 38 image pairs of general dynamic scenes. While HOPE-F focuses on highly structured indoor environments, AdelaideRMF introduces objects with complex, irregular 3D shapes and very few flat surfaces. This sharp contrast makes it an ideal testbed to ensure our algorithm generalizes effectively to less constrained scenarios..
- **KITTI [9]:** To qualitatively demonstrate the framework's robustness in highly dynamic, unconstrained outdoor scenarios, we employ sequences from the KITTI autonomous driving dataset. This benchmark involves significant ego-motion interacting with independently moving agents (e.g., cars, cyclists), effectively showcasing the algorithm's potential for real-world robotic perception and autonomous navigation.

7.2. Competitors

We benchmark our proposed *Semi-Calibrated T-Linkage* against the standard uncalibrated baseline and a selection of state-of-the-art multi-model fitting algorithms. These competitors were carefully chosen to represent different optimization strategies within the field:

- **T-Linkage (Uncalibrated) [18]:** This serves as our primary baseline. It is the consensus-based clustering method upon which our framework is built. Crucially, it operates in a purely uncalibrated setting, using continuous preference analysis but treating Fundamental matrices as independent algebraic entities without enforcing any consistency regarding the cameras' intrinsic parameters.
- **Progressive-X [2]:** This well-established algorithm takes a sequential "detect-and-remove" route. While our framework evaluates models globally, Progressive-X explicitly enforces spatial coherence during the fitting process. It operates under the practical assumption that true inliers of a specific motion are usually found near each other in the 2D image plane.
- **Progressive-X+ [3]:** An advanced evolution of Progressive-X. Instead of clustering in the data space, it performs clustering within the *consensus space*, representing the current state-of-the-art for sequential, spatially-aware geometric fitting.
- **Fast-CP [20]:** A highly efficient spectral clustering approach. Fast-CP ignores spatial heuristics entirely and instead uses Christoffel polynomials to draw fast hypothesis samples. By relying purely on the algebraic properties of the data, it achieves significant scalability and computational speed.

7.3. Evaluation Metrics

To quantitatively assess the quality of the motion segmentation, we employ several standard metrics from the multi-model fitting literature.

Misclassification Error (ME). The primary metric for motion segmentation is the Misclassification Error. Given a set of ground truth labels $\mathcal{G} = \{g_1, \ldots, g_N\}$ and the algorithm's predicted labels $\mathcal{L} = \{l_1, \ldots, l_N\}$ for the N data points, the ME measures the fraction of points assigned to an incorrect cluster. Because cluster identifiers are assigned arbitrarily by the algorithm, computing the true error requires finding the optimal one-to-one mapping between the predicted and ground truth label sets. Formally, this is defined as:

$$\text{ME} = \frac{1}{N} \min_{\pi \in \Pi} \sum_{i=1}^{N} \mathbb{I}(g_i \neq \pi(l_i)) \tag{41}$$

where Π represents the set of all possible label permutations, and $\mathbb{I}(\cdot)$ is the indicator function. This optimal permutation π is efficiently computed using the Hungarian algorithm [15].

Accuracy (ACC). Directly derived from the Misclassification Error, Accuracy quantifies the overall proportion of correctly classified correspondences after the optimal label permutation has been applied. It is simply defined as the complement of the ME:

$$\text{ACC} = 1 - \text{ME} = \frac{1}{N} \max_{\pi \in \Pi} \sum_{i=1}^{N} \mathbb{I}(g_i = \pi(l_i)) \tag{42}$$

False Negative Rate (FNR). In motion segmentation, realistic datasets are contaminated by gross outliers, typically assigned a specific label (e.g., $g_i = 0$). The FNR measures the proportion of true, valid inliers that the algorithm incorrectly discards as outliers. Let $\mathcal{V} = \{i \mid g_i \neq 0\}$ be the set of valid ground-truth inlier indices. The FNR is calculated as:

$$\text{FNR} = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \mathbb{I}(l_i = 0) \tag{43}$$

This metric is crucial for evaluating whether an algorithm achieves low classification error merely by aggressively rejecting difficult but valid matches.

