# Multi-body Self-Calibration

Andrea Porfiri Dal Cin
andrea.porfiridalcin@polimi.it
Giacomo Boracchi
giacomo.boracchi@polimi.it
Luca Magri
luca.magri@polimi.it
Politecnico di Milano

**Abstract**

One of the main assumptions behind Structure-from-Motion is that of a *rigid* scene, *i.e.*, the scene is static or composed of a single moving object. The rigidity constraint – typically encoded in the Kruppa equations – is at the core of self-calibration and enables Euclidean upgrading from uncalibrated images. In this work, we show how it is possible to improve self-calibration by considering a *dynamic* scene composed of multiple moving rigid objects. The rationale of our solution is that each rigid motion provides a useful constraint that can be used to better estimate the intrinsics of the camera. Specifically, we introduce a self-calibration method for a single camera that exploits motion segmentation to identify rigid motions. Our solution capitalizes on all the available epipolar relations to robustly initialize the camera parameters, which are then optimized through nonlinear refinement. Experiments on real-world data show that our approach is comparable to state-of-the-art self-calibration methods when the scene is static and improves performance in the case of dynamic scenes. The code and a dataset with images of dynamic scenes and ground truth intrinsics are available at https://github.com/andreadalcin/MultiBodySelfCalibration.

## 1 Introduction

Nowadays, Structure-from-Motion (SfM) [2, 12, 15, 44] is a mature technology that produces 3D reconstructions at a quality level that meets industrial standards. SfM algorithms work assuming that the scene is static. However, physical reality is far from static, as we live in a dynamic environment where multiple objects move independently. On the one hand, multiple motions challenge classical single-body SfM algorithms [24], as all motions but the dominant one act as outliers and strain the reconstruction. On the other hand, multiple motions provide information [7] that, when properly exploited, better constrain the reconstruction. In this work, we address camera self-calibration [6, 8], namely the problem of estimating the internal camera parameters from a collection of sparse images of a scene. This task is a key component in SfM pipelines. We demonstrate that self-calibration benefits from multiple motions in a dynamic scene and introduce a novel self-calibration algorithm that capitalizes on this information under the assumption that the motions are rigid and that the camera parameters remain fixed during the acquisition.

We address *multi-body self-calibration* as follows. We are given $n$ images $\mathcal{I} = \{I_1, \dots, I_n\}$ of a 3D scene acquired by a projective camera from $n$ different poses $(\mathsf{R}_i, \mathbf{t}_i)$. The camera intrinsics $\mathsf{K}$ are fixed and defined as:

$$\mathsf{K} = \begin{pmatrix} f_x & s & u \\ 0 & f_y & v \\ 0 & 0 & 1 \end{pmatrix} \tag{1}$$

where $f_x$ and $f_y$ are the focal lengths, $(u_0, v_0)$ the principal point and $s$ the camera skew. As usual, we assume zero skew ($s = 0$) and do not account for radial distortion, although it is possible to rectify images beforehand using existing self-calibration methods, e.g., via [35]. We consider 3D scenes that contain $m \geq 1$ independently moving rigid bodies $\mathcal{B} = \{\beta_1, \dots, \beta_m\}$. When $m = 1$, the scene is *static*; if $m > 1$ we say the scene is *dynamic*. We always refer to rigid motions and do not consider non-rigid scene deformations. Our goal is to recover the intrinsic parameters $\mathsf{K}$ given only the collection of images $\mathcal{I}$.

In practice, self-calibration is a difficult problem, as it involves solving a system of non-linear polynomial equations. This is hindered by noise and outliers, which unavoidably affect image correspondences, and by degenerate motions that yield indeterminate solutions of the system. Our intuition is to exploit multiple motions of bodies $\mathcal{B}$ to constrain self-calibration better, as more information can be inferred from images in $\mathcal{I}$ when the scene is dynamic rather than static. Unfortunately, the more individual motions, the fewer inliers support them, making fundamental matrices estimation more susceptible to noise and outliers. Moreover, moving objects are typically small compared to the camera field of view, and may yield degenerate 3D motions that do not provide valid constraints. Thus, it is not straightforward to apply self-calibration to dynamic scenes, as robustness must be considered.

**Contributions** To the best of our knowledge, we are the first to introduce a practical *multi-body* self-calibration method for dynamic scenes that deals with noise and outliers. The major contributions of this work can be summarized as follows:
*i*) Our approach capitalizes on *all* the rigid motions $\mathcal{B}$ in the dynamic scene to better constrain self-calibration, as opposed to classical approaches that treat non-dominant motions as outliers. Thus, self-calibration can be attained from fewer images, as theorized in [7].
*ii*) We introduce a *Motion Segmentation* tailored specifically to the problem. We exploit the rigidity constraints to recover fundamental matrices describing rigid motions and, at the same time, estimate the focal length. Instead, classical uncalibrated segmentation approaches are limited to fundamental matrices.
*iii*) Camera parameters are estimated by a non-linear optimization routine which we augment with multiple robustness layers. Specifically, we focus on the robustness of the initialization and on numerical stability.

## 2 Related Work

Since the introduction of self-calibration by [6], several methods have attained compelling results for *static* scenes. Early methods [13, 20, 41] have laid down the foundations to *directly* estimate the camera intrinsics by exploring both algebraic constraints (the Kruppa equations [6, 16, 19]) and geometric ones (Dual Image of Absolute Conic). As an alternative, *stratified* methods upgrade a projective 3D reconstruction to Euclidean [4, 26]. Recently, end-to-end Deep Learning approaches [3, 18, 21, 28] have been introduced to infer the focal length and the radial distortion of a camera from a single image.

