Reducing Drift in Structure From Motion Using Extended Features

Aleksander Holynski[1][*], David Geraghty[2], Jan-Michael Frahm[2], Chris Sweeney[3], Richard Szeliski[2]

[1]University of Washington   [2]Facebook   [3]Facebook Reality Labs

Abstract

Low-frequency long-range errors (drift) are an endemic problem in 3D structure from motion, and can often hamper reasonable reconstructions of the scene. In this paper, we present a method to dramatically reduce scale and positional drift by using extended structural features such as planes and vanishing points. Unlike traditional feature matches, our extended features are able to span non-overlapping input images, and hence provide long-range constraints on the scale and shape of the reconstruction. We add these features as additional constraints to a state-of-the-art global structure from motion algorithm and demonstrate that the added constraints enable the reconstruction of particularly drift-prone sequences such as long, low field-of-view videos without inertial measurements. Additionally, we provide an analysis of the drift-reducing capabilities of these constraints by evaluating on a synthetic dataset. Our structural features are able to significantly reduce drift for scenes that contain long-spanning man-made structures, such as aligned rows of windows or planar building facades.

1. Introduction

Structure from Motion (SfM) [41] and Simultaneous Localization and Mapping (SLAM) [16, 12, 11] algorithms serve as the foundation for a wide variety of applications in computer vision, including 3D reconstruction [1], 3D exploration of photo collections [46], and phone-based augmented reality [42]. Given a set of images as input, SfM and SLAM systems reconstruct the per-image camera locations and orientations, as well as a sparse set of 3D points.

Thanks to remarkable speed and accuracy improvements over the last decade [11], these camera-based tracking methods can be used to quickly reconstruct the layout or 3D model of a scene, simply by capturing a short video clip from multiple viewpoints. Creating an accurate model, however, requires highly accurate poses; but despite recent improvements, today's best algorithms still suffer from long-range *drift*, which results from the accumulation of

(a) Standard structure from motion

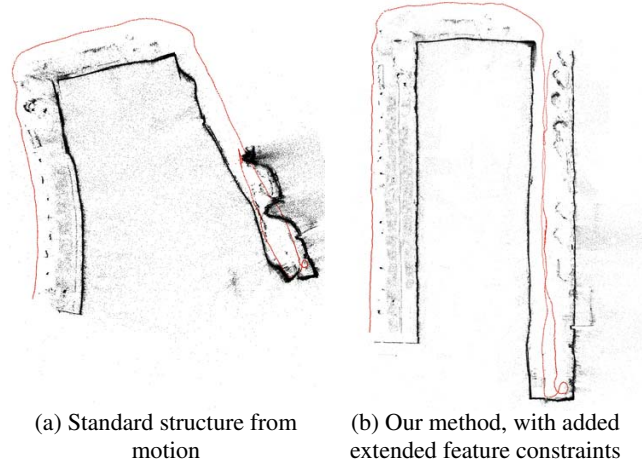(b) Our method, with added extended feature constraints

Figure 1: Standard SfM suffers from significant rotation and scale drift. In scenes with long-spanning structures, like an aligned row of windows wrapping around a building, or a long planar surface, our method can correct for this drift.

small estimation errors caused by noise in feature point location estimates and other unmodelled sources of error. Significant drift can result in bent or deformed reconstructions, such as the one shown in Figure 1a.

While all reconstructions contain some amount of (sometimes unnoticeable) drift, certain camera configurations are significantly more susceptible. For instance, narrow field-of-view cameras, like those in mobile phones, are much more prone to drift, since features persist for fewer frames when the camera is in motion. Since these feature tracks are the basis of motion estimation, if they are not observed from significantly different viewpoints, there are no direct constraints on the relative poses of those frames, and thus nothing to counteract the accumulation of drift error.

In fact, when trying to reconstruct objects which span a space larger than the field-of-view of the camera (such as a building or office-space), a drift-free reconstruction may be nearly impossible to capture. Those familiar with the limitations of these systems might attempt to choose camera paths that result in longer feature tracks, by keeping a greater distance from the subject, or fixating on single points; but often such paths are not feasible. For instance, to capture a building that is surrounded by trees, we're forced to capture the building from a very small distance. In such cases, our best

---

[*]This work was done while Aleksander was as an intern at Facebook. Author email address: holynski@cs.washington.edu.

option is to face the object and trace its perimeter, but doing so can result in significant drift (Fig. 1a).

In this paper, we demonstrate how tracking *extended virtual features* such as co-planar sets of points, which span non-overlapping image frames or sometimes even complete video sequences, can dramatically reduce accumulated drift, enabling us to reconstruct sequences such as those described above. We focus on extended features which are *structural*, namely vanishing points and oriented planes, which are common in man-made environments, such as cities and buildings. Note that our system does *not* make Manhattan World assumptions, i.e., that all dominant structural planes are orthogonal. It can handle Atlanta World scenes with structures at arbitrary orientations [40].

To demonstrate the effectiveness of such features, we incorporate them into Theia [49], a popular open-source SfM system, and evaluate on both synthetic and real-world captures of scenes with strong planar structures using low field-of-view cameras. We show that the addition of extended features drastically reduces drift in both cases, producing results that surpass in quality both Theia and COLMAP, two of the leading open-source SfM packages.

We show the utility of extended features applied to a global SfM framework. Global SfM systems, in contrast to *incremental* systems, solve for all camera poses at once, and are typically much faster, albeit less robust to errors in pairwise pose and correspondence. We show that for our drift-prone sequences, adding extended features to a global SfM system can produce higher quality reconstructions than an incremental SfM system at a fraction of the processing time. This is possible because our modifications preserve the efficiency of previous global methods, only adding a small number of extra parameters and a linear number (in the number of cameras) of additional constraints.

We begin the paper in Sec. 2 with a review of the previous literature, followed in Sec. 3 with a description of our global SfM baseline. Sec. 4 then describes the structural components of our algorithm, including the use of vanishing points for orientation estimation, local orthogonal plane fitting, and the integration of these planes into the global position solver. Sec. 5 presents our experimental results on both synthetic data and real-world mobile phone captures. We then wrap up with a discussion of our results.

2. Previous work

SfM and SLAM methods generally consist of the same set of components: identifying correspondences, pairwise pose estimation, and global integration.