Purity (PUR). Purity evaluates the homogeneity of the estimated clusters. It measures the extent to which each predicted cluster C_k contains data points originating from a single ground-truth class G_j. Formally, if we have K predicted clusters, the purity is defined as:

$$\text{PUR} = \frac{1}{N} \sum_{k=1}^{K} \max_j |C_k \cap G_j| \tag{44}$$

A high purity score indicates that the algorithm successfully avoids merging independent motions. However, purity does not penalize over-segmentation (i.e., splitting a single true motion into multiple smaller clusters).

Frobenius Norm (FROB). While the previous metrics evaluate point-wise classification, the Frobenius norm assesses the strict geometric accuracy of the estimated models. It measures the algebraic deviation between the estimated Fundamental matrix $\hat{F}$ and the ground-truth matrix F_{gt}. Because Fundamental matrices are projective entities defined only up to an arbitrary scale factor, both matrices must be L_2-normalized before comparison. Furthermore, since F and $-F$ represent the same epipolar geometry, we take the minimum distance across both signs:

$$\text{FROB}(\hat{F}, F_{gt}) = \min \left(\left\| \frac{\hat{F}}{\|\hat{F}\|_F} - \frac{F_{gt}}{\|F_{gt}\|_F} \right\|_F, \left\| \frac{\hat{F}}{\|\hat{F}\|_F} + \frac{F_{gt}}{\|F_{gt}\|_F} \right\|_F \right) \tag{45}$$

where $\| \cdot \|_F$ denotes the standard matrix Frobenius norm. A lower FROB score indicates that the estimated geometry closely aligns with the true physical camera configuration.

7.4. State-of-the-Art Comparison

Table 2 presents a comparative analysis with state-of-the-art methods on both the HOPE-F and AdelaideRMF benchmarks. The Misclassification Error (ME) is reported in terms of *Mean* ± *Standard Deviation*.

Comparative Results (ME)

Method	**HOPE-F** (ME ↓)	**AdelaideRMF** (ME ↓)
T-Linkage (Baseline)	24.43 ± 15.12	15.60 ± 12.23
Progressive-X [2]	22.78 ± 15.60*	12.85 ± 14.10*
Fast-CP [20]	24.59 ± 13.30*	$\mathbf{4.92 \pm 5.23}$*
Progressive-X+ [3]	43.23 ± 15.70*	6.57 ± 6.64*
Ours (Semi-Calibrated)	$\mathbf{22.31 \pm 13.74}$	10.49 ± 12.43

Table 2: Misclassification Error (%) reported as *Mean* ± *Std*. The proposed method achieves the lowest error on HOPE-F and outperforms both the baseline and Progressive-X on AdelaideRMF, demonstrating the effectiveness of the semi-calibrated focal length filter. (*) Values for competitors are taken from [13].

7.4.1. Discussion

The results highlight the effectiveness of the proposed semi-calibrated constraints. On the *HOPE-F* benchmark, which is characterized by synthetic scenes with varied focal lengths, our Semi-Calibrated T-Linkage achieves

the lowest mean error (22.31%). It outperforms both the uncalibrated baseline (24.43%) and highly optimized geometric methods like Fast-CP (24.59%) and Progressive-X (22.78%). Importantly, the reduction in standard deviation (from 15.12 to 13.74) compared to the baseline indicates that enforcing focal consistency not only improves average accuracy but also significantly stabilizes the estimation process across varying and challenging scene configurations.

On the *AdelaideRMF* dataset, which features generic dynamic objects, our method achieves a substantial improvement over the baseline, reducing the error from 15.60% to 10.49%. Notably, our approach outperforms the spatially-aware Progressive-X (12.85%), demonstrating that a global focal length consensus acts as a superior filter compared to local spatial coherence heuristics in unconstrained real-world scenarios. While highly specialized solvers like Fast-CP currently hold the lead on this specific dataset, the consistent performance gain we demonstrate over the standard uncalibrated baseline validates the proposed focal prior as a powerful, general-purpose tool for discriminating independent motions.

