Probabilistic Structure from Motion with Objects (PSfMO)

Paul Gay, Vaibhav Bansal, Cosimo Rubino, Alessio Del Bue
Visual Geometry and Modelling (VGM) Lab
Istituto Italiano di Tecnologia (IIT)
Via Morego 30, 16163 Genova, Italy
{paul.gay, vaibhav.bansal, cosimo.rubino, alessio.delbue}@iit.it

Abstract

This paper proposes a probabilistic approach to recover affine camera calibration and objects position/occupancy from multi-view images using solely the information from image detections. We show that remarkable object localisation and volumetric occupancy can be recovered by including both geometrical constraints and prior information given by objects CAD models from the ShapeNet dataset. This can be done by recasting the problem in the context of a probabilistic framework based on PPCA that enforces both geometrical constraints and the associated semantic given by the object category extracted by the object detector. We present results on synthetic data and extensive real evaluation on the ScanNet datasets on more than 1200 image sequences to show the validity of our approach in realistic scenarios. In particular, we show that 3D statistical priors are key to obtain reliable reconstruction especially when the input detections are noisy, a likely case in real scenes.

1. Introduction

Recovering 3D information from multiple-view images has been a long standing research topic in Computer Vision. Most of the efforts have been put in the robust computation of the 3D position of points extracted from 2D images correspondences [26]. This trend has led to impressive results showing accurate 3D points clouds obtained from realistic objects, even at the very large scale [1, 57, 22, 51]. Such description of the 3D world is however conveying a minimal semantic information: the point cloud provides just the localisation of the 3D points without the information associated, for instance, to the context of the scene it has been reconstructed from. Differently, on a higher semantic level, objects can provide an incredible amount of information to increase robustness when solving for geometry in multiple images and to improve performance of classification/recognition tasks [29, 19].

Applications such as 3D-aware scene understanding [52, 37, 54, 36], visual question and answering [59] and robot object manipulation [8, 49] strongly require to localise objects in 3D and thus motivate a joint geometric and semantic understanding. The usefulness of this problem is also stressed by the attempts of previous works injecting higher semantic in classical 3D reconstruction problems [4, 45, 55, 20, 14, 17]. This object-based geometric reasoning is now possible because of the accuracy and generalisation of modern objects detectors that can provide localization of several object classes in realistic scenarios [27, 15, 47, 46]. Still, one relevant and open problem is that most detectors provide this information onto the image plane only, without giving an indication of the object position and occupancy in 3D.

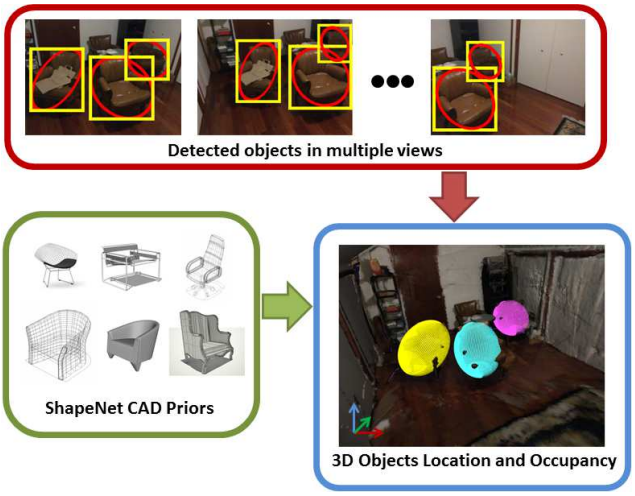

Figure 1. The figure shows the proposed framework. A semantic engine extracts and matches bounding boxes from objects in multiple frames using a CNN detector [47]. Given the object classes, we use the ShapeNet dataset [9] to create a realistic prior on the detected objects. Then, the Probabilistic Structure from Motion with Objects (PSfMO) method provides the metric localisation, occupancy and pose of object as a set of quadrics in the 3D space.

To this end, this paper proposes an efficient and robust

approach to recover 3D objects position and occupancy from multiple views using 3D object priors linked to the object categories of general purpose detectors. Fig. 1 shows the overall concept of our approach. We have a video processing engine that extracts and matches bounding boxes in multiple views for different object categories. Such association at each frame is made using each object appearance by proposing a variation of a tracking-by-detection approach [23]. We then construct a geometrical 3D prior given by the specific object classes. This step is done by processing CAD models in the ShapeNet [9] dataset thus giving a statistics about the object dimensions. We show that this very effective prior can be matched and merged with the semantic information obtained by the detector output.

Finally, ellipses and priors in multi-view are fed to a Probabilistic Structure from Motion with Objects (PSfMO) approach which outputs camera calibration and the objects position, occupancy and pose in 3D. Interestingly, this problem can be formalised as the factorization of a matrix storing all the 2D ellipses representation and then solving a 3D quadric reconstruction problem given the multi-view ellipses [12]. We show that this factorization can be derived as a Probabilistic Principal Component Analysis (PPCA) approach by injecting the priors computed from the ShapeNet dataset [9]. This formulation compares favourably against pure multi-view geometry approaches [11] since the prior counter the shortcomings of the inaccuracies given by the coarse localisation of the object detector.

1.1. Related Work

The proposed approach is a Structure from Motion (SfM) method that uses priors in order to obtain a reliable estimate of the geometry. There have been several examples in the literature providing probabilistic solution for SfM, mainly to improve the estimate of the 3D scene geometry. Forsyth et al. [21] recast the decomposition of the bilinear components in factorization, camera matrices and 3D points coordinates, as a Bayesian inference problem. The motivation is to encode in the prior the metric constraints involved in the problem, thus providing better results in the presence of degenerate configurations of points. In face modelling problems, the work of Solem and Kahl [53] used a learned shape model to aid the 3D inference over regions for which no 2D information is available. Del Bue et al. [18] used the information of the rigidity of some points to obtain reliable estimations of the 3D object structure with deforming objects. Information derived from object detections has already been used in SfM. The work described in [5] takes advantage of both semantic and geometrical properties associated with objects in the perspective case.

Another factorization problem that highly relies on priors is non-rigid SfM. This is due to the presence of objects 3D deformations that make the problem severely ill-posed. Torresani et al. [58] used Gaussian priors in a Probabilistic Principal Components Analysis (PPCA) framework together with a linear dynamic model over the deformation parameters. This framework is close to our method, however, our object representation enables us to build a better prior which is representative of the scene instead of a generic one. Similarly, [43] imposed a prior over temporal variations of the camera parameters combined with constraints over the proximity of projected 2D points and reconstructed 3D points. Again related to 3D points estimation, [16] defined a shape prior in a factorization based approach to help 3D reconstruction in case of degenerate motions. Akhter et al. [2] showed that a prior parametrization of the 3D trajectory motion can provide more efficient results. The work of Gotardo and Martinez [25] proposed a similar principle using DCT bases to represent the camera motion in order to regularise intrinsic and extrinsic parameters. Finally, [10] used a novel Procrustean Normal distribution to minimise geometrical deformations under an optimality criterion.

