Planar Structure-from-Motion with Affine Camera Models: Closed-Form Solutions, Ambiguities and Degeneracy Analysis

Toby Collins, *Member, IEEE,* and Adrien Bartoli

Abstract

Planar Structure-from-Motion (SfM) is the problem of reconstructing a planar object or surface from 2D images using motion information. This is well-understood with the perspective camera model and can be solved with Homography Decomposition (HD). However when the structure is small and/or viewed far from the camera the perspective effects diminish, and in the limit the projections become affine. In these situations HD fails because the problem itself becomes ill-posed. We propose a stable alternative using affine camera models. These have been used extensively to reconstruct non-planar structures, however the problem is fundamentally different with planar structures because the affine camera models one can use are more restricted and it is inherently more ambiguous and non-linear. We provide a general, accurate and closed-form solution for the orthographic camera model, which returns all metric structure solutions and camera poses. This does not require initialisation and optimises an objective function that is very similar to the reprojection error. In fact there is no clear benefit in refining its solutions with bundle adjustment, which is a significant result. We also present a new theoretical analysis that deepens our understanding of the problem. The main result is a complete geometric characterisation of degeneracies with the orthographic camera. We also show there can exist up to two metric structure solutions with four or more images (previously it was assumed to be unique), and we give the necessary and sufficient geometric conditions for disambiguation. Other theoretical results include showing that in the case of three images the optimal reconstruction (with respect to reprojection error) can usually be found in closed-form, and additional prior knowledge needed to solve with non-orthographic affine cameras.

Index Terms

Structure-from-Motion, factorization, stratification, critical motion, degeneracy, ambiguity, plane, orthographic, weak-perspective, para-perspective.

I. Introduction

A. Context and Motivation

The development and analysis of closed-form solutions to Structure-from-Motion (SfM) problems is an ongoing objective in computer vision that spans research over several decades [1], [2], [3], [4], [5], [6], [7], [8]. SfM is the problem of finding the 3D structure of a scene and the pose of cameras imaging the scene using 2D motion information in the camera images. The most common form of SfM is when motion comes from point correspondences, which can be

This research has received funding from the EU's FP7 ERC research grant 307483 FLEXABLE and Almerys Corporation

computed by matching pairs of views or by tracking points in a video. This is often the precursor to dense SfM where pixel-level photoconsistency is used to determine dense structure. SfM foremost requires a camera projection model. The Perspective projection model is the most common and can accurately model most real cameras. Affine projection models have also been used extensively and can be very accurate for certain scenes. They have been used for rigid SfM with non-planar structures [2], [9], [10], [3], [11], rigid multi-body and articulated objects [12], [13] and deformable objects [14], [15]. Traditionally affine projection models are used to simplify SfM because unlike perspective projection, they have linear projection functions. However there are several other important reasons to use them. The first is that one cannot solve the problem reliably with perspective projection for planar structures that are small and far from the camera, because the problem becomes badly-conditioned. The second reason is that the theoretical analysis of SfM with affine camera projection is important to understand degeneracies and ambiguities for both affine *and* perspective cameras. This is because perspective cameras behave like affine cameras when the perspective effects are small, and in the limit they become affine cameras. Any analysis of degeneracies or ambiguities with affine cameras is therefore important to complete our understanding of SfM with the perspective camera, because as the perspective effect diminishes and the effect of noise increases we will know which scene configurations become unstable or ambiguous. Our work is also relevant to nonrigid SfM, where the problem can be solved by dividing an object's surface into local planar regions and performing rigid SfM on each region [16], [15]. Because these regions need to be small one must work with affine or quasi-affine projections.

B. Existing Approaches

When the structure is non-planar it is possible to solve SfM with affine cameras uniquely up to scale and a global coordinate transform [9]. This can be done in closed-form using *factorisation-based stratification* [2], [9], [10], [3], [11], which works by stacking the correspondences to form a correspondence matrix with a theoretical rank of three and then computing the scene's affine reconstruction by a rank-three factorisation of the matrix. The factorisation is not unique but up to a non-singular 3×3 matrix known as the *upgrade matrix*. This can be resolved using orthogonality constraints from the camera projection matrices, and is found by solving a system of linear upgrade equations. However, these methods fail when the structure is planar because these upgrade equations cannot sufficiently constrain the problem (see Appendix A for details). Indeed, the nature of the problem is drastically different to SfM with non-planar structures. The types of affine projection models one can use are more restricted and the problem is inherently much more ambiguous.

Unlike affine cameras, solutions to planar SfM and perspective cameras are well established [17], [18], [19]. When the camera intrinsics are known the problem can be solved with only two images, which has a closed-form solution by factorising the inter-view homography matrix [17], [18]. This has a two-fold solution ambiguity in general which can be resolved with a third view. To date there are no closed-form solutions that are optimal in terms of reprojection error for a general number of images. Usually this is solved with local gradient-based optimisation with bundle adjustment [20], [21] which is initialised by computing relative poses between pairs of views and chaining them to a common coordinate system [21]. These solutions work well unless the plane's projection is quasi-affine (caused by the plane being small or viewed far from the camera), which makes the problem badly-conditioned.

Various affine camera models exist in the literature. The most common are the orthographic, weak-perspective and para-perspective (the latter being the most general model). We refer to the problem of Planar SfM with the Orthographic model by *PSfM-O*, the Weak-perspective model by *PSfM-W* and the Para-perspective model by *PSfM-PP*. It is possible to solve PSfM-O (up to discrete ambiguities) using only motion information [22]. However SfM-WP or PSfM-PP cannot be solved without additional information, because the reprojection constraints are insufficient to resolve all of the camera Degrees-of-Freedom (DoFs) (see §II-D). There are some previous methods that solve PSfM-O but only in special configurations. Solutions for the case of three views of three non-colinear points were presented in [23], [22]. However they are of limited practical use because they require the reprojection constraints to be satisfied exactly, and often fail to return real-valued solutions when there is noise. A closed-form solution to PSfM-O was presented in [15] that can handle noise and more than three views. This however solves a convex relaxation of the problem, and only when there are three points and four or more views. Furthermore it cannot handle cases where there is more than one structure solution. The algorithms in [23], [22], [15] also have flaws in their design because they introduce artificial degeneracies (see Appendix B).

C. Gauge Transforms, Ambiguities, Degeneracies, Well-posedness and Optimal Solutions

We review here several important SfM concepts. A *metric reconstruction* is a reconstruction of the scene up to scale and a global coordinate transform (also known as a gauge transform). For problems involving affine cameras the gauge transform is in general a rigid transform plus reflection [1]. An *ambiguity* occurs when there exists more than one metric reconstruction that can exactly satisfy the image measurements when noise is removed and the gauge transform has been fixed. Ambiguities can either be *continuous*, in which case there exist an infinite number of solutions or *discrete*, in which case there exist a finite number of solutions. We also divide ambiguities into *structure ambiguities* and *camera resection ambiguities*. A structure ambiguity is when there exists more than one structure solution and a camera resection ambiguity is when there exists more than one camera pose solution for a given structure solution. In general we can break down the causes of ambiguities into four groups. These are *critical structures*, *critical motion sequences*, *missing measurements* and *mixed*. Critical structures are when the ambiguity is caused by the structure being in a particular configuration. Critical motion sequences are when the ambiguity is caused by the camera poses being in a particular configuration. Missing measurement ambiguities are when the ambiguity is caused by one or more views having missing correspondences. Mixed ambiguities are when the ambiguity is caused by a particular combination of structure, camera poses and missing measurements.

Unlike most other SfM problems, planar SfM with affine camera models *always* has discrete ambiguities (see §II-D). Therefore the problem is never well-posed in the usual sense since it never has a unique solution. Instead, we say that the problem is *well-posed* if it can be solved up to a discrete number of solutions. Otherwise we say the problem is *degenerate*. An *artificial degeneracy* is when the problem is well-posed but a particular algorithm cannot solve it due to its design. If an algorithm is guaranteed to not introduce artificial degeneracies we call it a *Non-Artificially Degenerate Algorithm* (NADA). We say that a solution is *optimal* if it minimises the reprojection error. This is also the solution which is statistically optimal by assuming zero-mean IID Gaussian measurement noise. For point correspondences this noise model has been demonstrated many times to be a good choice [5].

D. Technical Contributions

We present a fast, closed-form and stratified approach to solve PSfM-O. The general process is as follows. First the plane's 2D affine structure is recovered from point correspondences. When all points have correspondences this can be done optimally in terms of reprojection error by a rank-two SVD. We then solve globally a set of non-convex upgrade constraints to determine all solutions for the plane's metric structure. Finally, for each solution the corresponding camera poses are recovered by an optimal plane-based pose estimation process. We present two variants of our approach, which serve two different purposes. The first, which we call *Approximate PSfM-O* solves the upgrade constraints in a least-squares sense, and is the method we use in practice. The second, which we call *Exact PSfM-O* solves the upgrade constraints exactly, and is mainly used to answer core theoretical questions. We list here some important properties of Approximate PSfM-O:

1) Approximate PSfM-O solves the general PSfM-O problem. This is when there are three or more views, the structure has three or more points and when there are missing correspondences.
2) Approximate PSfM-O generates all solutions when there is no exact physical interpretation of the data due to noise or modelling approximation error. This has not been achieved before for general scenes.
3) Extensive empirical evaluation shows that there is usually no noticeable benefit in using Bundle Adjustment (BA) to refine the solutions from Approximate PSfM-O. This is a remarkable result because Approximate PSfM-O does not optimize the full reprojection error, but rather an approximation of it. In the special case of three points our solutions are consistently more accurate than [15].

We also extend our approach to solve special cases of PSfM-W and PSfM-PP. Specifically, if within the set of views we know three or more views where the depth of the structure along the camera's projection direction is similar, PSfM-W can be approximated by an PSfM-O problem and solved/analysed with our approach. If we also have a perspective intrinsic calibration then PSfM-PP can also be approximated by an PSfM-O problem and solved/analysed with our approach.

E. Theoretical Contributions

Our theoretical contributions are presented in §II-E as eight new theorems. Our main contribution is to give the necessary and sufficient geometric conditions for PSfM-O to be degenerate given complete measurements (Theorem 1). To achieve this one must geometrically characterise all degeneracies and prove that the list is exhaustive. Our second main theoretical contribution is to show that for a general number of orthographic views there can exist up to two solutions for the plane's metric structure (previously it was assumed to be unique [15]). We give the necessary and sufficient geometric conditions to disambiguate structure with extra views in Theorem 2. We then extend these theorems to incomplete measurements in Theorems 3 and 4. Other theorems (Theorems 5 to 8) give some important theoretical guarantees for Exact PSfM-O, and extra knowledge necessary to solve with other affine cameras.

F. Paper Structure

In §II we give a notation guide, further background and the new theorems. In §III we present Exact PSfM-O and Approximate PSfM-O. We then generalise them to solve certain cases of PSfM-W and PSfM-PP. We finish §III with a table summarising the key differences between SfM with affine cameras and planar versus non-planar structures. In §IV we examine the performance of Exact PSfM-O and Approximate PSfM-O compared to other methods using simulated and real image data. In §V we conclude and discuss future research directions. In Appendices H and J we prove all theorems given in §II.

II. Background and New Theorems

A. Notation and Problem Setup

Vectors and matrices are in bold, scalars are in regular italic and sets are in upper-case calligraphic. We use $[\mathbf{A}]_{kl}$ to denote the element at row k, column l of a matrix $\mathbf{A}$, and $[\mathbf{A}]_{k\times l}$ to denote its top-left $k \times l$ submatrix. We use v_k to denote the k^{th} element of a vector $\mathbf{v}$. We use $\mathbf{0}_{K\times L}$ and $\mathbf{1}_{K\times L}$ to denote the all-zeros and all-ones $K \times L$ matrices and $\mathbf{I}_K$ to denote the $K\times K$ identity matrix. We use $\hat{\mathbf{A}}$ to denote an estimate of a matrix $\mathbf{A}$ from noisy measurements. Given a set of matrices $\mathcal{A} = \{\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_M\}$ where all $\mathbf{A}_i$ have the same width, we define $\operatorname{stack}(\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_M) = \left[\mathbf{A}_1^\top\, \mathbf{A}_2^\top \ldots \mathbf{A}_M^\top\right]^\top$ to be the operator that stacks $\mathcal{A}$ row-wise. We define $\operatorname{unstack}(\left[\mathbf{A}_1^\top\, \mathbf{A}_2^\top \ldots \mathbf{A}_K^\top\right]^\top) = \mathcal{A}$ to be the operator that unstacks the matrices. We use $s_k(\mathbf{A})$ to denote the k^{th} largest singular value of a matrix $\mathbf{A}$. We use $\lambda_k(\mathbf{F})$ to denote the k^{th} largest eigenvalue of a square symmetric matrix $\mathbf{F}$ and $\mathcal{V}_k(\mathbf{F})$ to denote the set of unit eigenvectors of $\mathbf{F}$ with eigenvalue $\lambda_k(\mathbf{F})$. We use $\mathcal{S}_{2\times 3}$ to denote the 2×3 Stiefel manifold (*i.e.* $\mathbf{M} \in \mathcal{S}_{2\times3} \Leftrightarrow \mathbf{M}\mathbf{M}^\top = \mathbf{I}_2$). We use $\mathcal{SS}_{2\times2}$ to denote the 2×2 sub-Stiefel manifold in $\mathcal{S}_{2\times3}$ (*i.e.* $\mathbf{A} \in \mathcal{SS}_{2\times2} \Leftrightarrow \exists \mathbf{M} \in \mathcal{S}_{2\times3}$ s.t. $\mathbf{A} = [\mathbf{M}]_{2\times2}$). We use $\mathcal{G}_{2\times2}$ to denote the Gramian of $\mathcal{SS}_{2\times2}$ (*i.e.* $\mathbf{G} \in \mathcal{G}_{2\times2} \Leftrightarrow \exists \mathbf{A} \in \mathcal{SS}_{2\times2}$ such that $\mathbf{A}^\top\mathbf{A} = \mathbf{G}$). The spectral definitions of $\mathcal{SS}_{2\times2}$ and $\mathcal{G}_{2\times2}$ are also used:

$$\begin{aligned}&\mathbf{A} \in \mathcal{SS}_{2\times2} \Leftrightarrow s_1(\mathbf{A}) = 1 \Rightarrow (s_2(\mathbf{A}) = |\det(\mathbf{A})|)\\ &\mathbf{G} \in \mathcal{G}_{2\times2} \Leftrightarrow (\mathbf{G} \in \mathcal{SS}_{2\times2},\ \ \mathbf{G} \succeq \mathbf{0}) \Leftrightarrow (s_1(\mathbf{G}) = \lambda_1(\mathbf{G}) = 1,\ \ s_2(\mathbf{G}) = \lambda_2(\mathbf{G}) = \det(\mathbf{G}))\end{aligned} \tag{1}$$

The scene geometry is illustrated in Figure 1. We use M to be the number of views in the scene indexed by $i \in \{1, 2, \ldots, M\}$. We use N to be the number of points in the scene indexed by $j \in \{1, 2, \ldots, N\}$. We define the *structure plane* to be the support plane of the structure points in world coordinates at $z = 0$. We use $\mathbf{S} \in \mathbb{R}^{3\times N}$ to be the unknown *structure matrix* that holds the j^{th} structure point $\mathbf{s}_j$ in its j^{th} column (with its third row being all-zeros). Without loss of generality we define the centroid of the structure points at the origin of world coordinates. We use $\mathbf{V} \in \{0, 1\}^{M\times N}$ to be a binary visibility matrix where $[\mathbf{V}]_{ij} = 1$ if we have a correspondence for point j in view i, and $[\mathbf{V}]_{ij} = 0$ otherwise.

The polar coordinates of a 3D vector $\mathbf{a} \in \mathbb{R}^3$ in world coordinates are given by:

$$\mathbf{a} = k\left[\sin\theta\cos\phi\ \ \sin\theta\sin\phi\ \ \cos\theta\right]^\top \tag{2}$$

where $\theta \in [0, \pi]$ is the inclination angle, $\phi \in [0, 2\pi]$ is the azimuth angle and $k = \|\mathbf{a}\|_2$ is the length. The inclination angle is the angle between $\mathbf{a}$ and the structure plane. The azimuth angle is the anticlockwise angle between the projection of $\mathbf{a}$ onto the structure plane and the x-axis of world coordinates.

B. Affine Cameras

The affine projection of a 3D point $\mathbf{s} \in \mathbb{R}^3$ in world coordinates to its 2D position $\mathbf{q} \in \mathbb{R}^2$ in image i is given by $\mathbf{q} = \mathbf{M}_i \operatorname{stack}(\mathbf{s}, 1)$ where $\mathbf{M}_i$ is the 2×4 projection matrix. There are a few ways to geometrically interpret an affine camera projection. A common way is to consider it as an orthographic camera whose image is warped by a 2D linear transform. This is equivalent to the decomposition

$$\mathbf{M}_i = k_i\mathbf{A}_i \begin{bmatrix} \mathbf{I}_2 & \mathbf{0}_{2\times 2} \end{bmatrix} \begin{bmatrix} \mathbf{R}_i & \mathbf{t}_i \\ \mathbf{0}_{1\times 3} & 1 \end{bmatrix} = k_i\mathbf{A}_i \left[[\mathbf{R}_i]_{2\times 3}\ [\mathbf{t}_i]_{2\times 1}\right] \tag{3}$$

The terms $\mathbf{R}_i \in \mathcal{S}_{3\times 3}$ and $\mathbf{t}_i \in \mathbb{R}^3$ are the extrinsic 3D rotation and translation of the orthographic camera. The term $k_i \in \mathbb{R}^+$ is the camera's magnification factor and $\mathbf{A}_i$ is a upper-triangular full-rank 2×2 matrix with $[\mathbf{A}_i]_{11} = 1$. An affine camera has a single *projection direction* $\mathbf{a}_i \in \mathbb{R}^3$ which is given in world coordinates by the nullspace of $[\mathbf{M}_i]_{2\times 3}$. This is equivalent to the third row of $\mathbf{R}_i$. An affine camera's *projection plane* is a plane whose normal is colinear to the projection direction.

Different affine cameras are obtained according to the DoFs of k_i and $\mathbf{A}_i$. Three main ones are the orthographic camera where $\mathbf{A}_i = \mathbf{I}_2$ and $k_i = k$ is fixed across all views for some $k \in \mathbb{R}^+$ (with five free DoFs), the weak-perspective camera where $\mathbf{A}_i = \mathbf{I}_2$ and k_i can vary between views (with six free DoFs), and the para-perspective camera where k_i and the two free terms in $\mathbf{A}_i$ can vary between views (with eight free DoFs). There exist other affine camera definitions, for example [11] proposed a seven DoF affine camera that was designed to best mimic perspective projection.

Unlike perspective cameras it is difficult to separate intrinsic from extrinsic parameters with affine cameras. Sometimes the combined term $k_i\mathbf{A}_i$ is called the *intrinsic matrix* [9], however this can be misleading because we cannot usually calibrate k_i. This is because unless the true camera's projection directions are perfectly parallel k_i increases as the depth of structure (which is unknown) decreases, so it is more like an extrinsic. For para-perspective cameras $\mathbf{A}_i$ can be considered as an intrinsic because it can be calibrated without knowing the depth of the structure. This has been done previously for non-planar structures using either a perspective intrinsic calibration or by self-calibration [9], [11].

C. Reprojection Error, Camera Resection and Resection Ambiguities with Planar Structures and Affine Cameras

a) Reprojection error and the camera resection problem: Following Equation (3), a structure point $\mathbf{s} \in \mathbb{R}^3$ in world coordinates on the plane $z = 0$ projects to image i with $\mathbf{q} = \mathbf{P}_i\text{stack}([\mathbf{s}]_{2\times 1}, 1)$ where $\mathbf{P}_i$ is a 2D-to-2D affine transform given by

$$\mathbf{P}_i \overset{\text{def}}{=} \begin{bmatrix} k_i\mathbf{A}_i[\mathbf{R}_i]_{2\times 2} & [\mathbf{t}_i]_{2\times 1} \end{bmatrix} \tag{4}$$

We can specify $\mathbf{P}_i$ with a particular camera model. For the three main types we have:

$$\mathbf{P}_i = \begin{cases} \begin{bmatrix} k[\mathbf{R}_i]_{2\times 2} & [\mathbf{t}_i]_{2\times 1} \end{bmatrix}, & \begin{bmatrix} k_i[\mathbf{R}_i]_{2\times 2} & [\mathbf{t}_i]_{2\times 1} \end{bmatrix}, & \begin{bmatrix} k_i \begin{bmatrix} 1 & \gamma_i \\ 0 & \beta_i \end{bmatrix} [\mathbf{R}_i]_{2\times 2} & [\mathbf{t}_i]_{2\times 1} \end{bmatrix} \\ \text{(Orthographic)} & \text{(Weak-Perspective)} & \text{(Para-Perspective)} \end{cases} \tag{5}$$

With the orthographic, weak and para-perspective cameras $\mathbf{P}_i$ has five, six and eight free DoFs respectively. Given a set of noisy correspondences $\{\hat{\mathbf{q}}_i^j\}$ and a visibility matrix $\mathbf{V}$, the reprojection error writes as follows:

$$E_{reproj} = \sum_{i=1}^{M}\sum_{j=1}^{N}[\mathbf{V}]_{ij}\|\mathbf{P}_i\text{stack}\left([\mathbf{s}_j]_{2\times 1}, 1\right) - \hat{\mathbf{q}}_i^j\|_2^2 \tag{6}$$

Camera resection is the problem of computing the camera matrices $\{\mathbf{P}_i\}$ given an estimate of the scene's structure $\{\mathbf{s}_j\}$. We say the resection is optimal if it minimises the reprojection error. We discuss resection here because it is equivalent to solving for each view the *plane-based pose estimation problem*, for which a number of solutions exist [24], [25], [26]. One should ask wither it is ever possible to recover the depth component of $\mathbf{t}_i$. For weak and para-perspective cameras it is possible when the magnification factor k_i can be recovered, because it is approximately inversely proportional to the depth of the camera. This can be seen by interpreting the weak and para-perspective cameras as linearised perspective cameras (see Appendix C). By contrast, one cannot estimate the depth of an orthographic camera because the magnification factor k is constant.

b) Plane-based pose estimation with weak and para-perspective cameras: Globally optimal solutions exist for weak-perspective cameras [24], [25]. This is done by first estimating $\mathbf{P}_i \approx \hat{\mathbf{P}}_i$ by the least-squares 2D affine transform from $\{\mathbf{s}_j\}$ to $\{\hat{\mathbf{q}}_j^i\}$. This is then exactly factored according to Equation (5). The factorisation is unique if and only if $\{\mathbf{s}_j\}$ is not colinear [24]. We then recover the full rotation matrix $\mathbf{R}_i$ from $[\mathbf{R}_i]_{2\times 2}$, which has two solutions, leading to a per-view two-fold camera resection ambiguity. Geometrically it is equivalent to an arbitrary reflection of the structure plane about each camera's projection plane [25]. Plane-based pose estimation with para-perspective cameras is only possible if two or more terms in $\{k_i, \beta_i, \gamma_i\}$ are known. If this is not the case the camera has seven or more unknown DoFs, which are not sufficiently constrained by the six coefficients of $\mathbf{P}_i$. The most common instance is when k_i is unknown but β_i and γ_i are known. This is because k_i varies inversely to depth, whereas β_i and γ_i can be computed directly from the 2D correspondences [25]. The problem can be solved in closed-form by factorising $\hat{\mathbf{P}}_i$ similarly to the weak-perspective camera.

c) Plane-based pose estimation with orthographic cameras: Optimal plane-based pose estimation with orthographic cameras is less easy than weak and para-perspective cameras, since it cannot be done by factorising $\hat{\mathbf{P}}_i$. This is because with noise $\hat{\mathbf{P}}_i$ has six DoFs, but the orthographic camera has only five view-dependent DoFs. Recall that in SfM with orthographic cameras k cannot be recovered because of the scale ambiguity (we cannot distinguish a larger structure from a larger k), so it must be fixed. We are then left with estimating the camera rotations and translations. We have not seen a globally-optimal solution to this before ([15] is the closest work but used gradient-based local optimisation with Levenberg-Marquardt). We have developed one using the fact that the problem is a small-scale Generalized Problem of Moments and can be solved globally with *e.g.* Gloptipoly [27]. We consider this a minor contribution and provide the details in Appendix D. Note that similarly to the weak and para-perspective cameras there are in general *two* optimal camera poses per view.