Correspondence. Both SfM and SLAM methods begin by establishing correspondences, which capture the motion observed between pairs of frames and serve as the foundation for pose estimation. Traditionally, the features used for correspondence are 2D point descriptors [20, 28, 53]. These descriptors analyze local image patches to determine salient keypoints and are invariant to geometric and radiometric transformations. Other types of features include line segments [2, 39, 30], vanishing points [44], 3D planes [10, 24, 39, 4, 3], and inertial measurements [26].

Pairwise pose estimation. Once the correspondences have been established and filtered, two-view or multi-view geometry estimation techniques [34, 37, 19, 13, 21, 22, 51] are used to recover relative camera pose between pairs of images, which, along with point correspondences, can be used to triangulate 3D points. These pairwise relative camera poses and 3D points are effectively *local* reconstructions consisting of two or three frames.

Global integration. In order to produce a global reconstruction, it is necessary to combine the local pose estimates. Techniques for robustly integrating these local reconstructions are usually divided into two categories: incremental and global.

Incremental methods [45, 55, 41, 1, 15, 49] have long dominated the state of the art, and can be found in the majority of open-source implementations, such as Bundler [45], VisualSfM [55], and COLMAP [41]. This form of global integration typically starts with a carefully selected two or three-frame reconstruction [46, 45, 41], and incrementally grows it by selecting and anchoring new frames to the already-reconstructed cameras and points. To account for the inconsistencies from the newly registered poses, each frame addition is typically followed by a number of filtering operations to verify and refine the pose [46, 41]. While these filtering operations cause the method to be very robust to outliers, they can also be very computationally expensive. To improve efficiency, incremental systems employ a number of optimizations to reduce incurred processing time [5, 47, 1, 56, 33]. SLAM methods [12, 32, 11, 38, 14] also belong to this category, as they incrementally track the camera pose by accumulating differential camera motion.

Global methods [17, 39, 44, 23, 31, 54, 30] estimate all camera poses simultaneously, making them efficient for large-scale problems. These methods are generally regarded as being less robust to outliers, as the lack of an incremental reconstruction precludes the ability to identify pairwise pose outliers by verifying against a global model. To improve robustness, global SfM methods utilize either groups of frames [31, 23, 18], observed 3D landmarks [44, 9], or pairwise geometry analysis [47] to further verify pairwise pose estimates before integrating globally.

Drift Mitigation. Incremental and global techniques both suffer from drift, particularly in long sequences without loop closures, caused by the accumulation of small relative pose errors. A number of avenues have been explored to mitigate these errors.

SfM systems will often perform global bundle adjustment [52] over all reconstructed frames and points. This process can reduce, *but not eliminate*, drift error, since the bundle-adjusted reconstruction can only be as good as

the correspondences used for optimization. If there are correspondence errors, the resulting reconstruction will inevitably contain some amount of drift.

Real-time SLAM methods [12, 32, 11, 38, 14] which cannot afford costly global bundle adjustment will often include inertial measurements [26] as a secondary source of motion information. While these inertial sensors have become commonplace in modern mobile phones, most captured or distributed videos do not retain inertial measurements, and thus this information is often unavailable.

For closed-loop sequences, a common tactic to mitigate drift is to perform *loop closure* or explicit matching and pairwise pose estimation between temporally distant frames that observe similar parts of the scene. While the constraints derived from these matches have the potential of reducing the drift in the reconstruction, in many cases, the drift error is simply redistributed to different parts of the reconstruction. This is because loop closure only adds constraints to a handful of images at the closure point, but does not apply direct constraints to the poses of images elsewhere.

Higher-dimensional geometric features, like vanishing points, have been shown to significantly reduce rotational drift in both SLAM [6] and SfM [44], since they are able to extract direct constraints between pairs of frames that do not observe the same part of the scene. For active depth sensors, dense normal statistics [48] provide a similar signal. These methods all provide strong constraints on rotational drift, but do not address *translational* drift.

Other geometric features, such as 3D lines, have shown potential for reducing translational drift. Micusik *et al.* [30] incorporate line segment endpoints into an incremental SfM system, allowing accurate reconstruction in the absence of dense point features. Similarly, Zhou et al. [58] propose a SLAM system that extracts and matches axis-aligned structural lines to apply constraints on the camera pose. Nurutdinova et al. [35] propose a generalization that includes arbitrary 3D curves in bundle adjustment. These methods rely on establishing correspondences between lines, either through endpoint matching or photometric comparison, techniques which, in real-world settings, are typically less reliable than point-based features.

Planar constraints have also been shown to resolve translational and scale drift. Szeliski and Torr [50] provide a theoretical introduction to the use of known planar structure as constraints in bundle adjustment. They show that for simple dataset of two cameras, the quality of reconstruction can be improved significantly by incorporating prior knowledge of planar scene structure into bundle adjustment. Extending upon this work, Rother [39] proposed a factorization-based reconstruction system which jointly reconstructs camera pose, points, lines, and planes. Similar to [50], results are only shown on small scenes with few images, since factorization-based approaches are typically quite sensitive to outliers, and do not easily scale to large real-world datasets. Additionally, both methods require a known reference plane, unlike our method, which automatically discovers and associates structural elements.

More recently, Liu et al. [27] demonstrated a SLAM framework that applies a piecewise-planar assumption in tracking, using homographies to efficiently track a camera under rapid motion. While the use of homographies enables fast and robust tracking, the method still suffers from significant drift over longer sequences, as planes are not associated across non-overlapping views.

Li et al. [25] show that additional constraints, such as coplanar sets of lines, found through vanishing point estimation and homography fitting, can further reduce positional drift. This method is reliant on long-spanning line segments in order to establish strong constraints, and requires mutually visible line endpoints for optimization. In contrast, our method does not require lines to be mutually visible in multiple views, and only relies on feature point correspondences, which are standard for SfM systems. Yang and Scherer [57] show that the boundaries between the semantic labels of surfaces such as walls, floor, and ceiling can also be used in constraining the camera position. This method relies on automatic semantic labeling, and thus does not easily extend to arbitrary planar structures. Cohen et al. [8] show that many of these same assumptions, such as Manhattan-oriented structure, symmetry, and repeating elements, can facilitate the fusion of disconnected or sparsely overlapping reconstructions.