7.4.2. Statistical Significance

To rigorously validate the performance gains over the baseline, we conducted a one-tailed *Wilcoxon Signed-Rank Test* [29] on the paired results for both datasets, analyzing both the Misclassification Error (ME) and the False Negative Rate (FNR).
On the *HOPE-F* benchmark, the improvements are highly statistically significant, yielding p-values of 6.05×10^{-5} for ME and 6.07×10^{-5} for FNR. These extremely low values confirm that the proposed geometric filtering systematically enhances performance in structured environments.
Crucially, the improvement is also statistically significant on the generic *AdelaideRMF* dataset. We obtained p-values of 7.45×10^{-3} for ME and 7.30×10^{-3} for FNR ($p < 0.01$). This consistent statistical significance across different metrics and dataset types reinforces the conclusion that the semi-calibrated constraint provides a robust and reliable gain in both segmentation accuracy and inlier recall.

7.5. Detailed Analysis and Ablation

To deeper investigate the source of the improvement, we performed a detailed comparison against the Baseline across all metrics and analyzed the impact of our chosen parameters.

7.5.1. Detailed Metrics

Table 3 reports a comprehensive evaluation using both geometric and clustering metrics. All classification metrics are reported in percentages.

Detailed Analysis vs Baseline

Method	ME ↓	ACC ↑	FNR ↓	PUR ↑	ARI ↑	FROB ↓
Dataset: HOPE-F						
T-Linkage (Baseline)	24.43 ± 15.12	75.57 ± 15.12	23.48 ± 15.11	86.92 ± 10.26	$\mathbf{65.15 \pm 17.82}$	0.028 ± 0.032
Ours (Semi-Calib)	$\mathbf{22.31 \pm 13.74}$	$\mathbf{77.69 \pm 13.74}$	$\mathbf{22.36 \pm 13.76}$	$\mathbf{87.15 \pm 9.70}$	63.85 ± 16.97	$\mathbf{0.027 \pm 0.028}$
Dataset: AdelaideRMF						
T-Linkage (Baseline)	15.60 ± 12.23	84.40 ± 12.23	15.71 ± 12.23	$\mathbf{94.02 \pm 8.76}$	$\mathbf{0.783 \pm 0.133}$	$\mathbf{0.0097 \pm 0.008}$
Ours (Semi-Calib)	$\mathbf{10.49 \pm 12.43}$	$\mathbf{89.51 \pm 12.43}$	$\mathbf{10.59 \pm 12.43}$	88.54 ± 11.91	0.774 ± 0.125	$\mathbf{0.0097 \pm 0.007}$

Table 3: Detailed comparison using full clustering and geometric metrics. Values are $Mean \pm Std$ (%). Note: **FROB** is reported in its original scale (geometric norm). Best results are highlighted in **bold**.

For the *HOPE-F* dataset, integrating the semi-calibrated filter raises the Accuracy (ACC) from 75.57% to 77.69% and brings Purity (PUR) up to 87.15%. More importantly, the lower Frobenius Norm (FROB) highlights a key advantage: once we discard the bad hypotheses, the remaining models are both better clustered and structurally much closer to the ground truth.
The performance shift is even sharper on *AdelaideRMF*. Here, Accuracy surges to 89.51% (up from 84.40%) while the Misclassification Error (ME) drops substantially. We also manage to bypass a well-known issue: whereas purely geometric methods often over-segment noisy scenes, our approach holds steady with an 88.54% Purity and a highly competitive FROB error of 0.0097. Ultimately, this confirms that demanding focal consistency successfully throws out physically impossible configurations while fully preserving the exactness of the recovered camera motions.

7.5.2. Parameter Sensitivity: Focal Consistency

A critical component of our pipeline is the rejection of hypotheses based on the scene-level focal length consensus. To determine the optimal threshold τ_{range}, we analyzed the distribution of the relative error between the estimated focal peak f_{peak} and the ground truth f_{gt} across the HOPE-F dataset.