D. Known Degeneracies and Ambiguities in Planar SfM with Affine Cameras

The problem is always degenerate if each camera has *six* or more unknown DoFs. Therefore it is degenerate with weak-perspective, para-perspective and axial symmetric cameras [11] unless we have some additional constraints. The reason is because there are insufficient reprojection constraints to resolve both metric structure and the camera poses, which is evident by parameter counting: For each view the structure plane-to-image transform $\mathbf{P}_i$ gives up to six constraints (because it is a 2D affine transform), so for M views we have $6M$ constraints. If each camera has six or more unknown DoFs then there are insufficient constrants to resolve their $6M$ DoFs *and* the plane's metric structure. The situation is different for orthographic cameras because they have only five view-dependent DoFs rather than six. This gives us redundancy for determining the plane's metric structure.

There is some knowledge about degenerate scenes with orthographic cameras but it is very incomplete. It has a trivial critical structure which is when the structure points are colinear [22]. This causes continuous camera resection ambiguities, since each camera is free to rotate about the structure line. A critical motion sequence was found in the related problem of Shape-from-Texture with orthographic cameras [28]. This happens when the camera projection directions all lie on a plane that is orthogonal to the structure plane. However the necessary and sufficient degeneracy conditions have not been established. Discrete structure ambiguities are known to exist in PSfM-O but have only been identified in three-view scenes. There are in general two structure solutions with three views [22], which leads to a maximum of 16 scene interpretations (2^3 possible camera poses for either structure). Unlike camera resection ambiguities, very little is known geometrically about these structure ambiguities. The geometric relationship between them has not been established, nor the requirement for disambiguating structure.

E. New Theorems

1) Full Geometric Characterisation of Degenerate Scenes in PSfM-O: Recall that a degenerate scene in PSfM-O is one where we cannot solve the problem up to a discrete number of solutions given complete measurements.

Theorem 1. *A scene is degenerate if and only if at least one of three geometric conditions are satisfied. The first condition is when structure is colinear. This is the only critical structure in PSfM-O. The second condition is a critical motion sequence which is when the camera projection directions lie on a plane that is orthogonal to the structure plane (Figure 1). Equivalently, in terms of polar coordinates, this is when all camera projection directions have the same azimuth. This is the only critical motion sequence in PSfM-O. The third condition is when there are fewer than three cameras whose projection directions are unique up to reflection about the structure plane and change of sign. There are no mixed degeneracies in PSfM-O between camera poses and structure.*

Corollary 1. *We can state Theorem 1 equally in terms of non-degenerate scenes by negating the implication using De Morgan's negation. A scene in PSfM-O is non-degenerate if and only if the structure points are non-colinear, at least one of the camera projection directions has a different azimuth to the other projection directions, and there are three or more cameras with projection directions that are unique up to reflection about the structure plane and change of sign.*

2) Structure Uniqueness in PSfM-O given Four or More Views and Complete Measurements:

Theorem 2. *Recall that PSfM-O with three views has at most two solutions for the plane's metric structure. Suppose the scene has three views $i \in \{1, 2, 3\}$, is non-degenerate and has two such solutions. Given an additional orthographic view $i = 4$ we can disambiguate structure if and only if the camera projection directions $\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \mathbf{a}_4\}$ are not in a special configuration when projected onto the structure plane. Specifically the outer product of $[\mathbf{a}_4]_{2\times1}$ must not be an affine combination of the outer products of $[\mathbf{a}_1]_{2\times1}$, $[\mathbf{a}_2]_{2\times1}$ and $[\mathbf{a}_3]_{2\times1}$. Formally, structure can be disambiguated if and only if:*

$$\nexists \alpha, \beta \in \mathbb{R} \text{ s.t. } [\mathbf{a}_4]_{2\times1}[\mathbf{a}_4]_{2\times1}^\top = \alpha[\mathbf{a}_1]_{2\times1}[\mathbf{a}_1]_{2\times1}^\top + \beta[\mathbf{a}_2]_{2\times1}[\mathbf{a}_2]_{2\times1}^\top + (1-\alpha-\beta)[\mathbf{a}_3]_{2\times1}[\mathbf{a}_3]_{2\times1}^\top \tag{7}$$

In general, given any number of additional orthographic views, structure can be disambiguated if and only if Equation (7) holds for at least one of the additional views.

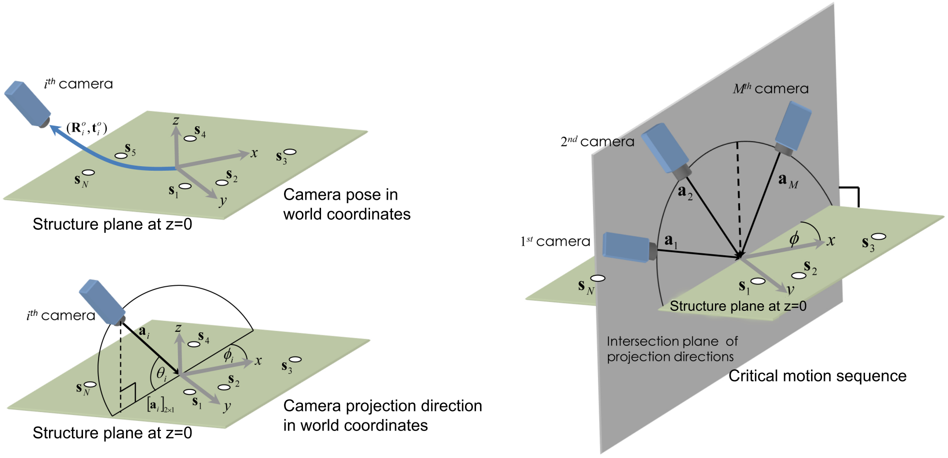

Fig. 1. Scene geometry with orthographic cameras and a planar structure (left top and left bottom) and the critical motion sequence (right). Vector $\mathbf{a}_i$ denotes the projection direction of the i^{th} camera. The critical motion sequence that causes the problem to be ill-posed (right).

3) Generalisation of Theorems 1 and 2 to Missing Measurements: PSfM-O problems with missing measurements can be partitioned into two types. The first (Type 1) are those where we can complete the rank-two correspondence matrix from the incomplete measurements. The second (Type 2) are those where we cannot. Type 1 problems are equivalent to those where we can compute the structure's 2D affine reconstruction, and Type 2 problems to those where we cannot.

Theorem 3. *Type 1 problems are degenerate if and only if the equivalent problem with complete measurements is degenerate. Therefore Type 1 problems do not have missing measurement degeneracies. Type 2 problems are always degenerate.*

Theorem 4. *Suppose the scene has three views, is non-degenerate and has two solutions for the plane's metric structure. If there is at least one additional view which has three or more correspondences that are non-colinear on the structure plane and Equation (7) holds, then we can disambiguate structure. If there is at least one additional view which has two correspondences and Equation (7) holds then it may be possible to disambiguate structure from the foreshortening effect. If all additional views have only one point correspondence then we cannot disambiguate structure.*

4) Theoretical Guarantees of Exact PSfM-O: Recall that Exact PSfM-O is our solution to PSfM-O for three views that satisfies a set of upgrade constraints exactly. Unlike previous solutions for three views [23], [22], Exact PSfM-O handles a general number of points (three or more) and has the following guarantees:

Theorem 5. *In the absence of noise Exact PSfM-O fails to find a metric reconstruction if and only if the scene configuration is degenerate. Therefore Exact PSfM-O is NADA.*

Theorem 6. *In the presence of noise, assume we have the structure's optimal affine reconstruction (which can be computed in closed-form when there are no missing measurements). If this can be upgraded by Exact PSfM-O to a metric reconstruction, then Exact-PSfM-O finds all optimal metric reconstructions.*

Theorem 5 is important for two reasons. The first is that Exact PSfM-O will never fail for problems that are theoretically solvable. The second is that it allows us to systematically characterise all degeneracies, by geometrically interpreting all inputs that cause Exact PSfM-O to fail. Theorem 6 is important because it tells us that the optimal solutions (those that minimise the reprojection error) for three views can be found in closed-form using Exact PSfM-O. This is not true in all cases: Exact PSfM-O will not have a solution if the affine reconstruction cannot be exactly upgraded to a metric reconstruction. In practice we find that the likelihood of Exact PSfM-O having solutions is typically between 80-90% of the time depending on the level of noise. Consequently, Exact PSfM-O is able to solve the problem optimally between 80-90% of the time.

5) Generalisation of Theorems to Other Affine Cameras:

Theorem 7. *Recall that we cannot solve SfM with affine cameras and planar structure if each camera has six or more unknown DoFs (see §II-D). Therefore without additional constraints we cannot solve with the weak-perspective, para-perspective or axially symmetric [11] affine cameras. From the affine camera interpretation in Equation (3) this is equivalent to saying that we cannot solve the problem if the magnification factors k_i are free DoFs and/or the terms in $\mathbf{A}_i$ are free DoFs. However if there exist dependencies it may be possible to solve the problem. Three interesting cases are as follows:*

- *Case 1: We can isolate a subset $\mathcal{I}'$ of three or more views where $\mathbf{A}_{i\in\mathcal{I}'}$ is known and $k_{i\in\mathcal{I}'}$ is assumed to be constant.*
- *Case 2: We can isolate a subset $\mathcal{I}''$ of five or more views where $\mathbf{A}_{i\in\mathcal{I}''}$ and $k_{i\in\mathcal{I}''}$ are unknown and assumed to be constant.*
- *Case 3: We can isolate three or more pairs of views where for each pair $(i, i') \in \{1, 2, \ldots, M\}^2$, $\mathbf{A}_i$ and $\mathbf{A}_{i'}$ are known and we assume $k_i = k_{i'}$.*

In Case 1, the problem of upgrading affine to metric structure is constrained only by the views in $\mathcal{I}'$. In the absence of noise this is exactly equivalent to upgrading with orthographic cameras using only the views in $\mathcal{I}'$.

Theorem 8. *Suppose the cameras have been intrinsically calibrated with the perspective camera model and we can isolate a subset $\mathcal{I}'$ of three or more views where the distance between the camera and a planar structure is far and approximately constant. We can solve this SfM problem with weak or para-perspective cameras because this is an instance of Case 1 in Theorem 7.*

III. Stratified Planar SfM with Orthographic Cameras

A. Upgrade Constraints and Upgrade Parameterisation

We first consider the case when correspondences are measured in all views. When there are missing correspondences the upgrade constraints are the same but the way to compute the scene's affine reconstruction is different (see §III-B). The upgrade constraints in PSfM-O act on a rank-two factorisation of the *correspondence matrix* $\hat{\mathbf{Q}} \in \mathbb{R}^{2M\times N}$. We build $\hat{\mathbf{Q}}$ by zero-centering the correspondences in each image (which eliminates translation) and stacking them row-wise. This is similar to the correspondence matrix used in non-planar reconstruction with the key difference that for planar structures it has a maximum theoretical rank of two whereas for non-planar structures it has a maximum theoretical rank of three. This is seen by factorising $\hat{\mathbf{Q}}$ using Equation (4):

$$\hat{\mathbf{Q}} = \text{stack}\left([\mathbf{P}_1]_{2\times 2},\ [\mathbf{P}_2]_{2\times 2}, \ldots, [\mathbf{P}_\mathrm{M}]_{2\times 2}\right)[\mathbf{S}]_{2\times N} + \varepsilon \tag{8}$$

where ε denotes correspondence noise. The two factors can be recovered up to noise and an unknown 2×2 upgrade matrix $\mathbf{X}$ using a rank-two SVD of $\hat{\mathbf{Q}}$: $\hat{\mathbf{Q}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$. This gives

$$\hat{\mathbf{Q}} \approx \mathbf{M}^A\mathbf{S}^A,\ \mathbf{M}^A \overset{\text{def}}{=} [\mathbf{U}]_{2M\times 2}[\mathbf{\Sigma}]^{1/2}_{2\times 2},\ \mathbf{S}^A \overset{\text{def}}{=} [\mathbf{\Sigma}]^{1/2}_{2\times 2}[\mathbf{V}]^\top_{N\times 2} \tag{9}$$

where $\text{stack}\left([\mathbf{P}_1]_{2\times 2},\ [\mathbf{P}_2]_{2\times 2}, \ldots, [\mathbf{P}_\mathrm{M}]_{2\times 2}\right) \approx \mathbf{M}^A\mathbf{X}$ and $[\mathbf{S}]_{2\times N} \approx \mathbf{X}^{-1}\mathbf{S}^A$. We now instantiate $\{\mathbf{P}_i\}$ with the orthographic camera to obtain constraints on $\mathbf{X}$. From Equation (5), in the absence of noise each view provides the following metric constraint:

$$[\mathbf{R}_i]_{2\times 2} = \mathbf{M}^A_i\mathbf{X} \Leftrightarrow \mathbf{M}^A_i\mathbf{X} \in \mathcal{SS}_{2\times 2} \Leftrightarrow s_1\left(\mathbf{M}^A_i\mathbf{X}\right) = 1 \tag{10}$$

where $\mathbf{M}^A_i$ denotes the i^{th} 2×2 sub-block of $\mathbf{M}^A$. We can also express this constraint in terms of the Gramian matrix $\mathbf{W} \overset{\text{def}}{=} \mathbf{X}\mathbf{X}^\top$. From Equation (1) we have $\mathbf{M}^A_i\mathbf{X} \in \mathcal{SS}_{2\times 2} \Leftrightarrow \mathbf{M}^A_i\mathbf{W}\mathbf{M}^{A\top}_i \in \mathcal{G}_{2\times 2}$. Therefore the equivalent constraint on $\mathbf{W}$ is

$$\lambda_1\left(\mathbf{M}^A_i\mathbf{W}\mathbf{M}^{A\top}_i\right) = 1 \tag{11}$$

We refer to Equation (11) as the *PSfM-O upgrade constraint*. This provides *one non-convex* equality constraint per view on a 2×2 positive definite upgrade matrix $\mathbf{W}$. By contrast, the upgrade constraint for non-planar structures (Equation (31)) provides *three linear* equality constraints per view on a 3×3 positive definite upgrade matrix.

Given $\mathbf{W}$ we can recover $\mathbf{X}$ from its Cholesky decomposition. We parameterise $\mathbf{W}$ with a three-vector $\mathbf{w}$ as follows:

$$\mathbf{W} = f(\mathbf{w}),\quad f(\mathbf{w}) \overset{\text{def}}{=} \begin{bmatrix} w_1 & w_2 \\ w_2 & w_3 \end{bmatrix},\ w_1 > 0,\ w_3 > 0,\ w_1w_3 - w_2^2 > 0 \tag{12}$$

where the inequality constraints enforce positive definiteness. We then recover $\mathbf{X}$ from $\mathbf{w}$ with

$$\mathbf{X} = \begin{bmatrix} \sqrt{w_1 - w_2^2/w_3} & w_2/\sqrt{w_3} \\ 0 & \sqrt{w_3} \end{bmatrix} \tag{13}$$

This gives $\mathbf{X}$ up to an arbitrary 2D unitary gauge transform $\mathbf{U} \in \mathcal{S}_{2\times 2}$ because for any $\mathbf{U} \in \mathcal{S}_{2\times 2}$, $\mathbf{W} = (\mathbf{X}\mathbf{U})(\mathbf{X}\mathbf{U})^\top$.

B. Missing Correspondences

When there are missing correspondences we cannot factorise $\hat{\mathbf{Q}}$ straighforwardly with the SVD. For non-planar structures the usual strategy is to apply heuristics to obtain an initial factorisation and then to refine it either by gradient-based optimisation or alternation. For planar structures we can use the fact that the 2D-to-2D inter-image transform $\mathbf{P}_{ij} \overset{\text{def}}{=} \mathbf{P}_j\,\text{stack}(\mathbf{P}_i, [0,0,1])^{-1}$ from view i to view j is a 2D affine transform, so it can be used to fill-in missing correspondences in view j by transferring correspondences from view i. By chaining views we can complete $\hat{\mathbf{Q}}$ and factor it with the SVD. This factorisation is not optimal in terms of reprojection error, and can usually be improved by gradient-based refinement. The specific algorithm we use is given in Appendix F.

C. Exact PSfM-O: An Exact Stratified Solution to PSfM-O for Three views

1) Method Overview: Exact PSfM-O solves PSfM-O with three views by satisfying the upgrade constraint in Equation (11) exactly. Given the 2D affine camera factor $\mathbf{M}^A$ (which is 6×2 for three views), Exact PSfM-O solves the following upgrade problem:

$$\boxed{\begin{array}{l}\text{Exact\_PSfM-O\_upgrade}(\mathbf{M}^A)\\ \text{find}\,\mathbf{w}\in\mathbb{R}^3\ \text{ s.t.}\\ \begin{cases}\lambda_1\left(\mathbf{M}_i^A f(\mathbf{w})\mathbf{M}_i^{A\top}\right)=1,\ \ \forall i\in\{1,2,3\} & (a)\\ f(\mathbf{w})\succ\mathbf{0} & (b)\end{cases}\end{array}} \tag{14}$$

The solution to problem (14) is given in Algorithm 1. In the absence of noise it has two solutions in general, and with noise it has either zero, one or two solutions. For each solution we recover the upgrade matrix $\mathbf{X}$ from $\mathbf{w}$ using Equation (13) and then estimate the plane's metric structure with $\hat{\mathbf{S}} = \text{stack}(\mathbf{X}^{-1}\mathbf{S}^A, \mathbf{0}_{1\times N})$. For each $\hat{\mathbf{S}}$ we then resect the cameras, which has in general 2^3 solutions due to two-fold camera pose ambiguity per view (see §II-C0c). The maximal number of solutions is therefore sixteen. The complete algorithm for Exact PSfM-O is given in Algorithm 2.