In this paper, we describe a technique that automatically extracts structural elements from a real-world video sequence, including vanishing points and planes, and automatically finds associations between non-overlapping observations of these elements in order to establish long-range constraints that reduce pose drift. Most similar to our work, Shariati et al. [43] jointly optimize for wall positions relative to the camera, but require depth and inertial sensors, and also assume Manhattan structure. Our method takes as input only a monocular video sequence, and easily extends to scenes without a global Manhattan coordinate frame.

3. Baseline global SfM pipeline

This section describes the baseline Theia system [49], shown as non-bolded boxes in Fig. 2. Section 4 describes our novel structural constraints, which are shown in bold. Theia is a point-based SfM system, using tracked point features to create local reconstructions, which establish consistent relative camera motion between pairs of frames. These local reconstructions are then used in rotation averaging [7] to estimate the global orientations of all the cameras (Sec. 3.2). Finally, the global camera centers are computed using the estimated rotations (Sec. 3.3), and the 3D points are triangulated to produce the final reconstruction.

Theia implements a number of global rotation and position solvers. For our baseline system, we chose the L1-
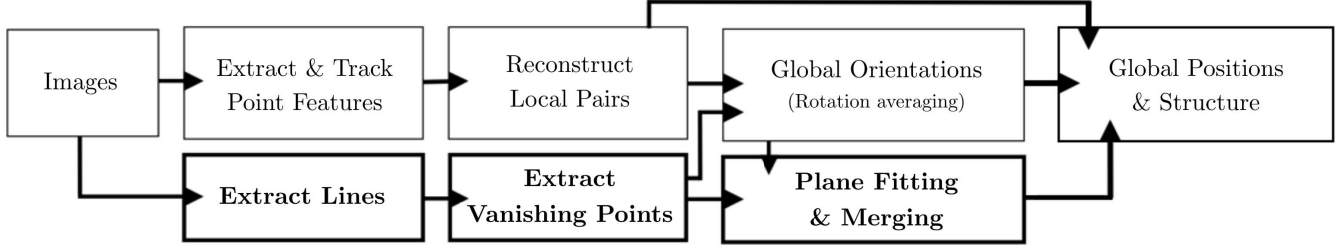
Figure 2: An overview of the global SfM system, showing the baseline components (non-bold) and our added structural constraints (bold).

IRLS rotation solver [7] and the LUD position solver [36]. These are the recommended default solvers in Theia and also consistently produced the best results on our datasets. Apart from two modifications, described in Sec. 3.4, the implementation of Theia is unchanged from the publicly available library. Here, we provide a quick review of the formulation used in both the rotation and position solvers.

3.1. Problem formulation and notation

We start with a set of *point tracks* relating 2D point features $\mathbf{x}_{ip}$ with camera (image) indices i and 3D point indices p. Since we assume known camera intrinsics, these points have already been centered w.r.t. the image center and their pixel coordinates divided by each camera's focal length f_i so that the $\hat{\mathbf{x}}_{ip} = (x_{ip}, y_{ip}, 1)$ correspond to metric (Euclidean) ray directions in each camera's frame.

The projection equations relating a 3D world coordinate $\mathbf{p}_p$ to its corresponding 2D projection $\mathbf{x}_{ip}$ in image i is then

$$\hat{\mathbf{x}}_{ip} \sim \mathbf{R}_i(\mathbf{p}_p - \mathbf{c}_i), \qquad (1)$$

where $\mathbf{R}_i$ is the camera's 3D orientation, $\mathbf{c}_i$ is its 3D position, and $\sim$ indicates similarity up to scale.

3.2. Global orientations (rotation averaging)

To solve for the global orientations of all cameras, global SfM systems typically use rotation averaging, which solves for the global camera orientations that best agree with pairwise rotation estimates by minimizing the consistency error

$$E_{\text{rot}}(\{\mathbf{R}\}) = \sum_{i,j} ||\mathbf{R}_j^{-1}\mathbf{R}_{ij}\mathbf{R}_i - \mathbf{I}||_1 \qquad (2)$$

where $\mathbf{R}_i$, $\mathbf{R}_j$ are the unknown global orientations of frames i and j, respectively, and $\mathbf{R}_{ij}$ is the known pairwise rotation estimate between the two frames.

We use Theia's robust L1-IRLS solver, proposed in [7], which first performs an L1 minimization and then refines the solution using iteratively reweighted least squares.

3.3. Global position estimation

Once the global orientations of all cameras have been estimated, all pairwise constraints are then integrated to estimate global camera centers. Theia uses the "least unsquared deviation" (LUD) global position estimator proposed by

[36], which formulates the optimization as:

$$E_{\text{pos}}(\{\mathbf{c}, s\}) = \sum_{i,j} ||s_{ij}\mathbf{t}_{ij} - \mathbf{c}_j + \mathbf{c}_i||_2 \qquad (3)$$

where $\mathbf{t}_{ij}$ is a known pairwise translation estimate (after rotating all cameras into the global coordinate system), $\mathbf{c}_i$ and $\mathbf{c}_j$ are the unknown camera centers, and s_{ij} is the unknown scaling coefficient for the pairwise reconstruction. These equations are optimized using a fast convex solver.

3.4. Extensions to handle degenerate configurations

In order for the baseline Theia to work on our sequences (narrow field-of-view videos), which often contain "degenerate cases" such as linear camera motions and single planes (building walls), we added the following extensions, which are described in more detail in our supplementary material:

1. The above global position estimation suffers from degeneracy for colinear camera motion (as described in [23, 31]). In order to resolve this, we adapt a simplified version of [31], applying additional constraints on the relative scales of pairwise reconstructions by comparing the triangulated depths of shared feature tracks.

2. In order to deal with planar scenes and pure rotations, for which the 5-point algorithm may produce degenerate configurations, we additionally estimate pairwise pose from a homography (using [29]), keeping whichever approach produces a larger number of inliers.

4. Structural constraints

In order to reduce the drift (global low-frequency errors and deformations) in our reconstructions, we exploit large-scale *structural constraints* such as vanishing points and planes. These can be thought of as *extended features* since they will often span many more frames than traditional point tracks, which come in and out of view.