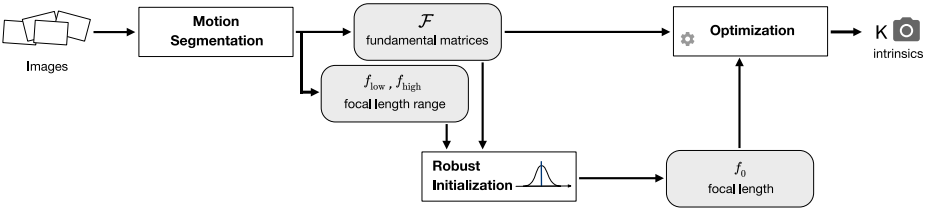


Figure 1: A graphical overview of the proposed multi-body self-calibration pipeline. The main steps of the algorithm are depicted (white boxes), as well as their inputs and outputs (grey boxes).

Self-calibration usually requires solving a system of non-linear polynomials. At a high level, we identify two streams of work. One line of research is rooted in numerical algebraic geometry and leverages homotopy continuation [43] or Gröbner basis. However, these methods suffer from high computational costs and are very sensitive to noise. A recent work [25] employs a consensus maximization [29] coupled with Branch-and-Bound to robustly sample algebraic varieties defined by the Kruppa equations and the Modulus constraint.

The second line of research tackles self-calibration as the optimization of a non-linear cost function [20, 23] derived from the rigidity of the scene. Our solution belongs to this category. Several strategies have been proposed to deal with noise or outliers from incorrect image matches and to limit the impact of degenerate motions, which make self-calibration ill-conditioned. For instance, it has been proposed to weigh the Kruppa constraints to reduce the influence of outlying fundamentals [5, 19, 23]. In [27], the authors discard image pairs that give rise to critical motions. Robustness is also pursued in [10], where interval analysis is introduced, and in [30], where a randomized multi-start approach is presented. A good initialization is crucial for the success of these approaches.

All the aforementioned works assume the 3D scene to be static. Self-calibration in *dynamic* scenes has not been explored significantly, as, in general, multi-body SfM [24] has not reached the maturity of its single-body counterpart. Practical 3D reconstruction pipelines that operate under this realistic assumption and provide accurate results are still missing, despite attempts in this direction, *e.g.*, [14]. Most recently, the problem of multi-body self-calibration has never been addressed. While more challenging, the multi-body scenario offers advantages that have been somewhat overlooked in the literature. A notable exception is [7], where Fitzgibbon and Zisserman show that the multi-body analysis allows for a Euclidean reconstruction in cases that are under-constrained for a static scene. However, the analysis is mainly theoretical and does not address robustness, as we did in this work.

# 3 Method

We propose a self-calibration method that robustly estimates the intrinsic parameters of a camera from a set of images depicting a scene containing multiple independently moving *rigid* bodies. Our method is structured in three steps (Fig. 1) that work together to achieve robustness towards noise and outliers:

***i*) Motion Segmentation** (Sec. 3.1) leverages robust multi-model fitting to segment the rigidly moving objects in image pairs. Given the assumption that the intrinsics are constant, we use the 6-point algorithm [34] to compute fundamental matrices $\mathcal{F}$ describing the rigid motions. In contrast to the classical 7-point algorithm, this allows us to fully exploit the rigidity constraints to derive an interval of tentative focal lengths. Remarkably, by reducing

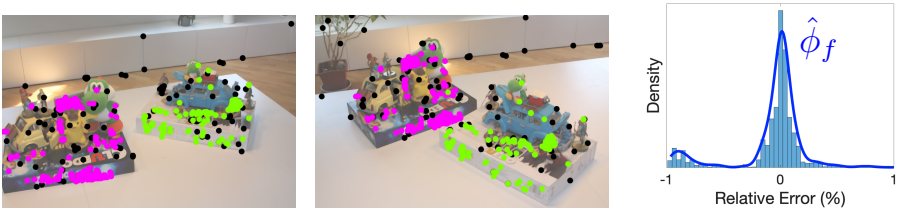


Figure 2: Left and middle: Motion Segmentation for a sample image pair. Motions are color coded, outliers are marked as black points. Right: KDE distribution $\hat{\phi}_f$ of focal lengths from Kernel Voting.

the minimal sample set to 6 matches, also robustness is improved.
***ii*) Robust initialization** (Sec. 3.2) derives epipolar constraints from $\mathcal{F}$ at step (i) to compute a distribution of focal length $\phi_{f_0}$. We exploit the redundancy of these constraints and the interval of focal lengths, to prune out bad initializations using Kernel Voting.
***iii)* Robust Optimization** (Sec. 3.3) refines the camera parameters. The optimization achieves robustness by sampling initial guesses from $\phi_{f_0}$ at step (ii) and exploiting subsets of $\mathcal{F}$ from step (i) to define the cost function. Specifically, we decouple the computation of the focal length from the optical center to achieve faster convergence and numerical stability. In addition, our method is parallelized to improve efficiency.

## 3.1 Motion Segmentation

In the first step, we estimate: *i*) a set of pairwise fundamental matrices $\mathcal{F} = \{\mathsf{F}_1, \dots, \mathsf{F}_k\}$ describing 3D motions in the scene, *ii*) an interval $[f_{\text{low}}, f_{\text{high}}] \subset \mathbb{R}$ of tentative focal lengths. Our *Motion Segmentation* leverages T-linkage [22], a multi-model fitting framework that exploits preference analysis to cluster correspondences according to their rigid motion.

As usual in SfM, we extract keypoints and establish matches $\mathsf{M}_{ij}$ between image pairs $(I_i, I_j)$. T-linkage takes the matches $\mathsf{M}_{ij}$ as input and automatically segments the motions by fitting multiple fundamental matrices. Specifically, T-linkage clusters the preference of data w.r.t. a pool of provisional models attained with random sampling of minimal sample sets (MSS). As opposed to classical segmentation methods, we exploit the constant parameters $\mathsf{K}$ across images and apply the 6-point algorithm [34] to sample fundamental matrices and their corresponding focal lengths. The quality of a fundamental matrix estimated with the 6-point algorithm correlates to the ratio of its two largest singular values [40], as a low ratio indicates critical motions. Thus, we discard models for which this ratio is below 0.9. The lower and upper limits $f_{\text{low}}, f_{\text{high}}$ are defined as the lower and upper quantiles respectively of the set of the focal lengths of the attained clusters. Finally, the fundamental matrices are fitted on the inliers of the clusters having enough matches (12 in our experiments) and refined by minimizing the Sampson distance. A clustering example is reported in Fig. 2, where matches are clustered according to the rigid motions in the scene.