Algorithm 1 (The solution to problem (14))

```
Require: $\mathbf{M}^A \in \mathbb{R}^{6\times 2}$                    ▷ the affine camera factor for three views
function Exact_PSfM-O_Upgrade($\mathbf{M}^A$)
    $\{\mathbf{M}_1^A, \mathbf{M}_2^A, \mathbf{M}_3^A\} \leftarrow \text{unstack}(\mathbf{M}^A)$
    $\mathcal{W} \leftarrow \emptyset$                                    ▷ the set of upgrade solutions
    $\mathbf{E}_i \leftarrow \mathbf{M}_i^{A\top}\mathbf{M}_i^A,\ i \in \{1,2,3\}$
    $\mathbf{A}_E \leftarrow \begin{bmatrix} [\mathbf{E}_1]_{11} & 2[\mathbf{E}_1]_{12} & [\mathbf{E}_1]_{22} & -\det(\mathbf{E}_1) \\ [\mathbf{E}_2]_{11} & 2[\mathbf{E}_2]_{12} & [\mathbf{E}_2]_{22} & -\det(\mathbf{E}_2) \\ [\mathbf{E}_3]_{11} & 2[\mathbf{E}_3]_{12} & [\mathbf{E}_3]_{22} & -\det(\mathbf{E}_3) \end{bmatrix}$
    $\mathbf{z} \leftarrow \text{null}(\mathbf{A}_E)\,\text{with}\,\mathbf{z}^\top\mathbf{z} = 1$
    $\text{stack}(\mathbf{w}', s') \leftarrow \mathbf{A}_E^\top(\mathbf{A}_E\mathbf{A}_E^\top)^{-1}\mathbf{1}_{3\times 1}$
    $a \leftarrow z_2^2 - z_1 z_3$
    $b \leftarrow z_4 - w_1' z_3 + 2w_2' z_2 - w_3' z_1$
    $c \leftarrow w_2'^2 + s' - w_1' w_3'$
    $\{\alpha_1, \ldots, \alpha_L\} \leftarrow \text{realRoots}(a\alpha^2 + b\alpha + c),\ 0 \le L \le 2$
    for $l = 1$ to $L$ do
        $\mathbf{w} \leftarrow \mathbf{w}' + \alpha_l[\mathbf{z}]_{3\times 1}$
        if $\big(f(\mathbf{w}) \succ 0$ and $\det(\mathbf{M}_i^A f(\mathbf{w})\mathbf{M}_i^{A\top}) \le 1,\ \forall i \in \{1,2,3\}\big)$ then, $\mathcal{W} \leftarrow \mathcal{W} \cup \{\mathbf{w}\}$
    return $\mathcal{W}$
```

2) Solving the upgrade: This is done by transforming the upgrade constraints to quadratic constraints with inequalities:

$$\lambda_1\left(\mathbf{M}_i^A f(\mathbf{w})\mathbf{M}_i^{A\top}\right) = 1 \Leftrightarrow \begin{cases} \lambda_1\left(\mathbf{M}_i^A f(\mathbf{w})\mathbf{M}_i^{A\top}\right) = 1 \,\text{or}\, \lambda_2\left(\mathbf{M}_i^A f(\mathbf{w})\mathbf{M}_i^{A\top}\right) = 1 \\ \lambda_2\left(\mathbf{M}_i^A f(\mathbf{w})\mathbf{M}_i^{A\top}\right) \le 1 \end{cases} \Leftrightarrow \begin{cases} \det(\mathbf{M}_i^A f(\mathbf{w})\mathbf{M}_i^{A\top} - \mathbf{I}_2) = 0 \\ \det(\mathbf{M}_i^A f(\mathbf{w})\mathbf{M}_i^{A\top}) \le 1 \end{cases} \tag{15}$$

Algorithm 2 (Exact PSfM-O)

```
Require: $\{\hat{\mathbf{q}}_i^j\}$    ▷ point correspondences with view index $i \in \{1,\dots,M\}$ and point index $j \in \{1,\dots,N\}$

function Exact_PSfM-O($\{\hat{\mathbf{q}}_i^j\}$)
    $(\mathbf{M}^A, \mathbf{S}^A) \leftarrow$ affineReconstruct2D($\{\hat{\mathbf{q}}_i^j\}$)    ▷ Gives the 2D affine scene reconstruction (see appendix F)
    $\mathcal{W} \leftarrow$ Exact_PSfM-O_Upgrade$\{\mathbf{M}^A\}$    ▷ the upgrade matrix solutions
    $\mathcal{U}, \mathcal{R}, \mathcal{T} \leftarrow \emptyset$    ▷ solutions to structure, rotation and translation respectively
    for $k = 1$ to size($\mathcal{W}$) do
        $\mathbf{w} \leftarrow \mathcal{W}_k$, $\mathbf{X} \leftarrow \begin{bmatrix} \sqrt{w_1 - \frac{w_2^2}{w_3}} & \frac{w_2}{\sqrt{w_3}} \\ 0 & \sqrt{w_3} \end{bmatrix}$, $\hat{\mathbf{S}} = \text{stack}(\mathbf{X}^{-1}\mathbf{S}^A, \mathbf{0}_{1\times N})$
        $\{[\hat{\mathbf{R}}_i]_{2\times 2}\} \leftarrow \text{unstack}(\mathbf{M}^A\mathbf{X})$
        $\left\{[\hat{\mathbf{t}}_i]_{2\times 1}\right\} \leftarrow \sum_{j=1}^N [\mathbf{V}]_{ij}\left([\hat{\mathbf{s}}_j]_{2\times 1} - \mathbf{q}_i^j\right) / \sum_{j=1}^N [\mathbf{V}]_{ij}$    ▷ $[\mathbf{V}]_{ij} = 1$ if point $j$ is measured in view $i$, otherwise $[\mathbf{V}]_{ij} = 0$
        $\mathcal{U}_k \leftarrow \hat{\mathbf{S}},\ \mathcal{R}_k \leftarrow \{[\hat{\mathbf{R}}_1]_{2\times 2}, [\hat{\mathbf{R}}_2]_{2\times 2}, [\hat{\mathbf{R}}_3]_{2\times 2}\},\ \mathcal{T}_k \leftarrow \left\{[\hat{\mathbf{t}}_1]_{2\times 1}, [\hat{\mathbf{t}}_2]_{2\times 1}, [\hat{\mathbf{t}}_3]_{2\times 1}\right\}$
    return $\mathcal{U}, \mathcal{R}, \mathcal{T}$
```

The first equivalence comes because $\lambda_2 \leq \lambda_1$. The second equivalence comes from the characteristic polynomial of $\mathbf{M}_i^A f(\mathbf{w})\mathbf{M}_i^{A\top}$ and using $\lambda_2\left(\mathbf{M}_i^A f(\mathbf{w})\mathbf{M}_i^{A\top}\right) = \det(\mathbf{M}_i^A f(\mathbf{w})\mathbf{M}_i^{A\top})$ (which comes from Equation (1) because $\mathbf{M}_i^A f(\mathbf{w})\mathbf{M}_i^{A\top} \in \mathcal{G}_{2\times 2}$). From Equation (15) a solution to problem (14) must satisfy:

$$\det(\mathbf{M}_i^A f(\mathbf{w})\mathbf{M}_i^{A\top} - \mathbf{I}_2) = 0,\ \forall i \tag{16}$$

The three quadratic constraints on $\mathbf{w}$ in Equation (16) actually have a special structure that reduces them to a single quadratic constraint in one variable α. This is then solved with at most two solutions. For each solution we compute the corresponding value of $\mathbf{w}$. We then test whether $f(\mathbf{w})$ is *(i)* positive definite and *(ii)* satisfies $\det\left(\mathbf{M}_i^A f(\mathbf{w})\mathbf{M}_i^{A\top}\right) \leq 1, \forall i$. If it does then it satisfies all constraints and hence is a solution to problem (14), and is put into the solution set $\mathcal{W}$. We give a full derivation of Algorithm 1 in Appendix G.

3) Camera Resection: Becase the upgrade constraints are satisfied exactly the camera poses are computed directly from the upgraded camera factor. For each view we have $[\hat{\mathbf{R}}_i]_{2\times 2} = \mathbf{M}_i^A\mathbf{X}$. The full rotation matrix can be completed from $[\hat{\mathbf{R}}_i]_{2\times 2}$ up to two solutions using orthonormality constraints (we give the Equation in Appendix D). The camera translation is given by the point centroid of the visible correspondences in the image.

D. Approximate PSfM-O: An Approximate Stratified Solution to PSfM-O for Three or More views

1) Method Overview: Approximate PSfM-O solves metric structure with three or more views by satisfying the upgrade constraints approximately in a least squares sense. It then solves the camera poses using our optimal method described in Appendix D. We give the full algorithm for Approximate PSfM-O in Algorithm 4. The upgrade constraints are satisfied with a least squares cost function $C : \mathbb{R}^3 \rightarrow \mathbb{R}^+$ based on Equation (16):

$$C(\mathbf{w}) \overset{\text{def}}{=} \sum_{i=1}^{M} \det{}^2(\mathbf{M}_i^A f(\mathbf{w}) \mathbf{M}_i^{A\top} - \mathbf{I}_2) \tag{17}$$

An important property of Approximate PSfM-O is that it finds *all* local minima of C and not just the global minimum. This is necessary to handle cases where we cannot uniquely resolve metric structure. In these cases there will be multiple upgrade matrices that can satisfy the upgrade constraints up to noise, which causes C (and the reprojection error) to be multi-modal. *The correct upgrade solution is therefore not necessarily the one that globally minimises* C. Instead, by computing all local minima of C we obtain a small number of locally-optimal upgrade matrices (which we will show is at most four). We can then verify each upgrade matrix *a posteriori* for how well it can explain the data. All upgrade matrices that can explain the data up to noise should be kept.

Our task is to compute all local minima of C over the domain $f(\mathbf{w}) \succ \mathbf{0}$. Because the domain is an open set we may not always find a local minimum. However if a local minimum exists then it is a local minimum of $C(\mathbf{w} \in \mathbb{R}^3)$. The problem is solved therefore by finding all local minima of $C(\mathbf{w} \in \mathbb{R}^3)$ then discarding those where $f(\mathbf{w}) \not\succ \mathbf{0}$. Because C is quartic in $\mathbf{w}$ its local minima are roots of a system of three third-order polynomials. We have developed a very fast method to find these by exploiting the particular form of C. Specifically we show that its critical points are roots of a univariate degree-seven polynomial, which can be found quickly with the SVD of an associated 7×7 companion matrix. There are either 1, 3, 5 or 7 real-valued critical points, and because $C(\mathbf{w})$ is a Sum-of-Squares polynomial it therefore has either 1, 2, 3 or 4 real-valued local minima.

For each local minimum $\tilde{\mathbf{w}}$ we keep it if $f(\tilde{\mathbf{w}}) \succ \mathbf{0}$ (which means it is a feasible upgrade solution) and we generate the associated upgrade matrix $\mathbf{X}$ using Equation (13). Next we compute its associated metric structure matrix $\hat{\mathbf{S}} = \text{stack}(\mathbf{X}^{-1}\mathbf{S}^A, \mathbf{0}_{1\times N})$. Unlike Exact PSfM-O the upgraded camera factor $\mathbf{M}^A\mathbf{X}$ is *not* guaranteed to be a metric camera factor because the metric constraints have been satisfied approximately. We deal with this by resecting with our optimal solution described in Appendix D. Note that an alternative solution is to correct the upgraded camera factor to the closest metric camera factor. However this correction involves minimising an algebraic error function (typically the Frobenius norm is used) so its solution is not optimal in terms or reprojection error. In practice we find this leads to worse pose estimates (typically by a few degrees). Because our resection method is closed-form there is no good reason to use the upgraded camera matrices.

We finish this section by showing how to find the local minima of C. We have tried to keep this light for the curious reader but it is not essential for understanding the overall algorithm.

2) Finding the Local Minima of C: We can rewrite $C(\mathbf{w})$ as follows:

$$\begin{gathered} C(\mathbf{w}) = \textstyle\sum_{i=1}^M \det^2\left(\mathbf{M}_i^A f(\mathbf{w}) \mathbf{M}_i^{A\top} - \mathbf{I}_2\right) = \left\|\mathbf{B}\,\text{stack}(\mathbf{w}, w_1w_3 - w_2^2) - \mathbf{1}_{M\times 1}\right\|_2^2 \\ \mathbf{E}_i \overset{\text{def}}{=} \mathbf{M}_i^{A\top}\mathbf{M}_i^A,\ \mathbf{B} \overset{\text{def}}{=} \begin{bmatrix} [\mathbf{E}_1]_{11} & 2[\mathbf{E}_1]_{12} & [\mathbf{E}_1]_{22} & -\det(\mathbf{E}_1) \\ [\mathbf{E}_2]_{11} & 2[\mathbf{E}_2]_{12} & [\mathbf{E}_2]_{22} & -\det(\mathbf{E}_2) \\ \vdots & \vdots & \vdots & \vdots \\ [\mathbf{E}_M]_{11} & 2[\mathbf{E}_M]_{12} & [\mathbf{E}_M]_{22} & -\det(\mathbf{E}_M) \end{bmatrix} \end{gathered} \tag{18}$$

Using the slack variable $s = w_1w_3 - w_2^2$, the local minima of $C(\mathbf{w} \in \mathbb{R}^3)$ are stationary points of the Lagrangian

$$L(\mathbf{w}, s, \nu) \overset{\text{def}}{=} \|\mathbf{B}\,\text{stack}(\mathbf{w}, s) - \mathbf{1}_{M\times 1}\|_2^2 + \nu\left(w_1w_3 - w_2^2 - s\right) \tag{19}$$

where ν is a Lagrange multiplier for the constraint $s = w_1 w_3 - w_2^2$. We now show that the stationary points of L are the roots of a degree-seven polynomial in ν. The stationary points are the solutions to

$$\begin{aligned} &\tfrac{\partial}{\partial(\mathbf{w},s,\nu)} L(\mathbf{w}, s, \nu) = \mathbf{0}_{5\times 1} \Leftrightarrow \begin{cases} \mathbf{H}\,\mathrm{stack}\,(\mathbf{w}, s) - \mathbf{B}^\top \mathbf{1}_{M\times 1} + \nu\,\mathrm{stack}\,(\mathbf{F}\mathbf{w}, 1) = \mathbf{0}_{4\times 1} & (a) \\ w_1 w_3 - w_2^2 - s = 0 & (b) \end{cases} \\ &\mathbf{H} \overset{\mathrm{def}}{=} \mathbf{B}^\top\mathbf{B},\ \mathbf{F} \overset{\mathrm{def}}{=} \begin{bmatrix} 0 & 0 & -1 \\ 0 & 2 & 0 \\ -1 & 0 & 0 \end{bmatrix} \end{aligned} \tag{20}$$

We decompose $\mathbf{H}$ with the QL decomposition to give $\mathbf{H} = \mathbf{Q}\mathbf{L}$ where $\mathbf{Q}$ is a 4×4 orthogonal matrix and $\mathbf{L}$ is a lower-triangular 4×4 matrix. Left-multiplying Equation (20-a) by $\mathbf{Q}^\top$ and re-substituting $s \leftarrow w_1 w_3 - w_2^2$ gives

$$\mathbf{L}\,\mathrm{stack}(\mathbf{w}, w_1 w_3 - w_2^2) - \mathbf{Q}^\top \mathbf{B}^\top \mathbf{1}_{M\times 1} + \nu \mathbf{Q}^\top\,\mathrm{stack}(\mathbf{F}\mathbf{w}, 1) = \mathbf{0}_{4\times 1} \tag{21}$$

Now consider only the first three rows of Equation (21). Because $\mathbf{L}_{3\times 4}\,\mathrm{stack}(\mathbf{w}, w_1 w_3 - w_2^2) = \mathbf{L}_{3\times 3}\mathbf{w}$ (since $\mathbf{L}$ is lower triangular) we have

$$[\mathbf{L}]_{3\times 3}\mathbf{w} - [\mathbf{Q}]_{4\times 3}^\top \mathbf{B}^\top \mathbf{1}_{M\times 1} + \nu [\mathbf{Q}]_{4\times 3}^\top\,\mathrm{stack}\,(\mathbf{F}\mathbf{w}, 1) = \mathbf{0}_{3\times 1} \tag{22}$$

Therefore given a solution to ν we can recover $\mathbf{w}$ by solving a linear system using Equation (22). This is given after rearrangement by

$$\begin{aligned} &\mathbf{w} = \det^{-1}(M(\nu))g(\nu) \\ &M(\nu) : \mathbb{R} \to \mathbb{R}^{3\times 3} \overset{\mathrm{def}}{=} [\mathbf{L}]_{3\times 3} + \nu [\mathbf{Q}]_{3\times 3}^\top \mathbf{F} \\ &g(\nu) : \mathbb{R} \to \mathbb{R}^3 \overset{\mathrm{def}}{=} \mathrm{adj}\,(M(\nu))\,[\mathbf{I}_3 | \mathbf{0}_{3\times 1}] \left(\mathbf{Q}^\top \mathbf{B}^\top \mathbf{1}_{M\times 1} - \nu \mathbf{Q}^\top [0\,0\,0\,1]^\top\right) \end{aligned} \tag{23}$$

where $\mathrm{adj}\,(M(\nu))$ is the adjoint of $M(\nu)$. We now re-introduce the constraint from the fourth row of Equation (20-a):

$$\nu = a - \mathbf{h}_4\,\mathrm{stack}\left(\mathbf{w}, w_1 w_3 - w_2^2\right) \tag{24}$$

where a is the fourth element of $\mathbf{B}^\top \mathbf{1}_{M\times 1}$ and $\mathbf{h}_4$ is the fourth row of $\mathbf{H}$. Multiplying both sides of Equation (24) by $\det(M(\nu))^2$ and substituting $\det(M(\nu))\mathbf{w} \leftarrow g(\nu)$ gives after simplification:

$$\det(M(\nu))^2\nu - \det(M(\nu))^2 a + \mathbf{h}_4\,\mathrm{stack}\,(\det(M(\nu))g(\nu), \det(g(\nu))) = 0 \tag{25}$$

Equation (25) defines a polynomial $p(\nu)$ in ν because $\det(M(\nu))$ and $g(\nu)$ are quadratic and cubic polynomials in ν respectively (and a and $\mathbf{h}_4$ are constant and known). The polynomial is non-homogeneous in general because $\det(g(\nu))$ is non-homogeneous in ν in general. The polynomial's highest order term is $\det(M(\nu))^2\nu$, which means it is a degree-seven polynomial in general. The roots $\{\nu_1, \ldots, \nu_L\}$ of $p(\nu)$ are computed efficiently by the SVD of the associated 7×7 companion matrix.

E. Generalising PSfM-O Solutions to Other Affine Cameras

Recall that we cannot solve planar SfM with affine cameras if the cameras have six or more unknown DoFs (see §II-D). Theorem 7 provides three cases where additional knowledge can be used to make the problem solvable. In Cases 1 and 2 we can isolate a subset of views $\mathcal{I}' \subseteq \mathcal{I}$ where the magnification factors $k_{i\in\mathcal{I}'}$ are assumed to be constant. In practice this occurs when the variation of the depth of the structure is small in those views. In Case 1 we also know the 'intrinsic' terms $\mathbf{A}_i$. From the decomposition of the affine camera in Equation (3) we can effectively convert the cameras in $\mathcal{I}'$ to orthographic cameras by transforming the points in $\mathcal{I}'$ by $\mathbf{q}_i^j \leftarrow \mathbf{A}_i^{-1}\mathbf{q}_i^j$. Because Theorem 7 tells us that the upgrade problem is only constrained by the views in $\mathcal{I}'$, we can solve the problem using Exact PSfM-O or Approximate PSfM-O only for the views in $\mathcal{I}'$. After this is solved the cameras can be resected (including those not in $\mathcal{I}'$), as discussed in II-C. In Case 2 $\mathbf{A}_{i\in\mathcal{I}'}$ is assumed constant and unknown: $\mathbf{A}_{i\in\mathcal{I}'} = \mathbf{A}$ for some $\mathbf{A}$. Here we must solve the metric upgrade matrix *and* $\mathbf{A}$. This can be considered an autocalibration problem that has not been looked at before, and is left to future work.

Algorithm 3 (The local minima of C)

```
Require: $\mathbf{M}^A \in \mathbb{R}^{2M\times 2}$, $M \geq 3$    ▷ the affine camera factor for $M$ views
function localMinimaOfC($\mathbf{M}^A$)
    $\{\mathbf{M}_1^A, \mathbf{M}_2^A, \dots, \mathbf{M}_M^A\} \leftarrow \mathrm{unstack}(\mathbf{M}^A)$
    $\mathcal{W} \leftarrow \emptyset$    ▷ the set of upgrade solutions
    $\mathbf{E}_i \leftarrow \mathbf{M}_i^{A\top}\mathbf{M}_i^A$
    $\mathbf{B} \leftarrow \begin{bmatrix} [\mathbf{E}_1]_{11} & 2[\mathbf{E}_1]_{12} & [\mathbf{E}_1]_{22} & -\det(\mathbf{E}_1) \\ [\mathbf{E}_2]_{11} & 2[\mathbf{E}_2]_{12} & [\mathbf{E}_2]_{22} & -\det(\mathbf{E}_2) \\ \vdots & \vdots & \vdots & \vdots \\ [\mathbf{E}_M]_{11} & 2[\mathbf{E}_M]_{12} & [\mathbf{E}_M]_{22} & -\det(\mathbf{E}_M) \end{bmatrix}$
    $\mathbf{H} \leftarrow \mathbf{B}^\top\mathbf{B},\ (\mathbf{Q},\mathbf{L}) \leftarrow \mathrm{lq}(\mathbf{H}),\ \mathbf{p} \leftarrow \mathbf{Q}^\top\mathbf{B}^\top\mathbf{1}_{M\times 1},\ \mathbf{F} \leftarrow \begin{bmatrix} 0 & 0 & -1 \\ 0 & 2 & 0 \\ -1 & 0 & 0 \end{bmatrix}$    ▷ lq denotes the LQ decomposition
    $\mathbf{c} \leftarrow P\left(\mathbf{L},\mathbf{Q},\mathbf{p}\right),\ \mathbf{c} \in \mathbb{R}^8$    ▷ computes coefficients of the degree 7 polynomial $p(\nu)$ in Equation (25).
    $\{\nu_1, \dots, \nu_L\} = \mathrm{realRoots}\left(\mathbf{c}\right),\ L \in \{1,3,5,7\}$    ▷ real roots of $p(\nu)$
    for $l = 1$ to $L$ do
        $\mathbf{M} \leftarrow \mathbf{L}_{3\times 3} + \nu_l\mathbf{Q}_{3\times 3}^\top\mathbf{F}$, $\mathbf{g} \leftarrow \mathrm{adj}\left(\mathbf{M}\right)[\mathbf{I}_3|\mathbf{0}_{3\times 1}]\left(\mathbf{Q}^\top\mathbf{B}^\top\mathbf{1}_{M\times 1} - \nu_l\mathbf{Q}^\top[0\,0\,0\,1]^\top\right)$, $\hat{\mathbf{w}} \leftarrow \det(\mathbf{M})^{-1}\mathbf{f}$
        if $\frac{\partial^2}{\partial\mathbf{w}^2}C(\hat{\mathbf{w}}) \succ \mathbf{0}$ then
            $\mathcal{W} \leftarrow \{\hat{\mathbf{w}}\}$    ▷ $\hat{\mathbf{w}}$ is a local minimum of $C$
    return $\mathcal{W}$
```

Algorithm 4 (Approximate PSfM-O)

```
Require: $\{\mathbf{q}_i^j\}$    ▷ point correspondences with view index $i \in \{1,\dots,M\}$ point index $j \in \{1,\dots,N\}$
$\{\mathbf{A}_i\}$    ▷ known affine projection matrix terms
function Approximate_PSfM-O($\{\mathbf{q}_i^j\}$)
    $(\mathbf{M}^A, \mathbf{S}^A) \leftarrow$ affineReconstruct2D($\{\mathbf{q}_i^j\}, \{\mathbf{A}_i\}$)    ▷ affine scene factorisation with planar structures (see appendix F)
    $\mathcal{W} \leftarrow$ localMinimaOfC$\{\mathbf{M}^A\}$
    $\mathcal{U}, \mathcal{R}, \mathcal{T} \leftarrow \emptyset$    ▷ solutions to structure, rotation and translation respectively
    $k \leftarrow 0$    ▷ the number of upgrade matrices/metric structure solutions
    for $l = 1$ to $\mathrm{size}(\mathcal{W})$ do
        $\mathbf{w} \leftarrow \mathcal{W}_l$
        if $f(\mathbf{w}) \succ \mathbf{0}$ then
            $k \leftarrow k + 1$, $\mathbf{X} \leftarrow \begin{bmatrix} \sqrt{w_1 - \frac{w_2^2}{w_3}} & \frac{w_2}{\sqrt{w_3}} \\ 0 & \sqrt{w_3} \end{bmatrix}$, $\hat{\mathbf{S}} = \mathrm{stack}(\mathbf{X}^{-1}\mathbf{S}^A, \mathbf{0}_{1\times N})$
            for $i = 1$ to $M$ do
                $\left([\hat{\mathbf{R}}_i]_{2\times 2}, [\hat{\mathbf{t}}_i]_{2\times 1}\right) \leftarrow \mathrm{orthographicResect}(\hat{\mathbf{S}}, \{\mathbf{q}_i^1, \mathbf{q}_i^2, \dots, \mathbf{q}_i^N\})$    ▷ optimal resection with orthographic camera (see Appendix D)
            $\mathcal{U}_k \leftarrow \hat{\mathbf{S}},\ \mathcal{R}_k \leftarrow \{[\hat{\mathbf{R}}_1]_{2\times 2}, [\hat{\mathbf{R}}_2]_{2\times 2}, \dots, [\hat{\mathbf{R}}_M]_{2\times 2}\},\ \mathcal{T}_k \leftarrow \left\{[\hat{\mathbf{t}}_1]_{2\times 1}, [\hat{\mathbf{t}}_2]_{2\times 1}, \dots, [\hat{\mathbf{t}}_M]_{2\times 1}\right\}$
    return $\mathcal{U}, \mathcal{R}, \mathcal{T}$
```

F. Summary of the Differences between Stratified SfM with Affine Cameras for Planar and Non-planar Scenes

We finish this section with a summary of the core differences between solving SfM with affine cameras by stratification for planar and non-planar structures (Table I). For non-planar structures results have been aggregated from [2], [9], [1].

