Revisiting Structure from Motion with 3D Reconstruction Priors

Daniel Korth, Matthias Nießner

Technical University of Munich

Abstract

Structure from Motion (SfM) has been a cornerstone of computer vision for decades, aiming to reconstruct 3D scene structure and camera parameters from 2D images. Traditionally, this problem is broken into several subproblems—image matching, triangulation, and global optimization—typically relying on 2D keypoints and reprojection error. Recently, 3D reconstruction models like DUSt3R have proven highly effective for a variety of 3D vision tasks. These models, based on feed-forward neural networks, regress dense 3D pointmaps from pairs of images in a shared coordinate system. In this work, we integrate modern 3D reconstruction priors into the incremental SfM pipeline. We propose a novel optimization method that aligns 3D pointmaps with scene structure and incorporates them as additional constraints in the optimization process. This allows us to enhance global optimization by combining both 2D keypoints and 3D pointmaps, resulting in improved robustness. We evaluate our approach on indoor scenes and demonstrate that it outperforms the baseline pipeline that relies solely on 2D constraints from reprojection error.

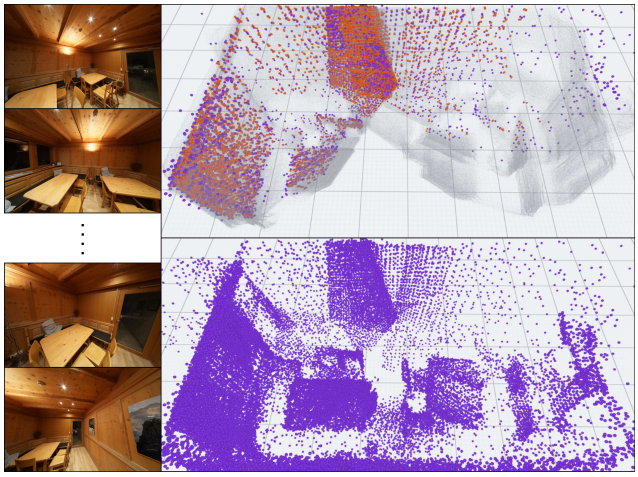

Figure 1. Left: Input images. **Top Right:** Partial reconstruction with aligned 3D pointmaps (gray) and their corresponding constraints (orange). **Bottom Right:** Final scene structure of our reconstruction.

1. Introduction

Reconstructing the 3D geometry from a set of images is a long-standing and fundamental problem in the computer vision community. The problem is usually solved using Structure from Motion (SfM), and can enable solutions for heritage preservation, robotics, augmented and virtual reality, and many other areas.

Traditional incremental SfM pipelines rely on detecting and matching local features [17] between images, estimating camera parameters, triangulating 3D points [12], and refining reconstruction through global optimization - most notably, optimizing for a reprojection error using Bundle Adjustment (BA) [27]. While effective, these pipelines are heavily handcrafted, often suffering from cumulative errors across stages and exhibiting sensitivity to challenging conditions such as low-texture regions or high outlier rates.

Recently, there has been a growing shift toward replacing traditional SfM components with deep learning-based alternatives. Early approaches focused on enhancing individual stages of the pipeline, such as feature detection [6], description [4], and image matching [20, 24]. Over time, researchers have progressively substituted more of the traditional pipeline, leading to end-to-end deep SfM methods [23, 28, 29]. As a final step, 3D foundation models go directly from images to dense scene structure using a large transformer trained on millions of images and corresponding 3D reconstructions. The seminal work in this field is DUSt3R [30], which regresses dense scene structure from image pairs in a common coordinate system. Through a global optimization in 3D space, they can be adapted to estimate camera poses. While promising, they entirely neglect the rich body of research in SfM and particularly the accuracy of keypoint-based methods. Instead of treating these networks as a final result, we instead view them as data-driven reconstruction prior and combine them with a traditional SfM pipeline relying on keypoint matches. This new perspective allows us to add 3D constraints into the

traditional SfM pipeline, which has so far been dominated solely by 2D constraints based on the reprojection error of detected keypoints.

Our focus lies on the global optimization stage, which is traditionally dominated by BA - an approach that minimizes reprojection error to jointly refine camera parameters and 3D structure. Although effective, BA is susceptible to outliers and ambiguities, especially under challenging conditions such as low triangulation angles.

To address these limitations, we augment the BA process with additional constraints based on pairwise pointmaps predicted by 3D reconstruction priors, such as DUSt3R [30]. These learned priors provide complementary 3D cues that can be aligned with the estimated scene structure, introducing additional geometric structure into the optimization. By incorporating them directly into the global optimization, our method effectively combines 2D and 3D constraints to refine our scene structure.

We show that leveraging these 3D constraints improves performance on sparse-view indoor scene reconstruction compared to a baseline method using BA.