4.1. Lines and vanishing points

In order to obtain a drift-free set of rotation estimates, we first compute for each frame (wherever possible) a vertical vanishing point and one or more horizontal vanishing points, as described in the supplemental material and illustrated in Fig. 3. Once the vanishing points have been found in frame i, we estimate a global rotation $\mathbf{R}_i^{\text{vp}}$, which maps one of the Atlanta world horizontal directions to the
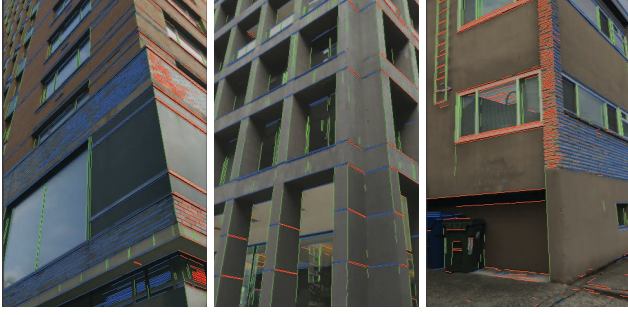
Figure 3: Examples of our edge detection and vanishing point estimation stage. Different colored edges indicate their membership in different vanishing points. Note that only two orthogonal vanishing points need to be detected to establish the coordinate frame.

dominant horizontal direction in the frame. Compared to the previous work by Sinha *et al.* [44], which performs a global matching step between vanishing directions of all frames, our approach is tailored for continuous video sequences where motion between frames is small, so the vanishing point associations can be chained through consecutive frames. This gives us a consistent *soft* (drift-free, but empirically lower-accuracy) global orientation constraint for every frame that has associated vanishing point data.

To take advantage of these estimated points, we initialize the solution for the global camera orientations using the vanishing points, $\mathbf{R}_i \leftarrow \mathbf{R}_i^{\text{vp}}$. In addition to the inter-frame constraints from pairwise pose estimates (2), we also add penalties on the difference between the current rotation estimates and the vanishing-point based rotations:

$$E_{\text{vp}}(\{\mathbf{R}\}) = \lambda_{\text{vp}} \sum_i W_{\text{vp}}(i) \|\mathbf{R}_i^{-1}\mathbf{R}_i^{\text{vp}} - \mathbf{I}\|_1. \quad (4)$$

These constraints are weighted using a regularization parameter λ_{vp} and a per-frame weighting function

$$W_{\text{vp}}(i) = \text{clamp}(1 - \frac{\Delta\theta}{\theta_{max}}, 0, 1) \quad (5)$$

where $\Delta\theta$ is the incremental rotation between frames $i - 1$ and i and θ_{max} is the largest tolerated incremental rotation. We use $\theta_{max} = 5$ degrees for all experiments. This weighting function makes the rotation estimates more robust to significant outliers in the vanishing point estimates, by lowering the weight of consecutive frames when they are very different. Further information on vanishing point estimation and integration is provided in the supplementary material.

4.2. Extended planes

While adding vanishing points as soft constraints on camera orientations can dramatically reduce orientation drift and increase reconstruction accuracy, positional drift still remains an issue. In order to address this, we exploit another major source of structural constraints, i.e., coplanar points arising from man-made structures such as buildings.

In this section, we describe how we identify coplanar 3D points in local pairwise reconstructions and then link these together into extended global planes that can be used as additional constraints in the pose estimation process. As mentioned before, in our current system, we restrict our attention to planes whose normals correspond to one of the dominant vanishing point directions.

Local plane fitting. We begin by discovering planes in each pairwise reconstruction. For each local reconstruction containing valid vanishing points, we use the pairwise pose estimate to perform two-view triangulation, resulting in a local 3D point cloud. We then perform a plane sweep along the three orthogonal vanishing directions associated with the base frame of the pairwise reconstruction. This results in a number of local planes π, parameterized as:

$$\mathbf{p} \cdot \hat{\mathbf{n}}_\pi^{ij} = d_{ij}^\pi, \quad (6)$$

where the $\mathbf{p}$ are the local 3D point inliers, and $\hat{\mathbf{n}}_\pi^{ij}$ is the local plane normal, and d_{ij}^π is the distance along the normal $\hat{\mathbf{n}}$ from the origin of pairwise estimate $i \to j$. We provide more details on plane fitting in the supplemental material.

Plane merging and constraints. Once local plane hypotheses have been generated for each pairwise reconstruction, we group these into global *extended* planes, which we then use to provide additional constraints on the local scales and global camera positions, as described below in Eq. 7. In order to link these local hypotheses, we rely on point correspondences. Since local planes are established by finding co-planar sets of 3D points, each local plane contains a set of *inlier* tracks which can be used to associate local planes with one another. We perform this association by greedily merging any two local planes which share a majority of tracks, and are associated with the same vanishing direction.

Once we have established estimates of which local plane hypotheses correspond to one another, we define constraints that encourage these local planes to coincide in the final 3D reconstruction. In order to integrate local plane hypotheses into global constraints on camera positions, we add scalar variables d_p to the linear system, where p is the index of the global extended plane (as opposed to the local plane index π). These variables define each global plane's location (distance to the world origin along the plane normal). This corresponds to one added constraint for each observation of a global plane p in a pairwise estimate ij:

$$E_{\text{pln}}(\{\mathbf{c}, s\}) = \sum_{i,j,p} W_{\text{p}}(ij, \pi) \|s_{ij} d_{ij}^\pi + \mathbf{c}_i \cdot \mathbf{n}_p - d_p\|_2 \quad (7)$$

where the s_{ij} is the unknown pairwise estimate scaling factor, $\mathbf{n}_p$ is the known plane orientation (normal vector), $\mathbf{c}_i$ is the unknown global camera position of the base camera in the pairwise transformation, and d_p is the unknown global distance of the plane from the world origin along $\mathbf{n}_p$. Fig. 4 provides a visualization of these terms.