	Non-planar structures	**Planar structures**
Maximal theoretical rank of the $2M \times N$ measurement matrix $\hat{\mathbf{Q}}$	3	2
Unknown upgrade matrix	$\mathbf{Y} \in \mathbb{R}^{3\times 3}$, $\mathrm{rank}(\mathbf{Y}) = 3$	$\mathbf{X} \in \mathbb{R}^{2\times 2}$, $\mathrm{rank}(\mathbf{X}) = 2$
	Orthographic Cameras	
Gauge transform	3D rotation and reflection	2D rotation and reflection
Upgrade constraint	$\tilde{\mathbf{M}}_i^A \mathbf{Y} \in \mathcal{S}_{2\times 3}$, $\tilde{\mathbf{M}}_i^A \in \mathbb{R}^{2\times 3}$ is known	$\mathbf{M}_i^A \mathbf{X} \in \mathcal{S}\mathcal{S}_{2\times 2}$, $\mathbf{M}_i^A \in \mathbb{R}^{2\times 2}$ is known
Number of equality constraints on upgrade matrix per view	3	1
Minimal number of views required for metric upgrade	3	3
Upgrade constraint complexity	Quadratic in $\mathbf{Y}$, linear in $\mathbf{Y}\mathbf{Y}^\top$	Quartic in $\mathbf{X}$, quadratic in $\mathbf{X}\mathbf{X}^\top$.
Number of distinct upgrade/structure solutions for three views with noise (up to gauge transforms)	0 or 1	0,1 or 2
Can the problem have a critical motion sequence?	No	Yes
Can the problem have discrete structure ambiguities?	No	Yes
Can the problem have critical structures?	Yes	Yes
Can the problem have camera resection ambiguities?	No	Yes (two-fold ambiguous for each view)
Can the upgrade solutions for three views be optimal in terms of reprojection error?	No	Yes
	Other Affine Cameras	
Can we solve with the weak and para-perspective cameras without additional information?	Yes	No
Can we solve with these cameras with some knowledge about the camera magnification factors?	Yes	Yes
If the magnification factors are constant and $\mathbf{A}_i$ is constant and unknown for all views, what is the complexity of upgrading & self calibration?	$\mathbf{A}_i$ can be trivially eliminated, and is quadratic in $\mathbf{Y}$, linear in $\mathbf{Y}\mathbf{Y}^\top$.	$\mathbf{A}_i$ cannot be trivially eliminated, and is 5 quadratic equations in 5 unknowns

TABLE I

SUMMARY OF THE DIFFERENCES BETWEEN STRATIFIED SFM WITH AFFINE CAMERAS FOR NON-PLANAR AND PLANAR STRUCTURES

IV. EMPIRICAL EVALUATION

We now evaluate the accuracy of Exact PSfM-O and Approximate PSfM-O compared to other methods with extensive simulation and real-data experiments.

A. *Method Comparison Summary*

The methods under comparison are as follows. **Exact PSfM-O**: Proposed exact solution (§III-C); **Approximate PSfM-O**: Proposed approximate solution (§III-D); **Approximate PSfM-O(LRE)**: Proposed approximate solution but returning at most one structure solution, which is the one that produces the Lowest Reprojection Error (Equation (6)); **TJK-CVPR10**: Solution from [15]; **TK-Factor**: Solution from [2]; **MC-CVIU09**: solution from [29] using the orthographic camera model; **MOVA**: A stratified method using the *Most Orthogonal Viewpoint Approximation* heuristic (details are provided in Appendix E). We summarise the applicability of the methods in Table II. The purpose for comparing Approximate PSfM-O(LRE) is to show how performance is affected in ambiguous cases when we force Approximate PSfM-O to select the structure solution that yields the lowest reprojection error. For the stratified methods (Exact PSfM-O, Approximate PSfM-O, Approximate PSfM-O(LRE), MOVA and TK-factorization) we use exactly the same method to compute the structures's affine reconstruction, given in Appendix F.

	Exact PSfM-O	Approximate PSfM-O	Approximate PSfM-O(LRE)	TJK-CVPR10	TK-Factor	MC-CVIU09	MOVA
Range of N	≥ 3	≥ 3	≥ 3	$= 3$	≥ 3	≥ 3	≥ 3
Range of M	$= 3$	≥ 3	≥ 3	≥ 4	≥ 3	≥ 3	≥ 1
Possible number of structure solutions	0,1,2	0,1,2,3,4	0,1	0,1	0,1	1	1
Are the solutions guaranteed to be planar?	Yes	Yes	Yes	Yes	No	No	Yes

TABLE II
PROPERTIES OF METHODS UNDER COMPARISON.

A difficulty with comparing all methods is that for a given test input some methods may be able to produce a metric structure solution but other methods may not (*e.g.* the stratified methods may not produce a valid upgrade matrix). This makes it hard to compute and compare accuracy statistics. We deal with this by applying a fall-back method, and a method reverts to the fall-back's solution if it does not produce a solution. The fall-back method should return a solution in all cases but is not necessarily the most accurate method. The fall-back method we use is MOVA.

A problem with TK-Factor and MC-CVIU09 is that they return a single solution to camera resection. Therefore for planar structures, even if they compute metric structure correctly their camera poses will be wrong approximately 50% of the time due to the two-fold ambiguity. To handle this fairly we resect the cameras in exactly the same way for *all* methods. This is done using our optimal method given in Appendix D. Note that this requires a planar estimate of metric structure which is not guaranteed by TK-Factor and MC-CVIU09. We deal with this by converting their structure solution $\hat{\mathbf{S}}$ to the closest planar solution before resecting using the rank-two SVD of $\hat{\mathbf{S}}$.

We also evaluate the gain in accuracy by refining the best solution among all methods with Orthographic camera Bundle Adjustment (OBA), which we denote by **Best+OBA**. This is done by taking the metric structure solution among all methods with the lowest error (see below) then resecting the cameras as described in Appendix D. Then structure and camera poses are jointly refined until convergence by minimising Equation (6) using Levenberg-Marquardt.

B. *Error Metrics*

We measure performance using four metrics. These are *(i) structure error*, *(ii) rotation error*, *(iii) translation error* and *(iv) success rate*. We use $\mathcal{U} = \{\hat{\mathbf{S}}_1, \dots \hat{\mathbf{S}}_K\}$ to denote the set of K metric structure solutions produced by a method.

If a method fails to compute a structure solution then $\mathcal{U} = \emptyset$. We use $\hat{\mathbf{S}}^{MOVA}$ to denote the structure solution produced by the fall-back method MOVA. We use $\mathbf{S}^{GT} \in \mathbb{R}^{2\times N}$, $\mathcal{R}^{\mathrm{GT}} \in \mathcal{SO}_3^M$ and $\mathcal{T}^{\mathrm{GT}} \in \mathbb{R}^{2\times M}$ to denote the ground truth structure, camera rotations and translations respectively.

1) Structure Error: $E_S \in \mathbb{R}^+$. Structure error is computed using the solution in $\mathcal{U}$ that is closest to ground truth up to a similarity transform. If a method does not return a structure solution the solution from the fall-back method is used (MOVA). Formally, we define it as:

$$E_S \overset{\mathrm{def}}{=} \begin{cases} \frac{1}{N}\sum_{j=1}^{N} \|\mathbf{S}_j^{\mathrm{GT}} - \hat{\mathbf{S}}'_j\|_2 & \text{if}\, \mathcal{U} \neq \emptyset \\ E_S^{MOVA} & \text{otherwise} \end{cases} \qquad \hat{\mathbf{S}}' \overset{\mathrm{def}}{=} \arg\min_{\hat{\mathbf{S}}\in\mathcal{U}} \|\mathrm{ABSOR}(\hat{\mathbf{S}}, \mathbf{S}^{\mathrm{GT}}) - \mathbf{S}^{\mathrm{GT}}\|_2^2 \tag{26}$$

The function $\mathrm{ABSOR}(\hat{\mathbf{S}}, \mathbf{S}^{\mathrm{GT}})$ aligns an estimate $\hat{\mathbf{S}}$ to $\mathbf{S}^{\mathrm{GT}}$ in the least squares sense over all 2D similarity transforms. This alignment is necessary to account for the gauge transform. Therefore $\hat{\mathbf{S}}'$ is the structure solution that has aligned best to $\mathbf{S}^{\mathrm{GT}}$. The value $E_S^{MOVA} \in \mathbb{R}^+$ is the structure error from MOVA.

2) Rotation and Translation Error: $E_R \in \mathbb{R}^+$, $E_T \in \mathbb{R}^+$. For each method we take the best structure solution $\hat{\mathbf{S}}'$, resect the cameras as described in Appendix D, then we measure the error of the camera poses. If a method has not produced a structure solution we use the camera poses from the fall-back method (MOVA). Let $\left(\hat{\mathbf{R}}_i^a, \hat{\mathbf{t}}_i^a\right)$ and $\left(\hat{\mathbf{R}}_i^b, \hat{\mathbf{t}}_i^b\right)$ be the camera pose estimates for view i (recall there are two due to the two-fold ambiguity). The rotation error is defined as follows:

$$E_R = \begin{cases} \frac{1}{M}\sum_{i=1}^{M} \min\left[\mathrm{ang}(\hat{\mathbf{R}}_i^a, \mathbf{R}_i^{\mathrm{GT}}),\, \mathrm{ang}(\hat{\mathbf{R}}_i^b, \mathbf{R}_i^{\mathrm{GT}})\right] & \text{if}\, \mathcal{U} \neq \emptyset \\ E_R^{MOVA} & \text{otherwise} \end{cases} \tag{27}$$

The function $\mathrm{ang}(\mathbf{R}, \mathbf{R}') : \mathcal{SO}_3^2 \to [0, 180]$ denotes the smallest angle in degrees of the rotation that maps $\mathbf{R}$ to $\mathbf{R}'$. Because there are two rotation estimates per view, the error of the one with the smallest angular error is used. The value $E_R^{MOVA} \in \mathbb{R}^+$ is the rotation error from MOVA.

We measure translation error as follows. For each view we determine which of the two pose estimates has the lowest rotation error, then measure the accuracy of its corresponding translation estimate $\hat{\mathbf{t}}_i \in \mathbb{R}^2$. This is defined as

$$E_T = \begin{cases} \frac{1}{M}\sum_{i=1}^{M} \left\|\hat{\mathbf{t}}_i - \mathbf{t}_i^{\mathrm{GT}}\right\|_2 & \text{if}\, \mathcal{U} \neq \emptyset \\ E_T^{MOVA} & \text{otherwise} \end{cases} \tag{28}$$

where $E_T^{MOVA} \in \mathbb{R}^+$ is the translation error from MOVA. We note here that the translation error is only useful for comparing methods when there are missing measurements. This is because when there are no missing measurements $\hat{\mathbf{t}}_i$ is the centroid of the point correspondences in view i, so it is the same for all methods. When there are missing measurements $\hat{\mathbf{t}}_i$ will depend on the particular rotation solution (see Appendix D), so it will differ depending on the method.

3) Success Rate: $E_{succ} \in [0, 100]$. We define the success rate as the percentage of instances for which a method produces a metric structure solution. The success rate of MC-CVIU09, MOVA and Best+OBA is 100%, so we only compare success rates for Exact PSfM-O, Approximate PSfM-O, TJK-CVPR10 and TK-factor. The success rate of Approximate PSfM-O(LRE) is the same as Approximate PSfM-O so we omit it from the results.

C. Simulation Experiments

We ran a large number of simulation experiments to test the accuracy of the methods in a variety of conditions. We generated ground truth structure matrices $\mathbf{S}^{\mathrm{GT}}$ by synthesizing N points positioned randomly on the structure plane $z = 0$

	N	M	σ_n	σ_k	γ
Experiment 1	$[3, 40]$	3	2	0	0
Experiment 2	$[3, 40]$	4	2	0	0
Experiment 3	$[3, 40]$	8	2	0	0
Experiment 4	$[3, 40]$	12	2	0	0
Experiment 5	3	4	2	$[0, 10]$	0
Experiment 6	3	8	2	$[0, 10]$	0
Experiment 7	50	3	2	5	$[0, 70]$
Experiment 8	50	8	2	5	$[0, 70]$

TABLE III

EXPERIMENTAL PARAMETERS USED IN EIGHT SIMULATION EXPERIMENTS.

in world coordinates. These were drawn within the square centred at the origin. The points were then normalised so the centroid was at the origin and the mean distance from the origin was 100 units. A set of M random rotations were then generated: $\mathcal{R}^{\text{GT}} = \{\mathbf{R}_1^{\text{GT}}, \ldots, \mathbf{R}_M^{\text{GT}}\}$ where $\mathbf{R}_i^{\text{GT}} \in \mathcal{SO}_3$ rotates the structure plane to camera coordinates. Similarly to [30] we randomly generated these using Euler angles where each angle was assigned with uniform probability in the range $[-80, +80]$ degrees. It was unnecessary to simulate variation in the camera translations because is has no effect on a method's accuracy (because the point correspondences are simply translated in the image), so we set $\mathbf{t}_i^{\text{GT}} = \mathbf{0}_{2\times 1}$ for all views. We also simulated variation of the camera magnification factors k_i because in real conditions the orthographic model may not hold perfectly due to variation in the scene's depths. For each view we assigned k_i with a random distribution $k_i \sim \mathcal{N}(1, \sigma_k/100)$, $\sigma_k \in \mathbb{R}$. We then projected the scene points for each view and perturbed them with IID zero-mean Gaussian noise with standard deviation σ_n. To simulate missing measurements we randomly removed $\gamma\%$ of the correspondences in each view. This was done while ensuring the scene's affine structure could still be recovered using Algorithm F.

We excluded from the evaluation all simulations that were badly conditioned, since they cannot be used to draw meaningful comparisons between the methods. This was done with the following policy. For a given simulation we first ran bundle adjustment initialised using the ground truth. If it converged far from the ground truth solution we assumed the problem was ill-conditioned and did not select it (we used a structure error threshold of 10%). We also tested whether the reprojection error had a local minimum at the point of convergence using the conditioning number of the residual error Jacobian matrix with a threshold of 1×10^{-7}. If so it was used for evaluation. We computed performance statistics over different values of the experimental parameters $\{N, M, \sigma_k, \sigma_n, \gamma\}$ by averaging over $T = 1000$ simulated scenes. We conducted eight experiments given in Table III.

1) Results: The results of experiments one to four are shown in Figure 2. Each column corresponds to one experiment and the six rows show different performance statistics across the methods (the success rate, mean and median structure error, mean and median rotation error and mean reprojection error). Results for Exact PSfM-O are shown only in the first column because it is only applicable when $M = 3$. TJK-CVPR10 is not present in experiment one and shown as a black cross in experiments two, three and four because it is applicable when $N = 3$ and $M > 3$ only. Results for translation error were not plotted because they were the same for all methods (because $\gamma = 0\%$).

With respect to success rate, when $N = 3$ points the success rate of TK-Factor is 0%. This is because when $N = 3$

the measurement matrix $\hat{\mathbf{Q}}$ never has a rank greater than two even with noise, so TK-Factor fails because it returns rank-deficient upgrade matrices that cannot be inverted. When $N > 3$, noise increases the rank of $\hat{\mathbf{Q}}$ beyond its theoretical rank, which means it is possible to find full-rank upgrade matrices using TK-Factor, which is about 65% of the time when $M = 4$ and 95% for $M = 12$. However the solutions from TK-Factor are poor compared to all other methods except MC-CVIU09, both of which perform worse than the fall-back method (*i.e.* MOVA). Structure and rotation errors for Exact PSfM-O, Approximate PSfM-O, Approximate PSfM-O(LRE) and Best+OBA tend to decrease with more points because the effect of noise reduces. There is virtually no difference between the accuracy of Approximate PSfM-O and Best+OBA across all statistics. For $M = 3$ there is a significant difference in the structure error from Approximate PSfM-O(LRE) and Approximate PSfM-O. *This is expected because for three views structure is not in general unique. Therefore the structure solution from Approximate PSfM-O that is closest to ground truth is not necessarily the correct one*. When $M = 4$ and beyond we see that the accuracy of Approximate PSfM-O(LRE) and Approximate PSfM-O is similar. There is slight deviation for $M = 4$, which can be explained by the fact that sometimes there can be multiple structure solutions when the cameras are in a particular configuration (see Theorem 2). With more views the likelihood of this occurring rapidly diminishes, which explains why they show the same error for $M > 4$.

The success rate of Approximate PSfM-O (and Approximate PSfM-O(LRE)) was approximately constant in all experiments and for all N at approximately 99.8%. For Exact PSfM-O, we see in experiment one a gradual improvement in success rate from 82% to 94% as the number of points increases. This suggests that as the influence of noise decreases the chances of being able to exactly upgrade the scene's affine structure to metric structure increases. Because the solution from Exact PSfM-O gives the optimal solution to PSfM-O (from Theorem 6), this indicates that *we can find the optimal solutions in closed-form for three views between 82% to 94% of the time* in these cases. The reason why Exact PSfM-O has worse performance than Approximate PSfM-O is because it fails more often, so we revert back to the fall-back method more often than with Approximate PSfM-O.

The results for experiments five and six are shown in the first two columns of Figure 3. For all methods we see a reduction in accuracy as the magnification factor standard deviation σ_k increases, which is due to increasing the modelling error. In experiment six we see a significant reduction in the success rate of TKJ-CVPR10 to 80.6% when $\sigma_k = 10\%$. There is also a very small drop in success rate of Approximate PSfM-O and Approximate PSfM-O(LRE) but only to 99.1% when $\sigma_k = 10\%$. The success rate of TK-Factor is 0% for all values of σ_k, which as discussed above is because it never computes an invertible upgrade matrix when $N = 3$. Unusually, we see that the mean structure error of bundle adjustment appears worse than our methods, however the median error is very similar. The problems are caused by the fact that for large k the data violate the noise model (it is no longer IID Gaussian), so we may not necessarily observe bundle adjustment giving the most accurate solutions. By contrast we see the mean errors of Approximate PSfM-O and Approximate PSfM-O(LRE) degrade gracefully with increased k. Similarly to the previous experiments we see that the accuracy of Approximate PSfM-O and Approximate PSfM-O(LRE) is indistinguishable when the number of views reaches eight, because the likelihood of there being discrete structure ambiguities diminishes considerably. As σ_k increases we see a greater difference in accuracy between TKJ-CVPR10 and Approximate PSfM-O, which indicates TKJ-CVPR10 cannot handle modelling approximation error as well as Approximate PSfM-O. The results for experiments seven and eight are shown in the last two columns of Figure 3. In these experiments we plot the translation error in the last two rows (recall that the translation error is only relevant when $\gamma > 0$). Again we see virtually no difference between bundle adjustment and Approximate PSfM-O. When $M = 8$ Approximate PSfM-O and Approximate PSfM-O(LRE) are indistinguishable.

In summary, these experiments show that under I. I. D. Gaussian measurement noise there is virtually no gain in the bundle adjustment solution compared to Approximate PSfM-O. This is an unusual and interesting result because bundle adjustment optimises *both* structure and camera poses with the full reprojection error. By contrast Approximate PSfM-O estimates structure by an algebraic upgrade function (Equation (17)). This result tells us something quite profound about the problem. *It indicates that the optimal metric structure is extremely similar to the optimal affine structure up to an upgrade transform, and Equation (17) does an excellent job for finding the transform (or transforms if the problem is ambiguous).*

D. Real-Data Experiments

In this section we present results using real image data. We add to the methods bundle adjustment with a perspective camera (with fixed and pre-calibrated intrinsic matrices), which we call **Best+PBA**. Similarly to Best+OBA we compute this by taking the best solution to structure across all methods, but then we resect the cameras with perspective planar PnP [24]. We then run bundle adjustment to jointly refine the structure (which we constrain to lie on the plane $z = 0$ in world coordinates) and perspective camera poses with Levenberg-Marquardt.