In summary, our contributions are:

- We present a hybrid SfM framework that integrates learned 3D reconstruction priors into the traditional SfM pipeline, improving reconstruction quality.
- We introduce a novel global optimization strategy that augments the traditional minimization of 2D reprojection error with an additional optimization in 3D space, by minimizing the distance between the scene structure and the points estimated by our reconstruction prior.
- We validate our approach through extensive experiments on indoor scenes, demonstrating improved results in camera pose estimation.

2. Related Work

2.1. Traditional SfM

The traditional SfM pipeline begins with a set of unordered images. Keypoints are detected, descriptors are extracted, and matches are established across image pairs [4, 17, 20]. These two-view correspondences are geometrically verified by epipolar geometry, and feature tracks are formed by merging matching keypoints across images. There are two main paradigms in SfM: incremental [21] and global [19]. Both rely on the same feature matching process but differ in how they estimate camera parameters and 3D scene points.

In incremental SfM, reconstruction starts from an image pair with a strong baseline to ensure robust initialization. Frames overlapping with the scene structure are registered using Perspective-n-Point [9]. Newly visible scene structure is triangulated [12] and followed by a global optimization to refine both cameras and scene structure, traditionally via Bundle Adjustment [27]. Robust estimators such as RANSAC [8] are used throughout the pipeline to mitigate outliers. Outliers are further filtered based on triangulation angles or reprojection error. This process is repeated until the reconstruction is completed or no more frames can be registered.

Global SfM [19], in contrast, estimates all camera poses simultaneously - typically via rotation averaging [11] and translation estimation from pairwise epipolar geometry - followed by triangulation and global BA.

A key component in both paradigms is BA, a nonlinear optimization that jointly refines camera poses and 3D points to minimize reprojection error. It is widely considered the core of any SfM pipeline. Among existing systems, COLMAP [21] remains the gold standard for traditional SfM, known for its robustness, accuracy, and ease of use.

2.2. Modern SfM Variants

Traditional SfM pipelines are heavily handcrafted and rely on many small subproblems, each of which can introduce errors that propagate through the pipeline. Recent efforts have sought to partially or fully replace components of the traditional pipeline using modern deep learning techniques. For example, SuperPoint [4] or DeDoDe [6] learn to detect and describe keypoints using neural networks. SuperGlue [20] employs a graph neural network to match keypoints between two images. Furthermore, detector-free matching avoids sparse keypoints by leveraging self- and cross-attention mechanisms [24]. Recent state-of-the-art methods like PixSfM [16] use feature-metric keypoint adjustment and feature-metric BA. Detector-Free SfM [13] proposes to first build a coarse model using detector-free matches and then to refine the map and poses with keypoint matches. Other works aim to fully replace the SfM pipeline with neural counterparts or alternative formulations. VGGSfM [29] builds a fully differentiable SfM pipeline that is trained end-to-end. FlowMap [23] and ACE0 [1] train self-supervised depth and scene coordinate regressors as part of the reconstruction process.

While these approaches try to *replace* traditional components with learned counterparts, we instead focus on *extending* the traditional pipeline with additional constraints from learned priors.

2.3. 3D Reconstruction Priors

Wang et al. [30] recently introduced DUSt3R, a 3D foundation model trained on a large-scale dataset to regress pairwise pointmaps in a common coordinate system. It estimates camera poses through a global optimization tech-

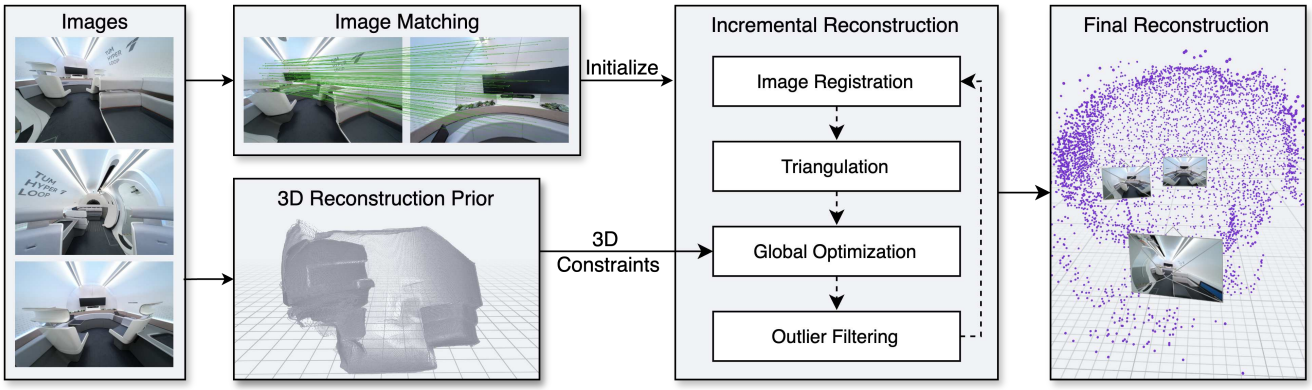

Figure 2. Method Overview. Starting with an unordered collection of images, we perform image matching to extract 2D keypoint constraints. Additionally, we extract 3D pointmaps from a reconstruction prior. We then perform incremental reconstruction, incorporating the 3D pointmaps as additional constraints into the global optimization. The final reconstruction includes both the scene structure and the camera parameters.