All these approaches deal with 2D point trajectories or matches in multi-view, only few works directly localise objects in a factorization framework. Previous methods attempted the joint reconstruction of different geometrical entities such as lines, curves and conics [44, 41, 6, 50, 7, 40, 32, 33, 42, 12, 34, 48]. However, even if these methods were able to obtain an inference of the 3D structure, none of them was aimed at obtaining an object based representation of the 3D world. Recently, the work of [11] proposed the SfM with Objects (SfMO). This method provides a solution to the localization of objects in a factorization framework by using the output of detectors only. However, even if the method is closed-form, it can lead to unreliable estimates, especially for the object occupancy, if the detector output is not accurate enough or if very few views are available.

Last but not least, two papers focusing on 3D object occupancy appeared at the same moment of this work [3, 31]. The first one learns a regression of the 3D bounding box parameters, while the second system uses sampling inside a EKF SLAM framework and exploits inertial sensor to obtain the object scale.

Related to previous work, this paper provides the following contributions:

- the probabilistic framework to include both object detections and shape priors to estimate of object position, occupancy and pose.
- an extensive experimental evaluation on approximately 1217 real sequences extracted from the ScanNet dataset.

Experimental evaluation shows that this framework can provide object estimates, especially with unreliable object detections extracted from real images.

2. Probabilistic Structure from Motion with Objects

We assume that different object have been detected in an image sequence and that detections between consecutive frames have been associated by tracking. After having detections matched for all the image sequence, we switch from a bounding box description by fitting an ellipse to the rectangular contour as shown in Fig. 1 (top figure). Such ellipse representation is certainly a coarse approximation of the real object shape but it is a computationally appealing function that allows the recovery of a better 3D position, occupancy and pose of the object.

Precisely, the localisation of objects in 3D is instantiated as a quadric (i.e. 3D ellipsoid) estimation from multiple 2D ellipses problem [11]. In this estimation framework there is no notion about the ratio of the different dimensions of the object, the method obtains the estimation of object position and occupancy from coarse multi-view relations only. We aim here to provide a novel probabilistic framework that can seamlessly include the priors into the multi-view geometrical problem.

2.1. Quadric reconstruction from conics

Before explaining our probabilistic approach, we first present the SfMO method and introduce the geometrical formulation of the problem. Consider a set of image frames indexed by $f = 1 \dots F$ representing a 3D scene under different viewpoints. A set of $i = 1 \dots O$ rigid objects in arbitrary positions are detected in each of the F images. Each object i in each image frame f is identified by an ellipse $\mathtt{D}_{fi}$ as described in the previous paragraph. The 3×3 matrix $\mathtt{D}_{fi}$ is the expression of a 2D conic in homogeneous coordinates. The aim of our problem is to find the 3D ellipsoid given by the 4×4 quadric $\mathtt{E}_i$ whose projections onto the image planes best fit the 2D ellipses $\mathtt{D}_{fi}$. This will solve for both the 3D localisation and occupancy of each object starting from image detections in different views.

Since the relationship between $\mathtt{D}_{fi}$ and $\mathtt{E}_i$ is not straightforward in the primal space, i.e. the Euclidean space of 3D points (2D points in the images), it is convenient to reformulate it in dual space, i.e. the space of the planes (lines in the images) [12]. In particular, the conics in 2D can be represented by the envelope of all the lines tangent to the conic curve, while the quadrics in 3D can be represented by the envelope of all the planes tangent to the quadric surface. Hence, the dual quadric is defined by the matrix $\mathtt{Q}_i = adj(\mathtt{E}_i)$, where adj is the adjoint operator, and the dual conic is defined by $\mathtt{C}_{fi} = adj(\mathtt{D}_{fi})$ [26].

Each quadric $\mathtt{Q}_i$, when projected onto the image plane, gives a conic denoted with $\mathtt{C}_{fi} \in \mathbb{R}^{3\times3}$. In this work, we assume an orthographic camera matrix, however, this formalisation can be adapted to more generic affine cameras [35].

The relationship between $\mathtt{Q}_i$ and $\mathtt{C}_{fi}$ is defined by the orthographic projection matrix $\mathtt{P}_f \in \mathbb{R}^{3\times4}$ as:

$$\mathtt{P}_f = \left[\begin{array}{c|c} \mathtt{R}_f & \mathbf{t}_f \\ \mathbf{0}_3^\top & 1 \end{array}\right] = \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (1)$$

where $\mathtt{R}_f \in \mathbb{R}^{2\times3}$ is an orthographic camera matrix such that $\mathtt{R}_f\mathtt{R}_f^\top = \mathtt{I}_{2\times2}$, the vector $\mathbf{t}_f$ is the camera translation and $\mathbf{0}_m$ denotes a vector of m zeros.

The dual conic $\mathtt{C}_{fi}$ and the dual quadric $\mathtt{Q}_i$ are defined up to an overall scale factor that can be arbitrarily fixed by setting the elements (3,3) of $\mathtt{C}_{fi}$ and (4,4) of $\mathtt{Q}_i$ to -1. After such normalisation, the relation between a dual quadric and its dual conic projections can be written as:

$$\mathtt{C}_{fi} = \mathtt{P}_f\mathtt{Q}_i\mathtt{P}_f^\top. \quad (2)$$

From each projection, it is possible to define a factorization problem by first vectorizing Eq. (2) by defining $\mathbf{v}_i = vech(\mathtt{Q}_i)$ and $\mathbf{c}_{fi} = vech(\mathtt{C}_{fi})$ as the vectorization of symmetric matrices $\mathtt{Q}_i$ and $\mathtt{C}_{fi}$ respectively[1].

Then, let us arrange the products of the elements of $\mathtt{P}_f$ and $\mathtt{P}_f^\top$ in a unique matrix $\mathtt{G}_f \in \mathbb{R}^{6\times10}$ as follows [28]:

$$\mathtt{G}_f = \mathtt{Y}(\mathtt{P}_f \otimes \mathtt{P}_f)\mathtt{W} \quad (3)$$

where $\otimes$ is the Kronecker product and matrices $\mathtt{Y} \in \mathbb{R}^{6\times9}$ and $\mathtt{W} \in \mathbb{R}^{16\times10}$ are two matrices such that $vech(\mathtt{X}) = \mathtt{Y}\, vec(\mathtt{X})$ and $vec(\mathtt{X}) = \mathtt{W}\, vech(\mathtt{X})$ respectively, where $\mathtt{X}$ is a symmetric matrix. We can rewrite Eq. (2) for all frames as:

$$\mathbf{c}_i = \mathtt{G}\mathbf{v}_i, \quad (4)$$

Where $\mathtt{c}$ and $\mathbf{v}_i$ are respectively a $6F$ and $10F$ dimension vectors containing the conics and the ellipsoids in all the frames, $\mathtt{G}$ is $6F \times 10$ matrix which contains the camera for each view.

To solve this factorization problem and perform the metric upgrade, the SfMO method uses the fact that the translation parameters of the quadrics can be solved separately. In practice it consists in solving a classical SfM problem where the 2D measurements correspond to the centres of the ellipses. The solution provides both the centre $\mathbf{t}_i^e$ of each ellipsoid $\mathbf{v}_i$ and the camera matrix $\mathtt{G}$.