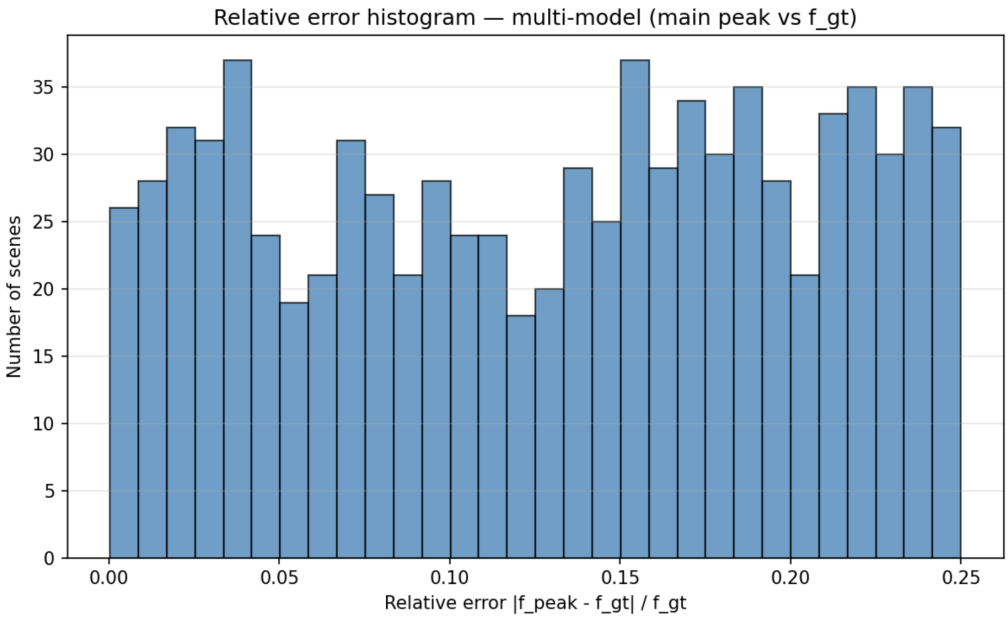


Figure 9: Distribution of the relative error $|f_{peak} - f_{gt}|/f_{gt}$ on HOPE-F. The variance highlights the challenge of focal estimation on regular scenes, justifying the choice of a tolerant threshold $\tau_{range} = 0.2$.

As illustrated in Figure 9, focal length estimation from regular scenes is inherently ambiguous due to frequent geometric degeneracies, resulting in a dispersed error distribution. The histogram reveals that a significant portion of valid scenes exhibits a relative error of up to 0.20. Based on this empirical evidence, selecting a strict threshold (e.g., < 0.05) would be detrimental, as it would aggressively reject correct matches and lower recall. Therefore, we set an intentionally permissive, globally fixed threshold $\tau_{range} = 0.2$. During the pipeline execution on any individual scene (as shown in the KDE plot of Figure 8), this global 20% tolerance is applied symmetrically around the locally estimated focal peak.

7.5.3. Ablation Study: Impact of Refinement

We validated the necessity of the non-linear refinement step (IRLS-Tukey) applied prior to the focal length solver. As reported in Table 4, omitting this geometric refinement leads to unstable focal estimates and a significantly higher Misclassification Error (28.45% vs 25.78%). This confirms that geometric stabilization is a strict prerequisite for robust consistency checks: raw algebraic estimates from noisy Minimal Sample Sets are often too degraded to allow for reliable focal length recovery.

Ablation Study (HOPE-F)

Configuration	ME (%) ↓
Raw-MSS + PoseLib	28.45
Refined-MSS + PoseLib (Ours)	**25.78**

Table 4: Impact of the geometric refinement step on the final Misclassification Error (ME).

7.5.4. Impact of the Number of Hypotheses (M)

A core hyperparameter in any hypothesize-and-cluster framework is the number of minimal sample sets, M, drawn during the initial generation phase. To investigate the scalability, data-efficiency, and sensitivity of our approach to this parameter, we evaluated the Misclassification Error (ME) as a function of M. We compared our Semi-Calibrated formulation against the standard Uncalibrated baseline across the AdelaideRMF dataset, varying M from 500 to 5000.