nique merging the pairwise pointmaps. The follow-up work, MASt3R [14], adapts the representation to regress pixel-wise feature vectors used for image matching on top. MASt3R-SfM [5] introduced an efficient way to leverage the output of DUSt3R and MASt3R to construct an end-to-end SfM pipeline. A common bottleneck is that these methods only work on pairs of input. Consequently, researchers have adapted the architecture to handle multiple images, all expressed in a common coordinate system [2, 26, 28]. Among them, VGGT [28] also adds additional regression heads for extrinsic and intrinsic parameters, depth maps, and 3D point tracks. These priors have been adapted, for example, to SLAM [18], Gaussian Splatting [7, 25] or dynamic scenes [10].

In contrast to these feed-forward 3D reconstruction systems, which reconstruct camera poses via optimization in 3D or directly regress them, our work explores how such 3D pointmaps can integrate into traditional SfM pipelines.

3. Preliminaries

Structure from Motion (SfM) aims to jointly estimate camera parameters (motion) and the 3D structure of a scene given a set of input images. Formally, given N RGB images denoted as $\mathcal{I} = \{I_i\}_{i=1}^{N}$, SfM recovers sparse scene structure $\mathcal{X} = \{x_k\}_{k=1}^{M}$ representing M 3D points $x_k \in \mathbb{R}^3$, and a set of camera parameters $\mathcal{H} = \{(K_i, T_i)\}_{i=1}^{N}$. Here, $K_i \in \mathbb{R}^{3\times3}$ is the intrinsic calibration matrix, and $T_i \in \mathbb{R}^{3\times4}$ encodes the extrinsic rotation and translation. A 3D point x_k is projected onto the i-th image plane using a camera projection function π, yielding 2D pixel coordinates $p_{i,k} = \pi(K_i, T_i, x_k) \in \mathbb{R}^2$.

Bundle Adjustment (BA) is a global optimization technique used to refine camera parameters and 3D structure by minimizing the reprojection error. Given 2D keypoint locations $y_{i,k} \in \mathbb{R}^2$ obtained from image feature matching, BA minimizes the following energy function:

$$E_{\text{BA}} = \sum_{i=1}^{N}\sum_{k=1}^{M} \|y_{i,k} - \pi(K_i, T_i, x_k)\|^2$$

To mitigate the effect of outliers, robust loss functions such as the Cauchy loss, defined as $f(x) = \log(1+x)$, are commonly applied to the residuals.

DUSt3R is a learned 3D reconstruction prior that takes an image pair (I_i, I_j) and produces dense, pixel-aligned pointmaps $\mathbb{X} \in \mathbb{R}^{W\times H\times3}$ along with associated confidence maps $\mathbb{C} \in \mathbb{R}^{W\times H\times1}$ in a common coordinate system. We define $\mathbb{X}^{i,j}$ as the 3D pointmaps from image I_i represented in the camera coordinate system of I_j.

4. Method

We propose a novel Point-to-Point (P2P) energy term incorporating 3D reconstruction priors from learned models such as DUSt3R into global optimization. We start by extracting pointmaps at keypoint locations determined from pairwise image matches. During reconstruction, within the global optimization step, we first align these pointmaps to the estimated scene structure and subsequently minimize a 3D Point-to-Point error alongside BA.

4.1. Pointmap Extraction

We start by matching images, resulting in $M_{i,j}$ corresponding keypoint location pairs $\{(y_{i,k}, y_{j,k})\}_{k=1}^{M_{i,j}}$ for images I_i and I_j, which are subsequently geometrically verified. We define the set of verified image pairs as $\mathcal{E} = \{(i,j) \mid 1 \le i < j \le N\}$.