To solve for the remaining parameters, namely the shape and the orientation, the ellipsoids are registered to the origin. Let us denote by $\mathbf{w}_i$ the 6 dimension vectorised quadric which has been registered from the non-translated quadric $\mathbf{v}_i$. It can be shown that with $\mathbf{w}_i$ and its corresponding centred conics $\mathbf{c}_i^r$, the factorization equation Eq. (4) becomes:

$$\mathbf{c}_i^r = \mathtt{G}^r\mathbf{w}_i \quad (5)$$

[1]The operator vec serialises all the elements of a generic matrix. The operator $vech$ vectorises the elements of the lower triangular part of a symmetric matrix, such that, given a symmetric matrix $\mathtt{X} \in \mathbb{R}^{n\times n}$, the vector $\mathbf{x}$, defined as $\mathbf{x} = vech(\mathtt{X})$, is $\mathbf{x} \in \mathbb{R}^g$ with $g = \frac{n(n+1)}{2}$.

where the $\mathbf{c}_i^r$ is a $(3F)$-vector containing all the centred conics for the object i observed in F frames and $\mathtt{G}^r$ is a $3F \times 6$ matrix obtained from $\mathtt{G}$. In practice, it consists simply in selecting the rows and the columns related to the quadric shape and disregarding the ones related to the translation (that are equal to zero given the centring). At this point, the SfMO method consists in multiplying by the pseudo inverse of $\mathtt{G}^r$ to obtain the remaining quadric parameters $\mathbf{w}_i$. Our PSfMO approach differs from SfMO on this last step where we use a probabilistic factorisation to include our object prior.

2.2. Object pose and shape decomposition

We seek to perform an optimization constrained by an apriori knowledge on the semi-axis lengths of each object. Let us first show how these lengths appear explicitly in the factorization problem defined by Eq. (4). A generic ellipsoid in dual space $\mathtt{Q}^*$ can be written as:

$$\mathtt{Q}^* = \mathtt{Z}\breve{\mathtt{Q}}^*\mathtt{Z}^\top \tag{6}$$

where $\breve{\mathtt{Q}}^*$ is an ellipsoid centred on the origin and with the axes aligned to the 3D coordinates and $\mathtt{Z}$ is an homogeneous transformation accounting for an arbitrary rotation and translation. Then $\mathtt{Z}$ and $\breve{\mathtt{Q}}^*$ can be written as:

$$\mathtt{Z} = \begin{bmatrix} \mathtt{R}(\boldsymbol{\theta}) & \mathbf{t} \\ \mathbf{0}_3^\top & 1 \end{bmatrix}, \quad \breve{\mathtt{Q}}^* = \begin{bmatrix} a^2 & 0 & 0 & t_1^e \\ 0 & b^2 & 0 & t_2^e \\ 0 & 0 & c^2 & t_3^e \\ 0 & 0 & 0 & -1 \end{bmatrix}, \tag{7}$$

where $\mathbf{t}^e = [t_1, t_2, t_3]^\top$ is the translation vector, $\mathtt{R}(\boldsymbol{\theta})$ is the rotation matrix given by the Euler angles $\boldsymbol{\theta} = [\theta_1, \theta_2, \theta_3]^\top$ and $a,\ b,\ c$ are the three semi-axes of the ellipsoid.

Given the registered quadric $\mathbf{w}_i$ used in Eq. 5, the translation $\mathbf{t}$ can be then decoupled from the quadric matrix so we can register the ellipsoids to the origin by setting $\mathbf{t}^e$ to $\mathbf{0}$ and removing this variable from the equations. Now, let us examine the expression of the registered quadric $\mathbf{w}_i$ in more details. This vector can be expressed in terms of the remaining six quadric parameters. Defining the vector $\mathbf{e} \in \mathbb{R}^6$ as $\mathbf{e} = [\theta_1, \theta_2, \theta_3, a, b, c]^\top$, we can evaluate a functional form of the vector $\mathbf{w}_i(\mathbf{e})$ as follow:

$$\mathbf{w}_i(\mathbf{e}) = \begin{bmatrix} r_{11}(\boldsymbol{\theta})^2 a^2 + r_{12}(\boldsymbol{\theta})^2 b^2 + r_{13}(\boldsymbol{\theta})^2 c^2 \\ r_{11}(\boldsymbol{\theta}) r_{21}(\boldsymbol{\theta}) a^2 + r_{12}(\boldsymbol{\theta}) r_{22(\boldsymbol{\theta})} b^2 + r_{13}(\boldsymbol{\theta}) r_{23}(\boldsymbol{\theta}) c^2 \\ r_{11}(\boldsymbol{\theta}) r_{31}(\boldsymbol{\theta}) a^2 + r_{12}(\boldsymbol{\theta}) r_{32}(\boldsymbol{\theta}) b^2 + r_{13}(\boldsymbol{\theta}) r_{33}(\boldsymbol{\theta}) c^2 \\ r_{21}(\boldsymbol{\theta})^2 a^2 + r_{22}(\boldsymbol{\theta})^2 b^2 + r_{23}(\boldsymbol{\theta})^2 c^2 \\ r_{21}(\boldsymbol{\theta}) r_{31}(\boldsymbol{\theta}) a^2 + r_{22}(\boldsymbol{\theta}) r_{32}(\boldsymbol{\theta}) b^2 + r_{23}(\boldsymbol{\theta}) r_{33}(\boldsymbol{\theta}) c^2 \\ r_{31}(\boldsymbol{\theta})^2 a^2 + r_{32}(\boldsymbol{\theta})^2 b^2 + r_{33}(\boldsymbol{\theta})^2 c^2 \end{bmatrix} \tag{8}$$

From now, we will again denote this vector by $\mathbf{w}_i$ to simplify notations. We observe that it is possible to decompose it in the following way:

$$\mathbf{w}_i = \mathtt{R}_i(\theta)\mathbf{l}_i, \tag{9}$$

where $\mathtt{R}_i(\boldsymbol{\theta})$ contains the orientation of the quadric and is of size 6×3:

$$\mathtt{R}_i(\boldsymbol{\theta}) = \begin{bmatrix} r_{11}(\boldsymbol{\theta})^2 & r_{12}(\boldsymbol{\theta})^2 & r_{13}(\boldsymbol{\theta})^2 \\ r_{11}(\boldsymbol{\theta}) r_{21}(\boldsymbol{\theta}) & r_{12}(\boldsymbol{\theta}) r_{22}(\boldsymbol{\theta}) & r_{13}(\boldsymbol{\theta}) r_{23}(\boldsymbol{\theta}) \\ r_{11}(\boldsymbol{\theta}) r_{31}(\boldsymbol{\theta}) & r_{12}(\boldsymbol{\theta}) r_{32}(\boldsymbol{\theta}) & r_{13}(\boldsymbol{\theta}) r_{33}(\boldsymbol{\theta}) \\ r_{21}(\boldsymbol{\theta})^2 & r_{22}(\boldsymbol{\theta})^2 & r_{23}(\boldsymbol{\theta})^2 \\ r_{21}(\boldsymbol{\theta}) r_{31}(\boldsymbol{\theta}) & r_{22}(\boldsymbol{\theta}) r_{32}(\boldsymbol{\theta}) & r_{23}(\boldsymbol{\theta}) r_{33}(\boldsymbol{\theta}) \\ r_{31}(\boldsymbol{\theta})^2 & r_{32}(\boldsymbol{\theta})^2 & r_{33}(\boldsymbol{\theta})^2 \end{bmatrix}, \tag{10}$$

where $\mathbf{l}_i = [a^2,\ b^2,\ c^2]^\top$ contains the three axes lengths. This provides the separation between the rotational pose and ellipse shape components making possible to impose the priors over the axes. In the following and unless stated otherwise, we will denote the orientation matrix with $\mathtt{R}_i$ to simplify the notation.