1) Reconstruction with a Textured Planar Surface: The first set of real-data experiments were performed with an unorganised collection of eight views of a textured sheet of A4 paper mounted on a flat surface (Figure 5). The views were taken with a Nikon D800 DSLR with a 120mm lens with fixed focal length and image resolution of 3680×2456. The pattern on the paper measured 23.0×20.5mm with an average distance to the camera of approximately 3.2m. We intrinsically pre-calibrated the camera with Bouguet's toolbox [31] which gave an effective focal length of 1.0068×10^4px and 1.0060×10^4px in the x and y axes respectively (which is approximately 2.27 times the image diagonal). Feature points were computed over the pattern with SIFT [32] using the VLFeat implementation [33], which gave on average 288.3 features per view. We computed ground truth camera poses using a digital image of the paper as a 2D template which we registered in 3D to each view using a direct approach based on the DIRT toolbox [34]. Correspondences and ground truth structure were determined by matching SIFT descriptors and computing the optimal positions of the features on the 2D template given the camera poses. For this we computed all inter-image homographies from the camera poses then computed putative correspondences between each pair by matching features with the closest SIFT descriptors. Correspondences were used if predicted by the homography to within 7 pixels. The correspondences were then chained to give 1842 unique points, and we then refined the points' positions on the 2D template by minimising the reprojection error using Levenberg-Marquardt. The average number of missing correspondence per view was 64.5%.

We measured the performance of each method across two dimensions. The first dimension was the number of views M which we varied from $M = 3$ to $M = 7$. For each M we ran the methods over all possible subsets of M views. We also measured how performance varied with smaller neighbourhoods of correspondences. The purpose was to investigate how methods perform as the number of point correspondences decreased. We performed this by taking each of the 1842 points in turn, and for each point we used only neighbouring correspondences that were within a neighbourhood radius r to that point. In our experiments we varied r between 15% and 60% of the whole pattern's size.

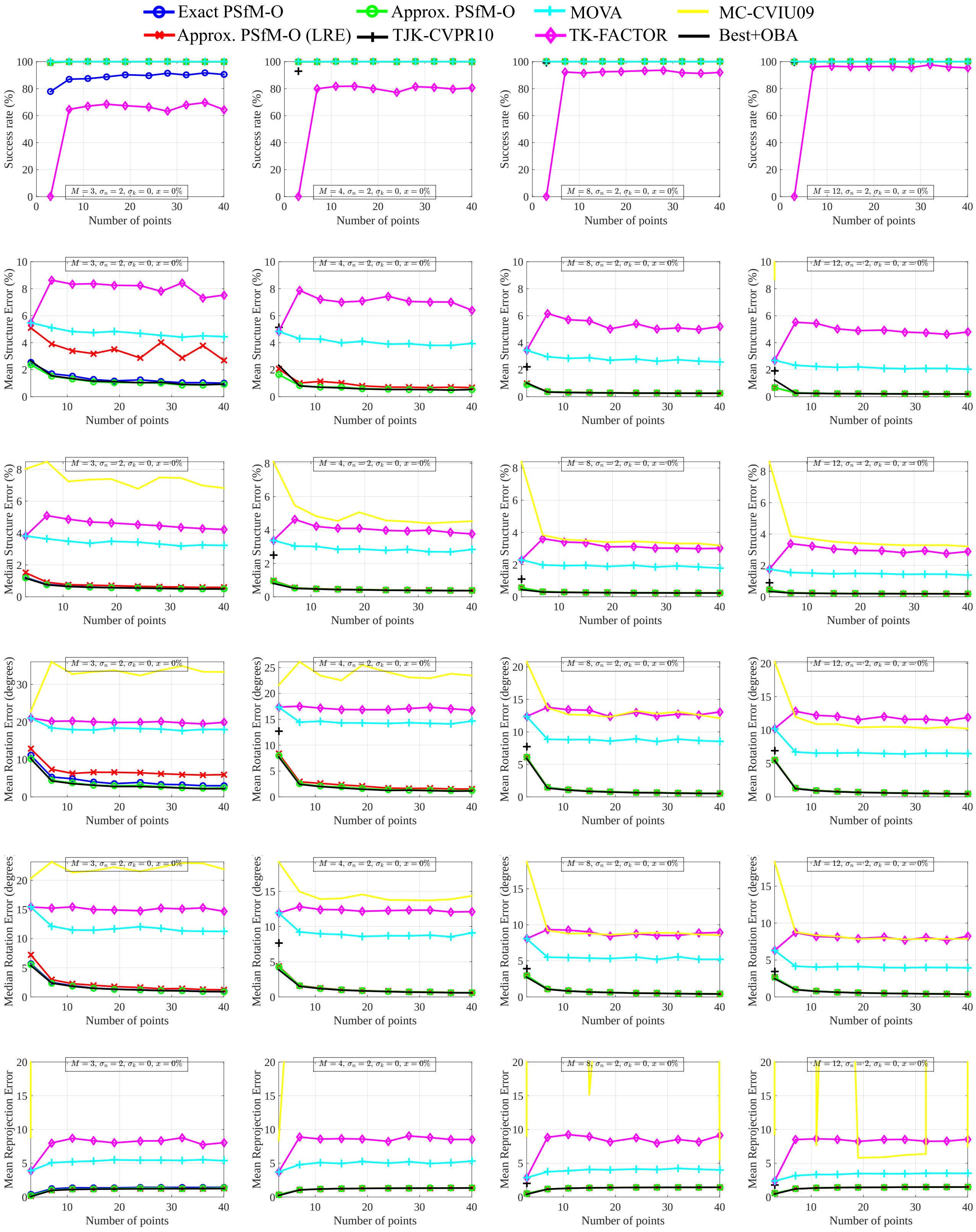

Fig. 2. Simulation experimental results: experiments one to four with one experiment per column. Best viewed in colour.

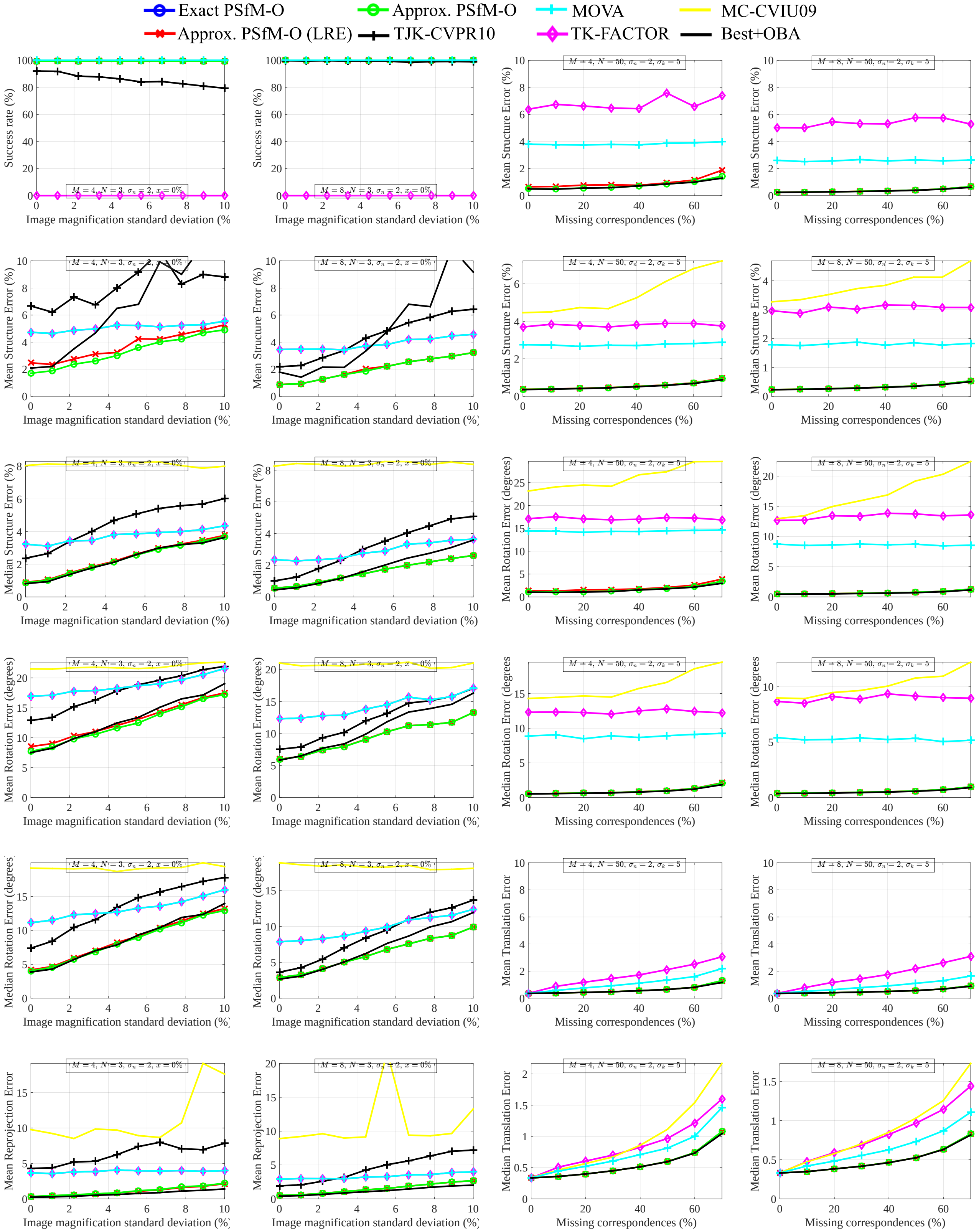

Fig. 3. Simulation experimental results: experiments five to eight with one experiment per column. Best viewed in colour.

The results are shown in Figure 4. Along each column we plot results for $M = 4$, $M = 5$ and $M = 8$ from left to right. Along each row we plot the mean and median structure error, mean and median rotation error, and success rate against the neighbourhood radius. We first inspect structure error. For all M the accuracy of Best+PBA is poor but tends to improve with a larger neighbourhood size. This is because for smaller neighbourhoods planar SfM with perspective cameras becomes badly-conditioned. The results for TK-FACTOR look better than they actually are, which is because its success rate is so low. Therefore very often it had to revert to the solution from MOVA. We again see that across all settings there is very little difference between Approximate PSfM-O and bundle adjustment. In terms of success rate Approximate PSfM-O never dropped below 99.92%.

2) Reconstruction from an Orbiting Image Sequence: The second real-data experiment involved reconstructing the top surface of a bottle cap from an ordered set of orbiting views (Figure 6). The image set consists of 18 1600×1800px views taken with an automatic compact camera with fixed zoom. Views 1, 10 and 18 are shown in Figure 6 (first row, left). We intrinsically pre-calibrated the camera with Bouguet's toolbox which gave an effective focal length of 1.26×10^4px in the x and y axes (which is approximately 6.32 times the image diagonal). Points were computed using the Harris detector on the first view of the bottle cap (using default parameters), which gave 137 points (Figure 6 (row three, left image)). We tracked the points in subsequent views using KLT. To reduce tracking drop-off, for each view $i \in \{2, \ldots 18\}$ we computed an approximate homography between view 1 and i, then back-warped image i to image 1, and then ran KLT on the back-warped image. The approximate homography was computed using SIFT feature matching and RANSAC. Outliers from KLT were detected by refitting the homography with RANSAC to the KLT matches and rejecting matches with a transport error beyond 5 pixels. We used the fact that the bottle cap's top surface is circular to rectify image 1 with a homography (Figure 6 (row three, left image)). Ground truth structure was computed by first computing the optimal 2D affine structure using Algorithm 6 and mapping the point' affine structure to the rectified image. The number of missing correspondences began at 0% in image 1 and rose to 56.2% in image 18.

Similarly to the previous experiment, we measured performance by randomly sampling sub-sets of views from the full collection of 18 views, whose size we varied from 3 to 10. For each size we drew 50 random subsets and computed performance statistics over the 50 subsets. The results for mean structure error, median structure error, mean reprojection error and success rate are shown in Figure 6, second row. We see that PSfM-O again performs very well and there is no significant difference with bundle adjustment. The success rate of Approximate PSfM-O (and Approximate PSfM-O(LRE)) was 100% in all cases. The success rate of TK-Factor generally reduced with more views, and for 10 views was 36%. The method with lowest reprojection error was Best+PBA, however this also produced much higher structure error. This tells us the perspective camera is unsuitable for solving this problem, due to the bottle cap being too small to reliably estimate structure *and* the perspective camera projection matrices. In the last two rows of Figure 6 we show the reconstructed points computed from each method with a subset size of 10. The ground truth structure is shown by the set of red circles. The dark circles show the reconstructed points from a method. Because for each method we performed 50 reconstructions, we overlaid all 50 reconstructions. One can see a good clustering of the points by Approximate PSfM-O and Approximate PSfM-O(LRE) about the ground truth positions.

We also ran a second experiment to test how the methods performed by adding views in sequential order, starting from the first three views. The purpose was to see how accuracy improved as the baseline of the image set increased. The

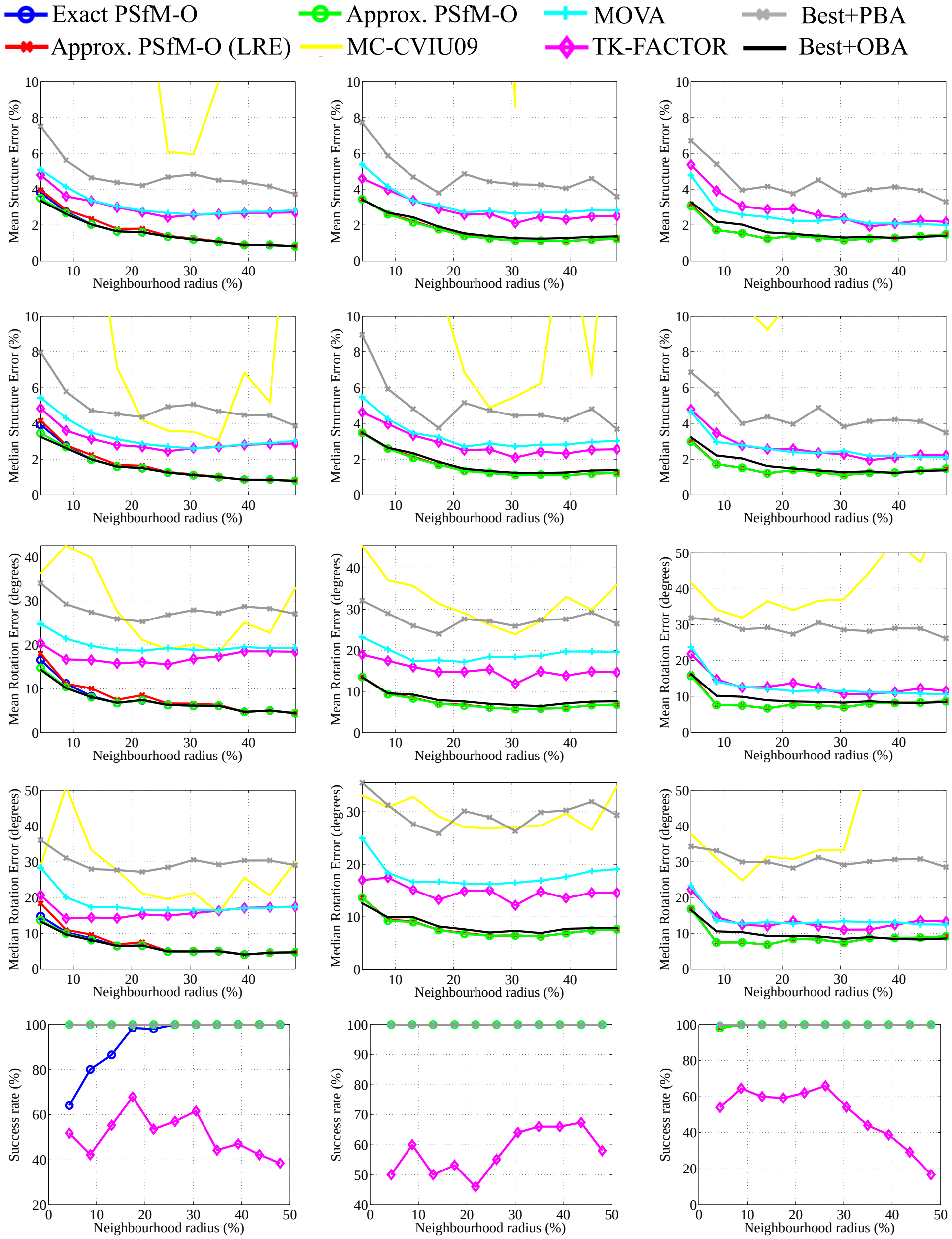

Fig. 4. Results on the image set shown in Figure 5. In the three columns we show results using three, five and eight of the views. In the rows we show the corresponding performance statistics.

results are shown in Figure 6, top-right. One can see a smooth reduction error of Approximate PSfM-O, Approximate PSfM-O(LRE) and Best+OBA as the number of views (and the the baseline) increased. This demonstrates again the accuracy of our method and that there is no significant gain in accuracy by refining our solutions with bundle adjustment.

V. Conclusion

We have presented a number of important technical and theoretical contributions for planar Structure-from-Motion with affine cameras. The problem is fundamentally different to SfM with non-planar structures, because the affine camera

Fig. 5. A real test set consisting of eight unorganised 3680 × 2456 views of a textured flat A4 sheet of paper.

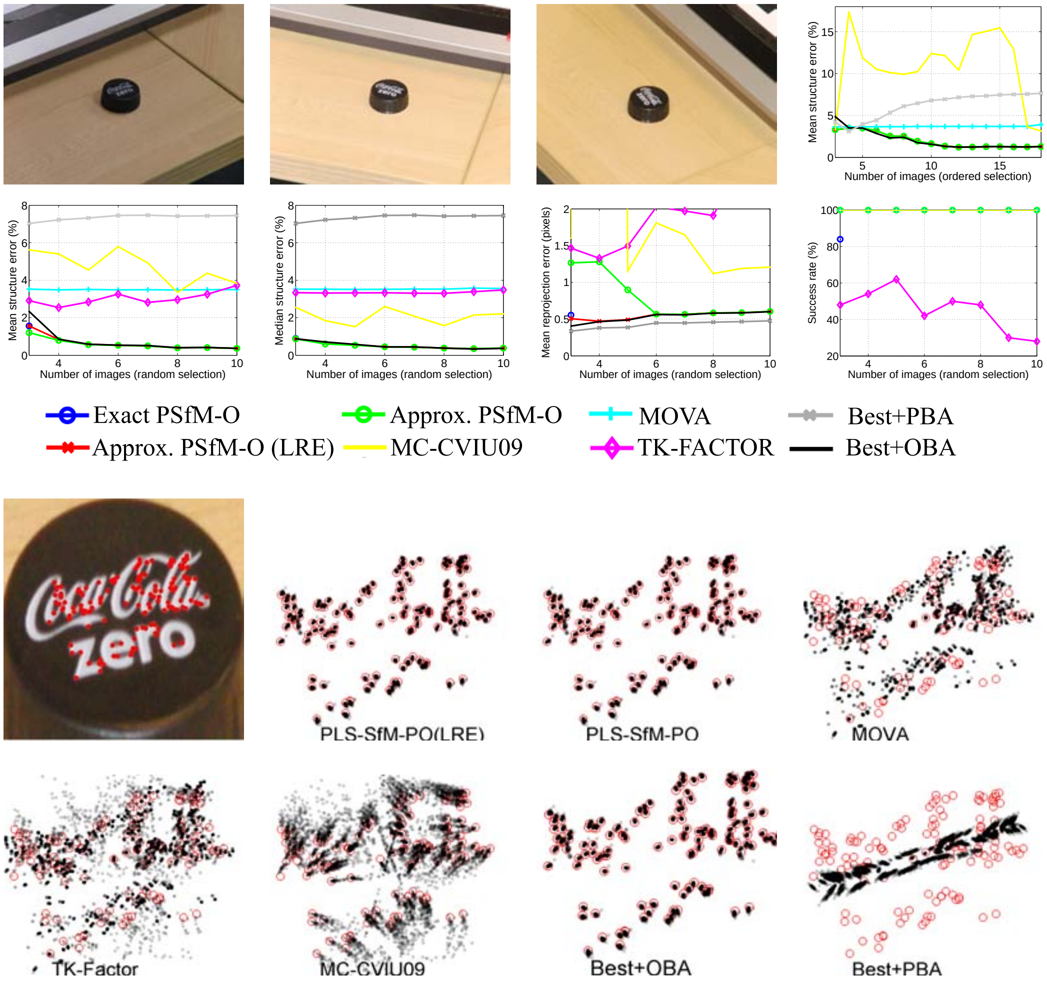

Fig. 6. Results for reconstructing the top section of a bottle top from an orbiting image sequence (best viewed in colour).

models one can use are more restricted, the upgrade constraints are non-linear and non-convex, and the problem is far more ambiguous. We have presented eight new theorems that significantly deepen our understanding of the problem. Our main theoretical result is a complete geometric characterisation of degeneracies with orthographic cameras (*i.e.* the PSfM-O problem). A key to achieving this was Exact PSfM-O, which is the first non-artificially degenerate algorithm to solve the PSfM-O problem. The second main theoretical result is to show that the PSfM-O problem can have discrete structure ambiguities with a general number of views, and to give the necessary and sufficient geometric conditions for disambiguation. We have also presented three cases when SfM may be solvable with other affine cameras, which necessarily requires additional knowledge.

Our main technical contribution is Approximate PSfM-O, which solves the PSfM-O problem in its most general form. Approximate PSfM-O handles cases when there exist discrete structure ambiguities, which is not true of previous algorithms. The solutions from Approximate PSfM-O tend to be very close to locally-optimal metric reconstructions. This has been demonstrated by extensive empirical evaluation which shows that the solutions are not significantly improved by running bundle adjustment. Because Approximate PSfM-O is stratified it does not optimise the reprojection error at all stages (it does this only at the affine reconstruction and camera resection stages). The fact that the results are very close in accuracy compared to fully optimising the reprojection error tells us two important things that were previously unknown. The first is that we can compute the plane's metric structure very well from the optimal affine reconstruction up to an unknown upgrade transform. The second is that our upgrade cost function does an excellent job of finding the transform (or multiple transforms if the problem is ambiguous) in closed-form, even for high levels of measurement noise. In the case of three views we have some theory to explain this phenomenon (Theorem 6). Specifically, if we can exactly upgrade an optimal affine reconstruction to a metric reconstruction then the upgraded reconstruction *is* the optimal metric reconstruction. In these cases running bundle adjustment will do absolutely nothing. Empirically we have found that we can do this for 80 to 90 percent of cases depending on noise (see Section 4c1). More theoretical analysis is required to study the precise relationship between optimising the full reprojection error and the stratified cost functions in our method for more than three views, and we leave this to future work. This is non-trivial and requires uncertainty propagation to analyse how error in the measurements propagate to errors in the upgrade cost function.