Next, we forward I_i and I_j through our 3D reconstruction prior to obtain pointmaps $\mathbb{X}^{i,j}$ and $\mathbb{X}^{j,j}$. For clarity, we

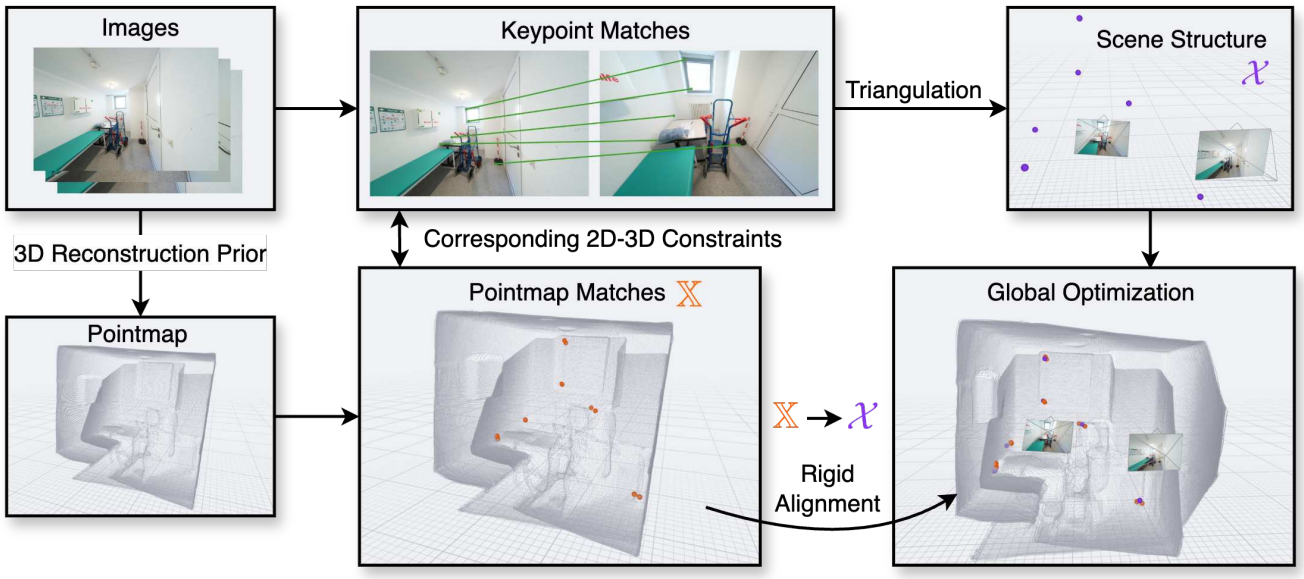

Figure 3. Visualization of our pointmap alignment process. Starting with an image collection, we match images to extract pairwise keypoint constraints. We pass all pairwise images through our 3D reconstruction prior to extract pointmaps. Using the provided keypoint matches (green), we extract the corresponding pointmap matches (orange). Next, we estimate a rigid alignment from the pointmap matches to their corresponding scene structure (purple). This serves as initialization to our global optimization. Note: Gray points are for visualization purposes only and are not used in the global optimization.

introduce the notation $\mathbb{X}^{i,e} := \mathbb{X}^{i,j}$ and $\mathbb{X}^{j,e} := \mathbb{X}^{j,j}$ for a given $e = (i, j) \in \mathcal{E}$, indicating that both pointmaps reside in the same coordinate system.

From these pointmaps, we sample the 3D locations at the matched pixel coordinates: $x_k^{i,e} = \mathbb{X}^{i,e}(y_{i,k})$ and $x_k^{j,e} = \mathbb{X}^{j,e}(y_{j,k})$, alongside their associated confidence scores $c_k^{i,e} = \mathbb{C}^{i,e}(y_{i,k})$ and $c_k^{j,e} = \mathbb{C}^{j,e}(y_{j,k})$.

Thus, for each geometrically verified pair e, we obtain a set of correspondences

$$\{(y_{i,k}, y_{j,k}, x_k^{i,e}, x_k^{j,e}, c_k^{i,e}, c_k^{j,e})\}_{k=1}^{M_e}$$

which constrain the scene structure. We then proceed with the classical SfM pipeline by initialization, image registration, triangulation, and further refinement.

4.2. Pointmap Alignment

Assume a partial reconstruction with scene structure $\mathcal{X}$. Every point $x_k \in \mathcal{X}$ can be observed in at least 2 input frames with corresponding keypoint locations, pointmaps and confidence maps have associated pointmap values $x_k^{i,e}$ and $x_k^{j,e}$ from the keypoint matches extracted above. Since pointmaps and scene structure live in different coordinate systems, we estimate a rigid transformation as follows:

$$T_e^*, s_e^* = \arg\min_{T_e, s_e} \sum_{k=1}^{M_e} \sum_{l \in \{i,j\}} \|x_k - s_e T_e(x_k^{l,e})\|^2$$

where T is parameterized by a rotation matrix and translation vector and M_e the numbers of scene points visible in both I_i and I_j. We solve the rigid alignment via Procrustes analysis [22] using RANSAC [8], as both x_k and $x_k^{i,j}$ can contain outliers. We repeat this process for every image pair observing at least 3 similar scene points yielding a set of all rigid transformations $\mathcal{T} = \{(T_e, s_e) \mid e \in \mathcal{E}\}$.