## 3.2 Robust Initialization

In this step, we compute a robust estimate $f_0$ of the focal length and its distribution $\phi_{f_0}$. This information serves as an initial guess for the following optimization step. As in [32, 37], we initialize the focal length by assuming unit aspect ratio, i.e., $f_0 = f_x = f_y$ in the calibration matrix $\mathsf{K}$, and separately refine $f_x, f_y$ in the subsequent optimization step (Sec. 3.3). Specifically, we initialize $f_0$ by exploiting a *hypothesize-and-verify* framework (Alg. 1). We test a

**Algorithm 1:** Robust initialization

**Input :** Fundamental matrices $\mathcal{F} = \{\mathsf{F}_k\}$, $f_{\text{low}}$, $f_{\text{high}}$, image width $w$ and height $h$, maximum opening angles $\mathcal{A} = \{\alpha_j\}$

**Output:** Distribution $\phi_{f_0} \sim \mathcal{N}(f_0, \sigma_{f_0}^2)$ modelling the focal length

1 $\mathcal{J} \leftarrow \varnothing$;
2 **foreach** $\alpha_j \in \mathcal{A}$ **do**
3  Compute $f_j$ from $\alpha_j$ using Eq. (2) ; /* Hypothesis step */
4  **if** $f_j \in (f_{low}, f_{high})$ **then**
5  **foreach** $\mathsf{F}_k \in \mathcal{F}$ **do**
6  Compute semi-calibrated matrix $\mathsf{G}_{j,k}$ from $f_j$ and $\mathsf{F}_k$ via Eq. (3) ;
7  $\mathsf{G} = \mathsf{G}/\|\mathsf{G}\|_F$ ;
8  Compute Kruppa Equations via SVD of $\mathsf{G}_{j,k}$ ;
9  $\hat{f}_{j,k} \leftarrow$ solution of Kruppa quadratic equation closest to $f_j$ ;
10  **if** $\hat{f}_{j,k}$ *solves Kruppa two linear equations* **then**
11  insert $\hat{f}_{j,k}$ in $\mathcal{J}$ ; /* Verification step */
12  **end**
13  **end**
14  **end**
15 **end**
16 $\hat{\phi}_f \leftarrow \mathsf{KDE}(\mathcal{J})$ ; /* Kernel Voting */
17 $f_0 \leftarrow$ maximal peak in $\hat{\phi}_f$ ; /* mean */
18 $\sigma_{f_0} \leftarrow \mathsf{Q}_n(\mathcal{J}, \mu)$ ; /* std */
19 $\phi_{f_0} \sim \mathcal{N}(f_0, \sigma_{f_0}^2)$ ;

set of feasible focal lengths within the limits $f_{\text{low}}$ and $f_{\text{high}}$ and verify these using the rigidity constraints in the fundamental matrices $\mathsf{F}_k \in \mathcal{F}$ from Motion Segmentation.

Following [32], in the *hypothesis* step, we parameterize the space of focal lengths in terms of the maximal opening angles $\mathcal{A} = \{\alpha_1, \ldots, \alpha_m\}$ of the camera. The motivation is that, unlike focal lengths, opening angles are independent of image resolution. In practice, we test $m = 100$ opening angles from 0.5 to 99.5 sampled with step 1.0. Each $\alpha_j \in \mathcal{A}$ (line 2) is converted into a focal value $f_j$ (line 3) using the image width $w$ and height $h$ via:

$$f_j = \frac{\max(w,h)}{2 \times \tan(\alpha_j/2)}, \quad \alpha_j = 0.5, \ldots, 99.5 \tag{2}$$

The range of opening angles is further restricted to those $\alpha_j \in \mathcal{A}$ that, when converted to $f_j$, fall within the interval $[f_{\text{low}}, f_{\text{high}}]$ obtained in Sec. 3.1 (line 4).

The *verification* step builds upon [37], which was conceived to estimate a focal length by the Kruppa equations. For each fundamental $\mathsf{F}_k \in \mathcal{F}$ (line 5), a "semi-calibrated fundamental" matrix $\mathsf{G}_{j,k}$ is derived by inverting a tentative calibration matrix built using $f_j$ (line 6):

$$\mathsf{G}_{j,k} = \begin{pmatrix} f_j & 0 & 0 \\ 0 & f_j & 0 \\ w/2 & h/2 & 1 \end{pmatrix} \mathsf{F}_k \begin{pmatrix} f_j & 0 & w/2 \\ 0 & f_j & h/2 \\ 0 & 0 & 1 \end{pmatrix} \tag{3}$$

Intuitively, $\mathsf{G}_{j,k}$ would coincide with the essential matrix if the unknown focal length of the camera were equal to the tentative $f_j$ and the principal point were located at the image center $(\frac{w}{2}, \frac{h}{2})$. $\mathsf{G}_{j,k}$ is scaled to unit Frobenius norm (line 7) to improve numerical stability and is decomposed via SVD to derive the simplified Kruppa equations reported in [37] (line 8).