REFERENCES

[1] S. Ullman, *The interpretation of visual motion*. The MIT press, 1979.

[2] C. Tomasi and T. Kanade, “Shape and motion from image streams under orthography: a factorization method,” *International Journal of Computer Vision*, vol. 9, pp. 137–154, 1992.

[3] C. J. Poelman and T. Kanade, “A paraperspective factorization method for shape and motion recovery.” *Pattern Analysis and Machine Intelligence*, vol. 19, no. 3, pp. 206–218, 1997.

[4] B. Triggs, “Linear projective reconstruction from matching tensors,” *Image and Vision Computing*, vol. 15, no. 8, pp. 617–625, 1997.

[5] R. Hartley and F. Kahl, “Critical configurations for projective reconstruction from multiple views,” *International Journal of Computer Vision*, vol. 71, no. 1, pp. 5–47, 2007.

[6] P. Sturm, “A case against kruppa's equations for camera self-calibration,” *Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1199–1204, 2000.

[7] R. Szeliski and S. B. Kang, “Shape ambiguities in structure from motion,” *Pattern Analysis and Machine Intelligence*, vol. 19, no. 5, pp. 506–512, 1997.

[8] X.-S. Gao, X.-R. Hou, J. Tang, and H.-F. Cheng, “Complete solution classification for the perspective-three-point problem,” *Pattern Analysis and Machine Intelligence*, vol. 25, no. 8, pp. 930–943, 2003.

[9] L. Quan, “Self-calibration of an affine camera from multiple views,” *International Journal of Computer Vision*, vol. 19, pp. 93–105, 1994.

[10] D. Weinshall and C. Tomasi, “Linear and incremental acquisition of invariant shape models from image sequences.” *Pattern Analysis and Machine Intelligence*, vol. 17, no. 5, pp. 512–517, 1995.

[11] K. Kanatani, Y. Sugaya, and H. Ackermann, “Uncalibrated factorization using a variable symmetric affine camera,” *IEICE Transactions*, vol. 90-D, no. 5, pp. 851–858, 2007.

[12] J. Costeira and T. Kanade, "A multi-body factorization method for independently moving objects," *International Journal of Computer Vision*, vol. 29, pp. 159–179, 1997.

[13] P. Tresadern and I. Reid, "Articulated structure from motion by factorization," in *Computer Vision and Pattern Recognition*, vol. 2, 2005, pp. 1110–1115.

[14] L. Torresani, A. Hertzmann, and C. Bregler, "Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors." *Pattern Analysis and Machine Intelligence*, vol. 30, no. 5, pp. 878–892, 2008.

[15] J. Taylor, A. D. Jepson, and K. N. Kutulakos, "Non-rigid structure from locally-rigid motion." in *Computer Vision and Pattern Recognition*, 2010, pp. 2761–2768.

[16] T. Collins and A. Bartoli, "Locally planar and affine deformable surface reconstruction from video," in *The International Workshop on Vision, Modeling, and Visualization (VMV)*, 2010.

[17] Z. Zhang and A. R. Hanson, "3D reconstruction based on homography mapping," in *In ARPA Image Understanding Workshop*, 1996.

[18] O. D. Faugeras and F. Lustman, "Motion and structure from motion in a piecewise planar environment," *International Journal of Artificial Intelligence and Pattern Recognition*, vol. 2, no. 3, pp. 485–508, 1988.

[19] E. Malis and M. Vargas, "Deeper understanding of the homography decomposition for vision-based control," INRIA, Research Report RR-6303, 2007.

[20] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon, "Bundle adjustment – a modern synthesis," in *Vision Algorithms: Theory and Practice*, 2000, pp. 298–375.

[21] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004.

[22] T. S. Huang and C. H. Lee, "Motion and structure from orthographic projections," *Pattern Analysis and Machine Intelligence*, vol. 11, no. 5, pp. 536–540, 1989.

[23] D. D. Hoffman and B. M. Bennett, "The computation of structure from fixed-axis motion: rigid structures," *Biological Cybernetics*, vol. 54, no. 2, pp. 71–83, 1986.

[24] T. Collins and A. Bartoli, "Infinitesimal plane-based pose estimation," *International Journal of Computer Vision*, vol. 109, no. 3, pp. 252–286, 2014.

[25] R. Horaud, F. Dornaika, B. Lamiroy, and S. Christy, "Object Pose: The Link between Weak Perspective, Paraperspective and Full Perspective," *International Journal of Computer Vision*, vol. 22, pp. 173–189, 1997.

[26] P. Sturm, "Algorithms for plane-based pose estimation," in *The International Conference on Computer Vision and Pattern Recognition, Hilton Head Island, South Carolina, USA*, 2000.

[27] D. Henrion, J.-B. B. Lasserre, and J. Lofberg, "GloptiPoly 3: moments, optimization and semidefinite programming," *Optimization Methods and Software*, vol. 24, no. 4-5, pp. pp. 761–779, 2009.

[28] A. Lobay and D. A. Forsyth, "Shape from texture without boundaries," *International Journal of Computer Vision*, vol. 67, pp. 71–91, 2006.

[29] M. Marques and J. Costeira, "Estimating 3D shape from degenerate sequences with missing data," *Computer Vision and Image Understanding*, vol. 113, no. 2, pp. 261–272, 2009.

[30] V. Lepetit, F. Moreno-Noguer, and P. Fua, "EPnP: An accurate O(n) solution to the PnP problem," *International Journal of Computer Vision*, vol. 81, pp. 155–166, 2009.

[31] J. Y. Bouguet, "A camera calibration toolbox for matlab." [Online]. Available: http://www.vision.caltech.edu/bouguetj/calib_doc/

[32] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[33] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms." [Online]. Available: http://www.vlfeat.org/

[34] A. Bartoli, "Groupwise geometric and photometric direct image registration," *Pattern Analysis and Machine Intelligence*, vol. 30, no. 12, pp. 2098–2108, 2008.

BIOGRAPHIES

Toby Collins received the MSc degree in Artificial Intelligence at the University of Edinburgh (first in class) in 2005. In 2006 he began his PhD in Computer Vision at the University of Edinburgh. Since 2009 he has been a full-time research fellow in ALCoV. His research interests include nonrigid shape analysis, registration and reconstruction, AR for deformable objects and computer assisted intervention.

Adrien Bartoli has held the position of Professor of Computer Science at Université d'Auvergne since fall 2009. He leads the ALCoV (Advanced Laparoscopy and Computer Vision) research group, member of CNRS and Université d'Auvergne, at ISIT. His main research interests include image registration and Shape-from-X for rigid and non-rigid environments, with applications to computer-aided endoscopy.

Appendix A

Why Existing Stratified Methods Cannot Solve Planar SfM with Affine Cameras

Assuming complete measurements, let $\tilde{\mathbf{q}}_i^j \overset{\text{def}}{=} \hat{\mathbf{q}}_i^j - \sum_{k=1}^{N} \hat{\mathbf{q}}_k^j$ be the zero centred correspondence of the j^{th} point in the i^{th} view. Measurements stack to form the $2M \times N$ measurement matrix $\hat{\mathbf{Q}}$ that factorises as follows:

$$\hat{\mathbf{Q}} = \begin{bmatrix} \tilde{\mathbf{q}}_1^1 & \tilde{\mathbf{q}}_1^2 & \cdots & \tilde{\mathbf{q}}_1^N \\ \tilde{\mathbf{q}}_2^1 & \tilde{\mathbf{q}}_2^2 & \cdots & \tilde{\mathbf{q}}_2^N \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\mathbf{q}}_M^1 & \tilde{\mathbf{q}}_M^2 & \cdots & \tilde{\mathbf{q}}_M^N \end{bmatrix} = \text{stack}\left([\mathbf{M}_1]_{2\times3},\ [\mathbf{M}_2]_{2\times3}, \ldots, [\mathbf{M}_M]_{2\times3}\right)\mathbf{S} + \varepsilon \tag{29}$$

where $\varepsilon \in \mathbb{R}^{2M\times N}$ denotes measurement noise. Existing stratified methods would first compute an affine reconstruction using the rank-three approximation of $\hat{\mathbf{Q}}$, then use orthogonality constraints to upgrade it to a metric reconstruction. The optimal affine reconstruction is computed from the SVD of $\hat{\mathbf{Q}}$: $\hat{\mathbf{Q}} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$, where $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices and $\boldsymbol{\Sigma}$ is the diagonal singular value matrix (sorted by decreasing magnitude). Ignoring noise we have

$$\text{stack}\left([\mathbf{M}_1]_{2\times3},\ [\mathbf{M}_2]_{2\times3}, \ldots, [\mathbf{M}_M]_{2\times3}\right) = [\mathbf{U}]_{2M\times3}[\boldsymbol{\Sigma}]_{3\times3}^{1/2}\mathbf{Y},\ \ \mathbf{S} = \mathbf{Y}^{-1}[\boldsymbol{\Sigma}]_{3\times3}^{1/2}[\mathbf{V}]_{N\times3}^\top \tag{30}$$

where $\mathbf{Y}$ is the unknown full-rank 3×3 upgrade matrix. The problem is solved by finding $\mathbf{Y}$ using metric upgrade constraints. For orthographic cameras we have $[\mathbf{M}_i]_{2\times3} \in \mathcal{S}_{2\times3} \Leftrightarrow \mathbf{U}_i\mathbf{Y} \in \mathcal{S}_{2\times3} \Leftrightarrow \mathbf{U}_i\mathbf{Y}\mathbf{Y}^\top\mathbf{U}_i^\top = \mathbf{I}_2$, where $\mathbf{U}_i$ denotes the i^{th} 2×3 sub-block of $[\text{U}]_{2M\times3}[\boldsymbol{\Sigma}]_{3\times3}^{1/2}$. This imposes linear equality constraints on the Gramian matrix $\mathbf{Z} \overset{\text{def}}{=} \mathbf{Y}\mathbf{Y}^\top$. When the structure is planar the maximum theoretical rank of $\mathbf{S}$ is two, so $[\boldsymbol{\Sigma}]_{33} = 0$, and the metric upgrade constraint becomes:

$$\mathbf{U}_i \begin{bmatrix} [\boldsymbol{\Sigma}]_{2\times2}^{1/2}[\mathbf{Z}]_{2\times2}[\boldsymbol{\Sigma}]_{2\times2}^{1/2} & \mathbf{0}_{2\times1} \\ \mathbf{0}_{1\times2} & 0 \end{bmatrix} \mathbf{U}_i^\top = \mathbf{I}_2 \tag{31}$$

Existing stratified methods would try to recover $\mathbf{Z}$ using Equation (31) by relaxing $\mathbf{Z} \succ \mathbf{0}$ to $\mathbf{Z}$ being symmetric, which makes this an unconstrained Linear Least Squares (LLS) problem. If the solution to $\mathbf{Z}$ is positive definite, $\mathbf{Y}$ would be recovered from its Cholesky decomposition (which is up to a 3D unitary gauge transform). However Equation (31) constrains only the top-left 2×2 submatrix of $\mathbf{Z}$. Because $\mathbf{Z}$ is positive definite this means only 3 of its 6 DoFs are constrained, so it cannot be recovered.

Appendix B

Artificial Degeneracies in Previous PSfM-O Methods

The method of [23] produces artificial degeneracies because there exist non-degenerate inputs which cause the coefficients in Equation (2.9) to be zero, which means a solution cannot be computed. An example is as follows:

$$\mathbf{M}_1 = \begin{bmatrix} -0.4893 & -0.0706 & -0.8693 & 0 \\ -0.6538 & -0.6299 & 0.4192 & 0 \end{bmatrix} \quad \mathbf{M}_2 = \begin{bmatrix} -0.5100 & 0.8452 & -0.1597 & 0 \\ -0.4637 & -0.4265 & -0.7766 & 0 \end{bmatrix}$$

$$\mathbf{M}_3 = \begin{bmatrix} -0.5265 & -0.5458 & -0.6518 & 0 \\ -0.4506 & -0.4710 & 0.7584 & 0 \end{bmatrix} \quad \mathbf{S} = \begin{bmatrix} 0 & 0.7492 & 0.7637 \\ 0 & 0.0392 & 0.5588 \\ 0 & 0 & 0 \end{bmatrix}$$

The method of [22] produces artificial degeneracies because it fails when the structure points are colinear in the views, which occurs when the structure plane's normal is orthogonal to the camera projection directions. However this is not not a necessary condition for a degeneracy (see Theorem 1). The method of [15] produces artificial degeneracies because it cannot solve the minimal case of three views, and fails with four or more views when there is discrete structure ambiguity (see Theorem 7 for the geometric conditions when this occurs).

APPENDIX C
AFFINE CAMERA DECOMPOSITION AND INTERPRETATION AS A LINEARIZED PERSPECTIVE CAMERA

Any affine camera can be interpreted as a linearised perspective camera. In Table IV we give the relationship between a perspective camera and its corresponding affine camera by linearising the pinhole projection function $\pi : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ about some 3D point $\mathbf{y} \in \mathbb{R}^3$ in camera coordinates. This also requires decomposing a full-rank affine projection matrix $\mathbf{M}$ according to Equation (3), which is given in Algorithm 5. We obtain different types of affine cameras with different parameterisations of $\mathbf{y}$. Specifically, we obtain an orthographic camera with $\mathbf{y} = \mathrm{stack}(0, 0, 1)$ (the linearisation is made on the optical axis at a depth 1). We obtain a weak-perspective camera with $\mathbf{y} = \mathrm{stack}(0, 0, d)$ (the linearisation is made on the optical axis at a depth d, which can vary between views). We obtain a para-perspective camera with $\mathbf{y} \in \mathbb{R}^3$ (the linearisation can be made anywhere in space).

Algorithm 5 affineCameraFactorisation: Factorises a 2×4 affine projection matrix according to Equation (3)

```
Require: $\mathbf{M} \in \mathbb{R}^{2\times 4}$
function affineCameraFactorisation($\mathbf{M}$)
    $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top \leftarrow \mathrm{svd}([\mathbf{M}]_{2\times 3}),\ \det(\mathbf{V}) = 1$
    $\mathbf{A}\mathbf{Q} \leftarrow \mathrm{lq}\left(\mathbf{U}\mathbf{\Sigma}\right)$                ▷ LQ decomposition
    $\mathbf{R} \leftarrow \begin{bmatrix} \mathbf{Q} & \mathbf{0}_{3\times 1} \\ \mathbf{0}_{1\times 3} & 1 \end{bmatrix} \mathbf{V}^\top$
    $[\mathbf{t}]_{2\times 1} \leftarrow \mathbf{M}\left[\mathbf{0}_{3\times 1}^\top\, 1\right]^\top$
    return $\mathbf{A}, \mathbf{R}, [\mathbf{t}]_{2\times 1}$
```

APPENDIX D
OPTIMAL PLANE-BASED POSE ESTIMATION WITH ORTHOGRAPHIC CAMERAS

For simplicity we drop the view index. We assume there are L point correspondences in the image with $3 \leq L \leq N$. We use $\mathbf{s}_0 \in \mathbb{R}^2$ and $\mathbf{q}_0 \in \mathbb{R}^2$ to denote the point centroids of the structure and image points respectively. We first center the points by subtracting their centroids, which eliminates translation and we are left with computing rotation. We define the centered point sets by $\{\tilde{\mathbf{s}}_j\}$ and $\{\tilde{\mathbf{q}}_j\}$ respectively, and define the unknown sub-rotation matrix $\mathbf{C} \overset{\mathrm{def}}{=} [\hat{\mathbf{R}}]_{2\times 2}$. We then have the problem

$$\begin{aligned} & \underset{\mathbf{C}\in\mathcal{SS}_{2\times 2}}{\arg\min} \textstyle\sum_{j=1}^{L} \|\mathbf{C}[\tilde{\mathbf{s}}_1, \tilde{\mathbf{s}}_2, \ldots \tilde{\mathbf{s}}_L] - [\hat{\mathbf{q}}_1, \hat{\mathbf{q}}_2, \ldots \hat{\mathbf{q}}_L]\|_2^2 \\ = \; & \underset{\mathbf{C}\in\mathcal{SS}_{2\times 2}}{\arg\min}\, \mathrm{trace}\left(\mathbf{C}\mathbf{Z}\mathbf{C}^\top - \mathbf{C}\mathbf{Y}\right) \end{aligned} \tag{32}$$

	The point $\mathbf{y} = d[\mathbf{v}^\top 1]^\top$ where π is linearised	Affine projection matrix $\mathbf{M} = k\mathbf{A}\,[[\mathbf{R}]_{2\times 3}\ [\mathbf{t}]_{2\times 1}]$	Projection direction
Orthographic	$d = 1$ $\mathbf{v} = \mathbf{0}_{2\times 1}$	$k = 1,\ \mathbf{A} \approx [\mathbf{K}]_{2\times 2}$ $\mathbf{R} \approx \mathbf{R}^p$ $\mathbf{t} \approx \mathbf{t}^p + \begin{bmatrix} [\mathbf{K}]^{-1}_{2\times 2}\mathbf{c} \\ 1 \end{bmatrix}$	$\mathbf{a}^\top = [0,0,1]^\top \mathbf{R}^p$
Weak-perspective	$d \in \mathbb{R}^+$ $\mathbf{v} = \mathbf{0}_{2\times 1}$	$k \approx 1/d,\ \mathbf{A} \approx [\mathbf{K}]_{2\times 2}$ $\mathbf{R} \approx \mathbf{R}^p$ $\mathbf{t} \approx \mathbf{t}^p + \begin{bmatrix} [\mathbf{K}]^{-1}_{2\times 2}\mathbf{c} \\ 1 \end{bmatrix}$	$\mathbf{a}^\top = [0,0,1]^\top \mathbf{R}^p$
Para-perspective	$d \in \mathbb{R}^+$ $\mathbf{v} \in \mathbb{R}^2$	$k \approx 1/d,\ \mathbf{A} \approx [\mathbf{K}]_{2\times 2}\hat{\mathbf{A}}$ $\mathbf{R} \approx \hat{\mathbf{R}}\mathbf{R}^p$ $\mathbf{t} \approx \hat{\mathbf{R}}\left(\mathbf{t}^p + \begin{bmatrix} [\mathbf{K}]^{-1}_{2\times 2}\mathbf{c} + \mathbf{v}_{2\times 1} \\ 1 \end{bmatrix}\right)$ $\left(\hat{\mathbf{A}}, \hat{\mathbf{R}}\right) = \text{afactor}([\mathbf{I}_2 \vert -\mathbf{v} \vert \mathbf{0}_{2\times 1}])$	$\mathbf{a}^\top = [0,0,1]^\top \hat{\mathbf{R}}\mathbf{R}^p$

TABLE IV

ORTHOGRAPHIC, WEAK-PERSPECTIVE AND PARA-PERSPECTIVE CAMERAS DEFINED BY A LINEARISED PERSPECTIVE CAMERA. THE TERMS $\mathbf{K} \in \mathbb{R}^{3\times 3}$, $\mathbf{c} \in \mathbb{R}^2$, $\mathbf{R}^p \in \mathcal{SO}_3$ AND $\mathbf{t}^p \in \mathbb{R}^3$ DENOTE THE PERSPECTIVE CAMERA'S INTRINSIC CALIBRATION MATRIX, PRINCIPAL POINT, ROTATION AND TRANSLATION RESPECTIVELY. THE FUNCTION $\pi(x,y,z) \overset{\text{def}}{=} 1/z\,[x,y]^\top$ IS A PINHOLE PROJECTION AND $\approx$ DENOTES AN APPROXIMATION UP TO FIRST-ORDER.

where

$$\begin{aligned} \mathbf{Z} &\overset{\text{def}}{=} [\mathbf{s}_1, \mathbf{s}_2, \ldots \mathbf{s}_L]^\top [\mathbf{s}_1, \mathbf{s}_2, \ldots \mathbf{s}_L] \\ \mathbf{Y} &\overset{\text{def}}{=} 2\,[\mathbf{s}_1, \mathbf{s}_2, \ldots \mathbf{s}_L]^\top [\hat{\mathbf{q}}_1, \hat{\mathbf{q}}_2, \ldots \hat{\mathbf{q}}_L] \end{aligned} \tag{33}$$

From the spectral definition of $\mathcal{SS}_{2\times 2}$ we have $\mathbf{C} \in \mathcal{SS}_{2\times 2} \Leftrightarrow \det(\mathbf{C}^\top\mathbf{C} - \mathbf{I}_2) = 0,\ \det(\mathbf{C}) \leq 1$. The problem is therefore a convex quadratic optimization on $\mathbf{C}$ subject to quadratic equality and inequality constraints. This can be cast as a Generalised Problem of Moments [27] and because it is a small scale low-order problem is exactly the type of problem that can be solved globally with Gloptipoly. Given $\mathbf{C}$ we reconstruct the two full rotation matrix solutions $\hat{\mathbf{R}}_a$ and $\hat{\mathbf{R}}_b$ as follows. The third column of $\hat{\mathbf{R}}_a$ is given by stack $\left([\mathbf{F}]_{11}^{1/2}, [\mathbf{F}]_{22}^{1/2}, \det(\mathbf{C})\right)$ where $\mathbf{F} \overset{\text{def}}{=} \mathbf{I}_2 - \mathbf{C}^\top\mathbf{C}$. The third column of $\hat{\mathbf{R}}_b$ is given by stack $\left(-[\mathbf{F}]_{11}^{1/2}, -[\mathbf{F}]_{22}^{1/2}, \det(\mathbf{C})\right)$. The third rows of $\hat{\mathbf{R}}_a$ and $\hat{\mathbf{R}}_b$ are then given by cross-producting their first two rows. Because point centroids are preserved by affine transforms the corresponding solutions to translation are $\hat{\mathbf{t}}_a = \hat{\mathbf{t}}_b = \mathbf{q}_0 - \mathbf{C}\mathbf{s}_0$.