4.3. Point-to-Point Energy

We now minimize the distance between the scene points and pairwise point maps. While we can align the scene structure to the pointmaps, directly optimizing the 3D constraints would lead to a trivial solution. We, therefore, minimize with respect to the rigid transformations T_e, s_e of the pairwise pointmaps to the scene structure, which we estimated in the last step. We define the Point-to-Point energy term across all image pairs $\mathcal{E}$ as:

$$E_{\text{P2P}} = \sum_{e \in \mathcal{E}} \sum_{l \in \{i,j\}} \sum_{k=1}^{M_e} \|x_k - s_e T_e(x_k^{l,e})\|^2$$

where we optimize for scene structure x_k and pairwise rigid transformation parameters T_e. The scale s_e is held fixed to avoid degenerate solutions such as global shrinkage or collapse of the reconstruction. We apply this energy term only to inliers as estimated by the RANSAC alignment. Furthermore, we add the confidence maps $c_k^{i,e}$ to weight residuals based on the pointmap confidence values.

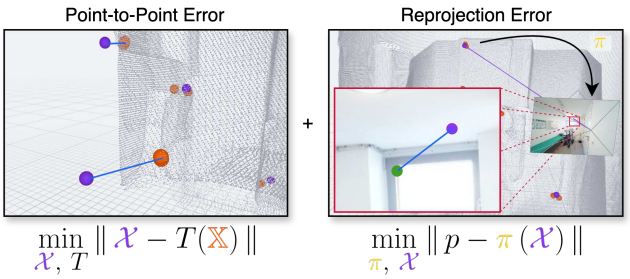

Figure 4. Visualization of our global optimization process. **Left:** Point-to-Point optimization, aligning scene structure $\mathcal{X}$ to pointmaps $\mathbb{X}$. T represents a rigid transformation applied to pairwise pointmaps. **Right:** Classical reprojection error, minimizing the distance in 2D screen space to the keypoint location.

4.4. Joint Energy Minimization

The Point-to-Point energy complements the standard reprojection loss in BA and provides additional geometric constraints derived from learned 3D reconstruction prior. We visualize the energies in Figure 4. Our final global optimization energy function then combines both terms to minimize

$$\mathcal{X}^*, \mathcal{H}^* = \arg\min_{\mathcal{X},\mathcal{H},\mathcal{T}} (E_{BA} + \beta E_{P2P})$$

Note that $\mathcal{T}$ serves only as an auxiliary optimization to ensure proper alignment of the pointmaps to constrain the scene structure accurately and will be thrown away after every global optimization in the incremental reconstruction.

4.5. Implementation Details

We implement a traditional incremental SfM pipeline around our above-mentioned method. We fix the gauge freedom by initializing the first camera with zero translation and identity rotation. Image registration is done using Perspective-n-Point [9] with RANSAC [8]. We perform multi-view triangulation using direct linear transformation [12]. We use a RANSAC inlier threshold of 0.1 for the pointmap alignment. To refine the structure and motion, we apply global optimization where we minimize $E_{BA} + \beta E_{P2P}$. We set $\beta = 0.01$ for our joint energy after rescaling the 3D error to align with the pixel error (i.e. scale the residuals to pixel distance). We perform global optimization with gradient descent and line search. After triangulation and global optimization, we filter scene structure based on triangulation angle and reprojection error. We remove scene points from the current visibility if they fail thresholds, but retriangulate them once additional images are registered. We use the intrinsics provided in our dataset and keep them fixed. We iterate until all images have been registered or no further images can be registered. Most of the code is written in PyTorch, and we occasionally use OpenCV.

5. Experiments

5.1. Experimental Setup

Baselines. Our baseline is the incremental pipeline described before, which we evaluate without incorporating the proposed Point-to-Point energy. We tune its hyperparameters - including all outlier filtering heuristics - and keep them fixed when introducing Point-to-Point energy, to ensure a fair and controlled comparison. Unless stated otherwise, we use MASt3R [14,15] with fast reciprocal matching for image matching and DUSt3R [30] with a DPT head and an input resolution of 512×336 for the 3D reconstruction prior. We limit pairwise matches to 256 to improve efficiency, with minimal performance loss. We benchmark our approach against modern 3D methods, including DUSt3R with their global optimization (DUSt3R+GO) in 3D [30], VGGT [28] - which directly regresses camera poses - and MASt3R-SfM [5].

Metrics. We evaluate reconstruction quality using three metrics. Following [5], we use: (1) Average Translation Error (ATE) — the mean Euclidean distance between ground-truth and estimated camera positions, computed after procrustes alignment [22] on normalized coordinates; and (2) Registration Rate — the percentage of successfully registered cameras. In addition, following [28], we report (3) AUC@30, which combines relative rotation accuracy and relative translation accuracy across image pairs at varying thresholds (up to 30 degrees).

