

Stable and Scalable Bundle Adjustment of Holistic 3D Structures

Shaohui Liu¹, Rémi Pautrat², Daniel Barath¹, Richard Hartley³, Viktor Larsson⁴, and Marc Pollefeys^{1,2}

¹ETH Zurich ²Microsoft Spatial AI Lab

³Australian National University ⁴Lund University

Abstract. Bundle Adjustment (BA) is a cornerstone of 3D computer vision and has benefited from decades of advances in sparse optimization and numerical methods. It was originally developed for jointly optimizing camera intrinsics, poses and sparse 3D points. While extensions incorporate lines and other primitives, integrating richer geometric structures such as parallelism, coplanarity, or wireframes often introduces significantly increased computational cost and reduced numerical stability. In this paper, we propose a unified framework that extends bundle adjustment to jointly optimize geometric features and higher-order relations. We first introduce a taxonomy that distinguishes scalable geometric *features* with direct 2D measurements (e.g., points and lines), from *groups* encoding higher-order relations (e.g., coplanarity, parallelism, etc.), where we show that groups can be modeled as camera-like entities within the bundle adjustment framework. Building on this formulation, we propose that both group constraints and cross-feature relations (i.e., point–line associations) can be expressed through 2D reprojection measurements. By formulating group-induced and cross-feature reprojection errors, we preserve the sparsity structure of classical point-based BA under Schur elimination, while avoiding direct 3D regularization that degrades the conditioning and stability. Experiments on both real-world and synthetic datasets demonstrate runtime performance comparable to classical point-only bundle adjustment, while producing significantly richer 3D structures and improved geometric accuracy.

1 Introduction

Bundle adjustment plays a central role in 3D reconstruction, jointly refining camera parameters and scene geometry by minimizing reprojection errors across all observations. Decades of research have produced highly efficient solvers that exploit the characteristic sparsity of the problem: because each 3D landmark couples only to its observing cameras, the Hessian over landmarks is block-diagonal, enabling Schur complement elimination that reduces the system to a compact camera-only problem. This structure has been the key to scaling bundle adjustment to large reconstructions with millions of points [2, 49, 68].

Yet real-world scenes, particularly man-made environments, contain geometric structures far richer than isolated point landmarks. Lines appear along

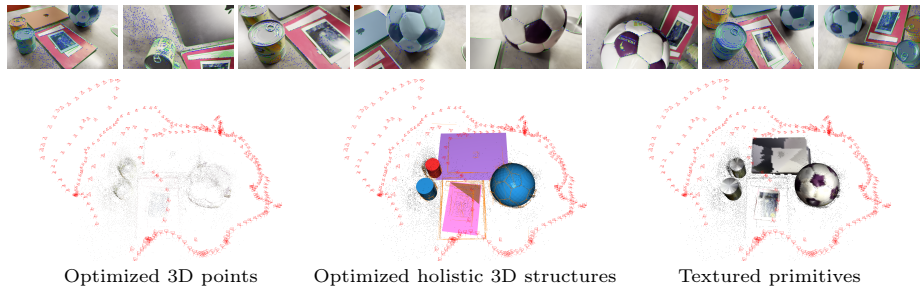


Fig. 1: Bundle adjustment with structural constraints. From multi-view images with detected points, lines, and groups (top), our proposed unified bundle adjustment framework is able to jointly optimize features (points/lines), camera poses, and structural constraints while preserving computational efficiency. This yields **richer optimizable sparse 3D reconstruction with geometric primitives** (planes, spheres, cylinders, etc.) beyond point clouds (left).

edges and boundaries, often in parallel configurations governed by vanishing points [47]. Surfaces form planes and other primitives that constrain the positions of points and lines alike. At finer scales, points and lines meet at junctions to form wireframe structures [23, 72, 80] that encode the geometric skeleton of buildings, furniture, and urban infrastructure. These structures provide powerful constraints that, if properly exploited, can improve reconstruction accuracy and completeness. However, incorporating them into bundle adjustment has remained a long-standing challenge due to two fundamental difficulties.

The first challenge is structural coupling. In classical bundle adjustment, the efficiency of the Schur complement relies on the fact that 3D landmarks are mutually independent: no residual couples two landmarks directly. Introducing geometric relations such as coplanarity, parallelism, orthogonality, or wireframe connectivity fundamentally violates this property. When two features share a geometric relation, their parameters become coupled, creating off-diagonal blocks in the landmark Hessian that destroy its block-diagonal structure. With many such associations, the efficient Schur elimination may break down entirely, leading to denser fill-in and increased computational cost.

The second challenge is the scale ambiguity coming with 3D constraints. Even when sparsity is handled, formulating structural constraints as direct 3D penalties (*e.g.*, point-to-plane distances [31, 54]) mixes 3D metric residuals with 2D pixel-space terms, degrading conditioning due to scale mismatch and introducing ad hoc weighting. Because such residuals lack a principled noise model, there is no natural way to anchor structural entities through well-observed features. Moreover, 3D penalties break the scale gauge ambiguity of bundle adjustment, biasing the optimizer toward shrinking the reconstruction to reduce the penalty.

In this work, we present a unified bundle adjustment framework that integrates geometric structures while addressing both challenges. We introduce a taxonomy on features and groups, which distinguishes geometric features with direct 2D measurements (points, lines) from groups encoding higher-order relations (vanishing points, planes, conics, etc.). This taxonomy provides a formu-

lation for incorporating diverse geometric entities under a single optimization scheme: virtual groups are added to the optimization problem and are placed alongside cameras in the non-eliminated block of the normal equations, absorbing the coupling between associated features and preserving the block-diagonal structure over the feature block.

Based on this formulation, we express both group and wireframe constraints as reprojection-based residuals in pixel space. Group-induced reprojection errors measure the pixel displacement from projecting a feature onto the group surface, while cross-feature reprojection errors use the observation of one feature to constrain another. This avoids direct 3D regularization and implicitly weights each constraint by the uncertainty (via Fisher information) of the involved features, naturally anchoring structural entities through more reliable associations.

We validate our framework through experiments on synthetic and real-world datasets. Our proposed framework is able to optimize on top of significantly richer reconstructions with holistic 3D structures (as in Fig. 1), while achieving runtime performance comparable to classical bundle adjustment and consistent improvements on camera poses and geometry. Code will be made available.

2 Related Work

Advances in Bundle Adjustment. Bundle adjustment [59] has been widely studied and applied in 3D reconstruction for decades. Sparse matrix techniques and the Schur complement [21] have enabled efficient solutions by exploiting the bipartite structure between cameras and 3D points. There have been many efforts to improve its scalability via parallel/distributed computing [2, 17, 69] and advanced solvers [24, 37]. Other advances also include robust cost functions [75], incremental formulation [25], and differentiable BA for end-to-end learning [55, 56]. Building upon these advances, several open-sourced libraries have emerged and become standard tools, notably including g2o [28], Ceres solver [1], GTSAM [14], etc. These libraries work together with sparse linear algebra libraries [11, 13] and provide a nice ecosystem for modern developments of geometric 3D vision, enabling multiple reliable systems on structure-from-motion [40, 49, 53, 68] and SLAM [16, 27, 41] across years. Recently, bundle adjustment is also proved to be effective in refining the poses and geometry of feedforward 3D models [10, 61–63, 66]. In this work, we introduce a unified framework extending bundle adjustment to handle holistic 3D structures with higher-level primitives and relations.