2.3. Object 3D prior

The decomposition in Eq. 9 shows that the shape of the ellipsoid is given by the square of the axis length $\mathbf{l}_i$ and we will now describe how to build a prior distribution on such axes. The prior is given by statistics on the dimensions of the objects collected from the ShapeNet dataset [9]. For each object i present in the dataset, we extract a 3 dimensional vector $\mathbf{h}_i = [a, b, c]^T$ containing the lengths along the main axes. These values have been normalised so that $|\mathbf{h}_i|^2 = 1$ and thus providing the information about the ratio of these axis. From the normalisation, we can deduce that $\mathbf{l}_i$ lies in a 2-dimensional space. For each object category c, we first use PCA to compute a 3x2 projection matrix $\mathtt{V}_c$ and a mean $\boldsymbol{\mu}_c$ to project each vector $\mathbf{l}_i$ in the 2D space. Secondly, we fit a 2D Gaussian parametrised by $(0, \Sigma_c)$. This Gaussian encodes our *a priori* information about the ratio of the different dimensions of the object. Fig. 2 shows an example in the case of the bottle category where we can notice that the horizontal dimension accounts for the ratio between x and y axes and the vertical one for the ratio between the bottle diameter and its height.

In order to apply the prior on test images, we must select one of the six possible correspondences between the axes of the estimated ellipsoids and the prior axes i.e. the rows of $\mathtt{V}_c$ and $\boldsymbol{\mu}_c$. In practice, we estimate the axes lengths for the 6 possible configurations and keep the one with the higher likelihood. Lastly, we must estimate a scale factor z to account for the scale difference between the normalised axes of the prior and the non normalised axis of the ellipsoids we wish to estimate. This last variable is estimated from the data as described in the following section.

The other parameters such as the camera translation and orientation are estimated by our method directly from the data using multi-view relations.

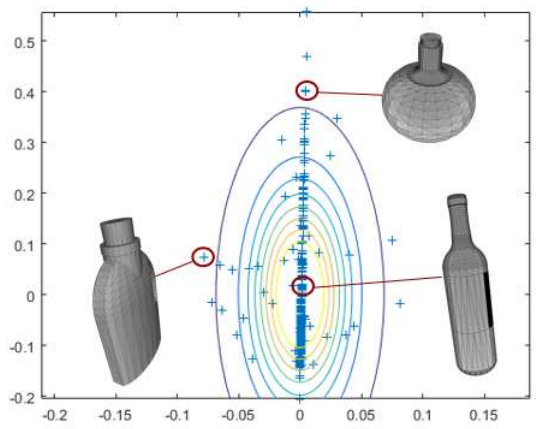

Figure 2. The contour lines are the Gaussian sections of the prior for the bottle category. Each blue cross corresponds to a CAD model extracted from the ShapeNet database.

2.4. Probabilistic PCA with shape priors

From a generative point of view, our Probabilistic Principal Component Analysis (PPCA)[2] model assumes that each centred conic vector $\mathbf{c}_i^r$ is obtained by sampling a 2-dimensional latent random vector $\mathbf{s}_i$ and an additive Gaussian noise ϵ such that:

$$\mathbf{c}_i^r = \mathtt{G}^r\, \mathtt{R}_i\, z_i\, (\mathtt{V}_{c_i}\mathbf{s}_i + \boldsymbol{\mu}_{c_i}) + \epsilon, \tag{11}$$

where $\epsilon \hookrightarrow N[0, \sigma \mathtt{I} d^{3F}]$. The hidden variable $\mathbf{s}_i$ encodes the axis lengths, and thus the ellipsoid shape, that we have to estimate. μ_{c_i} and $\mathtt{V}_{c_i}$ are used to re-project the $\mathbf{s}_i$ into the 3 dimensional axes length vector, where c_i is the category of the object i. Conversely from the standard PPCA, we have a latent additional variable z_i which accounts for the scale difference between the prior and the reconstructed ellipsoid. As in the decomposition of Eq. (9), $\mathtt{R}_i$ is the rotation matrix defined in Eq. 10, and $\mathtt{G}^r$ is the camera matrix defined in Eq. 5. Inference and parameter estimation with PPCA is usually done with an *EM* algorithm. In the *E-step*, we estimate sufficient statistics from the posterior distribution $P(\mathbf{s}_i|\mathbf{c}_i^r)$. In the *M-step*, we then compute the noise variance parameter σ.

The posterior over the latent variables can be written as:

$$P(\mathbf{s}_i, z_i|\mathbf{c}_i^r) \propto P(\mathbf{c}_i^r|\mathbf{s}_i, z_i)P(\mathbf{s}_i) \tag{12}$$

From Eq. 11, we can write that:

$$P(\mathbf{c}_i^r|\mathbf{s}_i, z_i) = N(x_i, \sigma \mathtt{I} d^{3F}), \tag{13}$$

where we used the notation $x_i = \mathtt{G}^r\mathtt{R}_i z_i(\mathtt{V}\mathbf{s}_i + \mu_{c_i})$. It has been said in Sec. 2.3 that $P(\mathbf{s}_i)$ is the Gaussian $N(0, \Sigma_{c_i})$ where c_i is the category of object i. Skipping the constants that do not depend on the latent variables, the logarithm of the posterior can be written as:

$$log(P(\mathbf{s}_i, z_i|\mathbf{c}_i^r)) \propto (\mathbf{x}_i - \mathbf{c}_i^r)\frac{1}{\sigma} Id^{3F}(\mathbf{x}_i - \mathbf{c}_i^r) + \mathbf{s}_i^T \frac{1}{\sigma}\mathtt{I} d^{3F}\mathbf{s}_i, \tag{14}$$

It can be noticed from this formula that the log likelihood is simply a sum of two terms. The first one accounts for the reprojection error with respect to the observed conics $\mathbf{c}_i^r$, and the second one refers to the prior. In particular, the noise covariance parameter σ can be seen as a trade-off parameter between the two terms which is automatically estimated from the data. Since this distribution is intractable, we resort to Markov Chain Monte Carlo [24] (MCMC) to estimate the expectation of the latent variables $\{\hat{z}, \hat{\mathbf{s}}_i\}$ under the posterior.