Data. Our evaluation is based on ScanNet++ [31], a widely used dataset of indoor scenes. While ground-truth camera poses are not provided, we rely on COLMAP [21] for pose estimation, which is considered highly reliable when supplied with dense, overlapping image sequences. We sample scenes from the recent ScanNet++ v2 release, which none of the methods were trained on, though DUSt3R and MASt3R were trained on earlier ScanNet++ versions and VGGT on ScanNet [3]. For each scene, frames are randomly selected for evaluation.

5.2. Experimental Results

We evaluate overall reconstruction quality for frame counts of 15, 20, and 25 on 30 randomly sampled scenes (Table 1). Our proposed Point-to-Point energy formulation matches the Baseline at 15 images and shows improved performance at 20 and 25 images, highlighting its effectiveness as a complementary optimization strategy. We also observe that our Baseline achieves strong results compared to prior approaches, including MASt3R-SfM [5] and VGGT [28], which we attribute to the use of exhaustive pairwise feature matching. In contrast, MASt3R-SfM omits exhaustive matching to maintain scalability, VGGT regresses poses without post-optimization, and DUSt3R+GO only optimizes in 3D space, limiting its accuracy.

Method	15 Images			20 Images			25 Images		
	ATE ↓	AUC@30 ↑	Reg. ↑	ATE ↓	AUC@30 ↑	Reg. ↑	ATE ↓	AUC@30 ↑	Reg. ↑
Baseline	**0.0181**	82.4	97.1	0.0117	86.6	98.0	0.0107	86.7	99.3
Baseline+Ours	0.0190	**83.5**	96.9	**0.0090**	**88.3**	98.7	**0.0074**	**90.8**	98.6
DUSt3R+GO	0.0234	80.8	**100**	0.0147	84.7	**100**	0.0134	85.2	**100**
VGGT*	0.0240	69.9	**100**	0.0192	71.4	**100**	0.0179	71.5	**100**
MASt3R-SfM	0.0211	76.3	**100**	0.0133	78.8	**100**	0.0118	78.8	**100**

Table 1. Camera pose estimation on ScanNet++ [31] with varying view counts (15, 20, 25). ATE (↓), AUC@30 (↑), and registration rate (↑). Metrics averaged over 30 scenes. *Feed-forward pose regression without further optimization.

5.3. Ablations

We perform ablation studies on the design choices of the Point-to-Point energy, on the 2D keypoint constraints, and on the 3D reconstruction priors underlying our method. All ablations are performed on 10 scenes with 15 randomly sampled frames.

Point-to-Point Energy (Table 2). We observe that naively incorporating the pointmaps as constraints in the energy formulation (+P2P) leads to significant instability due to the presence of outliers. This is reflected in the strong degradation across all reported metrics when applying the Point-to-Point energy directly. Particularly telling is the drop in the number of scene points (**#Pts**), reported as an additional metric. It suggests that the energy function moves scene points too far from their corresponding keypoint observations, leading to them being filtered after optimization. We discuss this further in the limitations section 5.4. To mitigate the effect of outliers, we propose selectively applying the energy only to inlier correspondences via RANSAC alignment (+*Inliers only*). This significantly improves performance and even surpasses the baseline. Furthermore, incorporating confidence maps estimates from the 3D reconstruction prior (+*Conf. Weight*) boosts performance. Accounting for this uncertainty allows the optimization to focus on more reliable correspondences, leading to improved reconstruction results.

Image Matching (Table 3). Even if the 3D reconstruction prior is perfect, we still need good triangulated image matches to effectively align and constrain the scene structure. We therefore evaluate our method with different image matching methods, both traditional and neural, using SIFT [17] and MASt3R [14]. Matches for SIFT are selected using nearest neighbor with Lowe ratio [17]. MASt3R matches are computed using fast reciprocal matching [14]. We find that MASt3R matches perform better than SIFT, which is to be expected, as they are more reliable and dense. Furthermore, adding our Point-to-Point energy consistently outperforms upon BA Baseline in both scenarios. However, the relative improvement for SIFT is smaller. This can be attributed to SIFT matches being computed at a higher resolution (1752×1168), whereas the pointmaps are limited to 512×336, making them less accurate.

Method	ATE ↓	AUC@30 ↑	Reg. ↑	#Pts ↑
Baseline	0.0159	80.6	95.3	1204
+P2P	0.0736	54.0	74.0	795
+*Inliers only*	0.0166	82.6	94.0	1224
+*Conf. Weight*	**0.0138**	**84.9**	**98.0**	**1260**

Table 2. Ablation study on design choices for our energy formulation. Metrics are averaged over 15 images from 10 different scenes in ScanNet++ [31].