Geometric Primitives in Optimization. While bundle adjustment is traditionally formulated over 3D points, extending it to general geometric primitives dates back to the 1990s [22]. However, most recent works that incorporate object-level primitives [29, 32, 42, 70, 73] directly optimize high-dimensional variables representing the primitives through 2D modeling of primitive observations. Such formulations typically ignore the structural relationships between primitives and their supporting features, which prevents seamless integration with classical bundle adjustment over 3D points.

Among geometric primitives, 3D lines have been extensively studied. In [4], Bartoli and Sturm formulate bundle adjustment over infinite 3D lines using an orthonormal parameterization of Plücker coordinates, which has since been adopted in many modern line-based optimization systems [35, 36, 71, 82]. These systems often further exploit structural cues such as vanishing points [47, 57], Manhattan-world assumptions [12] and their variants [30, 48], as well as learned monocular priors [26]. In practice, these systems can scale to thousands of lines, which motivates us to treat lines as feature-level entities similar to points.

Planes are another fundamental primitive, especially in man-made environments. Planar constraints have long been studied in geometric optimization, including formulations based on two-view homographies [18], auto-calibration [58], and explicit parameterization of 3D structures [5]. The use of homography constraints has recently been revisited in [78] for computational efficiency. However, this formulation requires enumerating over image pairs. Explicit parameterization of 3D structures [5] is conceptually appealing, but is sensitive to incorrect feature-plane associations, which are difficult to avoid in practice. Many recent works [8, 31, 52, 54] therefore introduce soft point-plane regularization terms in 3D, which have also been extended to line-plane relationships [3, 31]. However, unlike point-to-plane formulations in LiDAR or RGB-D SLAM [9, 20, 39, 79], which benefit from principled noise models on 3D measurements, monocular systems typically measure errors in 3D space at absolute scale with heuristic weighting. In contrast, our framework jointly optimizes multiple types of primitives using a unified residual formulation in pixel space, which implicitly models the uncertainty of different feature-primitive associations.

Modeling of Point-Line Associations. Wireframes are ubiquitous in man-made environments. The advent of deep neural networks has enabled reliable 2D wireframe prediction from single images [72, 80]. Wireframe representations have also proven beneficial for feature matching, either through junction-aware heuristics [38, 71] or learned approaches [44, 60]. Recent line-based reconstruction systems [35, 36] further incorporate point-line associations in optimization via 3D point-to-line distance regularization. However, these formulations introduce dense couplings between line segments that share junctions, disrupting the sparse structure and leading to expensive Schur complement operations, particularly with dense solvers. Our method addresses this limitation through a reformulation that preserves sparsity while fully exploiting wireframe constraints.

3 A Unified Framework of Holistic Bundle Adjustment

3.1 Review of Classical Bundle Adjustment

Bundle adjustment (BA) jointly optimizes camera parameters and 3D structure by minimizing reprojection errors across all observations [59]. Given N images, each with pose $[R_i \mid \mathbf{t}_i]$ and associated (potentially shared) camera intrinsics $K_{c(i)}$, we denote the full camera parameters of image i as $C_i = (K_{c(i)}, R_i, \mathbf{t}_i)$

and collect them in $\mathcal{C} = \{C_1, \dots, C_N\}$. Given M 3D points $\mathcal{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_M\}$, the classical BA objective is:

$$E(\mathcal{C}, \mathcal{X}) = \sum_{(i,j) \in \mathcal{O}} \|\pi(C_i, \mathbf{X}_j) - \mathbf{x}_{ij}\|^2, \quad (1)$$

where $\pi(C_i, \mathbf{X}_j)$ projects point $\mathbf{X}_j$ into camera C_i , $\mathbf{x}_{ij}$ is the corresponding 2D observation, and $\mathcal{O}$ is the set of all visibility pairs. Minimizing Eq. (1) via the Levenberg-Marquardt algorithm leads to normal equations with a characteristic block structure:

$$\begin{bmatrix} \mathbf{U} & \mathbf{W} \\ \mathbf{W}^\top & \mathbf{V} \end{bmatrix} \begin{bmatrix} \delta\mathcal{C} \\ \delta\mathcal{X} \end{bmatrix} = - \begin{bmatrix} \mathbf{g}_c \\ \mathbf{g}_x \end{bmatrix}, \quad (2)$$

where $\mathbf{U}$ and $\mathbf{V}$ are the camera-camera and point-point blocks of the approximate Hessian, $\mathbf{W}$ is the camera-point off-diagonal, and $\mathbf{g}_c$, $\mathbf{g}_x$ are the gradient vectors. A crucial property is that $\mathbf{V}$ is *block-diagonal*: each 3D point appears only in residuals involving its observing cameras, creating no direct coupling between different points. This enables efficient Schur complement elimination, where points are marginalized to yield a reduced camera-only system:

$$(\mathbf{U} - \mathbf{W}\mathbf{V}^{-1}\mathbf{W}^\top) \delta\mathcal{C} = -\mathbf{g}_c + \mathbf{W}\mathbf{V}^{-1}\mathbf{g}_x. \quad (3)$$

Since $\mathbf{V}^{-1}$ reduces to inverting small independent blocks, this elimination is both efficient and numerically stable. Modern solvers such as Ceres [1] exploit this structure extensively. However, real-world scenes often exhibit richer geometric relations such as parallelism, coplanarity, and wireframe connectivity. We next introduce a taxonomy that distinguishes features from groups and show how both can be incorporated while preserving efficient Schur elimination. Figure 2 presents an illustration of the main concepts introduced in this paper.

3.2 Features and Groups: A Taxonomy on Sparse 3D Structures

We propose a taxonomy that categorizes geometric entities into *features* and *groups* based on their role in the normal equations of bundle adjustment.

Features are geometric primitives with direct 2D measurements in each observing image. Points and lines are the canonical examples: a 3D point $\mathbf{X}_j$ is observed as a 2D keypoint $\mathbf{x}_{ij}$; a 3D line $\mathbf{L}_k$, parameterized in Plücker coordinates [6], is observed as a 2D line segment ℓ_{ik} . Adding lines to Eq. (1) yields:

$$E = \sum_{(i,j) \in \mathcal{O}_p} \|\pi_p(C_i, \mathbf{X}_j) - \mathbf{x}_{ij}\|^2 + \sum_{(i,k) \in \mathcal{O}_l} \|\rho(\pi_l(C_i, \mathbf{L}_k), \ell_{ik})\|^2, \quad (4)$$

where $\rho(\cdot, \cdot) \in \mathbb{R}^4$ computes the perpendicular distance vectors between the endpoints of the observed 2D segment and the projected 3D line. In the vanilla point-line bundle adjustment with reprojection errors, each feature contributes an independent diagonal block to $\mathbf{V}$ in the normal equation, since its residuals involve only itself and its observing cameras, without coupling the other features.

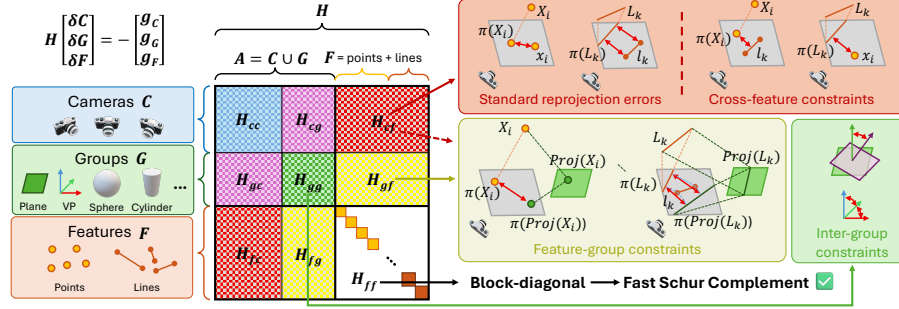


Fig. 2: Taxonomy of features, groups, and cross-feature associations. Left: Features $\mathcal{F}$ (points, lines) have direct 2D measurements; groups $\mathcal{G}$ (planes, VPs, 3D primitives) encode higher-order relations among features. **Middle:** approximate Hessian with block diagonal features $\mathbf{H}_{ff}$ matrix. **Right:** representations of the features, feature-group, cross-feature, and inter-group constraints in our optimization problem.