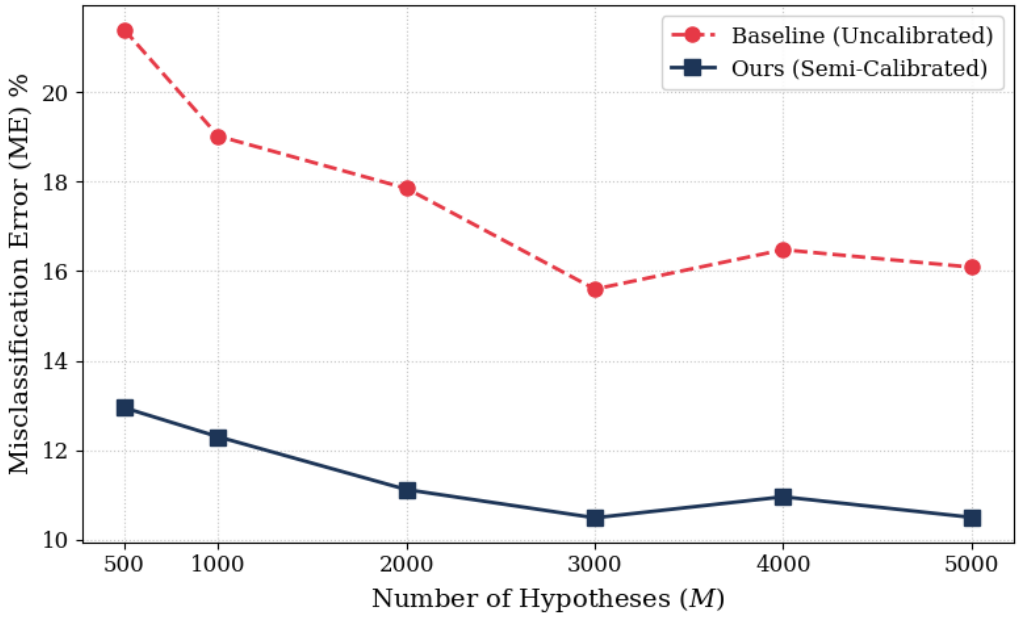


Figure 10: **Impact of Hypothesis Count on AdelaideRMF.** Misclassification Error (ME) plotted against the number of generated hypotheses M. The proposed semi-calibrated filter (blue squares) allows the algorithm to converge to a significantly lower error much faster than the uncalibrated baseline (red circles). The optimal trade-off between geometric accuracy and computational burden is empirically found at $M = 3000$.

As illustrated in Figure 10, a general downward trend in the error is observable for both methods as M increases. This behavior is theoretically expected: generating a larger pool of samples guarantees a denser and more comprehensive exploration of the geometric hypothesis space, increasing the probability of extracting all-inlier minimal sets for every independent motion in the scene.
However, a stark and critical difference emerges in the *convergence rate* and the absolute error margins between the two approaches. The uncalibrated baseline (red dashed line) exhibits a slow descent, requiring a massive number of samples ($M \geq 3000$) merely to stabilize around a 16% error rate. This happens because, in the uncalibrated setting, the hypothesis space is highly polluted; a large portion of the generated Fundamental matrices are algebraically valid but physically meaningless (spurious artifacts). Consequently, a massive M is needed just to ensure a sufficient absolute number of "good" models.
In stark contrast, our Semi-Calibrated T-Linkage (blue solid line) demonstrates exceptional data-efficiency. Even at a severely constrained setting of $M = 500$, our method yields an error ($\sim 13\%$) that is substantially lower than what the baseline achieves at its absolute best configuration ($M = 5000$, $\sim 16\%$). By preemptively discarding non-Euclidean models via focal consensus, our formulation drastically increases the *effective density* of physically valid motions within the surviving hypothesis set.

Justification for the Optimal Operating Point ($M = 3000$). A detailed inspection of the semi-calibrated curve reveals a distinct performance plateau. The error drops steeply until $M = 3000$, where it reaches a local minimum of 10.49%. Beyond this point, increasing the number of hypotheses to $M = 4000$ or $M = 5000$ does not yield any further benefit; the error marginally fluctuates and stabilizes back to roughly 10.50% at $M = 5000$.
This saturation indicates that at $M = 3000$, the algorithm has already gathered a structurally complete set of valid geometric models to optimally explain the scene. While setting $M = 5000$ might seem intuitively safer to maximize spatial coverage, it introduces a severe penalty in execution time. As formalized in Section 7.7, the continuous agglomerative clustering phase of T-Linkage is the most computationally expensive step, and its complexity scales directly with the dimension of the preference matrix. Feeding 5000 initial hypotheses into the pipeline substantially inflates the pairwise distance computations without providing any tangible gain in segmentation accuracy.
Therefore, based on this empirical evidence, we selected $M = 3000$ as the definitive hyperparameter for all our main experiments. This value represents the optimal "sweet spot": it guarantees state-of-the-art accuracy by fully saturating the geometric performance bounds, while strictly preventing unnecessary computational overhead during the preference-based clustering phase.