APPENDIX E
MOVA: THE FALLBACK METHOD

MOVA is an approximate stratified-based solution that uses the following heuristic: if the camera orientations are distributed randomly and independently then with a sufficiently large number of views there is likely to be one that has a fronto-parallel view of the structure plane (in the limit the probability reaches 1). If we do indeed have a fronto-parallel view then the points in its image give the metric structure up to noise. In reality we are never likely to have such a view, but we can approximate metric structure using the view that is *most* fronto-parallel. This is the view where $\det^2([\mathbf{R}_i]_{2\times 2})$ is largest. Because $\det([\mathbf{R}_i]_{2\times 2}) = \det(\mathbf{M}_i^A\mathbf{X}) = \det(\mathbf{M}_i^A)\det(\mathbf{X})$, it is also the view where $\det^2(\mathbf{M}_i^A)$ is largest, so it can be determined entirely from the affine structure matrix. Suppose this is given by view i^*. Assuming this view is fronto-parallel we have $\mathbf{M}^A_{i^*}\mathbf{W}\mathbf{M}^{A\top}_{i^*} \approx \mathbf{I}_2$, so we can approximate the upgrade matrix by $\mathbf{W} \approx \left(\mathbf{M}^A_{i^*}\right)^{-1}\left(\mathbf{M}^A_{i^*}\right)^{-\top}$. Metric structure can then be computed by factoring $\mathbf{W}$ as described in §III-A. We use MOVA as the fall-back solution because we can always compute metric structure unless for all views $\det(\mathbf{M}^A_i) = 0$ (in which case we cannot perform the matrix inversion). This happens when the structure plane normal is orthogonal to the projection direction in all views, and is a rare occurrence in practice.

APPENDIX F
ALGORITHM FOR COMPUTING PLANAR AFFINE STRUCTURE WITH GENERAL AFFINE CAMERAS AND MISSING MEASUREMENTS

The algorithm we use is given in Algorithm 6.

Algorithm 6 (Planar Affine Structure with General Affine Cameras from Point Correspondences)

```
Require: $\{\mathbf{q}_i^j\}$    ▷ point correspondences with view index $i \in \{1,\dots,M\}$ point index $j \in \{1,\dots,N\}$
function affineReconstruct2D($\{\mathbf{q}_i^j\}$))
    Construct a directed graph $\mathcal{G}$ of $M$ nodes with weighted edges $E(j,k) \in \mathbb{R}^{+M\times M}$. $E(j,k)$ is the conditioning number of the linear system for solving the Least Squares 2D affine transform from view $j$ to $k$ using points measured in both views.
    Compute the connected components of $\mathcal{G}$ and remove all views not connected to the largest component.
    Assign the root view $i^*$ to be the one with the shortest sum of paths from all other views.
    Compute 2D affine transforms $\mathbf{F}_i$ from $i$ to $i^*$ by chaining affine transforms along the shortest path to $i^*$.
    Transfer all measured points to the root view using $\mathbf{F}_i$. For each point $j \in \{1,\dots,M\}$ compute its median $\mathbf{s}_j \in \mathbb{R}^{2\times 1}$ in the root view.
    Initialise the affine structure $\mathbf{S}^A$ with $\mathbf{s}_j$ in its $j^{th}$ column.
    Compute Least Squares 2D affine transform $\mathbf{F}'_i$ mapping $\mathbf{S}^A$ to measured points in $i^{th}$ view.
    Jointly refine $\mathbf{F}'_i$ and $\mathbf{S}^A$ to minimise the affine reconstruction reprojection error using Levenberg-Marquardt.
    return $\mathbf{M}^A = \text{stack}\left([\mathbf{F}'_1]_{2\times2},\dots,[\mathbf{F}'_M]_{2\times2}\right), \mathbf{S}^A$
```

APPENDIX G
DERIVATION OF ALGORITHM 1

Equation (16) is a quadratic constraint on $\mathbf{w}$ that has the form

$$\begin{aligned} &a_i w_1 + b_i w_2 + c_i w_3 + d_i(w_1 w_3 - w_2^2) = 1 \\ &\mathbf{E}_i \overset{\text{def}}{=} \mathbf{M}_i^{A\top}\mathbf{M}_i^A,\ a_i \overset{\text{def}}{=} [\mathbf{E}_i]_{11},\ b_i \overset{\text{def}}{=} 2[\mathbf{E}_i]_{12},\ c_i \overset{\text{def}}{=} [\mathbf{E}_i]_{22},\ d_i \overset{\text{def}}{=} -\det(\mathbf{E}_i) \end{aligned} \tag{34}$$

We solve this system by introducing the determinant of $\mathbf{W}$ as an auxiliary variable $s \overset{\text{def}}{=} w_1 w_3 - w_2^2$. The following is then equivalent to Equation (16):

$$\mathbf{A}_E\,\text{stack}(\mathbf{w}, s) = \mathbf{1}_{3\times1},\ w_1 w_3 - w_2^2 - s = 0 \tag{35}$$

where $\mathbf{A}_E$ is a 3×4 matrix holding $[a_i\,b_i\,c_i\,d_i]$ in its i^{th} row. This system has three linear equations and one quadratic equation, so it must have either 0, 1 or 2 real solutions. We solve it by first using the linear constraints to find $\text{stack}(\mathbf{w}, s)$ up to a 1-DoF affine subspace: $\text{stack}(\mathbf{w}, s) = \text{stack}(\mathbf{w}', s') + \alpha\mathbf{z}$, where $\mathbf{z}$ is a unit nullvector of $\mathbf{A}_E$, α is an unknown scalar and $(\mathbf{w}' \in \mathbb{R}^3, s' \in \mathbb{R})$ is any solution to $\mathbf{A}_E\,\text{stack}(\mathbf{w}, s) = \mathbf{1}_{3\times1}$. We compute $(\mathbf{w}', s')$ with the Moore-Penrose pseudoinverse: $\mathrm{stack}(\mathbf{w}', s') = \mathbf{A}_E^\top(\mathbf{A}_E\mathbf{A}_E^\top)^{-1}\mathbf{1}_{3\times 1}$. We then resolve α with the quadratic constraint in Equation (35). This is given by all real solutions to $a\alpha^2 + b\alpha + c = 0$ where

$$a \overset{\text{def}}{=} z_2^2 - z_1 z_3,\ b \overset{\text{def}}{=} z_4 - w_1' z_3 + 2w_2' z_2 - w_3' z_1,\ c \overset{\text{def}}{=} w_2'^2 + 1 - w_1' w_3' \tag{36}$$

For each solution to α, $\mathbf{w}$ is given by $\mathbf{w} = \mathbf{w}' + \alpha[\mathbf{z}]_{3\times 1}$. We then test whether this solution satisfies $f\mathbf{w}) \succ \mathbf{0}$, which means it is a feasible upgrade solution and we can recover $\mathbf{X}$ (and hence metric structure) from it using Equation (13). We also test whether the inequality constraints in the right-hand side of Equation (15) are satisfied. If so, then we know that all constraints in problem (14) are satisfied.

Appendix H
Proof of Theorem 1

A. *Definitions and Theorem 1 in a Compact Form*

We define $(\tilde{\mathbf{M}}^A, \tilde{\mathbf{S}}^A)$ to be a 2D affine reconstruction computed from noise-free measurements, with the rank of $\tilde{\mathbf{S}}^A$ being equal to the rank of the metric structure matrix $\mathbf{S}$ (which is at most two). We refer to solving PSfM-O using $(\tilde{\mathbf{M}}^A, \tilde{\mathbf{S}}^A)$ as the *noise-free PSfM-O problem*:

$$\boxed{\begin{array}{l}\text{The noise-free PSfM-O problem:}\\ \text{find}\,\mathbf{w} \in \mathbb{R}^3 \ \text{ s.t.}\\ \begin{cases} \tilde{\mathbf{M}}_i^A f(\mathbf{w}) \tilde{\mathbf{M}}_i^{A\top} \in \mathcal{G}_{2\times 2},\ \ \forall i \in \{1, 2, \dots M\} & (a)\\ f(\mathbf{w}) \succ \mathbf{0} & (b)\end{cases}\end{array}} \tag{37}$$

The number of solutions to problem (37) gives the number of metric structure solutions.

Definition 1. An input $(\tilde{\mathbf{M}}^A, \tilde{\mathbf{S}}^A)$ is a *degenerate input* if and only if problem (37) has an infinite number of solutions.

Definition 2. An input $(\tilde{\mathbf{M}}^A, \tilde{\mathbf{S}}^A)$ is a *non-degenerate input* if and only if problem (37) has a finite number of solutions.

We use the following terms defined in §I-C: trivial camera degeneracies, critical motion sequences, structural degeneracies, mixed degeneracies, artificial degeneracies and Non-artificially Degenerate Algorithms (NADAs). We also define two more geometric entities:

Definition 3. The *camera Gramian matrix* $\mathbf{G}_i$ for view i:

$$\begin{aligned}\mathbf{G}_i \in \mathcal{G}_{2\times 2} \quad &\overset{\text{def}}{=} [\mathbf{R}_i]_{2\times 2}^\top [\mathbf{R}_i]_{2\times 2}\\ &= \mathbf{X}^\top \tilde{\mathbf{M}}_i^{A\top} \tilde{\mathbf{M}}_i^A \mathbf{X}\\ &= \mathbf{I}_2 - [\mathbf{a}_i]_{2\times 1}[\mathbf{a}_i]_{2\times 1}^\top\end{aligned} \tag{38}$$

The second line is because $[\mathbf{R}_i]_{2\times 2} = \tilde{\mathbf{M}}_i^A \mathbf{X}$ and the third line comes from the fact that $\mathbf{a}_i$ is the third row of $\mathbf{R}_i$ and $\mathbf{R}_i$ is unitary. The camera Gramian matrix is important as a tool for geometrically interpreting the problem's degeneracies.

Definition 4. The scalar D with $1 \leq D \leq M$ is the number of unique camera Gramian matrices in the scene.

From Equation (38) we have:

$$\mathbf{G}_i = \mathbf{G}_j \Leftrightarrow [\mathbf{a}_i]_{2\times 1} = \pm[\mathbf{a}_j]_{2\times 1} \Leftrightarrow \mathbf{a}_i = \begin{bmatrix} \pm\mathbf{I}_2 & \mathbf{0}_{2\times 1} \\ \mathbf{0}_{1\times 2} & \pm 1 \end{bmatrix} \mathbf{a}_j \tag{39}$$

This says that two camera Gramian matrices are equivalent if and only if the projection directions of the two cameras are the same up to a sign change and a reflection about the structure plane (recall the structure plane is defined in world coordinates on the plane $z = 0$). This means D is also the number of projection directions in the scene that are unique up to reflections about the structure plane and changes of sign.

The trivial camera degeneracy stated in Theorem 1 is equivalent to the condition $D < 3$. The critical motion sequence stated in Theorem 1 is when all projection directions lie on a plane which is orthogonal to the structures plane (Figure 1, right). This is equivalent to the condition:

$$\exists \mathbf{a} \neq \mathbf{0}_{2\times 1} \;\; \text{s.t.} \forall i \in \{1, 2, \ldots, M\}, \;\; [\mathbf{a}_i]_{2\times 1} \propto \mathbf{a} \tag{40}$$

Theorem 1 is then stated compactly as follows:

$$\left(\tilde{\mathbf{M}}^A, \tilde{\mathbf{S}}^A\right) \text{is degenerate} \;\Leftrightarrow\; \text{rank}(\mathbf{S}) < 2 \,\text{or}\, D < 3 \,\text{or}\,\text{Eq. (40) holds} \tag{41}$$

This states that there is no mixed degeneracy in PSfM-O and a structural degeneracy only occurs when $\text{rank}(\mathbf{S}) < 2$ (*i.e.* the structures points being co-linear).

The reverse implication of Equation (41) is easy to prove and given at the end of this section. The forward implication trivially holds when $\text{rank}(\tilde{\mathbf{S}}^A) < 2$ because $\text{rank}(\tilde{\mathbf{S}}^A) = \text{rank}(\mathbf{S})$ (from the definition of $\tilde{\mathbf{S}}^A$ at the beginning of this section). The forward implication when $\text{rank}(\tilde{\mathbf{S}}^A) = 2$ is however not easy to prove. We do this first for the minimal case of $M = 3$ views. The generalisation to $M > 3$ views then follows quite easily. To ease readability we use $\bar{\mathbf{S}}^A$ to denote a rank-two affine structure matrix.

B. *Proof of Forward Implication of Theorem 1 with $M = 3$ views*

When $M = 3$ we can solve problem (37) with Algorithm 1. We then prove the forward implication with a hypothetical syllogism:

$$\begin{aligned} &(\tilde{\mathbf{M}}^A, \bar{\mathbf{S}}^A)\,\text{is degenerate} \Rightarrow \text{Algorithm 1 fails} \\ &\text{Algorithm 1 fails} \Rightarrow D < 3 \text{ or Eq. (40) holds} \\ &\therefore (\tilde{\mathbf{M}}^A, \bar{\mathbf{S}}^A)\,\text{is degenerate} \Rightarrow \text{D} < 3 \text{ or Eq. (40) holds} \end{aligned} \tag{42}$$

The first line is true by definition (because all algorithms fail with a degenerate scene). Our task is to prove the second line. We do this with the following lemmas, recalling that in Exact PSfM-O $\mathbf{A}_E$ is the linear constraint matrix (see Equation (35)) and a, b and c are the three quadratic coefficients (see Equation (36)).

Lemma 1. *Exact PSfM-O_Upgrade* $\left(\tilde{\mathbf{M}}^{\mathrm{A}}\right)$ fails $\Leftrightarrow \text{rank}(\mathbf{A}_E) \leq 2 \,\text{or}\, a = b = c = 0$

Lemma 2. $\text{rank}(\mathbf{A}_E) \leq 2 \Rightarrow (a = b = c = 0)\,\text{does not hold}$

Lemma 3. $\text{rank}(\mathbf{A}_E) \leq 2$ and $\text{rank}(\mathbf{S}) = 2 \Rightarrow D < 3$ or Eq. (40) holds

Lemmas 1 and 2 tell us that the only time Exact PSfM-O fails is when the matrix $\mathbf{A}_E$ is rank-deficient. Lemma 3 tells us that when $\mathbf{A}_E$ is rank-deficient and $\mathrm{rank}(\mathbf{S}) = 2$ the right side of Equation (41) must hold, which completes the proof.

Proof of Lemma 1: Exact PSfM-O fails if and only if we cannot compute the upgrade matrix using Algorithm 1. This can happen for one of two reasons. The first is at Algorithm 1, line 6 and happens when $\mathbf{A}_E$ is rank-deficient (*i.e.* $\mathrm{rank}(\mathbf{A}_E) \leq 2$). This means we cannot compute $\mathbf{z}$ uniquely up to scale (*i.e.* we cannot compute a 1D affine subspace for the upgrade matrix). If however $\mathbf{A}_E$ is full-rank then we can always compute $\mathrm{stack}(\mathbf{w}', s')$ with line 7. The second place where Algorithm 1 may fail is at line 11 and happens when all coefficients in the quadratic equation are zero: $a = b = c = 0$. This means we cannot resolve α and so we cannot resolve where in the affine subspace the upgrade matrix exists. □

Proof of Lemma 2: We prove this by splitting the space of full-rank $\mathbf{A}_E$ matrices into two sets and showing that in either set $a = b = c = 0$ is contradicted. Set 1 is when $\det(\mathbf{E}_i) = 0 \, \forall i \in \{1, 2, 3\}$. Set 2 is the complement (when $\exists i \in \{1, 2, 3\}$, $\det(\mathbf{E}_i) \neq 0$).

Set 1: By definition in Set 1 the fourth column of $\mathbf{A}_E$ is all-zeros. Therefore $\mathrm{rank}\,(\mathbf{A}_E) = 3 \Rightarrow \mathbf{z} = \pm[0\,0\,0\,1]^\top$. However from Equation (36) this implies $b = \pm 1$ which contradicts $b = 0$.

Set 2: Without loss of generality let $\det(\tilde{\mathbf{M}}_1^A) \neq 0$, which implies $\tilde{\mathbf{M}}_1^A$ is full-rank. Because the affine reconstruction is up to an arbitrary full-rank 2D affine transform, the problem does not change by redefining the factors with $\tilde{\mathbf{M}}_i^A \leftarrow \tilde{\mathbf{M}}_i^A \left(\tilde{\mathbf{M}}_1^A\right)^{-1}$ and $\mathbf{S}^A \leftarrow \tilde{\mathbf{M}}_1^A \mathbf{S}^A$. Thus without loss of generality we can assume $\tilde{\mathbf{M}}_1^A = \mathbf{I}_2$. We then have

$$\begin{aligned} &(c = 0) \Rightarrow (w_1' w_3' - w_2'^2 = s') && (a) \\ &([1\,0\,1\,-1]\,\mathrm{stack}(\mathbf{w}', s') = 1) \Rightarrow w_1' + w_3' = s' && (b) \end{aligned} \tag{43}$$

Equation (43-a) comes from the definition of c in Equation (36). Equation (43-b) comes from the first linear constraint in Equation (35). When $c = 0$, this means the quadratic constraint in Equation (35) is satisfied by $s \leftarrow s'$ and $\mathbf{w} \leftarrow \mathbf{w}'$. By definition, $(\mathbf{w}', s')$ also satisfies the linear constraints in Equation (35), which means $s \leftarrow s'$ and $\mathbf{w} \leftarrow \mathbf{w}'$ is a solution to Equation (35), and is therefore a solution to Equation (16). Now because $\tilde{\mathbf{M}}_1^A = \mathbf{I}_2$, we have $\det\left(f(\mathbf{w}') - \mathbf{I}_2\right) = 0$, which implies either $\lambda_1(w(\mathbf{w}')) = 1$ or $\lambda_2(w(\mathbf{w}')) = 1$. However this is contradicted by Equation (43-b). To see this, we can eliminate s' from the right sides of Equations (43-a,b) to give:

$$\begin{aligned} &w_1' + w_3' = w_1' w_3' - w_2'^2 \\ &\Leftrightarrow \mathrm{trace}(w(\mathbf{w}')) = \det(w(\mathbf{w}')) \\ &\Leftrightarrow \lambda_1(w(\mathbf{w}')) + \lambda_2(w(\mathbf{w}')) = \lambda_1(w(\mathbf{w}'))\lambda_2(w(\mathbf{w}')) \end{aligned} \tag{44}$$

If $\lambda_1(w(\mathbf{w}')) = 1$ this means $1 + \lambda_2(w(\mathbf{w}')) = \lambda_2(w(\mathbf{w}'))$ which is false for all values of $\lambda_2(w(\mathbf{w}'))$. If $\lambda_2(w(\mathbf{w}')) = 1$ this means $1 + \lambda_1(w(\mathbf{w}')) = \lambda_1(w(\mathbf{w}'))$ which is false for all values of $\lambda_1(w(\mathbf{w}'))$. Therefore we have a contradiction. □

Proof of Lemma 3: From the definition of $\mathbf{A}_E$ in Equation (35) we have:

$$\begin{aligned} &\mathrm{rank}(\mathbf{A}_E) \leq 2 \Rightarrow \exists \alpha, \beta \in \mathbb{R}, \text{ s.t.} \forall \{i, j, k\} \in \mathrm{perm}(\{1, 2, 3\}) \\ &\begin{cases} \tilde{\mathbf{M}}_i^{A\top} \tilde{\mathbf{M}}_i^A = \alpha \tilde{\mathbf{M}}_j^{A\top} \tilde{\mathbf{M}}_j^A + \beta \tilde{\mathbf{M}}_k^{A\top} \tilde{\mathbf{M}}_k^A & (a) \\ \det(\tilde{\mathbf{M}}_i^A) = \alpha \det(\tilde{\mathbf{M}}_j^A) + \beta \det(\tilde{\mathbf{M}}_k^A) & (b) \end{cases} \end{aligned} \tag{45}$$

Equation (45-a) comes from the first three columns of $\mathbf{A}_E$, and Equation (45-b) comes from the fourth column. These equations impose linear constraints on the camera Gramian matrices:

$$\begin{array}{ll} \text{Eq. (45-a)} \Rightarrow (\mathbf{G}_i = \alpha\mathbf{G}_j + \beta\mathbf{G}_k) & (a) \\ \text{Eq. (45-b)} \Rightarrow (\det(\mathbf{G}_i) = \alpha\det(\mathbf{G}_j) + \beta\det(\mathbf{G}_k)) & (b) \end{array} \quad (46)$$

This comes by pre and post-multiplying Equation (45-a,b) by $\mathbf{X}^\top$ and $\mathbf{X}$ respectively and substituting in $\mathbf{G}_i$ using Equation (38). Because $\mathbf{G}_i \in \mathcal{G}_{2\times2}$, $\mathrm{trace}(\mathbf{G}_i) = \lambda_1(\mathbf{G}_i) + \lambda_2(\mathbf{G}_i) = 1 + \det(\mathbf{G}_i)$, so taking the trace of the right hand of Equation (46-a) gives:

$$1 + \det(\mathbf{G}_i) = \alpha(1 + \det(\mathbf{G}_j)) + \beta(1 + \det(\mathbf{G}_k)) \quad (47)$$

Subtracting Equation (46-b) from both sides of Equation (47) gives $\beta = 1 - \alpha$. We now take the right side of Equation (46-a) and substitute β by $(1-\alpha)$:

$$\mathbf{G}_i = \alpha\mathbf{G}_j + (1-\alpha)\mathbf{G}_k \Leftrightarrow [\mathbf{a}_i]_{2\times1}[\mathbf{a}_i]_{2\times1}^\top = \alpha[\mathbf{a}_j]_{2\times1}[\mathbf{a}_j]_{2\times1}^\top + (1-\alpha)[\mathbf{a}_k]_{2\times1}[\mathbf{a}_k]_{2\times1}^\top \quad (48)$$

The right part comes by substituting the camera Gramian matrices for the camera projection directions using Equation (38). For what projection directions does Equation (48) hold? Clearly the determinant of both sides of the second equality in Equation (48) must be zero. Taking the right hand side, after simplification we have:

$$\alpha(1-\alpha)\det\begin{bmatrix} [\mathbf{a}_j]_{2\times1} & [\mathbf{a}_k]_{2\times1} \end{bmatrix} = 0 \quad (49)$$

