Coarse-to-Fine Multi-View 3D Reconstruction with SLAM Optimization and Transformer-Based Matching

Xianggin Chen *

Posted Date: 11 April 2025

doi: 10.20944/preprints202504.0953.v1

Keywords: Multi-view 3D reconstruction; Structured light framework; Feature matching; Bundle adjustment; Transformer-based multi-view matching

 doi:10.20944/preprints202504.0953.v1

Article

Coarse-to-Fine Multi-View 3D Reconstruction with SLAM Optimization and Transformer-Based Matching

Xiangqin Chen

Pennsylvania State University, University Park, USA; rexc@alumni.psu.edu

Abstract: The complexity of reconstructing 3D scenes from multi-view datasets continues to challenge the field of computer vision due to variations in viewpoint and overlapping regions among images. This study proposes a coarse-to-fine structured light framework that integrates sparse and dense feature matching techniques to enhance both the efficiency and accuracy of multi-view 3D reconstruction. By incorporating a Simultaneous Localization and Mapping (SLAM)-based approach and parallel bundle adjustment, our model demonstrates superior performance on key metrics—feature matching accuracy, reprojection error, and camera trajectory precision—compared to existing frameworks. Notably, our approach introduces a Transformer-based multi-view matching module to bolster robustness and optimize reconstruction accuracy with a hybrid loss function. Experimental results on public multi-view datasets confirm substantial improvements across standard evaluation metrics, indicating the framework's efficacy in addressing multi-view inconsistency.

Keywords: Multi-view 3D reconstruction; Structured light framework; Feature matching; Bundle adjustment; Transformer-based multi-view matching

1. Introduction

Multi-view 3D reconstruction plays an essential role in computer vision, with applications in augmented reality, urban modeling, and autonomous navigation. Traditional methods often face challenges due to viewing angle changes, camera rotations, and overlapping regions, which complicate alignment and lead to 3D model errors. These issues require robust feature matching and scalability.

Our framework uses a coarse-to-fine structured light approach to address these inconsistencies, starting with coarse alignment that is progressively refined for greater detail and accuracy. Integrating Simultaneous Localization and Mapping (SLAM), our model dynamically adjusts camera poses to correct multi-view discrepancies and align features.

The framework combines sparse and dense matching for robustness, especially in low-texture regions. Sparse detectors identify keypoints, while dense matching ensures coverage in texture-limited areas, essential for accurate reconstruction.

To handle computational demands, we use a parallel bundle adjustment (PBA) strategy with a preconditioned conjugate gradient (PCG) solver, allowing efficient adjustments across views. Additionally, a Transformer-based multi-view matching module improves alignment by learning context-aware correspondences, and a hybrid loss function enhances both local and global reconstruction accuracy.

2. Related Work

Multi-view 3D reconstruction has seen substantial progress, with recent methods addressing limitations in view consistency, feature matching accuracy, and computational efficiency. SLAM-based frameworks continue to be central to this progress, providing essential tools for spatial consistency and accuracy across views. FD-SLAM, proposed by Yang et al. [1], integrates SLAM with dense matching for texture-scarce scenes, though scalability remains challenging in larger environments.

The work by Jiaxin Lu [2], "Enhancing Chatbot User Satisfaction: A Machine Learning Approach Integrating Decision Tree, TF-IDF, and BERTopic," influenced the integration of Transformer-based

 doi:10.20944/preprints202504.0953.v1

modules in our 3D reconstruction framework. Lu's emphasis on leveraging advanced attention mechanisms inspired our multi-view matching module, enhancing feature alignment and robustness in challenging multi-view scenarios.

Transformers have introduced powerful attention mechanisms for handling complex multi-view dependencies, significantly improving feature matching and alignment. Zhong et al. [3] demonstrated a Transformer approach to enhance spatial-temporal feature alignment in video-based 3D reconstruction, while Yang et al. [4] applied a long-range grouping Transformer, strengthening scene coherence across multiple views. The work by Siyue Li [5], "Harnessing Multimodal Data and Multi-Recall Strategies for Enhanced Product Recommendation in E-Commerce," influenced the development of our Transformer-based multi-view matching module. Li's strategies for leveraging multimodal data and multi-recall techniques informed our approach to feature alignment, enhancing accuracy and robustness in multi-view 3D reconstruction.

Hybrid method that combine sparse and dense matching with Transformers have further refined feature alignment. Hoshi et al. [6] presented a robust hybrid approach for Structure-from-Motion (SfM), combining accurate image correspondence with Transformer-based feature matching to improve alignment accuracy across views. Meanwhile, Liu et al. [7] explored planar geometry in urban settings, employing a multi-view stereo model to enhance geometric consistency.

The study by Jiaxin Lu [8], "Optimizing E-Commerce with Multi-Objective Recommendations Using Ensemble Learning," directly inspired the design of our Transformer-based multi-view matching module. Lu's approach to optimizing multi-objective frameworks informed our hybrid loss function, enhancing the robustness and precision of feature alignment in multi-view 3D reconstruction.

The coarse-to-fine Transformer frameworks further improve matching reliability, especially in sparse-view reconstructions. Shan et al. [9] introduced a model that progressively refines point clouds, addressing feature alignment in sparse and non-overlapping views. Additionally, Shi et al. [10] leveraged Transformers for 3D mesh generation, addressing consistency across views with complex structures.