Adaptive Sampling via J-Silhouette. While fixing $M = 3000$ yields optimal results empirically, a theoretically superior approach would involve determining the number of hypotheses adaptively. In the context of preference-based clustering, the *J-Silhouette* index [25] has been proposed as an internal validity metric to assess the quality of the segmentation without ground truth. By measuring the tightness of the preference vectors within a cluster relative to their separation from other clusters, the J-Silhouette could strictly guide the sampling process, stopping the generation of hypotheses only when the global clustering quality maximizes. However, we excluded this adaptive mechanism from our final pipeline due to its prohibitive computational cost. Calculating the Silhouette coefficient requires computing all pairwise distances between data points, leading to a complexity of $\mathcal{O}(N^2)$ for every evaluation step. Integrating this metric into the iterative sampling loop would introduce a bottleneck significantly heavier than the geometric estimation itself, negating the runtime efficiency gains provided by our semi-calibrated filtering and Min-Heap optimization. Consequently, the empirical selection of a fixed M represents the necessary trade-off between geometric completeness and computational tractability.

7.6. Qualitative Results

Visual inspection provides further insight into the robustness of our approach across the different benchmarks.

7.6.1. HOPE-F: Regular Structures

Figure 11 visually compares our method against the baseline using a challenging example from the HOPE-F benchmark. By displaying both views, the actual 3D movement of the objects becomes much easier to track. The scene itself is quite complex, packed with distinct rigid bodies like a large box and various cylindrical containers that are all moving independently of one another.

T-Linkage **Semi-Calibrated T-Linkage**

(a) View 1 (b) View 1

(c) View 2 (d) View 2

Figure 11: **Qualitative Comparison on HOPE-F.**

As observed in the left column of Figure 11, the standard T-Linkage fails to correctly delineate the independent motions across the sequence. It suffers from specific under-segmentation, erroneously merging the yellow cylindrical container and the adjacent blue-and-white can into a single, inconsistent cluster. This is visualized by the shared pink diamonds on objects that are clearly disjoint in their 3D movement.
In contrast, our method, shown in the right column (Figures 11b, 11d), successfully retrieves the correct multi-body structure and maintains label consistency between View 1 and View 2. The proposed semi-calibrated formulation effectively prunes outlier hypotheses, specifically those bridge Fundamental matrices that algebraically satisfy multiple motions but lack metric consistency. This prevents the erroneous merging of independent structures, resulting in the correct separation of the two cylinders as they move through the scene.

7.6.2. AdelaideRMF: Generic 3D Objects

To assess applicability beyond regular environments, we report qualitative results on AdelaideRMF in Figure 12. Our Semi-Calibrated T-Linkage successfully segments objects with complex 3D structures (bread, cube, chips), distinguishing them despite the lack of dominant planes. This supports our claim that the "constant intrinsics" assumption acts as a powerful global constraint for general multi-model fitting tasks.

(a) View 1

(b) View 2

Figure 12: **Qualitative Results on AdelaideRMF.** Segmentation of objects with complex 3D structures.

7.6.3. Real-World Application: KITTI

To demonstrate the practical utility in robotic perception, we refer to the qualitative results on the KITTI benchmark presented in Figure 1. Unlike the previous benchmarks, this scenario involves a moving camera (ego-motion) interacting with independently moving agents.
As shown in the teaser, our method successfully segments the scene into geometrically consistent clusters: the static environment (buildings, road) is correctly grouped into a single dominant motion (red circles), while the dynamic cyclist and the truck are isolated as distinct entities. This confirms the framework's robustness in handling ego-motion and unconstrained dynamic objects.

7.7. Computational Complexity and Runtime

We analyze the theoretical complexity to evaluate the efficiency trade-offs between our semi-calibrated approach and state-of-the-art uncalibrated methods. Let N be the number of correspondences, S the initial number of sampled hypotheses, and M_{out} the number of hypotheses surviving our focal consensus filter ($M_{out} \ll S$).