The Kruppa equations consist of one quadratic and two linear equations in the unknown $f_{j,k}^2$. Firstly, we solve the quadratic equation, and keep the root $\hat{f}_{j,k}$ closest to the initial guess

**Algorithm 2:** Optimization

**Input :** A set of fundamental matrices $\mathcal{F} = \{\mathsf{F}_i\}$, a distribution $\phi_{f_0}$ of the initial guess of $f_0$, the width $w$ and height $h$ of the images

**Output:** The intrinsic parameters of the camera $f_x, f_y, u, v$

1 $\mathcal{J} = \varnothing$ ;
2 **for** $i \leftarrow 1$ **to** *max-iters* **do**
3 $\quad f_0 \leftarrow$ sample $\phi_{f_0}$;
4 $\quad u_0 \leftarrow$ sample $\mathcal{N}(\frac{w}{2}, \frac{w^2}{36})$, $v_0 \leftarrow$ sample $\mathcal{N}(\frac{h}{2}, \frac{h^2}{36})$;
5 $\quad \mathcal{F}_{\text{min}} \leftarrow$ random-minimal-sample($\mathcal{F}$);
6 $\quad \tilde{f} \leftarrow$ levenberg-marquardt($\mathcal{F}_{\text{min}}, f_0, u_0, v_0$);
7 $\quad$ insert $\tilde{f}$ in $\mathcal{J}$;
8 **end**
9 $f \leftarrow$ maximal peak in $\mathsf{KDE}(\mathcal{J})$;
`/* Joint refinement of the intrinsics */`
10 **for** $\mathsf{F}_i \in \mathcal{F}$ **do**
11 $\quad \mathsf{R}[i] \leftarrow$ mendonça-cipolla-residual($\mathsf{F}_i, \frac{w}{2}, \frac{h}{2}$) ;
12 **end**
13 $\mathcal{F}_{\text{optim}} \leftarrow$ set of 3 $\mathsf{F}_i \in \mathcal{F}$ with lowest residual $\mathsf{R}[i]$;
14 $f_x, f_y, u, v \leftarrow$ levenberg-marquardt($\mathcal{F}_{\text{optim}}, f, \frac{w}{2}, \frac{h}{2}$);

$f_j$ (line 9) if it also satisfies the Kruppa linear equations (line 10). We consider only the real part of $\hat{f}_{j,k}$ and discard solutions with an imaginary part greater than a tolerance at $10^{-6}$.

As opposed to the original Sturm's method, we do not require prior information about the camera, such as a fixed guess of the focal length. This is because our voting scheme explores the space of feasible focal lengths identified by the motion segmentation. In this way, we obtain several tentative values $\{\hat{f}_{j,k}\}$, one for each fundamental matrix $\mathsf{F}_k$ and hypothesized focal length $f_j$. Wrong guesses $f_j$ or small perturbations on $\mathsf{F}_k$ impact the Kruppa equations and yield noisy or severely wrong focals $\hat{f}_{j,k}$. However, we observe a general agreement between $\hat{f}_{j,k}$ that concentrates around the genuine solution. This is shown in Fig. 2 (right), where the frequency of $\hat{f}_{j,k}$ reaches its peak as the relative error w.r.t. the genuine solution tends to zero.

As in [17], we use a Kernel Voting scheme to identify the best $f_0$, and apply a Kernel Density Estimator (KDE) to derive the distribution $\hat{\phi}_f(x)$ of $\hat{f}_{j,k}$ (line 16) using a Gaussian kernel $\mathcal{K}$ with bandwidth $h$ at 5% of the median of $\mathcal{J} = \{f_{j,k}\}$. The distribution is given by:

$$\hat{\phi}_f(x) = \sum_{\hat{f}_{j,k} \in \mathcal{J}} \frac{\mathcal{K}(\hat{f}_{j,k} - x)}{h}. \tag{4}$$

We set the highest peak of $\hat{\phi}_f$ as the estimate of the focal length $f_0$ (line 17). Since guesses of the focal length concentrate around the genuine solution with few exceptions, we expect $\hat{\phi}_f$ to be unimodal. As depicted in Fig. 2 (right), we fit a Gaussian distribution $\phi_{f_0}$ with mean $f_0$ and standard deviation $\sigma_{f_0}$ given by the $\mathsf{Q}_n$ scale estimator [31] (line 18), and use this in the following step of our algorithm.

## 3.3 Robust Optimization

In this step, we refine all internal parameters $f_x, f_y, u, v$ by minimizing the Mendoça-Cipolla cost function [23] derived from the fundamental matrices in $\mathcal{F}$. In practice, since errors in

$\mathcal{F}$ are frequent due to image noise, outlying matches, and critical motions [38], we design our optimization to be robust. We achieve robustness in two main steps described in Alg. 2. The first step (lines 1-9) assumes $f_0 = f_x = f_y$ and builds upon randomized multi-start [30] to further improve the estimated focal length $f$. The idea is to sample initial points from the distribution $\phi_{f_0}$ of the guessed focal length to avoid local minima and numerical instabilities. At each iteration, a random subset of the fundamental matrices $\mathcal{F}$ is involved in the cost function to reduce the effect of outliers. In the second step (lines 10 - 14), a further refinement is performed to jointly estimate the focal lengths $f_x, f_y$ and the principal point $(u, v)$.

**Randomized multi-start optimization** is structured as a loop: at each iteration, we sample a focal length $f_0$ from $\phi_{f_0}$ (line 3) and a principal point $u_0$ and $v_0$ from $\phi_{u_o}$ and $\phi_{v_0}$ respectively (line 4). We assume that the coordinates of the principal point follow a Gaussian distribution $\phi_{u_0} \sim \mathcal{N}(\frac{w}{2}, (\frac{w}{6})^2)$ centered in the image center. We set $\sigma_{u_0}$ so that $3\sigma_{u_0} = \frac{w}{2}$ and 99% of sampled values fall within $[0, w]$. Similarly, $\phi_{v_0} \sim \mathcal{N}(\frac{h}{2}, (\frac{h}{6})^2)$. At each iteration, we sample a minimal sample set (MSS) $\mathcal{F}_{\text{min}}$ of three fundamentals from $\mathcal{F}$ (line 5), which allow to uniquely compute $\mathsf{K}$ in case the intrinsic parameters are constant [27]. Similarly to [30], the number of iterations of the multi-start is set so that each possible MSS of $\mathcal{F}$ is chosen with a 95% probability. Note that each iteration is independent and thus parallelized.