The last element to estimate is the matrix containing the ellipsoid orientation $\mathtt{R}_i$. It has been shown in Eq. (10) that this matrix is constructed from the three Euler angles $\boldsymbol{\theta}$. One solution would be to include them as additional variables in the posterior, but we observe that the MCMC estimation becomes unreliable. Instead, we propose to estimate these angles by solving the following optimisation problem:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\hat{\mathbf{x}}_i - \mathbf{c}_i^r\|^2, \tag{15}$$

where $\hat{x}_i = \mathtt{G}^r\mathtt{R}_i\hat{z}_i(\mathtt{V}\hat{\mathbf{s}}_i + \mu)$ we have explicitly written the dependencies between $\mathtt{R}_i$ and $\boldsymbol{\theta}$. We optimise the cost function with non-linear Least Squares by parametrising rotations with quaternions. From the solution $\hat{\boldsymbol{\theta}}$, we can compute the orientation matrix $\mathtt{R}_i$ and obtain our quadric estimate as:

$$\hat{\mathbf{w}}_i = \mathtt{R}_i(\hat{\boldsymbol{\theta}})\, \hat{\mathbf{s}}_i\, \hat{z} \tag{16}$$

In the *M-step*, we use our estimate of $\hat{\mathbf{w}}_i$ to estimate the noise covariance parameter σ given by:

$$\hat{\sigma} = \frac{1}{3F}\sum_{i=1}^{O}\{||\mathbf{c}_i^r||^2 - 2\hat{\mathbf{w}}_i^T\mathtt{G}^r\mathbf{c}_i^r + \mathrm{trace}(\hat{\mathbf{w}}_i^T\hat{\mathbf{w}}_i\mathtt{G}^{rT}\mathtt{G}^r)\}. \tag{17}$$

We provide an overview of the method in Alg. 1.

Initialisation. Given the ellipses in multiple views, we first apply the SfMO method to obtain the camera matrix $\mathtt{G}^r$ and estimation of the ellipsoid orientations θ. We use these values as an initialisation for the PSfMO method.

3. Experiments

Our experiments aim at evaluating the proposed PSfMO approach against the prior-less method SfMO [11]. We test the behaviour of both methods in presence of noise in a synthetic setting and show qualitative results on real examples

[2] For a more general and extensive description we refer to [56].

Algorithm 1 PSfMO algorithm

Require: : Ellipses from multi-view images object detections $\mathtt{C}$ and initial noise covariance σ.

1: Initialisation: $(\mathtt{G}^r, \hat{\theta}) \leftarrow$ SfMO $(\mathtt{C})$;
2: **while** not converged **do**
3: estimate (z, s_i) by optimizing the posterior (14);
4: estimate quadric orientation $\hat{\theta}$ (15);
5: compute the centred quadric $\hat{\mathbf{w}}_i$ (16);
6: estimate $\hat{\sigma}$ (17);
7: **end while**

in two freely available datasets. We collect CAD models and defined priors for the categories of objects present in our image sequences (bottles, monitors, chairs, cars, potted plant and coffee cup) for which hundreds of examples were available in the ShapeNetCore dataset for each category. Some examples of the priors have been shown in Fig. 2.

3.1. Experiments on synthetic data

We generated ellipsoids, randomly placed inside a cube of side 20 units. We generated a set of 5 objects drawn randomly from the categories present in our CAD collection. The length of the ellipsoid axes for each object category c were generated from Gaussian distributions obtained with the statistics $\hat{\boldsymbol{\mu}}_c$ and $\hat{\Sigma}_c$ that we computed from the CAD collection.

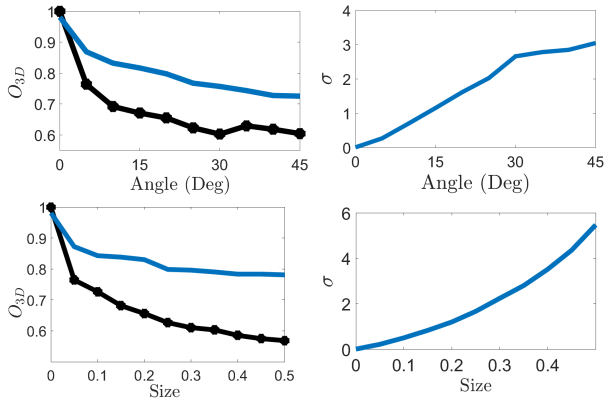

Figure 3. Top images: Comparison of the volume overlap $O_{3\mathrm{D}}$ and noise variance σ given rotation error of the ellipse for the SfMO (black dotted line) and the PSfMO (blue line) methods. Bottom images: values for $O_{3\mathrm{D}}$ and σ in the case of size errors in the ellipsoid.

A set of 20 camera views were generated with a camera trajectory computed such that azimuth and elevation angles span the range $[0°, 10°]$. Given the orthographic camera matrix projection $\mathtt{P}_f$ at each camera frame, ground-truth ellipses were calculated from the exact projections of the 3D quadrics.

The ellipses were corrupted with two types of errors on rotation and size. To impose such errors, the axes length l_1, l_2 and the orientation α of the first axis were perturbed as follows:

$$\hat{l}_j = l_j \left(1 + \nu^l\right), \qquad \hat{\alpha} = \alpha + \nu^\alpha, \tag{18}$$

where $\nu_j^{\mathbf{t}}$, ν^α and ν^l are random variables with uniform PDF and mean value equal to zero, and $\bar{l} = (l_1 + l_2)/2$.

In order to highlight the specific impact of each error, they were evaluated separately. Error magnitudes were set tuning the boundary values of the uniform PDFs of $\nu_j^{\mathbf{t}}$, ν^α and ν^l. In detail, for each kind of error, we considered 10 different values of $\nu_j^{\mathbf{t}}$, ν^a and ν^l, with uniform spacing, and we applied the resulting error realisations to the ellipse reprojections related to all the ellipsoids. We run 100 trials for each setup, described by the number of objects and error on ellipses. The accuracy of the estimated 3D object position and pose was measured by the volume overlap $O_{3\mathrm{D}}$ given by the intersection between ground-truth (GT) and estimated (ES) ellipsoids respectively:

$$O_{3\mathrm{D}} = \frac{1}{N} \sum_{i=1}^{N} \frac{\mathcal{Q}_i \cap \tilde{\mathcal{Q}}_i}{\mathcal{Q}_i \cup \tilde{\mathcal{Q}}_i}, \tag{19}$$

where $\mathcal{Q}_i$ and $\tilde{\mathcal{Q}}_i$ denote the volume of GT and ES ellipsoids in the dual space respectively thus the metric evaluating an algebraic error. It might happen that the quadrics obtained by the methods do not correspond to a valid ellipsoid. In this case, we consider the test as failed. In Fig. 3, we reported the results for both methods SfMO and PSfMO in terms of 3D overlap $O_{3\mathrm{D}}$.