Complexity Comparison

Method	Complexity	Strategy
T-Linkage [18]	$\mathcal{O}(N^2S + N^3)$	Global / Agglomerative
Progressive-X / X+ [2, 3]	$\mathcal{O}(N \log N + SN)$	Sequential / Greedy
Fast-CP [20]	$\mathcal{O}(SN)$	Batch / Ranked
Semi-Calibrated T-Linkage (Ours)	$\mathcal{O}(SN + N^2 M_{out})$	Global / Optimized

Table 5: Complexity comparison of motion segmentation frameworks.

7.7.1. Discussion

As summarized in Table 5, state-of-the-art methods like Progressive-X and Fast-CP achieve asymptotic linearity ($\mathcal{O}(N)$), explaining their speed advantage. However, they rely on greedy decisions (sequential or ranked selection) which can suffer from early commitment to incorrect models. Our method retains the theoretical advantage of global competition among models (avoiding greedy pitfalls) but renders it computationally tractable. The transition from the naive $\mathcal{O}(N^3)$ baseline complexity to a heap-based approach dominated by $\mathcal{O}(N^2 \cdot M_{out})$ ensures that efficiency is primarily bound by the number of valid geometric models rather than the overhead of invalid algebraic artifacts.

7.7.2. Runtime Analysis

We also evaluated how our pruning mechanism affects the overall execution time. By discarding approximately 10% of physically inconsistent hypotheses early on, the algorithm deals with a significantly smaller preference matrix during the distance calculations. This directly accelerates the core clustering phase, trimming its runtime down to 0.72s, compared to the 1.04s required by the baseline. Naturally, the initial IRLS refinement and focal length extraction introduce some computational overhead that offsets the time saved during clustering. Ultimately, however, our total pipeline runs in roughly the same amount of time as the standard baseline, while delivering a substantial boost in both accuracy and recall.

7.8. Limitations and Future Directions

The primary boundary of our current framework is the strict "semi-calibrated" assumption ($f_1 \approx f_2$). While valid for typical rigid scene sequences, this constraint limits applicability in scenarios with significant zoom variations or heterogeneous sensor setups, where the global consensus on a single scalar becomes ill-defined.
Future research will address this by relaxing the equality constraint, moving towards a density estimation of the focal ratio $\gamma = f_2/f_1$ to handle zooming cameras while maintaining rejection capabilities for degenerate configurations.
Finally, we highlight the *modularity* of the proposed contribution. The hypothesis generation and filtering pipeline is solver-agnostic and decouples the geometric consistency check from the clustering logic. A promising direction is the integration of this "physical filter" into real-time, sequential frameworks or energy-based methods. This would allow transferring the robustness of semi-calibrated constraints to faster optimization engines, paving the way for consistent motion segmentation in real-time robotic perception.

8. Conclusions

This work has revisited the multi-model fitting problem through the lens of semi-calibrated geometry. While the assumption of constant, unknown intrinsics is a cornerstone of static Structure-from-Motion, its potential as a regularizer in the *multi-body domain* has remained largely unexplored. Our investigation demonstrates that enforcing a scene-level consensus on the focal length is not merely an additional constraint, but a fundamental mechanism to bridge the gap between algebraic fitting and consistency. By explicitly modeling the focal length, our *Semi-Calibrated T-Linkage* effectively discriminates between geometrically valid motions and the spurious algebraic artifacts that plague standard solvers. Experimental evidence on both synthetic (HOPE-F) and real-world (AdelaideRMF) datasets suggests that the notorious instability of fundamental matrix estimation is

often a byproduct of unconstrained optimization. Notably, by pruning physically impossible configurations, our approach not only improves the baseline but also outperforms state-of-the-art geometric solvers like Progressive-X on real-world benchmarks, achieving robust segmentation without requiring training data or heuristics based on spatial proximity.

References

[1] Federica Arrigoni, Luca Magri, and Tomas Pajdla. On the usage of the trifocal tensor in motion segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 514–530. Springer, 2020.

[2] Daniel Barath and Jiri Matas. Progressive-X: Efficient, anytime, multi-model fitting algorithm. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3779–3787, 2019.

[3] Daniel Barath, Denys Rozumnyi, Ivan Eichhardt, Levente Hajder, and Jiri Matas. Progressive-X+: Clustering in the consensus space. *CoRR*, abs/2103.13875, 2021.

[4] Stéphane Bougnoux. From projective to euclidean space under any practical situation, a criticism of self-calibration. In *Proceedings of the Sixth International Conference on Computer Vision (ICCV)*, pages 790–798, 1998.