Matches	Method	ATE ↓	AUC@30 ↑	Reg. ↑
SIFT+NN	Baseline	0.0243	73.3	**64.0**
	+Ours	**0.0228**	**73.8**	**64.0**
MASt3R	Baseline	0.0159	80.6	95.3
	+Ours	**0.0138**	**84.9**	**98.0**

Table 3. Ablation study on different image matching methods (2D constraints). NN stands for nearest neighbor, MASt3R matches are computed using fast reciprocal matching [14]. Metrics are averaged over 10 ScanNet++ [31] scenes, each with 15 images.

Pointmaps (Table 4). Lastly, we ablate the different 3D reconstruction priors. Here, we compare the seminal work DUSt3R [30] with the current state-of-the-art VGGT [28]. VGGT can work with two or more images in the same coordinate system. We experiment with both extremes - pairwise extraction and full multi-view pointmaps. We find that all methods can outperform the BA baseline. It is however

3D Reconstruction Prior	ATE ↓	AUC@30 ↑	Reg. ↑
Baseline (No Prior)	0.0159	80.6	95.3
DUSt3R	0.0138	**84.9**	**98.0**
VGGT	0.0137	82.61	96.7
VGGT-MV	**0.0110**	84.06	97.3

Table 4. Ablation study on different 3D reconstruction priors. VGGT-MV extracts multi-view pointmaps instead of pairwise ones. Metrics are averaged over 10 ScanNet++ [31] scenes, each with 15 images.

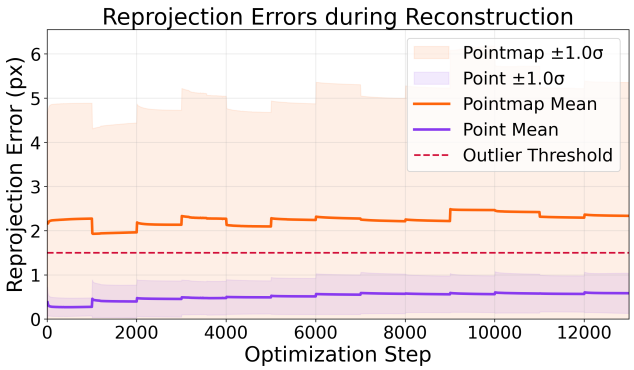

Figure 5. Visualization of the mean reprojection error and its standard deviation (std) for both the scene structure (Point) and the corresponding 3D pointmaps throughout the optimization process. Jumps indicate the addition of a new camera. Pointmaps are above the outlier threshold, so the energy terms must be properly weighted to avoid filtering out too much scene structure and destabilizing the reconstruction.

unclear which 3D reconstruction prior works best, and further investigation is necessary to determine the effect of different priors.

5.4. Limitations & Future Work

While adding 3D constraints in the global optimization can lead to better results, there is still some fundamental problem in the formulation of our energy. Since the rest of the pipeline - i.e., image registration, triangulation, and outlier filtering - solely relies on 2D keypoints, we cannot harness the full power of the 3D constraints. Assume perfect 3D constraints, i.e. all scene points should be where the 3D pointmaps are and we fully optimize for that. This would heavily destabilize the pipeline, as these points don't agree with our 2D constraints. We visualize this hypothetical scenario and the resulting reprojection error in Figure 5. Moving the scene structure to the pointmaps would cause most points to be identified as outliers and filtered from the reconstruction based on the reprojection error. Therefore, we hope to incorporate these 3D constraints into other parts of the pipeline to relax strict outlier filtering or to allow refinement of 2D keypoint locations by leveraging the 3D constraints (as e.g. in PixSfM [16], where they refine the 2D keypoint locations using feature-metric optimization).
Furthermore, the current pipeline is very expensive, performing matching and extracting pointmaps for all pairs of images, which limits scalability. Approaches like MASt3R-SfM [5] form a sparse scene graph to perform pairwise matching on, which greatly improves scalability. Furthermore, instead of computing pairwise pointmaps, we can compute for all connected components in the sparse scene graph using a multi-view 3D reconstruction prior like VGGT [28]. The VGGT-MV results in Table 4 indicate promising results for this avenue of future work.
Lastly, one can improve the alignment of pointmaps to the scene structure. Especially when moving to larger and/or outdoor scenes with bigger foreground/background differences, current 3D reconstruction priors seem to struggle and are often only coherent within a semantic group, but not between them [25]. Therefore, exploring semantic alignment or non-rigid alignment techniques, such as in SPARS3R [25] appear promising.

6. Conclusion

We revisit the traditional Structure from Motion pipeline, which relies on keypoints and reprojection error, and integrate modern 3D reconstruction priors into the global optimization, enabling direct optimization in 3D space. We propose a novel energy formulation that minimizes the distance between 3D scene structure and pointmaps from our prior, alongside the traditional reprojection error. Our approach enhances the global optimization with 3D constraints, yielding improved results for sparse-view reconstructions of indoor scenes. However, further research is needed to improve scalability for larger image collections and better integrate pointmaps into other components of the pipeline.