In the absence of noise, SfMO retrieves nearly perfect ellipsoids thanks to the closed-form solution. In this case, including prior information can only degrade slightly or produce similar performances. However, PSfMO is able to retrieve better ellipsoids when the ellipse orientations and sizes are corrupted. This effect is more important for the size error. This effect is expected, our prior information has a particular effect on the estimation of the axes lengths. We also observe that as the noise increases, the estimation of σ produces higher values, meaning that the method relies more on the prior. This indicates that the automatic trade-off between data and prior is effective and working as expected.

3.2. Experiments on ScanNet dataset

We provide exhaustive experimental evaluation over the ScanNet dataset [13] which consists of 1500 indoors RGBD scans annotated with 3D camera poses, surface reconstructions, and mesh segmentation related to several object categories. Such annotations enable us to construct 3D ground-truth (GT) data by fitting ellipsoids to each object's mesh so evaluating the accuracy of PSfMO in realistic environments.

In more details, the GT ellipsoids are computed as the one that include all the labelled 3D points in the mesh and have minimal volume. Given the available camera matrices, we can then reproject each object's mesh onto the image plane so localising the object as a 2D point cloud. Then, we draw a bounding box around the obtained 2D point cloud and consider this as a putative object detection for our system. From the bounding box, we finally construct an ellipse by taking the centre of the bounding box, its axes lengths as the width and high of the box and its orientation aligned to the axes of the box.

We also take into account occlusions when computing the bounding boxes. As occluded objects are typically missed by object detectors, we impose that at least 80% of the object surface should be visible, i.e. not occluded by another object, in the current frame to be detected. Moreover, we only keep images when more than three objects are appearing and we set the minimum length of a sequence to three frames. This experimental protocol provides us with 1217 sequences overall.

As with the synthetic experiments, we compare the SfMO and the proposed PSfMO with the O_{3D} overlap measure. As it is in standard in 3D reconstruction, we use the Procrustes method to align the centres $\mathbf{t}_i^e$ of the estimated ellipsoids with the centres of the GT ellipsoids and then proceed with the error evaluation.

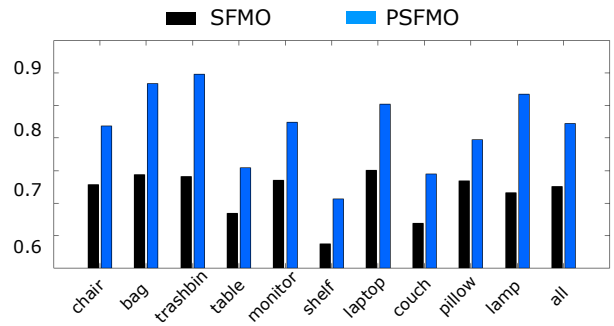

Figure 4. Each couple of bar shows the accuracy for both SfMO and PSfMO methods for the ten object categories the most present in our dataset.

In Fig 4, we show the O_{3D} for the ten object categories most represented in our sequences (83% of the objects) as well as the average for all categories. The results confirm the findings of the synthetic experiments: The PSfMO method shows the best performances by using the prior built on the ShapeNet dataset. The average O_{3D} for all the 1217 sequences is 0.72 for SfMO and 0.82 for PSfMO.

We can see that the prior is especially useful for category with low intra-class variability. For instance the dimensions of a trash bin category is rather constrained as x and y-axes have usually equal lengths. On the opposite, categories such as table and couch have higher variability.

We have also evaluated the camera estimation by using the GT perspective camera matrices provided by the ScanNet dataset. To recover camera poses in the affine case we solved a PnP problem [38] given the affine camera intrinsic, the 3D GT ellipsoid centres and the 2D noisy conic centres. We then align the estimated camera poses with the given GT poses. The relative average translation error (normalised by the sequence length) is 20% and the rotation error is 32 degree on average. These errors might be relevant for some applications and future work on better camera estimation could improve the overall results.

The figure 5 shows an illustration of the results. We can see that SfMO has difficulties when estimating the ellipsoid axis which is more parallel to the optical camera axis. By using prior knowledge, PSfMO retrieves more accurate ellipsoids and constrain the shape to the right occupancy. Due to the ambiguity of the problem, there are many ellipsoid configurations which can project closely to the observed ellipses. Our PSfMO method selects the one which is the closest to the prior.

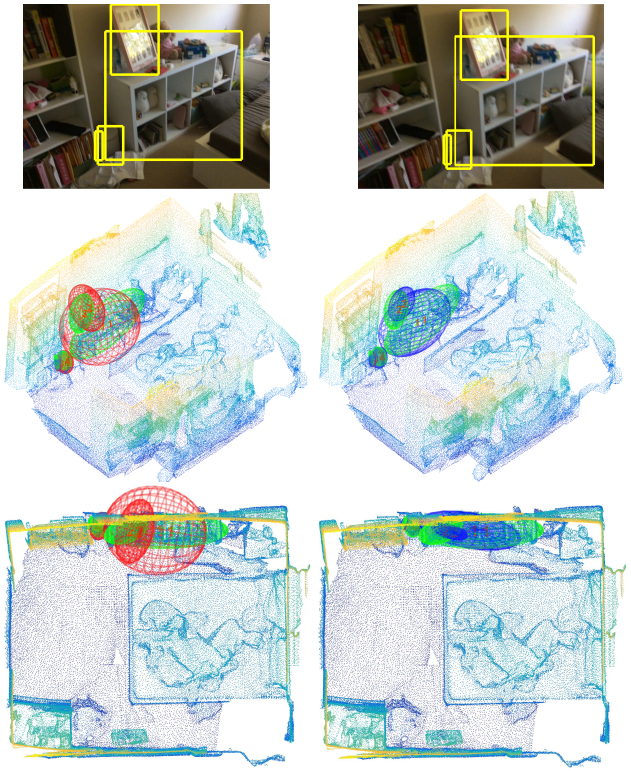

Figure 5. Illustration of results with ScanNet data. The top rows are color frames with object detections. The midddle and bottom rows display the scan with the ellipsoids. The green are the ground-truth, the red ones on the left are obtained with SfMO and the blue ones on the right with our method PSfMO.

Similar results can be noticed in Fig. 6 where the camera has a minimal baseline and rotation, thus making more challenging the accurate estimation of the objects size in the scene. To notice that the PSfMO solution provides a

remarkable alignment over the room furniture.

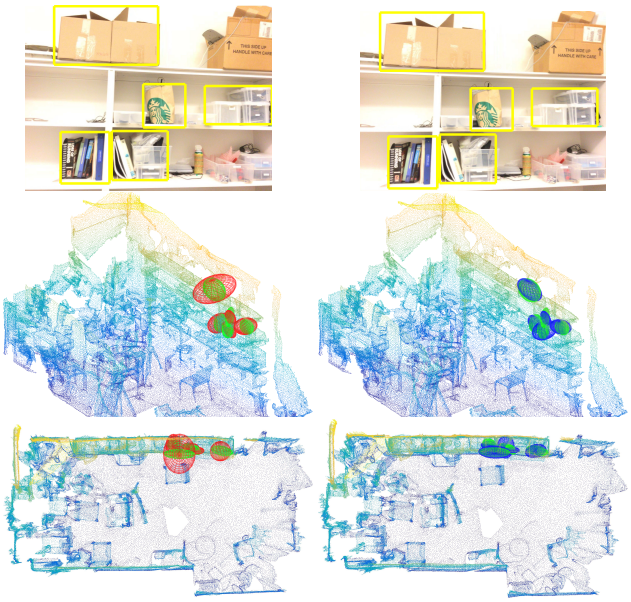

Figure 6. Illustration of results with ScanNet data. The top rows are color frames with object detections. The middle and bottom rows display the scan with the ellipsoids. The green are the ground-truth, the red ones are obtained with SfMO and the blue ones with PSfMO.