The optimization scheme is based on Levenberg-Marquardt and takes as input the initial guess of the intrinsics and $\mathcal{F}_{\text{min}}$ to minimize the Mendoça-Cipolla cost function (line 6). This cost encourages the essential matrix $\mathsf{E}_k = \mathsf{K}_k^\top \mathsf{F}_k \mathsf{K}_k$ derived from the $k$-th fundamental matrix $\mathsf{F}_k$ to have identical non-zero singular values $\sigma_{(1,k)}$ and $\sigma_{(2,k)}$:

$$\mathcal{C}(\mathsf{K}_k, i = 1, \dots, n) = \min \frac{1}{|\mathcal{F}|} \sum_{k=1}^{|\mathcal{F}|} \frac{\sigma_{(1,k)} - \sigma_{(2,k)}}{\sigma_{(2,k)}}. \tag{5}$$

Since cost function (5) shows excellent convergence, as demonstrated in [9], we limit the iterations of the Levenberg-Marquardt to 100 to achieve a balance between accuracy and running time. This optimization returns a refined focal length $\tilde{f}$, which is recorded in a set $\mathcal{J}$ (line 7). The most likely focal length $f$ is then estimated from $\mathcal{J}$ by fitting a KDE and finding its maximum (line 9), similarly to Sec. 3.2. The bandwidth for the KDE is set to 5% of the median of the input focal lengths.

**Joint refinement** of the focal lengths $f_x, f_y$ and the principal point $(u, v)$ allows full camera calibration. For all $\mathsf{F}_i \in \mathcal{F}$, we evaluate the Mendonça-Cipolla cost function with $f$ from the KDE as the focal length and the image center as the principal point (lines 10 - 11). The three fundamental matrices with the lowest residuals are chosen to derive the cost function of the final Levenberg-Marquardt routine (line 13), which is initialized starting from focal length $f$ and principal point $(\frac{w}{2}, \frac{h}{2})$ (line 14). This Levenberg-Marquardt routine is also limited to 10 iterations, as the intrinsics are assumed to be close to the optimum at this stage.

# 4 Experiments

We demonstrate the effectiveness of the proposed self-calibration algorithm in two sets of experiments. Sec. 4.1 focuses on evaluating our robust initialization step without considering the subsequent optimization. Isolating the performance assessment to the initialization is important for two reasons: *i*) the optimization may compensate for poor initializations, *ii*) since the optimization step is interchangeable, this evaluation is insightful for other optimization schemes that rely on a robust initialization. Then, in Sec. 4.2, we evaluate our

complete pipeline. Specifically, we show that our approach is comparable to a state-of-the-art self-calibration method [25] on static scenes and improves performance on dynamic ones. Remarks about the impact of critical motions and computational times conclude the section.

**Datasets** For *static* scenes (moving object $m = 1$) , we consider the popular SfM benchmarks in [36]. For *dynamic* scenes (moving objects $m > 1$), we consider the *traffic* video sequences in the `Hopkins155` dataset [42] characterized by 2 or 3 motions. For better accuracy, we refined the intrinsics using COLMAP [33] by providing manually annotated masks to segment the motions. We extract images from videos with a 5-frame interval to have a sufficient baseline between consecutive views. Given the scarcity of multi-body *calibrated* datasets, we also introduce a new dataset, called `Amiibo dataset`, which comprises three static scene sequences and three dynamic scenes with either two or three independent motions. The dataset is available at [1].

**Performance metrics** to assess the calibration performance are the percentage error on the estimated focals $(\tilde{f}_x, \tilde{f}_y)$ w.r.t. the real $(f_x, f_y)$ and the estimated principal point $(\tilde{u}, \tilde{v})$ w.r.t. the real $(u, v)$:

$$\text{err}_\text{f} = \frac{1}{2}\left(\left|\frac{\tilde{f}_x - f_x}{f_x}\right| + \left|\frac{\tilde{f}_y - f_y}{f_y}\right|\right) \times 100\,, \quad \text{err}_\text{uv} = \frac{1}{2}\left(\left|\frac{\tilde{u} - u}{u}\right| + \left|\frac{\tilde{v} - v}{v}\right|\right) \times 100\,. \tag{6}$$

## 4.1 Evaluation of Robust Initialization

We evaluate the accuracy $\text{err}_\text{f}$ and the precision $\text{std}_\text{f}$ of our robust initialization step on real datasets against two baselines: *i*) *vanilla*, the average of the focal lengths computed from each fundamental matrix using Sturm's method [37], *ii*) *vanilla with Kernel Voting*, where the *vanilla* method is equipped with Kernel Voting to select the most likely focal length.

Fig. 3 reports the performance on static (top) and dynamic scenes (middle). Our initialization outperforms the alternatives in all tests, both in precision and accuracy. We show in Fig. 3 (bottom) three examples of the estimated density $\phi_f$ for a static (S1-Amiibo) and two dynamic (M1-Amiibo and cars4) scenes. In the dynamic case, the distribution exhibits multiple peaks due to motions close to degenerate or with small support resulting in poor fundamental matrices. Nonetheless, our method also identifies the correct peak in these challenging cases. We also perform synthetic experiments (in the Supplementary Material) that confirm the advantages of our approach on less-noisy distributions as well.

## 4.2 Evaluation of the complete pipeline

We compare our self-calibration against the non-robust baseline [23] of Mendonça-Cipolla (M&C) and a state-of-the-art self-calibration based on Consensus Maximization and Branch-and-Bound (BnB) [25]. All methods are evaluated on both static and dynamic scenes. For *dynamic* scenes ($m > 1$), we ran experiments on two configurations: *i*) *Single-body*, where, as usually done, we consider only the dominant motion, *ii*) *Multi-body*, where we provide the competing methods, which are natively limited to static scenes, all the epipolar constraints recovered by our motion segmentation, enabling a fair comparison.

**Static scenes:** Tab. 1 (top-left) reports experiments on *static* scenes. In both accuracy metrics $\text{err}_\text{f}$ and $\text{err}_\text{uv}$, our method outperforms *M&C* and produces comparable results to *BnB*, outperforming it in 4 of 6 datasets for $\text{err}_\text{f}$ and in 3 of 6 datasets for $\text{err}_\text{uv}$.