Groups encode higher-order geometric relations among features. Vanishing points (VPs) encode parallelism among 3D lines, planes encode coplanarity of points and lines, and other primitives such as spheres, cylinders, cuboids, and object priors constrain features to lie on a parametric manifold. Unlike features, groups generally lack direct per-image 2D measurements in the classical sense. Instead, groups enforce consistency on its associated features: a VP group enforces directional consistency among its associated lines, while a plane group enforces coplanarity among its associated features.

There are two common approaches to incorporating such group constraints into bundle adjustment. The first is explicit reparameterization [5], where features are constrained to lie exactly on a structure, *e.g.*, reducing a 3D point on a plane from 3 to 2 degrees of freedom. However, this requires perfect feature-group associations prior to optimization, which is challenging in practice. The second introduces soft regularization terms [31, 54] that penalize deviations from structural constraints. This can be done either through direct pairwise or higher-order constraints between features (*e.g.*, multiple points should be coplanar), or by introducing explicit group parameters with soft feature-group regularizations.

Computational Structure. The two soft regularization approaches have different implications for the structure of the normal equations. Direct constraints between features introduce off-diagonal blocks in the feature Hessian, breaking the block-diagonal property that enables efficient Schur elimination. For example, enforcing parallelism between two lines without a VP parameter creates a residual depending on both line parameters, filling in off-diagonal blocks and thus coupling them in the Hessian.

In contrast, introducing explicit group parameters avoids this coupling (Fig. 2, middle). Each feature-group residual involves exactly one feature and one group, analogous to how a camera-feature residual involves one feature and one camera. The key is to place group parameters alongside cameras in the non-eliminated block. Given group parameters $\mathcal{G} = \{G_1, \dots, G_P\}$, we form an augmented set

$\mathcal{A} = \mathcal{C} \cup \mathcal{G}$ and rewrite Eq. (2) as:

$$\begin{bmatrix} \mathbf{H}_{aa} & \mathbf{H}_{af} \\ \mathbf{H}_{fa} & \mathbf{H}_{ff} \end{bmatrix} \begin{bmatrix} \delta\mathcal{A} \\ \delta\mathcal{F} \end{bmatrix} = - \begin{bmatrix} \mathbf{g}_a \\ \mathbf{g}_f \end{bmatrix}. \quad (5)$$

Since each residual involves at most one feature, $\mathbf{H}_{ff}$ remains block-diagonal. The Schur complement eliminates features to yield a reduced system over cameras and groups: $\mathbf{S} = \mathbf{H}_{aa} - \mathbf{H}_{af}\mathbf{H}_{ff}^{-1}\mathbf{H}_{fa}$.

Note that the categorization of entities into features and groups reflects computational considerations, not just geometric ones. A line could in principle be treated as a group enforcing collinearity among its associated points, but lines typically number in the thousands, and placing them in the non-eliminated block would make the reduced system $\mathbf{S}$ prohibitively large. Similarly, a VP could act as a feature if 2D direction measurements are available, and a plane normal could serve as a feature when estimated from monocular cues. However, such 2D measurements tend to be inaccurate, and groups like planes and VPs are typically few enough (*i.e.*, $|\mathcal{G}| \ll |\mathcal{F}|$) to remain in the non-eliminated block without significant overhead. Nonetheless, when available, 2D measurements of groups are still compatible with our framework, as they contribute only to $\mathbf{H}_{aa}$ and do not affect the block-diagonal structure of $\mathbf{H}_{ff}$.

Inter-Group Relations. A further benefit of placing groups in $\mathcal{A}$ is that cross-group and camera-group constraints can be added at no cost to the Schur complement structure, since they involve only parameters within $\mathcal{A}$ and do not affect $\mathbf{H}_{ff}$. For instance, we employ orthogonality and parallelism constraints between VP pairs and between plane normal pairs to enforce Manhattan world priors, which contribute only to $\mathbf{H}_{aa}$, or more specifically $\mathbf{H}_{gg}$ in Fig. 2.

Cross-Feature Associations (Wireframes). Beyond group-feature relations, features of different types can be directly associated with each other. In architectural and urban scenes, points frequently lie on lines at junctions, forming wireframe structures. These point-line associations encode geometric constraints that are complementary to group constraints: while groups relate features to a shared structural entity, cross-feature associations relate features directly to each other. Unlike group constraints, this coupling between features cannot be resolved through parameter ordering alone, as both entities belong to $\mathcal{F}$. We address this challenge in Sec. 3.4.

3.3 Group Constraints in Bundle Adjustment

Having established that groups can be placed alongside cameras to preserve the block-diagonal structure (Sec. 3.2), we now design the group-feature residuals. For directional groups such as VPs, angular residuals between the line direction $\mathbf{d}_k$ (from the Plücker representation) and the VP direction $\mathbf{v}_g$ are scale-free and straightforward to integrate:

$$r_v = |\mathbf{d}_k \times \mathbf{v}_g|. \quad (6)$$

Positional groups such as planes are less straightforward. A natural formulation uses direct 3D residuals; for example, coplanarity can be enforced via the

point-to-plane distance $r_{3D} = \mathbf{n}^\top \mathbf{X}_j - d$, where $\mathbf{n}$ and d denote the plane normal and offset. However, such 3D constraints are problematic. Because reprojection errors are invariant to global scale, the optimizer can reduce 3D penalties by shrinking the reconstruction. In SfM, scale is typically fixed by constraining the translation of the initial image pair, which does not prevent the remaining structure from collapsing towards it. Moreover, balancing 3D and 2D terms depends on scene scale: proper weighting would require the 3D covariance of the features, which depends on viewing configurations and cannot be recovered from 2D observations alone. In contrast, 2D pixel noise is well characterized and approximately isotropic, making reprojection-based formulations easier to calibrate.

Group-Induced Reprojection Error. Motivated by the discussions above, we propose to formulate positional group constraints through pixel-space residuals that measure the reprojection difference caused by the group projection (Fig. 2, right). For a plane group G_g with normal $\mathbf{n}$ and offset d , we let the group-induced residual for an associated point $\mathbf{X}_j$ observed in camera C_i be:

$$r_g^{(p)} = \pi_p(C_i, \mathbf{X}_j) - \pi_p\left(C_i, \text{Proj}_{G_g}(\mathbf{X}_j)\right) \in \mathbb{R}^2, \quad (7)$$

where $\text{Proj}_{G_g}(\mathbf{X}_j) = \mathbf{X}_j - \frac{\mathbf{n}^\top \mathbf{X}_j - d}{\|\mathbf{n}\|^2} \mathbf{n}$ projects the 3D point onto the plane surface. If the point lies on the plane, $\text{Proj}_{G_g}(\mathbf{X}_j) = \mathbf{X}_j$ and the residual vanishes. Otherwise, it measures the pixel displacement between the original and group-constrained reprojections. The same principle applies to associated lines: the residual is the difference between the standard line reprojection residual and that of the plane-projected line:

$$r_g^{(l)} = \rho(\pi_l(C_i, \mathbf{L}_k), \ell_{ik}) - \rho\left(\pi_l\left(C_i, \text{Proj}_{G_g}(\mathbf{L}_k)\right), \ell_{ik}\right) \in \mathbb{R}^4. \quad (8)$$

More generally, any geometric entity that defines a projection operator from 3D space onto a constraint surface can serve as a group in our framework, with the group-induced reprojection error following the same pattern. In all cases, each residual involves exactly one feature (in $\mathcal{F}$) and at most one group plus one camera (in $\mathcal{A}$), preserving the block-diagonal structure of $\mathbf{H}_{ff}$.