The number of global planes p (and hence extra scalar unknowns) is a small constant number per scene, and therefore the number of added constraints to the system is lin-
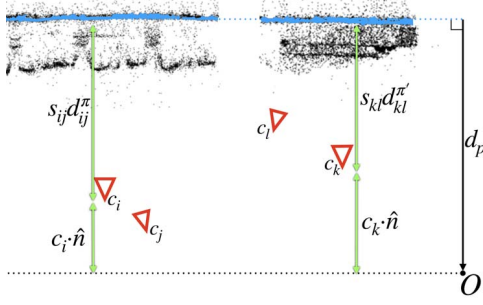
Figure 4: An illustrated example of our plane constraints described in Eq. 7. In the global coordinate frame, two pairwise local reconstructions $i \to j$ and $k \to l$ have their own local distances d_{ij}^{π} and $d_{kl}^{\pi'}$ between the base cameras and their respective local planes π and π'. When these planes correspond to the same global plane p (blue), our plane constraints will align these two locally estimated planes in the global reconstruction by encouraging each local plane distance to be consistent with a single global distance from the origin d_p. This example shows the result of optimization, where our planar loss has been minimized such that $d_p = s_{kl}d_{kl}^{\pi'} + c_k\hat{n} = s_{ij}d_{ij}^{\pi} + c_i\hat{n}$, and $E_{\mathsf{pln}} = 0$.

ear in the number of cameras. We weigh each of these constraints by the support of the global plane in the local pairwise reconstruction, i.e. by the number of inlier tracks. More formally, we define a weighting function

$$W_{\mathsf{p}}(ij, \pi) = \min\left(\frac{S(ij, \pi)}{S_{\max}}, 1\right) * \lambda_{\mathsf{p}} \qquad (8)$$

where $W_p(ij, \pi)$ defines the weight for plane π in pairwise estimate $i \to j$, $S(ij, \pi)$ returns the number of inliers in the local plane estimate, $S_{\max}$ defines the minimum number of inliers to receive full weight, and λ_{p} is the plane weighting coefficient (the maximum weight a plane can have). We use $S_{\max} = 10, \lambda_{\mathsf{p}} = 50$ in all our experiments. In Sec. 5, we show how adding plane constraints dramatically improves reconstructions of both synthetic and real-world scenes.

5. Evaluation

In this section, we present our experimental results on both a synthetic scene, where we know the ground truth results and can hence quantitatively measure accuracy, as well as some real-world hand-held videos sequences.

5.1. Synthetic scene

In order to test our algorithm's (and its variants') performance, we constructed a simple synthetic 3D scene consisting of a three-story building with regularly spaced windows (Fig. 6a). We render this scene from a camera path translating along the length of the building, with small regular fluctuations in elevation. We bypass traditional feature extraction, and instead establish feature tracks by projecting the 3D window corners into each view. We then add synthetic 2D Gaussian noise to these tracks before passing them to the reconstruction algorithm, in order to simulate the correspondence errors which cause drift. Vanishing points are

seattle1   more_half   seattle2   atlanta1   seattle3

Figure 5: Our real world datasets used for evaluation consist of five handheld video sequences of scenes with man-made structures. Here we show sample video frames and the approximate building facade traced on satellite imagery.

extracted normally, by running our vanishing point estimation on the rendered images. Fig. 6c shows that adding Gaussian noise to the 2D point tracks results in accumulated pose error, consisting of both rotational and positional (scale) drift. Fig. 6d shows that introducing vanishing point constraints significantly reduces the visible rotational drift (seen as bending in the reconstruction). The remaining errors, caused either by incorrect translation directions or incorrect scale estimates, are shown to be virtually eliminated by introducing global plane constraints (Fig. 6e).

In Fig. 7, we show quantitative results, comparing the positional drift in reconstructions with and without our proposed structural constraints. We measure drift by reconstructing the synthetic sequence shown in Fig. 6 and comparing the reconstructed poses against ground-truth. For this experiment, we use a sample sequence with 350 frames. Since the reconstructions have scale, rotation, and translation (gauge) ambiguities, we align the reconstructed poses to the ground truth by solving a similarity transformation.

5.2. Real-world videos

Typical SLAM and SfM benchmarks use static (or global shutter) cameras with wide fields-of-view and contain widely varying viewpoints of the scene. These configurations are chosen because they are the least susceptible to drift — tracks will be longer, feature positions more precise, and triangulation angles wider, thus reducing the effects of low-frequency pose error. In this paper, we focus instead on the converse: configurations which are *most* susceptible, i.e. low-field of view, rolling-shutter, handheld videos, where objects are only seen from a small range of viewpoints. These types of captures also more accurately reflect the type of sequence that might be captured by a layperson with a handheld camera-phone.
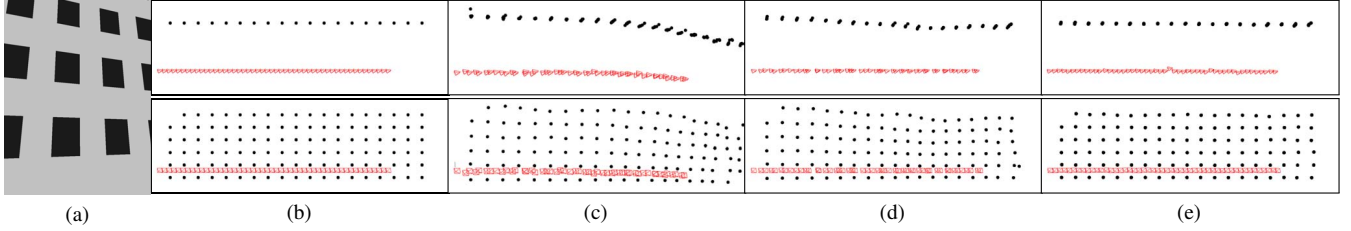
| (a) | (b) | (c) | (d) | (e) |

Figure 6: Synthetic test case: **(a)** a sample frame from the sequence. The sequence contains three rows of "windows" on a single plane. **(b)** Top-down and frontal views of the noise-free reconstruction produced by Theia without added structural constraints: red frusta are shown for the reconstructed camera positions (moving left to right), black points denote the reconstructed 3D window corners. **(c)** After introducing noise to the 2D point tracks, the reconstruction produced by Theia exhibits both rotation and scale drift. **(d)** Once vanishing point constraints constraints are added, the rotational drift (bending) is reduced, but scale drift is still present, seen as irregularities in the frame-to-frame camera translations, and non-planarity in the reconstructed points. **(e)** Once planar constraints are added to the global position estimation, the scale drift is eliminated.
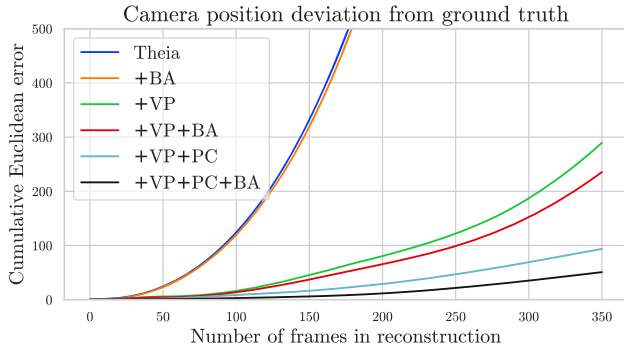

Figure 7: **Quantitative evaluation**: A comparison of positional drift across variants of Theia with global bundle adjustment (BA), our vanishing point constraints (VP), and our planar constraints (PC). One unit along the vertical axis is equal to the baseline between a pair of cameras. We see that bundle adjustment decreases error across all reconstructions, but cannot fully correct for the positional drift even after the rotational drift has been removed.