References

[1] Eric Brachmann, Jamie Wynn, Shuai Chen, Tommaso Cavallari, Áron Monszpart, Daniyar Turmukhambetov, and Victor Adrian Prisacariu. Scene coordinate reconstruction: Posing of image collections via incremental learning of a relocalizer. In *ECCV*, 2024. 2

[2] Yohann Cabon, Lucas Stoffl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jerome Revaud, and Vincent Leroy. Must3r: Multi-view network for stereo 3d reconstruction, 2025. 3

[3] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 5

[4] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection

and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 1, 2

[5] Bardienus Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jerome Revaud. Mast3r-sfm: a fully-integrated solution for unconstrained structure-from-motion, 2024. 3, 5, 7

[6] Johan Edstedt, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. DeDoDe: Detect, Don't Describe — Describe, Don't Detect for Local Feature Matching. In *2024 International Conference on 3D Vision (3DV)*. IEEE, 2024. 1, 2

[7] Zhiwen Fan, Kairun Wen, Wenyan Cong, Kevin Wang, Jian Zhang, Xinghao Ding, Danfei Xu, Boris Ivanovic, Marco Pavone, Georgios Pavlakos, Zhangyang Wang, and Yue Wang. Instantsplat: Sparse-view gaussian splatting in seconds, 2024. 3

[8] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 2, 4, 5

[9] Xiao-Shan Gao, Xiao-Rong Hou, Jianliang Tang, and Hang-Fei Cheng. Complete solution classification for the perspective-three-point problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(8):930–943, 2003. 2, 5

[10] Jisang Han, Honggyu An, Jaewoo Jung, Takuya Narihira, Junyoung Seo, Kazumi Fukuda, Chaehyun Kim, Sunghwan Hong, Yuki Mitsufuji, and Seungryong Kim. D^2ust3r: Enhancing 3d reconstruction with 4d pointmaps for dynamic scenes. *arXiv preprint arXiv:2504.06264*, 2025. 3

[11] Richard Hartley, Jochen Trumpf, Yuchao Dai, and Hongdong Li. Rotation averaging. *International journal of computer vision*, 103:267–305, 2013. 2

[12] Richard I Hartley and Peter Sturm. Triangulation. *Computer vision and image understanding*, 68(2):146–157, 1997. 1, 2, 5

[13] Xingyi He, Jiaming Sun, Yifan Wang, Sida Peng, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Detector-free structure from motion. *CVPR*, 2024. 2

[14] Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r, 2024. 3, 5, 6

[15] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. *arXiv preprint arXiv:2406.09756*, 2024. 5

[16] Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, and Marc Pollefeys. Pixel-perfect structure-from-motion with featuremetric refinement. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5987–5997, 2021. 2, 7

[17] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999. 1, 2, 6

[18] Riku Murai, Eric Dexheimer, and Andrew J. Davison. MASt3R-SLAM: Real-time dense SLAM with 3D reconstruction priors. *arXiv preprint*, 2024. 3

[19] Linfei Pan, Daniel Barath, Marc Pollefeys, and Johannes Lutz Schönberger. Global Structure-from-Motion Revisited. In *European Conference on Computer Vision (ECCV)*, 2024. 2

[20] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 1, 2

[21] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 5

[22] Peter H. Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966. 4, 5

[23] Cameron Smith, David Charatan, Ayush Tewari, and Vincent Sitzmann. Flowmap: High-quality camera poses, intrinsics, and depth via gradient descent. In *3DV*, 2025. 1, 2

[24] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. *CVPR*, 2021. 1, 2

[25] Yutao Tang, Yuxiang Guo, Deming Li, and Cheng Peng. Spars3r: Semantic prior alignment and regularization for sparse 3d reconstruction. *arXiv preprint arXiv:2411.12592*, 2024. 3, 7

[26] Zhenggang Tang, Yuchen Fan, Dilin Wang, Hongyu Xu, Rakesh Ranjan, Alexander Schwing, and Zhicheng Yan. Mv-dust3r+: Single-stage scene reconstruction from sparse views in 2 seconds. *arXiv preprint arXiv:2412.06974*, 2024. 3

[27] Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment—a modern synthesis. In *Vision Algorithms: Theory and Practice: International Workshop on Vision Algorithms Corfu, Greece, September 21–22, 1999 Proceedings*, pages 298–372. Springer, 2000. 1, 2

[28] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 1, 3, 5, 6, 7

[29] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Vggsfm: Visual geometry grounded deep structure from motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21686–21697, 2024. 1, 2

[30] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20697–20709, June 2024. 1, 2, 5, 6

[31] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023. 5, 6, 7