Statistical Interpretation: Group Anchoring through Implicit Weighting. Beyond avoiding scale ambiguity, the reprojection difference formulation provides a clean mechanism for anchoring the group surface through well-observed features. Concretely, let $\boldsymbol{\delta} = \mathbf{X}_j - \text{Proj}_{G_g}(\mathbf{X}_j) \in \mathbb{R}^3$ denote the displacement of a point from the group surface, and let $J_i = \partial \pi_p / \partial \mathbf{X}|_{C_i, \mathbf{X}_j} \in \mathbb{R}^{2 \times 3}$ be the projection Jacobian in view i . To first order, the group-induced residual (Eq. (7)) linearizes as:

$$r_g^{(p)} \approx J_i \boldsymbol{\delta}. \quad (9)$$

The effective cost of the group constraint, summed over all N_j views observing the point, is therefore:

$$E_g \approx \sum_i \|J_i \boldsymbol{\delta}\|^2 = \boldsymbol{\delta}^\top \left(\sum_i J_i^\top J_i \right) \boldsymbol{\delta}. \quad (10)$$

The matrix $\sum_i J_i^\top J_i$ is the Fisher information matrix of the point’s 3D position, *i.e.*, the inverse of its covariance $\Sigma_{\mathbf{X}}^{-1}$ [19]. Notably, this is the same matrix that arises from the standard reprojection residuals used to triangulate the point, so the group constraint inherits a geometrically consistent uncertainty model. The group-induced cost is thus equivalent to a Mahalanobis distance $\delta^\top \Sigma_{\mathbf{X}}^{-1} \delta$. We want to note that under the common isotropic noise assumption, our group residual is equivalent to 3D point-to-plane distance with 3D point covariance computed from relinearized reprojection Jacobians at each iteration, which is closely related to the reduction perspective of implicit models in [59].

This directly determines how the group surface is estimated: well-observed features (with large Fisher information) produce large residuals when off-surface, anchoring the group to their positions, while poorly observed features contribute little regardless of displacement. The group parameters are therefore driven by reliable features, which in turn benefit from the structural consistency enforced by the group constraint. As shown in our experiments, this consistently improves both pose and geometry accuracy.

Note that our formulation cannot recover features with degenerate viewing geometry onto the group surface. A 3D formulation could push such features onto the surface, but since degenerate directions are precisely those where reprojection is insensitive, cameras do not benefit from this recovery. Importantly, such features also pose minimal risk to group estimation: their weak Fisher information produces small residuals that have little influence on the group parameters. **Sparsity Analysis on $\mathbf{H}_{cg}$.** While the group-induced reprojection residual involves camera, feature, and group parameters simultaneously, it does not worsen the sparsity pattern of the reduced system compared to a 3D formulation. A 3D point-to-plane residual $r = \mathbf{n}^\top \mathbf{X} - d$ has no direct dependence on camera parameters, contributing only to $\mathbf{H}_{gf}$ and $\mathbf{H}_{ff}$. However, after Schur complement elimination of features, indirect camera-group coupling arises in $\mathbf{H}_{cg}$ wherever a camera observes a feature associated with a group. Our 2D formulation creates this coupling directly, but the sparsity pattern is identical – a nonzero entry in $\mathbf{H}_{cg}$ exists if and only if the camera sees some feature associated with the group. The structure of the reduced system $\mathbf{S}$ is thus unchanged.

3.4 Cross-Feature Constraints in Bundle Adjustment

As discussed in Sec. 3.2, wireframe constraints couple features directly in $\mathbf{H}_{ff}$, breaking its block-diagonal structure. For example, enforcing a point-on-line constraint in 3D as in [36] introduces a residual that depends on both feature parameters, and the 3D point-to-line distance suffers from the same scale-dependent weight calibration issues as discussed in Sec. 3.3. We therefore propose a reformulation that preserves sparsity.

Cross-Feature Reprojection Error. Our key observation is that wireframe constraints can be reformulated as reprojection errors where only *one* feature participates as an optimization variable, while the *2D observation* of the other feature serves as a fixed measurement. Given a wireframe edge $(\mathbf{X}_j, \mathbf{L}_k)$, we define two complementary residuals across their respective visibility sets.

For an image i where line $\mathbf{L}_k$ is observed as ℓ_{ik} but point $\mathbf{X}_j$ is *not* directly detected, the point-to-line residual measures the distance from the projected point to the observed line:

$$r_{p \rightarrow l} = d(\pi_p(C_i, \mathbf{X}_j), \ell_{ik}), \quad (11)$$

where $d(\cdot, \cdot)$ is the perpendicular point-to-line distance in pixels. The optimization variables are C_i (camera, in $\mathcal{A}$) and $\mathbf{X}_j$ (feature, in $\mathcal{F}$); the 2D line observation ℓ_{ik} enters as a constant measurement. Intuitively, if the point lies on the line, it should project near the observed line in any view where the line is visible, even in views where the point itself is not detected. As with the group-induced reprojection error (Sec. 3.3), this 2D formulation avoids scale-dependent 3D distances and implicitly weights the constraint by observation geometry, with well-observed features providing stronger anchoring signals (Eq. (10)).

Conversely, for an image i where point $\mathbf{X}_j$ is observed as $\mathbf{x}_{ij}$ but line $\mathbf{L}_k$ is not detected, the line-to-point residual is:

$$r_{l \rightarrow p} = d(\mathbf{x}_{ij}, \pi_l(C_i, \mathbf{L}_k)), \quad (12)$$

where $\mathbf{x}_{ij}$ is a constant measurement and $\mathbf{L}_k$ is the optimization variable. In images where both features are directly observed, we omit the cross-feature terms as their standard reprojection residuals (Eq. (4)) already provide the constraint. Note that this cross-feature reprojection error formulation inherits the same statistical interpretation of the classical reprojection error and thus naturally supports more advanced noise models.

Sparsity Analysis. Each cross-feature residual involves exactly one feature and one camera. The Jacobian of $r_{p \rightarrow l}$ has nonzero blocks only for C_i and $\mathbf{X}_j$; that of $r_{l \rightarrow p}$ only for C_i and $\mathbf{L}_k$. No off-diagonal blocks arise between point j and line k in $\mathbf{H}_{ff}$, preserving its block-diagonal structure. The constraint is mediated through the 2D observation of the partner feature, decoupling the features while still enforcing their geometric relation. Note that cross-feature constraints extend the effective visibility graph, making the reduced system $\mathbf{S}$ denser in camera-camera blocks, but efficient Schur elimination remains preserved.

3.5 Discussions and Applications

Convergence and Termination. A practical benefit of formulating constraints in pixel space is simplified convergence monitoring. When mixing 3D regularization terms (in metric units) with 2D reprojection errors (in pixels), the residuals have different magnitudes and convergence dynamics, complicating cost-based termination criteria such as the relative cost change $|\Delta E|/E$. With our unified pixel-space formulation, the total cost has a consistent interpretation and such criteria apply directly. For the remaining angular residuals (*e.g.*, VP constraints in Eq. (6) and cross-group angular constraints), we isolate their contribution and monitor convergence based on the reprojection terms only.