SEATTLE1		SEATTLE2		SEATTLE3		ATLANTA1		
Pos.	Orient.	Pos.	Orient.	Pos.	Orient.	Pos.	Orient.	
73.51	12.63°	45.04	8.70°	48.06	10.23°	36.22	15.71°	Theia
30.01	12.60°	11.63	8.56°	46.85	9.19°	32.02	14.26°	Theia+BA
16.98	**1.72°**	25.00	**2.01°**	40.07	2.70°	18.11	2.18°	Theia+VP
7.42	**1.72°**	3.61	**2.01°**	4.22	2.70°	6.50	2.18°	Ours
1.91	**1.72°**	**3.34**	2.49°	**3.16**	**1.03°**	**2.98**	**1.63°**	Ours+BA
64.38	17.67°	29.38	6.74°	82.80	24.10°	31.84	15.12°	COLMAP

Table 1: **Loop closure error** (lower is better): for sequences which end at approximately the same location, we can compute the *loop closure error* by duplicating the first image at the end of the sequence, and measuring the error between the two reconstructed views in both position and orientation. This effectively measures the amount of drift over the entire sequence. Position errors are divided by the median baseline to show the deviation in number of frames. The shown configurations are defined in Fig. 8.

With this in mind, we captured five handheld video sequences of man-made environments (Fig. 5). We used smartphone cameras in portrait orientation, and mostly captured sequences walking along building facades. These sequences are intended to showcase difficult configurations which typically result in drift, since point tracks do not persist for many frames, and therefore cannot establish long-term constraints. In Fig. 8, we show a comparison to our baseline method, as well as to COLMAP [41], a popular incremental SfM system. We can see that the baseline methods (Theia and COLMAP) both exhibit significant drift, and the addition of our constraints to Theia results in a system that produces drift-free reconstructions very quickly (a fraction of the time needed for COLMAP). It is important to note that none of these results use any form of explicit loop closure. This is intended to more visibly demonstrate the effects of drift, as automatic loop closure is not always an option (for sequences without enough loop overlap) and, when it is, does not always remove the effects of drift. It is also possible for automatic loop closure to introduce other sources of error, such as false matches between distant views, especially in sequences like ours, which contain significant repetitive structure. In the supplemental material, we show experiments with automatic loop closure enabled.

In order to quantify the amount of drift for the complete-building sequences, we copied the first frame in our sequence as the last frame, but ran both Theia and COLMAP without loop closure, and measured the error between the two reconstructed frames. The quantitative errors in both orientation and position are shown in Table 1. While adding vanishing point constraints dramatically reduces orientation errors, the positional error is even further reduced by adding the plane constraints. Performing a final bundle adjustment on this solution even further reduces the error.

Note that our added constraints do not modify the correspondences, and thus the constraints in bundle adjustment are unchanged. However, our constraints improve the initialization to bundle adjustment, providing reconstructions often much closer to the true solution. This causes bundle adjustment to converge more quickly, and reduces the likelihood of converging to local minima. Since planar constraints are not enforced in bundle adjustment, bundle adjustment could theoretically reintroduce drift, but we have not observed this in any of our experiments.
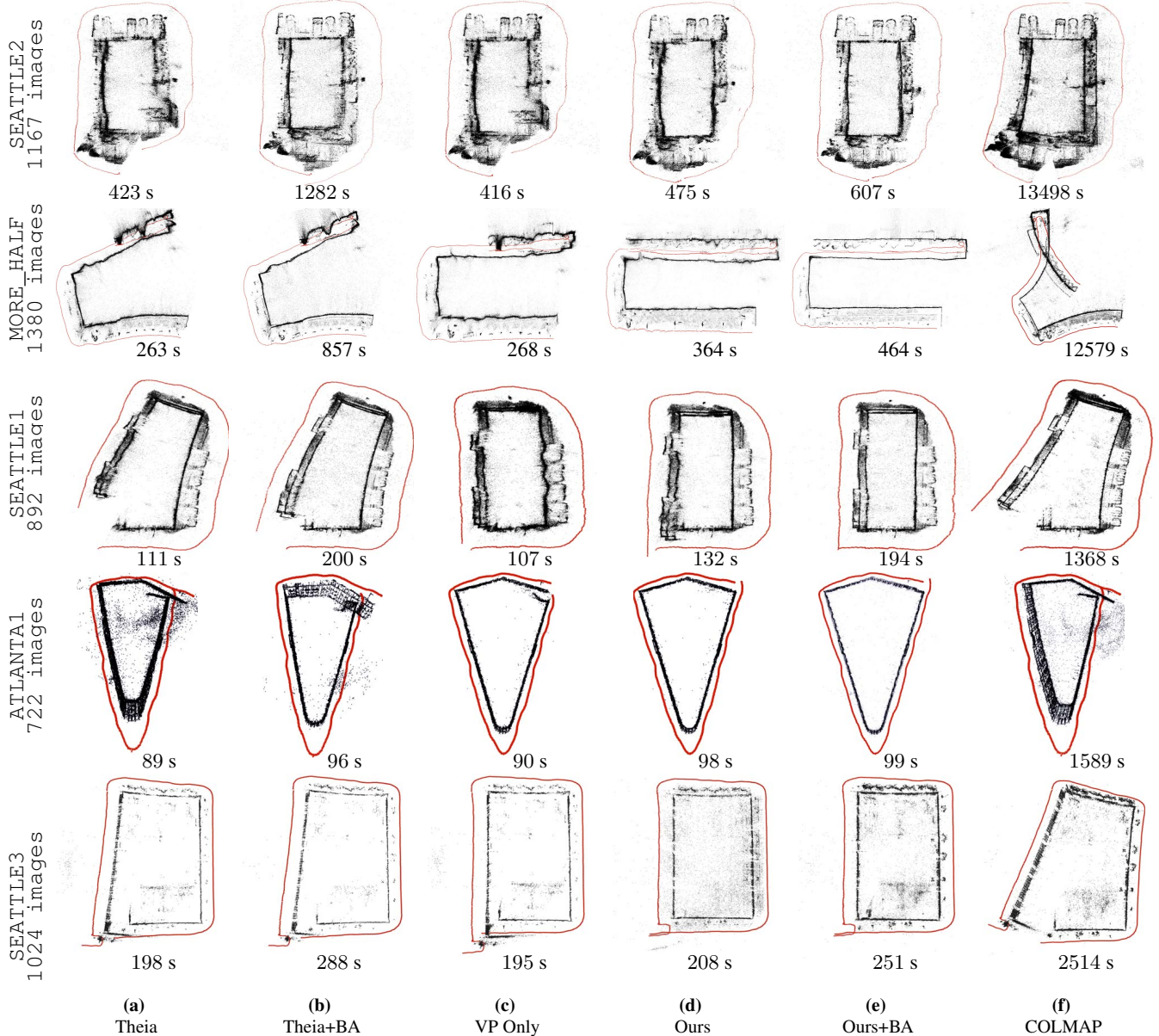
Figure 8: **Qualitative comparison:** top-down views of the real-world datasets using different reconstruction methods: **(a)** Theia without added structural constraints or global bundle adjustment (BA), **(b)** Theia without added constraints but with BA, **(c)** Theia with added vanishing point constraints (VP), but no planar constraints (PC) or BA **(d)** Theia with both VP and PC, but without BA **(e)** Theia with VP, PC, and BA, **(f)** COLMAP, with BA. None of the reconstructions use any form of explicit loop closure; all examples use window-based feature matching with a window size of 100 frames. See supplemental for a discussion about loop closure. Inset values indicate reconstruction time, excluding feature matching, measured on a MacBook Pro with a 6-core 2.9 GHz Intel processor and 32 GB of memory.