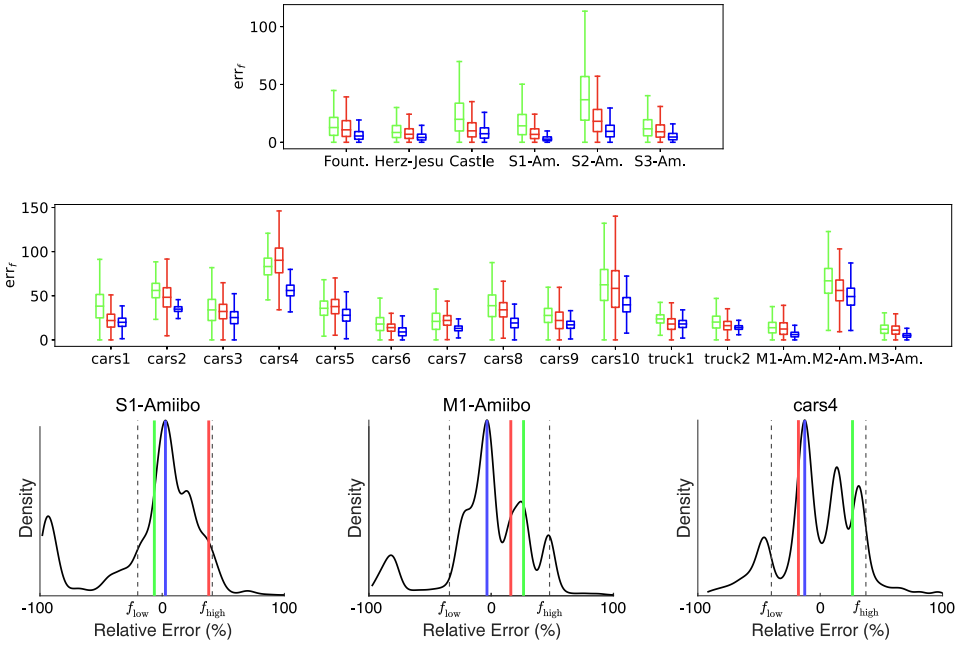


Figure 3: Our method vs. alternative initialization strategies. Vanilla in green, Vanilla with Kernel Voting in red, and ours in blue. *Top*: box plot of $\text{err}_\text{f}$ on static datasets. *Middle*: box plot of $\text{err}_\text{f}$ on dynamic datasets. *Bottom*: Kernel density function $\hat{\phi}_f$ (black) for the *static* S1-Amiibo, the *multi-body* M1-Amiibo and cars4 datasets. $x$-axes are scaled to represent the relative error (%) $\text{err}_\text{f}$ of the focal length. Lower and upper limits $f_\text{low}, f_\text{high}$ are reported. Our method attains a focal length error of 22px on S1-Amiibo, 48px on M1-Amiibo, and 126px on cars4.

**Dynamic scenes:** Tab. 1 (bottom) reports the results on dynamic scenes for single and multi-body configurations. The advantages of exploiting all motions in the scene are apparent, as neither *Ours* or *BnB* in their Single-body configuration match the accuracy in $\text{err}_\text{f}$ and $\text{err}_\text{uv}$ attained by *Ours* in its Multi-body configuration. In addition, results show that our method in Multi-body shows the most significant uplift in accuracy w.r.t. its Single-body counterpart and is consistently the most accurate in this configuration. Interestingly, when Mendonça-Cipolla is provided with all the motions, it shows a regression in performance due to its sensitivity to outlying fundamental matrices. This demonstrates that while multiple motions provide useful constraints, exploiting them is a hard problem that may introduce more noise and outliers, requiring an ad-hoc robust method. Synthetic experiments in the Supplementary Material also confirm these results.

Multi-body results on the Amiibo dataset are generally better than those observed on the `Hopkins155` sequences. This can be ascribed to the different characteristics of the two datasets. In the Amiibo sequences: i) well-textured figurines provide many keypoints, ii) viewpoint changes generally yield well-conditioned motions, iii) our method's initial focal length guess is closer to the ground truth due to the camera having an aspect ratio close to 1, iv) high-resolution images help to achieve more accurate results.

**Critical Motions** [33] hinder the self-calibration of a camera with a fixed focal length in configurations where either: *i*) the optical axes are parallel to each other or *ii*) the camera center lies on a circumference centered at the intersection of the optical axes [39]. Tests on