Structure-from-Motion Pipeline. Our bundle adjustment framework integrates naturally into incremental structure-from-motion. After registering each

new image and triangulating features and groups, the system performs local bundle adjustment with the full set of constraints, including features, groups, and wireframe edges. Image selection leverages both point and line visibility, and camera registration uses both point and line correspondences. This yields an end-to-end SfM pipeline where geometric structures participate from the earliest stages of reconstruction. We also want to note that, the sparsity analysis and Schur elimination strategy extend directly to global positioning [43, 81] in global SfM, as cameras, features, and groups share the same computational structure.

Discussion: Robustness to Noisy Upstream Associations. Our work focuses on optimization once structural priors are available, rather than solving the problem of adaptive model selection. We therefore employ an structure association strategy that favors precision over recall and optimize these residuals with robust losses. Importantly, robust thresholding becomes easier in our pixel-level formulation than in mixed 2D/3D formulations, especially for SfM with its scale ambiguity. Nonetheless, the reliance on reasonably good quality of associations is an important practical limitation of our framework. We also want to note the distinction between measurement noise and association reliability, where the former is handled in a principled way through how our residuals are formulated, while the latter is closely related to model selection and benefits from future advances in estimating confidences of the extracted structural constraints.

4 Experiments

Implementation Details. Our framework is agnostic to the choice of feature and group detection and matching methods. We use ALIKED [77] / DeepLSD [45] and LightGlue [34] / GlueStick [44] for point / line detection and matching. Vanishing points are detected with JLinkage [57], and planes with a custom segmentation method fitting planes to predicted depth and surface normals of MoGe-2 [64, 65]. Images are resized so that the longer side is at most 800 pixels, to satisfy model constraints. Within each image, feature-group associations are obtained during detection (e.g., lines assigned to VPs, points and lines to plane segments), while point-line associations are derived from spatial incidences. Across images, group matchings are established either by voting over matched features or by IoU checks of warped segmentation masks via dense matching [15, 76]. To construct 3D constraints from 2D observations, we aggregate feature-group and cross-feature associations across images through voting following [36]. Group constraints count images where a feature observation is associated with a group, while wireframe constraints count images where a point-line pair forms a junction. Constraints are instantiated when the number of such images exceeds 3, and subsequently optimized with a robust Cauchy loss. Please refer to more details in our supplementary material.

Our proposed BA is built on top of Ceres solver [1] with analytical Jacobians. The SfM pipeline follows COLMAP [49], which automatically selects the numerical solver depending on number of images. Runtime experiments are conducted on a workstation with an 8-core Intel Core i7-10700K CPU at 3.80 GHz.

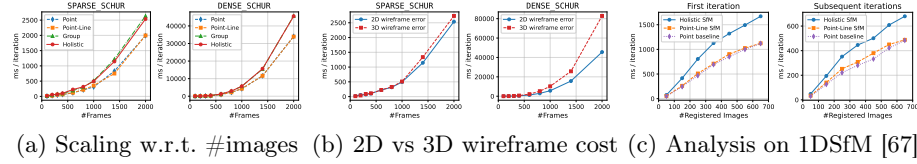


Fig. 3: Bundle adjustment runtime analysis. (a) Per-iteration cost scaling with number of images. (b) Comparison between 2D vs 3D wireframe cost. (c) First iteration vs. subsequent iterations of BA on 1DSfM, averaged across 3 scenes.

Table 1: Runtime and reconstruction statistics on 1DSfM [67]. The *Point baseline* is equivalent to COLMAP with only one change: replacing the gradient-norm termination criterion with the relative cost change based criterion. The two other methods use the same criterion as the point baseline. We report runtime, BA statistics, reconstruction size (#Reg: number of registrations, #pts: 3D points, #lns: lines, #grps: group primitives), and mean reprojection error over 3D point observations (in pixels).

Scene	Method	Runtime & Convergence				Reconstruction statistics				
		SfM (min)	ms/iter	iter/BA	#BA	#Reg	#pts/#obs	#lns	#grps	err (px)
Tower of London	COLMAP reference	63.6	—	16.2	194	935 / 1322	97k/731k	—	—	1.08
	Point baseline	46.4	132.5	4.8	196	909 / 1322	99k/731k	—	—	1.04
	Point-Line SfM	54.6	158.9	4.8	205	927 / 1322	99k/732k	2.9k	—	1.02
	Holistic SfM (Ours)	59.2	214.0	11.8	226	963 / 1322	99k/731k	2.8k	117	1.03
Madrid Metropolis	COLMAP reference	45.3	—	21.7	225	742 / 1061	59k/462k	—	—	1.13
	Point baseline	31.5	136.3	5.4	233	752 / 1061	65k/476k	—	—	1.07
	Point-Line SfM	34.8	153.5	5.6	249	753 / 1061	63k/465k	2.1k	—	1.05
	Holistic SfM (Ours)	37.6	192.1	9.5	253	750 / 1061	65k/469k	2.1k	42	1.06
Gendarmenmarkt	COLMAP reference	68.8	—	14.6	234	943 / 1173	85k/846k	—	—	1.16
	Point baseline	43.5	281.5	4.2	248	950 / 1173	89k/847k	—	—	1.10
	Point-Line SfM	45.7	274.7	4.5	266	947 / 1173	90k/851k	3.8k	—	1.10
	Holistic SfM (Ours)	63.5	394.3	11.3	281	951 / 1173	88k/846k	3.7k	91	1.10

4.1 Benchmarking Holistic Bundle Adjustment

Synthetic Benchmark. To analyze scaling behavior in isolation, we construct synthetic BA problems by varying the number of images from 100 to 2000. For each configuration, we generate points, lines, planes, and wireframe edges with counts scaling proportionally. Each feature is observed with an average track length of 4-8 images. Please refer to details in supplementary materials.

We compare four BA configurations of increasing complexity: *Point* (point-only), *Point-Line* (points and lines), *Groups* (adding group constraints from Sec. 3.3), and *Holistic* (further adding wireframe constraints from Sec. 3.4). Figure 3(a) shows per-iteration cost as a function of problem sizes. All bundle adjustment configurations scale similarly under both *SPARSE_SCHUR* ($\sim n^{1.6-1.9}$) and *DENSE_SCHUR* ($\sim n^{2.6-2.8}$), indicating that our formulation preserves the asymptotic complexity of classical BA. With *DENSE_SCHUR*, runtime is strictly determined by the size of the reduced system, with *Point* and *Point-Line*, *Groups* and *Holistic* having nearly identical cost. This supports our analysis that groups behave similarly to camera variables in the normal equations. Figure 3(b) compares the per-iteration runtime of our 2D reprojection-based wireframe formulation with a 3D endpoint-distance formulation. With *SPARSE_SCHUR*, the 2D formulation is moderately faster as it avoids direct coupling between features. With *DENSE_SCHUR*, the gap widens because the 3D formulation produces significantly denser fill-in in the reduced system.

Table 2: Geometry evaluation on Hypersim [46] (8 scenes). Points and lines are measured against ground-truth meshes.

	Points				Lines					
	Inlier (%) $\uparrow$			Med. $\downarrow$ (mm)	Length Recall (m) $\uparrow$			Precision (%) $\uparrow$		
	@1mm	@5mm	@10mm		@1mm	@5mm	@10mm	@1mm	@5mm	@10mm
Per-point/per-line refinement	9.3	31.7	45.6	15.10	72.1	222.5	301.2	21.3	59.4	77.3
+ Group constraints	11.5	35.6	48.5	13.80	93.2	248.8	315.5	27.3	65.5	81.0
+ Group and wireframe constraints	11.9	36.3	48.9	13.55	93.3	249.5	315.7	27.3	65.5	81.0

Table 3: Point reconstruction quality on ETH3D [50] (13 scenes). Reported against laser scans using the official ETH3D multi-view evaluation benchmark.