6. Discussion and Conclusions

In this paper, we have presented a method for efficiently constraining global reconstructions in scenes with man-made structures. We show that the detection and linking of *extended structural features* such as planes and vanishing points, which can span frames that may not have overlapping views of the scene, can be used as powerful additional constraints in structure from motion, enabling the reconstruction of sequences that are particularly susceptible to drift. We demonstrate that these constraints significantly reduce accumulated drift over longer sequences, allowing fast-but-brittle global reconstruction algorithms to rival — and often surpass — the accuracy of costly bundle-adjusted incremental systems in scenes with strong structural elements. We present comparisons to state-of-the-art structure-from-motion systems for both synthetic and real-world data.

References

[1] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski. Building Rome in a day. In *International Conference on Computer Vision (ICCV)*, pages 72–79, 2009. 4321, 4322

[2] C. Baillard, C. Schmid, A. Zisserman, and A. Fitzgibbon. Automatic line matching and 3D reconstruction of buildings from multiple views. In *ISPRS Conference on Automatic Extraction of GIS Objects from Digital Imagery*, volume 32, pages 69–80, 1999. 4322

[3] A. Bartoli and P. Sturm. Constrained structure and motion from multiple uncalibrated views of a piecewise planar scene. *International Journal of Computer Vision*, 52(1):45–64, 2003. 4322

[4] A. Bartoli, P. Sturm, and R. Horaud. *A projective framework for structure and motion recovery from two views of a piecewise planar scene*. PhD thesis, INRIA, 2000. 4322

[5] B. Bhowmick, S. Patra, A. Chatterjee, V. M. Govindu, and S. Banerjee. Divide and conquer: Efficient large-scale structure from motion using graph partitioning. In *Asian Conference on Computer Vision (ACCV)*, pages 273–287, 2014. 4322

[6] F. Camposeco and M. Pollefeys. Using vanishing points to improve visual-inertial odometry. In *2015 IEEE international conference on robotics and automation (ICRA)*, pages 5219–5225. IEEE, 2015. 4323

[7] A. Chatterjee and V. Madhav Govindu. Efficient and robust large-scale rotation averaging. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 521–528, 2013. 4323, 4324

[8] A. Cohen, T. Sattler, and M. Pollefeys. Merging the unmatchable: Stitching visually disconnected sfm models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2129–2137, 2015. 4323

[9] Z. Cui and P. Tan. Global structure-from-motion by similarity averaging. In *International Conference on Computer Vision (ICCV)*, pages 864–872, 2015. 4322

[10] P. E. Debevec, C. J. Taylor, and J. Malik. Modeling and rendering architecture from photographs: A hybrid geometry and image-based approach. In *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, volume 96, pages 11–20, 1996. 4322

[11] J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(3):611–625, 2018. 4321, 4322, 4323

[12] J. Engel, T. Schöps, and D. Cremers. LSD-SLAM: Large-scale direct monocular SLAM. In *European Conference on Computer Vision (ECCV)*, pages 834–849, 2014. 4321, 4322, 4323

[13] O. D. Faugeras, Q.-T. Luong, and S. J. Maybank. Camera self-calibration: Theory and experiments. In *European Conference on Computer Vision (ECCV)*, pages 321–334, 1992. 4322

[14] C. Forster, Z. Zhang, M. Gassner, M. Werlberger, and D. Scaramuzza. Svo: Semidirect visual odometry for monocular and multicamera systems. *IEEE Transactions on Robotics*, 33(2):249–265, 2016. 4322, 4323

[15] J.-M. Frahm, P. Fite-Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y.-H. Jen, E. Dunn, B. Clipp, S. Lazebnik, et al. Building Rome on a cloudless day. In *European Conference on Computer Vision (ECCV)*, pages 368–381, 2010. 4322

[16] J. Fuentes-Pacheco, J. Ruiz-Ascencio, and J. M. Rendón-Mancha. Visual simultaneous localization and mapping: a survey. *Artificial Intelligence Review*, 43(1):55–81, 2015. 4321

[17] V. M. Govindu. Combining two-view constraints for motion estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001. 4322

[18] V. M. Govindu. Robustness in motion averaging. In *Asian Conference on Computer Vision (ACCV)*, pages 457–466, 2006. 4322

[19] R. M. Haralick, H. Joo, C.-N. Lee, X. Zhuang, V. G. Vaidya, and M. B. Kim. Pose estimation from corresponding point data. In *Machine Vision for Inspection and Measurement*, pages 1–84. Elsevier, 1989. 4322

[20] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey vision conference*, 1988. 4322