| | | | Single-body on *static* datasets | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Ours | | BnB [25] | | M&C [25] | | Ours | | BnB [25] | | M&C [25] | |
| Dataset | $m$ | $\lvert\mathcal{I}\rvert$ | $\text{err}_\text{f}$ | $\text{err}_\text{uv}$ | $\text{err}_\text{f}$ | $\text{err}_\text{uv}$ | $\text{err}_\text{f}$ | $\text{err}_\text{uv}$ | $\text{err}_\text{f}$ | $\text{err}_\text{uv}$ | $\text{err}_\text{f}$ | $\text{err}_\text{uv}$ | $\text{err}_\text{f}$ | $\text{err}_\text{uv}$ |
| Fountain | 1 | 25 | 0.87 | 1.12 | **0.84** | **1.01** | 12.47 | 11.02 | - | - | - | - | - | - |
| Herz-Jesu | 1 | 7 | **1.04** | **1.21** | 1.10 | 1.28 | 18.09 | 11.54 | - | - | - | - | - | - |
| Castle | 1 | 20 | **0.91** | 1.18 | 0.94 | **1.09** | 12.31 | 11.46 | - | - | - | - | - | - |
| S1-Amiibo | 1 | 15 | **0.66** | **3.98** | 0.68 | **3.98** | 15.12 | 8.92 | - | - | - | - | - | - |
| S2-Amiibo | 1 | 12 | 2.87 | 6.02 | **2.74** | **6.00** | 18.67 | 18.93 | - | - | - | - | - | - |
| S3-Amiibo | 1 | 16 | **0.49** | **3.89** | 0.51 | 3.92 | 7.48 | 12.02 | - | - | - | - | - | - |
| | | | Single-body on *dynamic* datasets | | | | | | Multi-body on *dynamic* datasets | | | | | |
| cars1 | 2 | 4 | 17.41 | 1.28 | 16.85 | 1.21 | 67.81 | 74.31 | **12.45** | **1.21** | 16.48 | 1.21 | 69.01 | 73.92 |
| cars2 | 2 | 6 | 33.48 | 3.03 | 34.28 | 2.97 | 81.47 | 79.31 | **15.18** | **3.14** | 19.02 | 2.91 | 88.13 | 77.16 |
| cars3 | 3 | 4 | 21.76 | 6.21 | 21.01 | 6.49 | 156.87 | 74.01 | **7.82** | **2.54** | 9.10 | 2.86 | 184.09 | 102.42 |
| cars4 | 2 | 11 | 40.62 | 6.81 | 38.29 | 7.02 | 62.01 | 48.99 | **18.36** | **6.41** | 20.48 | 7.14 | 75.88 | 38.26 |
| cars5 | 3 | 7 | 23.49 | 4.21 | 22.01 | 4.01 | 76.01 | 58.12 | **11.78** | **4.01** | 15.71 | **4.01** | 72.45 | 61.34 |
| cars6 | 2 | 6 | 7.38 | 1.83 | 7.38 | 2.01 | 41.91 | 38.74 | **5.30** | **1.89** | **5.30** | 1.91 | 67.30 | 59.34 |
| cars7 | 2 | 6 | 8.79 | 1.92 | 8.67 | 1.85 | 38.54 | 41.27 | **6.51** | **1.85** | 8.46 | 1.98 | 45.87 | 31.28 |
| cars8 | 2 | 6 | 19.28 | 2.87 | 21.01 | 2.65 | 102.48 | 87.01 | **9.01** | **2.81** | 12.49 | 2.98 | 98.06 | 79.61 |
| cars9 | 3 | 12 | 15.99 | 2.68 | 14.87 | 2.41 | 27.61 | 19.47 | **8.12** | **2.42** | 9.34 | 2.67 | 38.91 | 65.81 |
| cars10 | 3 | 6 | 31.62 | 12.84 | 31.91 | 12.62 | 91.45 | 78.34 | **15.82** | **7.42** | 16.53 | 8.01 | 89.45 | 62.01 |
| truck1 | 2 | 6 | 14.78 | 6.84 | 13.98 | 7.28 | 68.12 | 72.13 | **4.87** | **1.87** | 6.92 | 2.32 | 72.58 | 45.61 |
| truck2 | 2 | 4 | 13.32 | 7.21 | 13.38 | 6.58 | 68.58 | 71.20 | **4.24** | **2.31** | 7.18 | 3.01 | 74.69 | 89.30 |
| M1-Amiibo | 2 | 10 | 3.29 | 4.28 | 3.21 | 4.22 | 43.29 | 51.32 | **1.27** | **3.86** | 2.84 | 4.04 | 62.88 | 50.14 |
| M2-Amiibo | 3 | 4 | 39.28 | 36.71 | 39.85 | 37.21 | 89.41 | 92.01 | **4.97** | **3.71** | 7.21 | 4.28 | 93.12 | 92.86 |
| M3-Amiibo | 2 | 4 | 3.20 | 5.89 | 3.28 | 5.71 | 41.28 | 39.62 | **1.37** | **3.96** | 2.19 | 4.27 | 53.76 | 46.12 |

Table 1: Self-calibration comparison on real data in Single-body and Multi-body. Number of rigid motions $m$ in the dataset, total number of images processed $|\mathcal{I}|$, and relative errors $\text{err}_\text{f}$ and $\text{err}_\text{uv}$. Tests were repeated 10 times, the average error is reported. Considering $f_x, f_y$ independently, relative errors $\text{err}_{f_x}$ and $\text{err}_{f_y}$ are equally distributed after accounting for the aspect ratio. This also applies to $\text{err}_u, \text{err}_v$. *Top*: *static* scenes. *Bottom*: *dynamic* scenes.

the KITTI dataset [11] show that when the forward motion is dominant (case *i*), our method, BnB [25] and COLMAP all fail to calibrate, whereas calibration is achieved when motions are varied. Refer to the Supplementary Material for calibration results on critical motions.

**Running time** of our method is dominated by the Motion Segmentation, though this also applies to BnB and M&C when used in their multi-body configuration. Without considering the segmentation, our method achieves running times in the same order of magnitude as BnB (60% slower on average). At the same time, the non-robust M&C is faster by two orders of magnitude. A comparison with exact timings is available in the supplementary material.

# 5 Conclusions

We have extended self-calibration of a single moving camera with constant but unknown intrinsic parameters to the case of dynamic scenes, *i.e.*, 3D scenes composed of multiple bodies moving rigidly. Results confirm that having constraints from multiple motions is beneficial, and demonstrate that our multi-body self-calibration solution can successfully address robustness issues exacerbated in dynamic setups. Future work will focus on applying the multi-body analysis to a wider extent, tackling the case of non-constant intrinsics and integrating it into SfM pipelines. We hope our solution can be a step forward for the development of practical techniques to address the challenges of multi-body SfM.

# References


[1] https://github.com/andreadalcin/MultiBodySelfCalibration.

[2] Sameer Agarwal, Noah Snavely, Ian Simon, Steven M. Seitz, and Richard Szeliski. Building rome in a day. In *IEEE International Conference on Computer Vision*, 2009.

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

[4] Manmohan Chandraker, Sameer Agarwal, David Kriegman, and Serge Belongie. Globally optimal algorithms for stratified autocalibration. *International Journal of Computer Vision*, 90(2):236–254, 2010.

[5] Gabriella Csurka, Cyril Zeller, Zhengyou Zhang, and Olivier D Faugeras. Characterizing the uncertainty of the fundamental matrix. *Computer Vision and Image Understanding*, 68(1):18–36, 1997.

[6] O.D. Faugeras, Q.T. Luong, and S.J. Maybank. Camera self-calibration: Theory and experiments. In *Proceedings of the European Conference on Computer Vision*, pages 321–334, 1992.