	Accuracy (%) $\uparrow$				Completeness (%) $\uparrow$				F1 (%) $\uparrow$			
	@1cm	@2cm	@5cm	@10cm	@1cm	@2cm	@5cm	@10cm	@1cm	@2cm	@5cm	@10cm
Per-point refinement	45.17	58.70	75.53	84.57	0.14	0.74	4.10	11.09	0.29	1.46	7.66	19.25
+ Group and wireframe constraints	47.05	60.59	76.84	85.40	0.15	0.76	4.18	11.21	0.30	1.50	7.81	19.47

Real-World Benchmark. We further evaluate three large-scale scenes from 1DSfM [67] within a full incremental SfM pipeline (Tab. 1). As discussed in Sec. 3.5, we use the relative cost change as the BA termination criterion, which empirically stops earlier than COLMAP, whose termination relies on the gradient norm. Our method registers a comparable or slightly larger number of images while maintaining similar reprojection errors. The higher iteration count per BA call (e.g., 11.8 vs. 4.8 on Tower of London) reflects the richer constraint set where structural constraints generally require more iterations for convergence, but the overall pipeline runtime remains $\sim 1.3\times$ of the point baseline, substantially lower than the $2\text{--}4\times$ overhead over COLMAP reported in [35]. Figure 3(c) shows the BA cost breakdown on 1DSfM. The first iteration of each BA call performs symbolic analysis (sparsity pattern computation and ordering), while subsequent iterations require only numeric factorization. This symbolic overhead is amortized across iterations, after which the steady-state numeric cost dominates.

4.2 Applications

Geometry Evaluation. To study the benefits of group and wireframe constraints on the resulting 3D maps, we evaluate reconstruction quality against ground-truth geometry on the first eight scenes of Hypersim [46], following the protocol of [35]. Points and lines are measured against ground-truth meshes. For points, we report the inlier ratio (percentage within threshold) and median distance. For lines, length recall is the average correct length in meters across scenes, while precision is the fraction of total reconstructed length within threshold. The results, reported in Tab. 2, show that group constraints consistently improve the accuracy of reconstructed 3D points and substantially increase both the recall and precision of line segments. Incorporating additional wireframe constraints further improves performance by anchoring points to structural lines, thereby introducing stronger geometric consistency. We also evaluate our structural constraints on ETH3D [50] using the official multi-view benchmark in Tab. 3. Our

Table 4: Evaluating the effectiveness of bundle adjustment in full structure-from-motion pipelines. Camera pose accuracy is measured by relative pose AUCs.

	Hypersim (8 scenes)			ScanNet++ (20 scenes)		
	AUC@3°	AUC@5°	AUC@10°	AUC@3°	AUC@5°	AUC@10°
COLMAP	89.3	90.4	92.2	84.0	86.0	87.5
Point-Line SfM	92.2	93.2	94.0	84.7	86.7	89.2
Holistic SfM (Ours)	92.5	93.4	94.2	87.4	89.4	90.6

	ETH3D (11 scenes)			7Scenes (7 scenes)		
	AUC@3°	AUC@5°	AUC@10°	AUC@3°	AUC@5°	AUC@10°
COLMAP	50.2	66.4	77.6	22.9	48.7	75.5
Point-Line SfM	49.4	67.1	79.6	23.1	48.9	76.8
Holistic SfM (Ours)	50.4	68.1	80.4	24.3	50.8	78.4

Table 5: Ablation of the group-induced 2D reprojection error vs direct 3D constraints via geometry evaluation on ETH3D and SfM evaluation on ScanNet++.

Method	ETH3D Acc. (%) $\uparrow$			ScanNet++ AUC $\uparrow$			
	@1cm	@5cm	@10cm	@1°	@3°	@5°	@10°
Point BA	45.17	75.53	84.57	66.1	84.0	86.0	87.5
Holistic BA w/. 3D group error ($w=1$)	45.17	75.55	84.58	67.4	85.0	87.1	89.1
Holistic BA w/. 3D group error ($w=10$)	45.24	75.61	84.63	68.1	85.0	89.0	90.5
Holistic BA w/. 3D group error ($w=100$)	45.72	76.12	84.88	<u>68.3</u>	87.8	89.9	91.1
Holistic BA w/. 3D group error ($w=1000$)	<u>46.69</u>	<u>76.71</u>	<u>85.15</u>	67.1	85.1	87.1	88.3
Holistic BA (Ours)	47.05	76.84	85.40	69.4	<u>87.4</u>	<u>89.4</u>	<u>90.6</u>

method improves both accuracy and completeness across all thresholds, demonstrating the effectiveness of group constraints in enhancing feature reconstruction by enforcing structural consistency.

Structure from Motion. As discussed in Sec. 3.5, we integrate our bundle adjustment (BA) framework into a full incremental SfM pipeline and evaluate camera pose accuracy on four datasets: Hypersim [46], ScanNet++ [74], ETH3D [50], and 7Scenes [51]. We use the first eight scenes of Hypersim as in [35] and adopt the splits from [33] for the remaining datasets. Our holistic bundle adjustment consistently improves pose accuracy over prior methods across all datasets, covering both indoor and outdoor scenes (Tab. 4). The improvement is particularly pronounced on ScanNet++, which contains indoor environments with rich geometric structures such as parallel and coplanar surfaces.

Ablation study. We report results in Tab. 5, comparing our 2D group-induced reprojection error with direct 3D plane-distance constraints under different weights. The 3D formulation requires careful tuning: small weights underuse structural information, while large weights ($w=1000$) degrade pose accuracy on ScanNet++. In contrast, our 2D formulation requires no tuning and achieves the best or competitive results on both geometry (ETH3D) and pose (ScanNet++) metrics.

Mapping Geometric Primitives. Beyond quantitative improvements, our framework produces semantically richer maps with explicit geometric primitives. Example holistic 3D reconstructions are shown in Fig. 4. While planes and van-

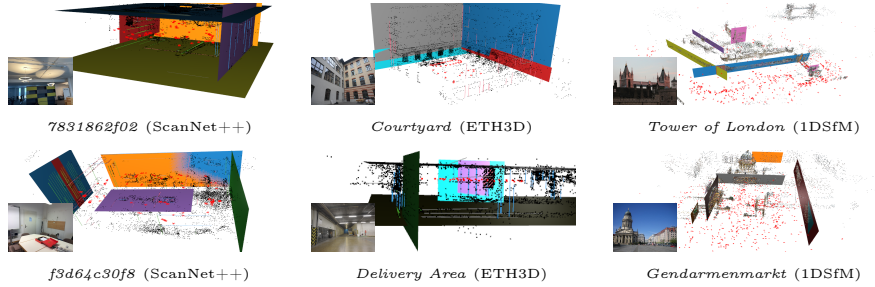


Fig. 4: Visualizations of bundle adjusted holistic 3D structures. Parallel lines are colored the same. Note that lines and planes are optimized as infinite entities in the bundle adjustment, and their extents in the figures are only for illustration purpose.

ishing points are the most common structural groups in real-world scenes, our formulation is general and supports other primitives. In Fig. 1, we show bundle adjustment on a scene containing planes, spheres, and cylinders, where groups are detected via SAM3 [7] with semantic prompts (football→sphere, can→cylinder, laptop/postcard→plane). This demonstrates its flexibility and shows the potential of combining geometric abstraction with modern visual-language models. We believe our framework opens up opportunities to further integrate geometric reasoning with visual-language understanding.