[21] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, 2003. 4322

[22] R. I. Hartley. Lines and points in three views and the trifocal tensor. *International Journal of Computer Vision*, 22(2):125–140, 1997. 4322

[23] N. Jiang, Z. Cui, and P. Tan. A global linear method for camera pose registration. In *International Conference on Computer Vision (ICCV)*, pages 481–488, 2013. 4322, 4324

[24] F. Lang and W. Förstner. Surface reconstruction of manmade objects using polymorphic mid-level features and generic scene knowledge. *International Archives of Photogrammetry and Remote Sensing*, 31:415–420, 1996. 4322

[25] H. Li, J. Yao, J.-C. Bazin, X. Lu, Y. Xing, and K. Liu. A monocular slam system leveraging structural regularity in manhattan world. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2518–2525. IEEE, 2018. 4323

[26] M. Li and A. I. Mourikis. High-precision, consistent EKF-based visual-inertial odometry. *International Journal of Robotics Research*, 32(6):690–711, May 2013. 4322, 4323

[27] H. Liu, G. Zhang, and H. Bao. Robust keyframe-based monocular slam for augmented reality. In *2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 1–10. IEEE, 2016. 4323

[28] D. G. Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision (ICCV)*, volume 2, pages 1150–1157, 1999. 4322

[29] E. Malis and M. Vargas. Deeper understanding of the homography decomposition for vision-based control. Technical Report 6303, INRIA, 2007. 4324

[30] B. Micusik and H. Wildenauer. Structure from motion with line segments under relaxed endpoint constraints. *International Journal of Computer Vision*, 124(1):65–79, 2017. 4322, 4323

[31] P. Moulon, P. Monasse, and R. Marlet. Global fusion of relative motions for robust, accurate and scalable structure from motion. In *International Conference on Computer Vision (ICCV)*, pages 3248–3255, 2013. 4322, 4324

[32] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015. 4322, 4323

[33] K. Ni, D. Steedly, and F. Dellaert. Out-of-core bundle adjustment for large-scale 3D reconstruction. In *International Conference on Computer Vision (ICCV)*, 2007. 4322

[34] D. Nistér. An efficient solution to the five-point relative pose problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(6):756–770, 2004. 4322

[35] I. Nurutdinova and A. Fitzgibbon. Towards pointless structure from motion: 3D reconstruction and camera parameters from general 3D curves. In *International Conference on Computer Vision (ICCV)*, pages 2363–2371, 2015. 4323

[36] O. Ozyesil and A. Singer. Robust camera location estimation by convex programming. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2674–2683, 2015. 4324

[37] M. Pollefeys, F. Verbiest, and L. Van Gool. Surviving dominant planes in uncalibrated structure and motion recovery. In *European Conference on Computer Vision (ECCV)*, pages 837–851, 2002. 4322

[38] T. Qin, P. Li, and S. Shen. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics*, 34(4):1004–1020, 2018. 4322, 4323

[39] C. Rother. Linear multi-view reconstruction of points, lines, planes and cameras using a reference plane. In *International Conference on Computer Vision (ICCV)*, 2003. 4322, 4323

[40] G. Schindler and F. Dellaert. Atlanta world: An expectation maximization framework for simultaneous low-level edge grouping and camera calibration in complex man-made environments. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004*. IEEE, 2004. 4322

[41] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4321, 4322, 4327

[42] T. Schöps, J. Engel, and D. Cremers. Semi-dense visual odometry for AR on a smartphone. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 145–150, 2014. 4321

[43] A. Shariati, B. Pfrommer, and C. J. Taylor. Simultaneous localization and layout model selection in manhattan worlds. *IEEE Robotics and Automation Letters*, 4(2):950–957, 2019. 4323

[44] S. N. Sinha, D. Steedly, and R. Szeliski. A multi-stage linear approach to structure from motion. In *ECCV 2010 Workshop on Reconstruction and Modeling of Large-Scale 3D Virtual Environments*, 2010. 4322, 4323, 4325

[45] N. Snavely. Bundler: Structure from motion (SfM) for unordered image collections. *Code available at http://phototour.cs.washington.edu/bundler/*, 2010. 4322

[46] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3D. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, 25(3):835–846, 2006. 4321, 4322

[47] D. Steedly, I. Essa, and F. Dellaert. Spectral partitioning for structure from motion. In *International Conference on Computer Vision (ICCV)*, 2003. 4322

[48] J. Straub, G. Rosman, O. Freifeld, J. J. Leonard, and J. W. Fisher, III. A mixture of Manhattan frames: Beyond the Manhattan world. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 4323

[49] C. Sweeney, T. Hollerer, and M. Turk. Theia: A fast and scalable structure-from-motion library. In *ACM International Conference on Multimedia*, pages 693–696, 2015. 4322, 4323

[50] R. Szeliski and P. H. Torr. Geometrically constrained structure from motion: Points on planes. In *European Workshop on 3D Structure from Multiple Images of Large-Scale Environments*, pages 171–186. Springer, 1998. 4323

[51] P. H. Torr and A. Zisserman. Robust parameterization and computation of the trifocal tensor. *Image and Vision Computing*, 15(8):591–605, 1997. 4322

[52] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon. Bundle adjustment—a modern synthesis. In *International Workshop on Vision Algorithms*, pages 298–372, 1999. 4322

[53] T. Tuytelaars, K. Mikolajczyk, et al. Local invariant feature detectors: a survey. *Foundations and Trends in Computer Graphics and Vision*, 3(3):177–280, 2008. 4322

[54] K. Wilson and N. Snavely. Robust global translations with 1DSfM. In *European Conference on Computer Vision (ECCV)*, pages 61–75, 2014. 4322

[55] C. Wu. VisualSFM: A visual structure from motion system. *Code available at http://ccwu.me/vsfm/*, 2011. 4322

[56] C. Wu, S. Agarwal, B. Curless, and S. M. Seitz. Multicore bundle adjustment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3057–3064, 2011. 4322

[57] S. Yang and S. Scherer. Monocular object and plane slam in structured environments. *IEEE Robotics and Automation Letters*, 4(4):3145–3152, 2019. 4323

[58] H. Zhou, D. Zou, L. Pei, R. Ying, P. Liu, and W. Yu. Structslam: Visual slam with building structure lines. *IEEE Transactions on Vehicular Technology*, 64(4):1364–1375, 2015. 4323