[5] Ondrej Chum, Tomas Werner, and Jiri Matas. Two-view geometry estimation unaffected by a dominant plane. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 772–779, 2005.

[6] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2432–2443, 2017.

[7] Ehsan Elhamifar and René Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(11):2765–2781, 2013.

[8] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[9] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361, 2012.

[10] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, 2003.

[11] Hossam Isack and Yuri Boykov. Energy-based geometric multi-model fitting. *International Journal of Computer Vision (IJCV)*, 97(2):123–147, 2012.

[12] Florian Kluger, Eric Brachmann, Hanno Ackermann, and Carsten Rother. Consac: Robust multi-model fitting by conditional sample consensus. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4634–4643, 2020.

[13] Florian Kluger and Bodo Rosenhahn. PARSAC: Accelerating robust multi-model fitting with parallel sample consensus. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[14] Viktor Kocur, Daniel Barath, and Viktor Larsson. Robust self-calibration of focal lengths from the fundamental matrix. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5220–5229, 2024.

[15] Harold W Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

[16] H Christopher Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293(5828):133–135, 1981.

[17] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[18] Luca Magri and Andrea Fusiello. T-linkage: A continuous relaxation of J-linkage for multi-model fitting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3954–3961, 2014.

[19] David Nistér. An efficient solution to the five-point relative pose problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 26(6):756–770, 2004.

[20] Bengisu Özbay, Octavia I. Camps, and Mario Sznaier. Fast two-view motion segmentation using Christoffel polynomials. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 13690, pages 1–19. Springer, 2022.

[21] David W Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.

[22] Charles V Stewart. Robust parameter estimation in computer vision. *SIAM Reviews*, 41(3):513–537, 1999.

[23] Henrik Stewénius, Christopher Engels, and David Nistér. Recent developments on direct relative orientation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 60:284–294, 2006.

[24] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning motion patterns in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3386–3394, 2017.

[25] Roberto Toldo and Andrea Fusiello. Robust multiple structures estimation with J-linkage. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 537–547, 2008.

[26] Roberto Tron and René Vidal. A benchmark for the comparison of 3-D motion segmentation algorithms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[27] René Vidal, Yi Ma, and Shankar Sastry. Generalized principal component analysis (GPCA). *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 27(12):1945–1959, 2005.

[28] Wikipedia contributors. Pinhole camera model. `https://en.wikipedia.org/wiki/Pinhole_camera_model`, 2024.

[29] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[30] Hoi Sim Wong, Tat-Jun Chin, Jin Yu, and David Suter. Dynamic and hierarchical multi-structure geometric model fitting. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1044–1051, 2011.

[31] Xun Xu, Loong Fah Cheong, and Zhuwen Li. Motion segmentation by exploiting complementary geometric models. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 2859–2867. Computer Vision Foundation / IEEE Computer Society, 2018.

Abstract in lingua italiana

La segmentazione del movimento a due viste, ovvero il compito di partizionare corrispondenze sparse in molteplici moti rigidi, risulta intrinsecamente complessa in uno scenario puramente non calibrato. Gli approcci standard trattano spesso i modelli di moto come entità geometriche indipendenti, trascurando un forte prior geometrico disponibile in molti scenari pratici: le immagini sono frequentemente acquisite da un singolo sensore con parametri intrinseci costanti, seppur ignoti. In questo articolo, affrontiamo tale problema operando in un *regime semi-calibrato*. Proponiamo un adattamento robusto dell'algoritmo T-Linkage, un metodo di clustering basato sulle preferenze, introducendo una fase di filtraggio delle ipotesi di moto basata sulla stima della lunghezza focale. Invece di affidarci alla totalità dei moti candidati generati, sfruttiamo l'ipotesi di parametri intrinseci costanti per identificare e scartare i modelli geometricamente incoerenti prima della fase di clustering. Ampi esperimenti condotti sui benchmark HOPE-F e AdelaideRMF dimostrano che l'utilizzo di questo vincolo semi-calibrato migliora significativamente l'accuratezza della segmentazione rispetto alla baseline non calibrata.

Parole chiave: Segmentazione del Movimento,Fitting Multi-Modello,Geometria Semi-Calibrata