5 Conclusion

We present a unified bundle adjustment framework that integrates higher-order geometric structures while preserving efficiency. By placing group parameters alongside cameras and formulating all constraints as reprojection-based residuals, we maintain efficient Schur elimination and avoid the scale ambiguity and conditioning issues of 3D regularization. Experiments demonstrate comparable runtime with significantly richer reconstructions and improved accuracy.

Acknowledgements. Viktor Larsson was supported by ELLIIT and the Swedish Research Council (Grant No. 2023-05424). We sincerely thank Alexander Veicht, Paul-Edouard Sarlin, Eric Dexheimer, Xinyue Zhang, Linfei Pan, Elisabetta Fedele, Johannes Schönberger, Lixin Xue, and anonymous reviewers for their helpful feedbacks.

References

1. Agarwal, S., Mierle, K.: Ceres solver. <http://ceres-solver.org>
2. Agarwal, S., Snavely, N., Seitz, S.M., Szeliski, R.: Bundle adjustment in the large. In: ECCV (2010)
3. Bai, X., Cui, H., Shen, S.: Consistent 3d line mapping. In: ECCV (2024)
4. Bartoli, A., Coquerelle, M., Sturm, P.: A framework for pencil-of-points structure-from-motion. In: ECCV (2004)

5. Bartoli, A., Sturm, P.: Constrained structure and motion from multiple uncalibrated views of a piecewise planar scene. *International Journal of Computer Vision* **52**(1), 45–64 (2003)
6. Bartoli, A., Sturm, P.: Structure-from-motion using lines: Representation, triangulation, and bundle adjustment. *Computer Vision and Image Understanding (CVIU)* **100**(3), 416–441 (2005)
7. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719* (2025)
8. Chen, C., Geneva, P., Peng, Y., Lee, W., Huang, G.: Monocular visual-inertial odometry with planar regularities. In: *ICRA*. pp. 6224–6231 (2023)
9. Chen, D., Wang, S., Xie, W., Zhai, S., Wang, N., Bao, H., Zhang, G.: Vip-slam: An efficient tightly-coupled rgb-d visual inertial planar slam. In: *ICRA*. IEEE (2022)
10. Chen, W., Zhang, G., Wimbauer, F., Wang, R., Araslanov, N., Vedaldi, A., Cremers, D.: Back on track: Bundle adjustment for dynamic scene reconstruction. In: *ICCV* (2025)
11. Chen, Y., Davis, T.A., Hager, W.W., Rajamanickam, S.: Algorithm 887: Cholmod, supernodal sparse cholesky factorization and update/downdate. *ACM Transactions on Mathematical Software (TOMS)* **35**(3), 1–14 (2008)
12. Coughlan, J., Yuille, A.L.: The manhattan world assumption: Regularities in scene statistics which enable bayesian inference. In: *NeurIPS* (2000)
13. Davis, T.A.: Algorithm 1000: Suitesparse: Graphblas: Graph algorithms in the language of sparse linear algebra. *ACM Transactions on Mathematical Software (TOMS)* **45**(4), 1–25 (2019)
14. Dellaert, F.: Factor graphs and gtsam: A hands-on introduction. Georgia Institute of Technology, Tech. Rep **2**(4) (2012)
15. Edstedt, J., Sun, Q., Bökmann, G., Wadenbäck, M., Felsberg, M.: Roma: Robust dense feature matching. In: *CVPR* (2024)
16. Engel, J., Schöps, T., Cremers, D.: Lsd-slam: Large-scale direct monocular slam. In: *ECCV* (2014)
17. Fan, T., Ortiz, J., Hsiao, M., Monge, M., Dong, J., Murphey, T.D., Mukadam, M.: Daba: Decentralized and accelerated large-scale bundle adjustment. *The International Journal of Robotics Research* **44**(10-11), 1892–1919 (2025)
18. Faugeras, O.D., Lustman, F.: Motion and structure from motion in a piecewise planar environment. *International Journal of Pattern Recognition and Artificial Intelligence* **2**(03), 485–508 (1988)
19. Förstner, W., Wrobel, B.P.: *Photogrammetric computer vision* (2016)
20. Geneva, P., Ekenhoff, K., Yang, Y., Huang, G.: Lips: Lidar-inertial 3d plane slam. In: *IROS*. IEEE (2018)
21. Hartley, R., Zisserman, A.: *Multiple view geometry in computer vision*. Cambridge university press (2003)
22. Hartley, R.I.: An object-oriented approach to scene reconstruction. In: *1996 IEEE International Conference on Systems, Man and Cybernetics. Information Intelligence and Systems (Cat. No. 96CH35929)*. vol. 4, pp. 2475–2480. IEEE (1996)
23. Huang, K., Wang, Y., Zhou, Z., Ding, T., Gao, S., Ma, Y.: Learning to parse wireframes in images of man-made environments. In: *CVPR* (2018)
24. Jeong, Y., Nister, D., Steedly, D., Szeliski, R., Kweon, I.S.: Pushing the envelope of modern methods for bundle adjustment. *IEEE transactions on pattern analysis and machine intelligence* **34**(8), 1605–1617 (2011)

25. Kaess, M., Johannsson, H., Roberts, R., Ila, V., Leonard, J., Dellaert, F.: isam2: Incremental smoothing and mapping with fluid relinearization and incremental variable reordering. In: ICRA. IEEE (2011)
26. Ke, Z., Tan, B., Xia, G.S., Shen, Y., Xue, N.: Interacted planes reveal 3d line mapping. arXiv preprint arXiv:2602.01296 (2026)
27. Klein, G., Murray, D.: Parallel tracking and mapping for small ar workspaces. In: 2007 6th IEEE and ACM international symposium on mixed and augmented reality. pp. 225–234. IEEE (2007)
28. Kümmerle, R., Grisetti, G., Strasdat, H., Konolige, K., Burgard, W.: g 2 o: A general framework for graph optimization. In: ICRA. IEEE (2011)
29. Laidlow, T., Davison, A.J.: Simultaneous localisation and mapping with quadric surfaces. In: 2022 International Conference on 3D Vision (3DV). pp. 1–9. IEEE (2022)
30. Li, H., Zhao, J., Bazin, J.C., Kim, P., Joo, K., Zhao, Z., Liu, Y.H.: Hong kong world: Leveraging structural regularity for line-based slam. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(11), 13035–13053 (2023)
31. Li, X., He, Y., Lin, J., Liu, X.: Leveraging planar regularities for point line visual-inertial odometry. In: IROS. IEEE (2020)
32. Liao, Z., Wang, W., Qi, X., Zhang, X., Xue, L., Jiao, J., Wei, R.: Object-oriented slam using quadrics and symmetry properties for indoor environments. arXiv preprint arXiv:2004.05303 (2020)
33. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)
34. Lindenberger, P., Sarlin, P.E., Pollefeys, M.: Lightglue: Local feature matching at light speed. In: ICCV (2023)
35. Liu, S., Gao, Y., Zhang, T., Pautrat, R., Schönberger, J.L., Larsson, V., Pollefeys, M.: Robust incremental structure-from-motion with hybrid features. In: ECCV (2024)
36. Liu, S., Yu, Y., Pautrat, R., Pollefeys, M., Larsson, V.: 3d line mapping revisited. In: CVPR (2023)
37. Lourakis, M.I., Argyros, A.A.: Sba: A software package for generic sparse bundle adjustment. ACM Transactions on Mathematical Software (TOMS) **36**(1), 1–30 (2009)
38. Ma, J., Wang, X., He, Y., Mei, X., Zhao, J.: Line-based stereo slam by junction matching and vanishing point alignment. IEEE Access **7** (2019)
39. Ma, L., Kerl, C., Stückler, J., Cremers, D.: Cpa-slam: Consistent plane-model alignment for direct rgb-d slam. In: ICRA (2016)
40. Moulon, P., Monasse, P., Perrot, R., Marlet, R.: OpenMVG: Open multiple view geometry. In: International Workshop on Reproducible Research in Pattern Recognition (2016)
41. Mur-Artal, R., Montiel, J.M.M., Tardos, J.D.: Orb-slam: a versatile and accurate monocular slam system. IEEE Transactions on Robotics **31**(5), 1147–1163 (2015)
42. Nicholson, L., Milford, M., Sünderhauf, N.: Quadricslam: Dual quadrics from object detections as landmarks in object-oriented slam. IEEE Robotics and Automation Letters **4**(1), 1–8 (2018)
43. Pan, L., Baráth, D., Pollefeys, M., Schönberger, J.L.: Global structure-from-motion revisited. In: ECCV (2024)
44. Pautrat, R., Suárez, I., Yu, Y., Pollefeys, M., Larsson, V.: Gluestick: Robust image matching by sticking points and lines together. In: ICCV (2023)