We also tested the proposed algorithm on the KINECT dataset [5] to demonstrate its generality to other datasets. In this experiment, automatic detection and tracking are used but no ground-truth is available. It is composed by five sequences, each one showing a different office desk, with about $10 - 15$ objects, from a variable number of frames. Bounding boxes associated to each object are also provided. Given all the extracted detections in each frame, we use a modified tracking by detection method [23] to associate the bounding boxes among different frames. This algorithm computes a distance matrix using patch appearance and associate detections using the Hungarian method for bipartite matching. We relaxed the part associated to the smoothness of the object trajectory because we might have consistent camera motion among consecutive frames thus causing the corresponding consecutive bounding boxes to be far apart.

When using all the images from the sequences, we observed that both SfMO and PSfMO have similar results. However, if we reduce the number of views, we observe that the prior has a major role. Illustrative results for both the SfMO and the PSfMO methods are shown in Fig. 7 for the five sequences of the dataset. Since the angles spanned by the cameras views is narrow, SfMO failed at estimating the ellipsoid axis which is close to the optical camera axis. This behaviour is a similar as the one observed in the ScanNet experiments.

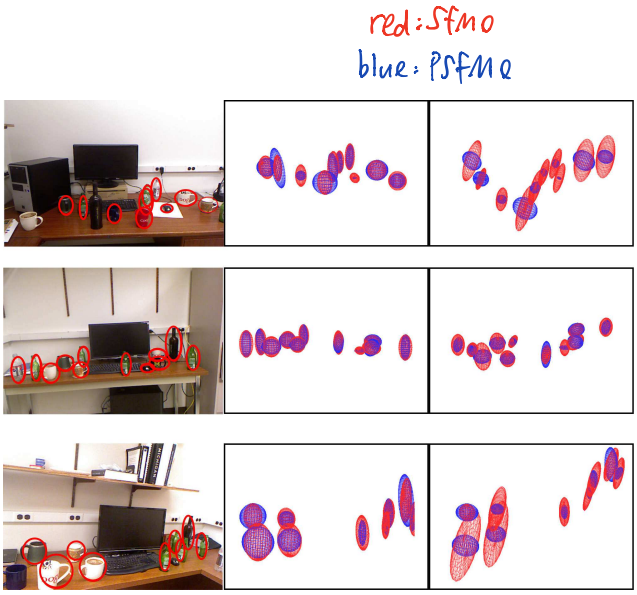

Figure 7. Each row corresponds to one sequence from the KINECT dataset. From left to right: An image frame with the ellipses obtained from the bounding box detections, a frontal view and a top view of the estimated 3D ellipsoids. The red ellipsoids are given by SfMO while the blue ones from PSfMO.

Acknowledgements

We thanks Marco Crocco for support and discussions about the SfMO approach, the authors of the ScanNet dataset for their assistance and James Stuart for its useful comments.

4. Conclusion

We presented a probabilistic factorization method to estimate objects occupancy and position from bounding box detections in multiple views. The main contribution of the method is its ability to use simple statistics collected from CAD datasets to achieve higher robustness to noise and be able to provide good results even if very few views are available. As a future works and perspectives, different kind of prior information could be tested. First, a classifier with coarse pose information would provide constraint on the ellipsoid orientation. Then higher-level semantic using scene context [30] and relation between objects [39] could be extracted to guide the ellipsoid centre estimation.

References

[1] S. Agarwal, N. Snavely, S. M. Seitz, and R. Szeliski. *Bundle Adjustment in the Large*, pages 29–42. Springer Berlin Heidelberg, 2010. 1

[2] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade. Trajectory space: A dual representation for nonrigid structure from motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(7):1442–1456, 2011. 2

[3] M. Arsalan, A. Dragomir, and F. John. 3d bounding box estimation using deep learning and geometry. In *Computer*

Vision and Pattern Recognition, pages 1271–1278. IEEE, 2017. 2

[4] S. Y. Bao, M. Bagra, Y.-W. Chao, and S. Savarese. Semantic structure from motion with points, regions, and objects. In *Computer Vision and Pattern Recognition*, pages 2703–2710. IEEE, 2012. 1

[5] S. Y. Bao and S. Savarese. Semantic structure from motion. In *Computer Vision and Pattern Recognition*, pages 2025–2032. IEEE, 2011. 2, 8

[6] A. Bartoli and P. Sturm. Structure-from-motion using lines: Representation, triangulation, and bundle adjustment. *Computer Vision and Image Understanding*, 100(3):416–441, 2005. 2

[7] R. Berthilsson, K. Åström, and A. Heyden. Reconstruction of general curves, using factorization and bundle adjustment. *International Journal of Computer Vision*, 41(3):171–182, 2001. 2

[8] G. Biegelbauer and M. Vincze. Efficient 3d object detection by fitting superquadrics to range image data for robot's object manipulation. In *Robotics and Automation, 2007 IEEE International Conference on*, pages 1086–1091. IEEE, 2007. 1

[9] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 1, 2, 4

[10] J. Cho, M. Lee, and S. Oh. Complex non-rigid 3d shape recovery using a procrustean normal distribution mixture model. *International Journal of Computer Vision*, 117(3):1–21, 2015. 2

[11] M. Crocco, C. Rubino, and A. Del Bue. Structure from motion with objects. In *Computer Vision and Pattern Recognition*, pages 782–788. IEEE, 2016. 2, 3, 5

[12] G. Cross and A. Zisserman. Quadric reconstruction from dual-space geometry. In *International Conference on Computer Vision*, pages 25–31. IEEE, 1998. 2, 3

[13] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. *arXiv preprint arXiv:1702.04405*, 2017. 6

[14] A. Dame, V. A. Prisacariu, C. Y. Ren, and I. Reid. Dense reconstruction using 3d object shape priors. In *Computer Vision and Pattern Recognition*, pages 1288–1295. IEEE, 2013. 1

[15] T. Dean, M. A. Ruzon, M. Segal, J. Shlens, S. Vijayanarasimhan, and J. Yagnik. Fast, accurate detection of 100,000 object classes on a single machine. In *Computer Vision and Pattern Recognition*, pages 1814–1821. IEEE, 2013. 1

[16] A. Del Bue. A factorization approach to structure from motion with shape priors. In *Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 2

[17] A. Del Bue. Adaptive non-rigid registration and structure from motion from image trajectories. *International Journal of Computer Vision*, 103:226–239, June 2013. 1