This holds if either $\alpha = 0$, $\alpha = 1$ or $[\mathbf{a}_j]_{2\times1} \propto [\mathbf{a}_k]_{2\times1}$. If $\alpha = 0$ then $\mathbf{G}_i = \mathbf{G}_k$, which implies $D < 3$ (*i.e.* a trival camera degeneracy). Similarly we have a trivial camera degeneracy when $\alpha = 1$. In the third case we have $[\mathbf{a}_j]_{2\times1} \propto [\mathbf{a}_k]_{2\times1}$. From Equation (48) we therefore have $[\mathbf{a}_i]_{2\times1}[\mathbf{a}_i]_{2\times1}^\top \propto [\mathbf{a}_j]_{2\times1}[\mathbf{a}_j]_{2\times1}^\top \propto [\mathbf{a}_k]_{2\times1}[\mathbf{a}_k]_{2\times1}^\top$, which implies Equation (40). □

C. Proof of Theorem 1 (Forward Implication for Arbitrary M)

If the degeneracy is caused by the scene's structure (*i.e.* $\mathrm{rank}(\mathbf{S}) < 2$) then the scene is degenerate no matter the value of M. Consider instead $\mathrm{rank}(\mathbf{S}) = 2$. The forward implication of Equation (41) is proved by showing that $D \geq 3$ implies Equation (40) holds. When the scene is degenerate there cannot exist a subset $\mathcal{I} \in \{1, 2, \ldots, M\}^3$ of three views that is non-degenerate. This is because if there were such a subset then we could compute the upgrade matrix using only the views in $\mathcal{I}$ and the problem would be solved. When $D \geq 3$ we have at least three camera projection directions that are distinct up to sign changes and reflections about the structure plane. For all subsets of three views these camera projection directions must lie on a plane that is orthogonal to the structure plane. It therefore follows that *all* of the camera projections must lie on this plane. This is equivalent to Equation (40) holding. □

D. Proof of Equation (41) (reverse implication)

Proof that $D < 3 \Rightarrow \left(\tilde{\mathbf{M}}^A, \tilde{\mathbf{S}}^A\right)$ is degenerate: By definition $D < 3$ means there are fewer than three unique camera Gramian matrices. Because $\mathbf{G}_i = [\mathbf{R}_i]_{2\times2}^\top[\mathbf{R}_i]_{2\times2}$ and $\mathbf{G}_j = [\mathbf{R}_j]_{2\times2}^\top[\mathbf{R}_j]_{2\times2}$ by definition, $(\mathbf{G}_i = \mathbf{G}_j) \Leftrightarrow ([\mathbf{R}_i]_{2\times2} = [\mathbf{R}_j]_{2\times2}\mathbf{U}) \Leftrightarrow (\mathbf{M}_i^A\mathbf{X} = \mathbf{M}_j^A\mathbf{X}\mathbf{U})$ for some 2D unitary matrix $\mathbf{U}$. This means $\mathbf{M}_i^A\mathbf{X} \in \mathcal{SS}_{2\times2} \Leftrightarrow \mathbf{M}_j^A\mathbf{X} \in \mathcal{SS}_{2\times2}$ (*i.e.* the upgrade constraint is satisfied by view i if and only if it is satisfied by view j). Therefore given view i, view j provides no extra constraints on the upgrade matrix. Therefore when $D \leq 3$ there are fewer than three constraints on $\mathbf{X}$ (which has 3DoFs). □

Proof that Equation $(40) \Rightarrow \left(\tilde{\mathbf{M}}^A, \tilde{\mathbf{S}}^A\right)$ is degenerate*:* Without loss of generality we rotate world coordinates about the z-axis so that $\mathbf{a} = [1, 0]^\top$ (*i.e.* the azimuths of all cameras is now zero). The camera rotations are now of the form

$$\begin{aligned} \mathbf{R}_i &= \begin{bmatrix} R_{2D}\left(\psi_i\right) & \mathbf{0}_{2\times 1} \\ \mathbf{0}_{1\times 2} & 1 \end{bmatrix} \begin{bmatrix} \cos(\theta_i) & 0 & \sin(\theta_i) \\ 0 & 1 & 0 \\ -\sin(\theta_i) & 0 & \cos(\theta_i) \end{bmatrix} \\ [\mathbf{R}_i]_{2\times 2} &= R_{2D}\left(\psi_i\right) \begin{bmatrix} \cos(\theta_i) & 0 \\ 0 & 1 \end{bmatrix} \end{aligned} \tag{50}$$

The 2D rotation matrix $R_{2D}\left(\psi_i\right) \in \mathcal{SS}_{2\times 2}$ denotes a rotation of the camera's image about the camera projection direction by an angle ψ_i. The angle θ_i is the inclination angle of the i^{th} camera's projection direction (see Figure 1). Given any factorisation $\hat{\mathbf{Q}} = \mathrm{stack}([\mathbf{R}_1]_{2\times 2}, \ldots, [\mathbf{R}_M]_{2\times 2})^\top \mathbf{S}_{2\times N}$, consider the alternative factorisation $[\mathbf{R}'_i]_{2\times 2} \leftarrow [\mathbf{R}_i]_{2\times 2} \begin{bmatrix} d & 0 \\ 0 & 1 \end{bmatrix}$ and $\mathbf{S}'_{2\times N} \leftarrow \begin{bmatrix} \frac{1}{d} & 0 \\ 0 & 1 \end{bmatrix} \mathbf{S}_{2\times N}$ for some scalar d. For all $0 \leq d < 1$ we have $[\mathbf{R}'_i]_{2\times 2} \in \mathcal{SS}_{2\times 2}$, so the alternative factorisation is metric. If there exists a non-zero inclination angle $\theta_i \neq 0$ then there are an infinite number of alternative metric factorisations, so the scene is degenerate. By contrast, if there does not exist a non-zero inclination angle then all cameras have the same projection direction (which is orthogonal to the structure plane), which implies $D = 1$, and from the first paragraph the scene is also degenerate. □

Proof that $\mathrm{rank}(\tilde{\mathbf{S}}) < 2 \Rightarrow \left(\tilde{\mathbf{M}}^A, \tilde{\mathbf{S}}^A\right)$ is degenerate*:* When $\mathrm{rank}(\tilde{\mathbf{S}}) < 2$ points in world coordinates are colinear. This means the camera rotations cannot be fixed because we can rotate each camera about an axis of rotation that is colinear with the points in world coordinates and the image measurements do not change. □

APPENDIX I
PROOF OF THEOREM 2

Without loss of generality let $\mathbf{w}_1$ and $\mathbf{w}_2$ be the two upgrade solutions using views 1,2 and 3. From Equation (34) we have $\mathbf{B}\,\mathrm{stack}(\mathbf{w}, s) = \mathbf{1}_{M\times 1}$, where $\mathbf{B}$ is an $M \times 4$ matrix with each row being $[a_i\, b_i\, c_i\, d_i]$, and $s = \det(f(\mathbf{w}))$. We use the following lemma:

Lemma 4. *Given four or more views we can disambiguate* $\mathbf{w}_1$ *and* $\mathbf{w}_2$ *if and only if* $\mathbf{B}$ *is full-rank.*

Proof. We first prove the reverse implication. When $\mathbf{B}$ is full-rank we can relax the quadratic constraint $s = \det(f(\mathbf{w}))$ to $s \in \mathbb{R}$, and we can then solve $\mathbf{w}$ and s uniquely by inverting the linear system: $\mathrm{stack}(\mathbf{w}, s) = (\mathbf{B}^\top\mathbf{B})^{-1}\mathbf{B}^\top\mathbf{1}_{M\times 1}$. We prove the reverse implication by showing that if we cannot disambiguate $\mathbf{w}_1$ and $\mathbf{w}_2$ then $\mathbf{B}$ is rank-deficient. Let $s_1 = \det(f(\mathbf{w}_1))$ and $s_2 = \det(f(\mathbf{w}_2))$. If we cannot disambiguate $\mathbf{w}_1$ and $\mathbf{w}_2$ then $\mathbf{B}\,\mathrm{stack}(\mathbf{w}_1, s_1) = \mathbf{1}_{M\times 1}$ and $\mathbf{B}\,\mathrm{stack}(\mathbf{w}_2, s_2) = \mathbf{1}_{M\times 1}$. Because $\mathbf{w}_1$ and $\mathbf{w}_2$ are distinct, this implies $\mathbf{B}$ has a nullspace which implies $\mathbf{B}$ is rank-deficient. □

We now use the following lemma:

Lemma 5. *Given a fourth view,* $\mathbf{B}$ *is full-rank if and only if Equation (7) holds.*

Proof. Because we have computed upgrade solutions using the first three views, the first three rows of $\mathbf{B}$ must be linearly independent (otherwise the PSfM-O problem using the first three views would be degenerate, see Appendix H-B). Therefore $\mathbf{B}$ is rank deficient if and only if its fourth column is a linear combination of its first three columns. From the definition of $\mathbf{B}$ this means

$$\begin{aligned}&\operatorname{rank}(\mathbf{B}) < 4 \Leftrightarrow \exists \alpha, \beta, \gamma \in \mathbb{R}\,\text{s.t.}\\ &\begin{cases}\mathbf{E}_4 = \alpha \mathbf{E}_1 + \beta \mathbf{E}_2 + \gamma \mathbf{E}_3 & (a)\\ \det(\mathbf{E}_4) = \alpha \det(\mathbf{E}_1) + \beta \det(\mathbf{E}_2) + \gamma \det(\mathbf{E}_3) & (b)\end{cases}\end{aligned} \tag{51}$$

Pre and post-multiplying Equation (51-a,b) by $\mathbf{X}$ and $\mathbf{X}^\top$, and using $\mathbf{G}_i = \mathbf{X}\mathbf{E}_i\mathbf{X}^\top$ gives

$$\begin{aligned}&\operatorname{rank}(\mathbf{B}) < 4 \Leftrightarrow \exists \alpha, \beta, \gamma \in \mathbb{R}\,\text{s.t.}\\ &\begin{cases}\mathbf{G}_4 = \alpha \mathbf{G}_1 + \beta \mathbf{G}_2 + \gamma \mathbf{G}_3 & (a)\\ \det(\mathbf{G}_4) = \alpha \det(\mathbf{G}_1) + \beta \det(\mathbf{G}_2) + \gamma \det(\mathbf{G}_3) & (b)\end{cases}\end{aligned} \tag{52}$$

We then take the trace of both sides of Equation (52-a), substitute $\operatorname{trace}(\mathbf{G}_i) \leftarrow 1 + \det(\mathbf{G}_i)$, and then subtract Equation (52-b), which gives $\alpha + \beta + \gamma = 1$. We then substitute $\gamma \leftarrow (1 - \alpha - \beta)$ into Equation (52-a) and substitute $\mathbf{G}_i \leftarrow \mathbf{I}_2 - [\mathbf{a}_i]_{2\times1}[\mathbf{a}_i]_{2\times1}^\top$, which gives

$$\begin{aligned}&\operatorname{rank}(\mathbf{B}) < 4 \Leftrightarrow \nexists \alpha, \beta \in \mathbb{R}\,\text{s.t.}\\ &[\mathbf{a}_4]_{2\times1}[\mathbf{a}_4]_{2\times1}^\top = \alpha[\mathbf{a}_1]_{2\times1}[\mathbf{a}_1]_{2\times1}^\top + \beta[\mathbf{a}_2]_{2\times1}[\mathbf{a}_2]_{2\times1}^\top + (1-\alpha-\beta)[\mathbf{a}_3]_{2\times1}[\mathbf{a}_3]_{2\times1}^\top\end{aligned} \tag{53}$$

The proof is completed by negating the implication in Equation (53). □

The proof of Theorem 2 is completed by generalising the result to $M > 4$ views. Lemma 5 tells us that if we have a fourth view for which Equation (7) holds then we can determine the correct structure solution. However, if Equation (7) does not hold then the rank of $\mathbf{B}$ stays at three. From Lemma 4 this means the fourth view provides no extra constraints on the solution. We can therefore only determine the correct structure solution when we have at least one additional view for which Equation (7) holds. □

APPENDIX J
PROOF OF THEOREMS 3 TO 8

A. Theorems 3 and 4

Theorems 3 has two parts. For the first part, the forward implication holds trivially because if a Type 1 problem is degenerate then the equivalent problem with full measurements is degenerate, because by definition Type 1 problems are those where we can complete the rank-2 measurement matrix from the incomplete measurements. The reverse implication holds because when a system with complete measurements is degenerate then if we remove any of the measurements the problem is still degenerate. For the second part, because we cannot compute the scene's 2D affine reconstruction with a Type 2 problem (by definition) then we cannot compute the scene's 2D metric reconstruction (because all metric reconstructions are affine reconstructions).

Theorem 4 has three parts. The first part holds trivially because if there are three or more correspondences that are non-colinear on the structure plane, then the structure-plane-to-image 2D affine transform is fully-determined. Thus any additional point correspondences are redundant, so the disambiguation problem is equivalent to disambiguating with complete measurements. The second part holds because if we have two distinct metric structures the Euclidean distance between two points will in general be different. Let $d_1 > 0, d_2 > 0$ be the Euclidean distance between the two points for structure solutions 1 and 2. Without loss of generality we assume $d_1 \leq d_2$. Due to image foreshortening with an orthographic camera the true Euclidean distance $d \in \{d_1, d_2\}$ must be equal to or exceed the Euclidean distance d_I between their positions in the image (*i.e.* $d_I \leq d$). We can disambiguate structure if and only if there exists an additional image with $d_I > d_1$. When this happens we know $d \neq d_1$ (by contradiction), so structure solution 2 is correct. By contrast suppose structure solution 1 is correct, so $d = d_1$. In this case we cannot disambiguate structure because the inequality $d_I \leq d$ is satisfied by both $d = d_1$ (because d_1 is the true distance) and $d = d_2$ (because $d_2 \geq d_1$). □

B. Theorem 5

Theorem 5 requires proving Algorithm 1 fails $\Leftrightarrow (\tilde{\mathbf{M}}^A, \tilde{\mathbf{S}}^A)$ is degenerate. The forward implication has been proved by the second line of Equation (42) and the reverse implication has been proved by Equation (41). □

C. Theorem 6

Let $r_A \in \mathbb{R}$ denote the reprojection error of the scene's optimal affine reconstruction and let $r_M \in \mathbb{R}$ denote the reprojection error of a metric reconstruction. There exists no metric reconstruction with $r_M < r_A$ because if we impose metric constraints on the cameras we reduce the solution space. Let r'_M denote the reprojection error of a metric reconstruction found by Exact PSfM-O by upgrading the optimal affine reconstruction. Because solutions from Exact PSfM-O exactly transforms the affine reconstruction to a metric reconstruction we have $r'_M = r_A$. Therefore it is not possible to find a better metric reconstruction of the scene, otherwise it would have a lower reprojection error than r_A. This means all solutions from Exact PSfM-O must be optimal metric reconstructions.

The last part of theorem 6 follows because Exact PSfM-O is NADA. Concretely, when Exact PSfM-O has no solution, this means we cannot turn the optimal affine reconstruction into a metric reconstruction by transforming it with an upgrade matrix. Therefore for any upgrade matrix $\tilde{\mathbf{X}}$ the reconstructed camera factor $\tilde{\mathbf{M}}^A\tilde{\mathbf{X}}$ cannot be a metric camera factor. The individual camera factors $\tilde{\mathbf{M}}_i^A\tilde{\mathbf{X}}$ must therefore be corrected *a posteriori* to make them members of $\mathcal{SS}_{2\times 2}$. However the corrected solution will not be optimal because no matter how the correction is performed the upgraded structure factor $\hat{\mathbf{S}} = \tilde{\mathbf{X}}^{-1}\mathbf{S}^A$ will not be optimal in terms of reprojection error. □

D. Theorem 7

Using the definition of the general affine camera in Equation (3), the generalisation of the PSfM-O upgrade constraint to PSfM-PP is as follows:

$$\mathbf{M}_i^A\mathbf{X} = k_i\mathbf{A}_i\mathbf{R}_i \Leftrightarrow \lambda_1\left(k_i^{-2}\mathbf{A}_i^{-1}\mathbf{M}_i^A f(\mathbf{w})\mathbf{M}_i^{A\top}\mathbf{A}_i^{-\top}\right) = 1 \Leftrightarrow \begin{cases} \det\left(\mathbf{M}_i^A w(\mathbf{w})\mathbf{M}_i^{A\top} - k_i^2\mathbf{A}_i\mathbf{A}_i^\top\right) = 0 & (a) \\ \det\left(\mathbf{M}_i^A w(\mathbf{w})\mathbf{M}_i^{A\top}\right) \leq k_i^2 \det^2(\mathbf{A}_i) & (b) \end{cases} \tag{54}$$

The first equivalence comes from $k_i^{-1}\mathbf{A}_i^{-1}\mathbf{M}_i^A\mathbf{X} \in \mathcal{SS}_{2\times 2}$ and the second equivalence comes from rewriting this constraint similarly as Equation (15). Therefore in PSfM-PP each view provides one equality constraint on $\mathbf{w}$ (which recall has three DoFs).

For Case 1 we divide the views into two disjoint sets: $\mathcal{I}'$ and $\mathcal{J} \overset{\text{def}}{=} [1, 2, \ldots, M] \setminus \mathcal{I}'$ with $\text{size}(\mathcal{I}') \geq 3$. The views in $\mathcal{J}$ provide no constraints on metric structure. To see this, for each view $i \in \mathcal{J}$ we have to solve camera resection by decomposing $\mathbf{A}_i^{-1}\mathbf{M}_i^A\mathbf{X}$ into $k_i\mathbf{R}_i$. This provides no constraints on $\mathbf{X}$ (and hence no constraints on $\mathbf{W}$) because for all $(\mathbf{A}_i, \mathbf{X})$ we can compute the decomposition by $k_i = \sigma_1$ and $\mathbf{R}_i = \frac{1}{\sigma_1}\mathbf{A}_i^{-1}\mathbf{M}_i^A\mathbf{X}$ with $\sigma_1 \overset{\text{def}}{=} s_1(\mathbf{A}_i^{-1}\mathbf{M}_i^A\mathbf{X})$. Therefore only the views in $\mathcal{I}'$ are relevant for constraining structure. Because $\mathbf{A}_{i\in\mathcal{I}'}$ is known and $k_{i\in\mathcal{I}'}$ is constant we can effectively convert all views in $\mathcal{I}'$ to orthographic views by transforming the points with $\mathbf{A}_i^{-1}$. This has the effect of undoing the 'intrinsic' component of the camera matrices (*i.e.* $\mathbf{A}_i$). Now, because k_i is assumed to be constant for all views in $\mathcal{I}'$ this is exactly the same as using orthographic cameras with a common magnification factor $k = k_i$. Therefore we have converted the problem to PSfM-O where we only consider the views in $\mathcal{I}'$. It therefore follows that structure is solvable if and only if the equivalent PSfM-O problem is solvable, and the geometric conditions for ensuring this are given by Theorem 1.

For Case 2 we divide the views into two disjoint sets: $\mathcal{I}''$ and $\mathcal{J}' \overset{\text{def}}{=} [1, 2, \ldots, M] \setminus \mathcal{I}''$. Similarly to Case 1 the views in $\mathcal{J}'$ provide no constraints on structure, so we need only consider the views in $\mathcal{I}''$. From Equation (54-a). For all $i \in \mathcal{I}''$ we can write the combined term $k_i^2\mathbf{A}_i\mathbf{A}_i^\top$ as an unknown positive definite matrix $\mathbf{V}$ parameterised with $\mathbf{V} = v(\mathbf{v} \in \mathbb{R}^3) = \begin{bmatrix} v_1 & v_2 \\ v_2 & v_3 \end{bmatrix}$. Equation (54) provides one homogeneous quadratic constraint on six unknowns (*i.e.* $\mathbf{w}$ and $\mathbf{v}$). This means that to obtain a metric reconstruction we require the size of $\mathcal{I}''$ to be at least 5.

For Case 3, let $\{p_1, p_2, \ldots p_P\}$ denote the set of view pairs with $p_{l\in\{1,2,\ldots P\}} \in \{1, 2, \ldots M\}^2$. Let view i be a view that does not belong to a view pair (*i.e.* there is no other view that has the same magnification factor as k_i). From the same reasoning as Case 1, view i provides no constraints on structure. Therefore to determine structure we need to only deal with the views in $\{p_1, p_2, \ldots p_P\}$. To fix the scene's scale ambiguity we can arbitrarily set the magnification factor of the first pair to 1, which means the number of unknowns is $P + 2$ (including three unknowns for $\mathbf{w}$). The number of equality constraints from Equation (54-a) is $2P$, which means to have the necessary number of equations we must have $P \geq 2$ (which means we require 4 or more views). □

E. Theorem 8

Recall in Table IV the relationship between a perspective camera and its corresponding affine camera by linearising the projection function about some 3D point $\mathbf{y}_i \in \mathbb{R}^3$ in camera coordinates, where i is the view index. We first take the problem with para-perspective cameras. This requires calibrating $\mathbf{y}_i$, $i \in \mathcal{I}'$ and the camera extrinsics. We can calibrate $\mathbf{y}_i$ using the perspective camera's intrinsic matrix $\mathbf{K}_i$. This is done by setting $\mathbf{y}_i$ so that the camera is the closest approximation to the perspective camera. From [25] we known that this is when $\mathbf{y}_i$ is at the structure's 3D centroid $\mathbf{c}_i \in \mathbb{R}^3$ in view i. Let us parameterise this by $\mathbf{y}_i = d_i\text{stack}(\mathbf{v}_i, 1)$, where $d_i \in \mathbb{R}$ is the depth of the centroid and $\mathbf{v}_i \in \mathbb{R}^2$ is its direction. The closest first-order approximation of $\mathbf{v}_i$ is known, and given by the vector passing through the centroid of the structure points in image i [25]. What is unknown is d_i and the camera extrinsics. From Table IV we know that d_i is, to first-order, inversely proportional to the camera's magnification factor k_i. Therefore if d_i is approximately constant then so is k_i. Therefore by definition we are in Case 1.

Exactly the same argument follows for weak-perspective cameras. The only difference is that $\mathbf{y}_i$ is constrained to lie on the optical axis (by the definition of the weak perspective camera). The depth of $\mathbf{y}_i$ is calibrated by setting it to the average depth of the structure in view i [25]. □