45. Pautrat, R., Barath, D., Larsson, V., Oswald, M.R., Pollefeys, M.: Deeplsd: Line segment detection and refinement with deep image gradients. In: CVPR (2023)
46. Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M.A., Paczan, N., Webb, R., Susskind, J.M.: Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In: ICCV (2021)
47. Rother, C.: A new approach to vanishing point detection in architectural environments. *Image and Vision Computing* **20**(9-10), 647–655 (2002)
48. Schindler, G., Dellaert, F.: Atlanta world: An expectation maximization framework for simultaneous low-level edge grouping and camera calibration in complex man-made environments. In: CVPR (2004)
49. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: CVPR (2016)
50. Schops, T., Schonberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: CVPR (2017)
51. Shotton, J., Glocker, B., Zach, C., Izadi, S., Criminisi, A., Fitzgibbon, A.: Scene coordinate regression forests for camera relocalization in RGB-D images. In: CVPR (2013)
52. Shu, F., Wang, J., Pagani, A., Stricker, D.: Structure plp-slam: Efficient sparse mapping and localization using point, line and plane for monocular, rgb-d and stereo cameras. In: ICRA (2023)
53. Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: exploring photo collections in 3d. In: ACM SIGGRAPH (2006)
54. Taguchi, Y., Jian, Y.D., Ramalingam, S., Feng, C.: Point-plane slam for hand-held 3d sensors. In: 2013 IEEE international conference on robotics and automation. pp. 5182–5189. IEEE (2013)
55. Tang, C., Tan, P.: Ba-net: Dense bundle adjustment network. In: International Conference on Learning Representations (ICLR) (2019)
56. Teed, Z., Deng, J.: Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *neurips* **34** (2021)
57. Toldo, R., Fusiello, A.: Robust multiple structures estimation with j-linkage. In: ECCV (2008)
58. Triggs, B.: Autocalibration from planar scenes. In: ECCV (1998)
59. Triggs, B., McLauchlan, P.F., Hartley, R.I., Fitzgibbon, A.W.: Bundle adjustment—a modern synthesis. In: Vision Algorithms: Theory and Practice: International Workshop on Vision Algorithms Corfu, Greece, September 21–22, 1999 Proceedings (2000)
60. Ubingazhibov, A., Pautrat, R., Suárez, I., Liu, S., Pollefeys, M., Larsson, V.: Light-gluestick: a fast and robust glue for joint point-line matching. In: Computer Vision and Pattern Recognition Workshops (CVPRW) (2025)
61. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: CVPR (2025)
62. Wang, J., Chen, M., Zhang, S., Karaev, N., Schönberger, J., Labatut, P., Bojanowski, P., Novotny, D., Vedaldi, A., Rupprecht, C.: Vggt- ω . In: CVPR (2026)
63. Wang, J., Karaev, N., Rupprecht, C., Novotny, D.: Vggsfm: Visual geometry grounded deep structure from motion. In: CVPR (2024)
64. Wang, R., Xu, S., Dai, C., Xiang, J., Deng, Y., Tong, X., Yang, J.: Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In: CVPR (2025)
65. Wang, R., Xu, S., Dong, Y., Deng, Y., Xiang, J., Lv, Z., Sun, G., Tong, X., Yang, J.: Moge-2: Accurate monocular geometry with metric scale and sharp details. In: arXiv (2025)

66. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: π^3 : Permutation-equivariant visual geometry learning. arXiv preprint arXiv:2507.13347 (2025)
67. Wilson, K., Snavely, N.: Robust global translations with 1dsfm. In: ECCV (2014)
68. Wu, C.: Visualsfm: A visual structure from motion system. <http://www.cs.washington.edu/homes/ccwu/vsfm> (2011)
69. Wu, C., Agarwal, S., Curless, B., Seitz, S.M.: Multicore bundle adjustment. In: CVPR. IEEE (2011)
70. Wu, Y., Zhang, Y., Zhu, D., Deng, Z., Sun, W., Chen, X., Zhang, J.: An object slam framework for association, mapping, and high-level tasks. *IEEE Transactions on Robotics* **39**(4), 2912–2932 (2023)
71. Xu, K., Hao, Y., Yuan, S., Wang, C., Xie, L.: Airlam: An efficient and illumination-robust point-line visual slam system. *IEEE Transactions on Robotics* **41**, 1673–1692 (2025)
72. Xue, N., Wu, T., Bai, S., Wang, F., Xia, G.S., Zhang, L., Torr, P.H.: Holistically-attracted wireframe parsing. In: CVPR (2020)
73. Yang, S., Scherer, S.: Cubeslam: Monocular 3-d object slam. *IEEE Transactions on Robotics* **35**(4), 925–938 (2019)
74. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: Scannet++: A high-fidelity dataset of 3d indoor scenes. In: ICCV (2023)
75. Zach, C.: Robust bundle adjustment revisited. In: ECCV (2014)
76. Zhang, Y., Keetha, N., Lyu, C., Jhamb, B., Chen, Y., Qiu, Y., Karhade, J., Jha, S., Hu, Y., Ramanan, D., Scherer, S., Wang, W.: Ufm: A simple path towards unified dense correspondence with flow. In: arXiv (2025)
77. Zhao, X., Wu, X., Chen, W., Chen, P.C., Xu, Q., Li, Z.: Aliked: A lighter keypoint and descriptor extraction network via deformable transformation. *IEEE Transactions on Instrumentation and Measurement* **72**, 1–16 (2023)
78. Zhou, L., Liu, J., Zhai, F., Ai, P., Ren, K., Mao, Y., Huang, G., Meng, Z., Kaess, M.: Efficient bundle adjustment for coplanar points and lines. In: ICRA. IEEE (2023)
79. Zhou, L., Wang, S., Kaess, M.: A fast and accurate solution for pose estimation from 3d correspondences. In: ICRA (2020)
80. Zhou, Y., Qi, H., Zhai, Y., Sun, Q., Chen, Z., Wei, L.Y., Ma, Y.: Learning to reconstruct 3d manhattan wireframes from a single image. In: ICCV (2019)
81. Zhuang, B., Cheong, L.F., Lee, G.H.: Baseline desensitizing in translation averaging. In: CVPR (2018)
82. Zuo, X., Xie, X., Liu, Y., Huang, G.: Robust visual slam with point and line features. In: IROS (2017)