[18] A. Del Bue, X. Llado, and L. Agapito. Non-rigid metric shape and motion recovery from uncalibrated images using priors. In *Computer Vision and Pattern Recognition*, volume 1, pages 1191–1198. IEEE, 2006. 2

[19] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert. An empirical study of context in object detection. In *Computer Vision and Pattern Recognition*, pages 1271–1278. IEEE, 2009. 1

[20] N. Fioraio and L. Di Stefano. Joint detection, tracking and mapping by semantic bundle adjustment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1538–1545, 2013. 1

[21] D. A. Forsyth, S. Ioffe, and J. Haddon. Bayesian structure from motion. In *International Conference on Computer Vision*, volume 1, pages 660–665. IEEE, 1999. 2

[22] J.-M. Frahm, P. Fite-Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y.-H. Jen, E. Dunn, B. Clipp, S. Lazebnik, et al. Building rome on a cloudless day. In *ECCV 2010*, pages 368–381. Springer, 2010. 1

[23] A. Geiger, M. Lauer, C. Wojek, C. Stiller, and R. Urtasun. 3d traffic scene understanding from movable platforms. *Transactions on Pattern Analysis and Machine Intelligence*, 36(5):1012–1025, 2014. 2, 8

[24] J. Goodman and J. Weare. Ensemble samplers with affine invariance. *Communications in applied mathematics and computational science*, 5(1):65–80, 2010. 5

[25] P. F. Gotardo and A. M. Martinez. Computing smooth time trajectories for camera and deformable shape in structure from motion with occlusion. *Transactions on Pattern Analysis and Machine Intelligence*, 33(10):2051–2065, 2011. 2

[26] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, 2003. 1, 3

[27] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 1

[28] H. V. Henderson and S. Searle. Vec and vech operators for matrices, with some uses in jacobians and multivariate statistics. *Canadian Journal of Statistics*, 7(1):65–81, 1979. 3

[29] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. *International Journal of Computer Vision*, 80(1):3–15, 2008. 1

[30] H. Izadinia, F. Sadeghi, and A. Farhadi. Incorporating scene context and object layout into appearance modeling. In *Computer Vision and Pattern Recognition*, pages 232–239. IEEE, 2014. 8

[31] D. Jingming and S. Xiaohan, Fei Stefano. Visual-inertial-semantic scene representation for 3d object detection. In *Computer Vision and Pattern Recognition*, pages 1371–1378. IEEE, 2017. 2

[32] M. Kaess, R. Zboinski, and F. Dellaert. Mcmc-based multiview reconstruction of piecewise smooth subdivision curves with a variable number of control points. In *European Conference on Computer Vision*, pages 329–341. Springer, 2004. 2

[33] F. Kahl and J. August. Multiview reconstruction of space curves. In *International Conference on Computer Vision*, pages 1017–1024. IEEE, 2003. 2

[34] F. Kahl and A. Heyden. Affine structure and motion from points, lines and conics. *International Journal of Computer Vision*, 33(3):163–180, 1999. 2

[35] K. Kanatani and Y. Sugaya. Factorization without factorization: complete recipe. *Memoirs of the Faculty of Engineering*, 38(1&2):61–72, 2004. 3

[36] B.-S. Kim, P. Kohli, and S. Savarese. 3d scene understanding by voxel-crf. In *International Conference on Computer Vision*, pages 1425–1432, 2013. 1

[37] H. S. Koppula, R. Gupta, and A. Saxena. Learning human activities and object affordances from rgb-d videos. *The International Journal of Robotics Research*, 32(8):951–970, 2013. 1

[38] V. Lepetit, F. Moreno-Noguer, and P. Fua. Epnp: An accurate o (n) solution to the pnp problem. *International Journal of Computer Vision*, 81:155–166, 2009. 7

[39] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision*, pages 852–869. Springer, 2016. 8

[40] F. Mai and Y. Hung. 3d curves reconstruction from multiple images. In *Digital Image Computing: Techniques and Applications*, pages 462–467. IEEE, 2010. 2

[41] D. Matinec and T. Pajdla. Line reconstruction from many perspective images by factorization. In *Computer Vision and Pattern Recognition*, volume 1, pages 491–497. IEEE, 2003. 2

[42] I. Nurutdinova and A. Fitzgibbon. Towards pointless structure from motion: 3d reconstruction and camera parameters from general 3d curves. In *International Conference on Computer Vision*, volume 1, pages 2363–2371. IEEE, 2015. 2

[43] S. I. Olsen and A. Bertoli. Implicit non-rigid structure-from-motion with priors. *Journal of Mathematical Imaging and Vision*, 2008. 2

[44] L. Quan and T. Kanade. A factorization method for affine structure from line correspondences. In *Computer Vision and Pattern Recognition*, pages 803–808. IEEE, 1996. 2

[45] N. D. Reddy, P. Singhal, V. Chari, and K. M. Krishna. Dynamic body vslam with semantic constraints. In *Intelligent Robots and Systems*, pages 1897–1904. IEEE, 2015. 1

[46] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Computer Vision and Pattern Recognition*, June 2016. 1

[47] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 1

[48] L. Reyes and E. Bayro Corrochano. The projective reconstruction of points, lines, quadrics, plane conics and degenerate quadrics using uncalibrated cameras. *Image and Vision Computing*, 23(8):693–706, 2005. 2

[49] R. B. Rusu, N. Blodow, Z. C. Marton, and M. Beetz. Close-range scene segmentation and reconstruction of 3d point cloud maps for mobile manipulation in domestic environments. In *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, pages 1–6. IEEE, 2009. 1

[50] G. Schindler, P. Krishnamurthy, and F. Dellaert. Line-based structure from motion for urban environments. In *3D Data Processing, Visualization, and Transmission, Third International Symposium on*, pages 846–853. IEEE, 2006. 2

[51] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In *Computer Vision and Pattern Recognition*, 2016. 1

[52] T. Shao, W. Xu, K. Zhou, J. Wang, D. Li, and B. Guo. An interactive approach to semantic modeling of indoor scenes with an rgbd camera. *ACM Transactions on Graphics (TOG)*, 31(6):136, 2012. 1

[53] J. E. Solem, F. Kahl, and A. Heyden. Visibility constrained surface evolution. In *Computer Vision and Pattern Recognition*, 2005. 2

[54] S. Song, S. P. Lichtenberg, and J. Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 567–576, 2015. 1

[55] M. Sung, V. G. Kim, R. Angst, and L. Guibas. Data-driven structural priors for shape completion. *ACM Transactions on Graphics (TOG)*, 34(6):175, 2015. 1

[56] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999. 5

[57] R. Toldo, R. Gherardi, M. Farenzena, and A. Fusiello. Hierarchical structure-and-motion recovery from uncalibrated images. *Computer Vision and Image Understanding*, 140:127 – 143, 2015. 1

[58] L. Torresani, A. Hertzmann, and C. Bregler. Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors. *Transactions on Pattern Analysis and Machine Intelligence*, 2008. 2

[59] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. v. d. Hengel. Visual question answering: A survey of methods and datasets. *arXiv preprint arXiv:1607.05910*, 2016. 1