[7] Andrew W Fitzgibbon and Andrew Zisserman. Multibody structure and motion: 3-d reconstruction of independently moving objects. In *Proceedings of the European Conference on Computer Vision*, pages 891–906. Springer, 2000.

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

[9] A. Fusiello. A new autocalibration algorithm: Experimental evaluation. In *Computer Analysis of Images and Patterns*, volume 2124, pages 717–724. Springer Berlin Heidelberg, 2001.

[10] A. Fusiello, A. Benedetti, M. Farenzena, and A. Busti. Globally convergent autocalibration using interval analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(12):1633–1638, December 2004.

[11] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11): 1231–1237, 2013.

[12] R. Gherardi, R. Toldo, V. Garro, and A. Fusiello. Automatic camera orientation and structure recovery with samantha. In *3D Virtual Reconstruction and Visualization of Complex Architectures*, pages 38–5, 2011.

[13] Anders Heyden and Kalle Astrom. Euclidean reconstruction from constant intrinsic parameters. In *Proceedings of the International Conference on Pattern Recognition*, volume 1, pages 339–343. IEEE, 1996.

[14] Petr Hruby and Tomas Pajdla. Reconstructing small 3d objects in front of a textured background. *arXiv preprint arXiv:2105.11352*, 2021.

[15] Heinly Jared, Johannes L Schonberger, Enrique Dunn, and Jan-Michael Frahm. Reconstructing the world in six days. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[16] E. Kruppa. Zur Ermittlung eines Objektes aus zwei Perspektiven mit innerer Orientierung. *Sitz.-Ber. Akad. Wiss., Wien, math. naturw. Kl., Abt. IIa.*, 122:1939–1948, 1913.

[17] Hongdong Li. A simple solution to the six-point two-view focal-length problem. In *Proceedings of the European Conference on Computer Vision*, pages 200–213. Springer, 2006.

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

[19] Manolis I.A. Lourakis and Rachid Deriche. Camera self-calibration using the singular value decomposition of the fundamental matrix. In *Proceedings of the Asian Conference on Computer Vision*, volume I, pages 403–408, January 2000.

[20] Q.-T. Luong and O. Faugeras. Self-calibration of a moving camera from point correspondences and fundamental matrices. *International Journal of Computer Vision*, 22 (3):261–289, 1997.

[21] Manuel López, Roger Marí, Pau Gargallo, Yubin Kuang, Javier Gonzalez-Jimenez, and Gloria Haro. Deep single image camera calibration with radial distortion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11809–11817, 2019.

[22] 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*, pages 3954–3961, June 2014.

[23] P.R.S. Mendonça and R. Cipolla. A simple technique for self-calibration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages I:500–505, 1999.

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

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

[26] Marc Pollefeys and Luc Van Gool. A stratified approach to metric self-calibration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 407–412. IEEE, 1997.

[27] Marc Pollefeys, Reinhard Koch, and Luc Van Gool. Self-calibration and metric reconstruction inspite of varying and unknown intrinsic camera parameters. *International Journal of Computer Vision*, 32(1):7–25, 1999.

[28] Georgy Ponimatkin, Yann Labbé, Bryan Russell, Mathieu Aubry, and Josef Sivic. Focal length and object pose estimation via render and compare. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3825–3834, 2022.

[29] Thomas Probst, Danda Pani Paudel, Ajad Chhatkuli, and Luc Van Gool. Convex relaxations for consensus and non-minimal problems in 3d vision. In *Proceedings of the International Conference on Computer Vision*, pages 10233–10242, 2019.

[30] Houman Rastgar, Eric Dubois, and Liang Zhang. Random sampling nonlinear optimization for camera self-calibration with modeling of intrinsic parameter space. In *Advances in Visual Computing*, pages 189–198, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[31] Peter J. Rousseeuw and Christophe Croux. Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88(424):1273–1283, 1993.

[32] Torsten Sattler, Chris Sweeney, and Marc Pollefeys. On sampling focal length values to solve the absolute pose problem. In *Proceedings of the European Conference on Computer Vision*, pages 828–843. Springer, 2014.

[33] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[34] Henrik Stewenius, David Nister, Fredrik Kahl, and Frederik Schaffalitzky. A minimal solution for relative pose with unknown focal length. In *cvpr*, pages 789–794, 2005.

[35] Henrik Stewénius, David Nistér, Fredrik Kahl, and Frederik Schaffalitzky. A minimal solution for relative pose with unknown focal length. *Image and Vision Computing*, 26 (7):871–877, 2008.

[36] C. Strecha, W. Von Hansen, L. Van Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[37] P. Sturm. On focal length calibration from two views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume II, pages 145–150, 2001.

[38] Peter Sturm. Critical motion sequences for the self-calibration of cameras and stereo systems with variable focal length. *Image and Vision Computing*, 20(5-6):415–426, 2002.

[39] Peter Sturm, ZL Cheng, Peter CY Chen, and Aun Neow Poo. Focal length calibration from two views: method and analysis of singular cases. *Computer Vision and Image Understanding*, 99(1):58–95, 2005.

[40] Akihiko Torii, Zuzana Kukelova, Martin Bujnak, and Tomas Pajdla. The six point algorithm revisited. In *Computer Vision – ACCV 2010 Workshops*, volume 6469, pages 184–193. Springer Berlin Heidelberg, 2011.

[41] B. Triggs. Autocalibration and the absolute quadric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 609–614, 1997.

[42] 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*, pages 1–8. IEEE, 2007.

[43] Jan Verschelde. Algorithm 795: Phcpack: A general-purpose solver for polynomial systems by homotopy continuation. *ACM Transactions on Mathematical Software (TOMS)*, 25(2):251–276, 1999.

[44] Siyu Zhu, Runze Zhang, Lei Zhou, Tianwei Shen, Tian Fang, Ping Tan, and Long Quan. Very large-scale global sfm by distributed motion averaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4568–4577, 2018.