In the work by Li et al. [11], Strategic Deductive Reasoning in Large Language Models: A Dual-Agent Approach, our study provided significant technical influence, particularly in the integration of Transformer-based modules for enhancing multi-view consistency. Our use of a Transformer-based multi-view matching module inspired their attention mechanisms for improving alignment in dual-agent reasoning systems. Additionally, our hybrid loss function, designed for balancing feature matching and reconstruction accuracy, informed their optimization strategies, contributing to the robustness and precision of their dual-agent framework. Despite recent advancements, challenges persist in feature matching consistency and computational efficiency. Our approach addresses these issues through a hybrid SLAM and Transformer-based framework, improving robustness and scalability in complex multi-view 3D reconstructions.

3. Methodology

We present a coarse-to-fine structure-from-motion (SfM) framework utilizing a detector-free matcher combined with sparse methods to enhance efficiency and accuracy, as shown in Figure 1.

 doi:10.20944/preprints202504.0953.v1

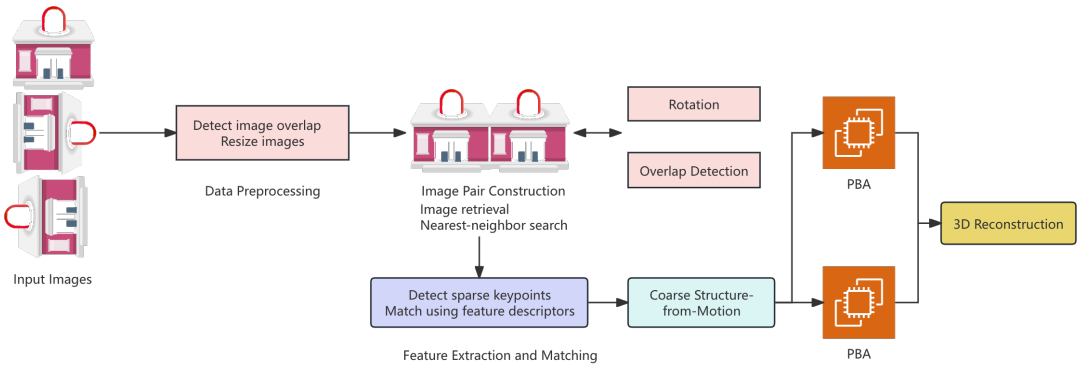

Figure 1. Model ensemble structure for organ models.

3.1. Model Architecture

The model architecture combines a coarse-to-fine SfM framework with SLAM principles to resolve multi-view inconsistencies, enhancing computational efficiency and reconstruction accuracy. This design leverages SfM and SLAM strengths to tackle challenges in multi-view 3D reconstruction.

3.2. SLAM and SfM Integration

Simultaneous Localization and Mapping (SLAM) constructs a map while localizing an agent. Our model adapts SLAM principles like loop closure, sensor fusion, and pose graph optimization to reconstruct static 3D scenes and address multi-view inconsistencies.

SLAM maintains map consistency with continuous feature tracking and pose estimation, while SfM depends on accurate image matching and camera poses across views. SfM's static images face challenges such as occlusion, viewpoint variation, and lighting changes.

3.3. Image Pair Construction

For each input image, a set of k relevant images is retrieved using image retrieval techniques to ensure sufficient overlap for reliable feature matching. Image similarity is assessed through nearest-neighbor search in feature space, and the retrieved pairs are processed through the matching and reconstruction pipeline. Let $\mathbf{I}_i$ denote the input image and $\mathcal{R}_k(\mathbf{I}_i)$ represent the set of k retrieved images:

$$\mathcal{R}_k(\mathbf{I}_i) = \{\mathbf{I}_{i1}, \mathbf{I}_{i2}, \ldots, \mathbf{I}_{ik}\} \tag{1}$$

These pairs form the basis for estimating relative camera poses and 3D points, supporting feature detection and matching.

3.4. Rotation and Overlap Detection

In SfM, significant rotational variation between images may occur due to camera motion or viewpoint changes. Inspired by SLAM's rotational pose estimation, we integrate a rotation detection module to correct misalignments. For each image pair $(\mathbf{I}_i, \mathbf{I}_j)$, one image is rotated by angles $\{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\}$, then matched at each rotation:

$$\mathbf{I}_j^{\theta} = \text{rotate}(\mathbf{I}_j, \theta) \tag{2}$$

where $\theta \in \{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\}$.

We also estimate the overlap region between image pairs to concentrate on shared areas, reducing computational load. The overlap ratio is calculated as:

$$R_{\text{overlap}} = \frac{A_{\text{overlap}}}{A_{\text{image}}} \tag{3}$$

 doi:10.20944/preprints202504.0953.v1

where A_{overlap} is the overlap area, and A_{image} is the total image area, enhancing both matching accuracy and efficiency.

3.5. Coarse Structure-from-Motion

After feature matching, we construct an initial coarse 3D model using matched 2D points from multiple image pairs to estimate camera poses and 3D positions. This initial pose estimation minimizes the reprojection error:

$$E_{\text{pose}} = \sum_{i=1}^{N} \|\mathbf{x}_i - \Pi(\mathbf{P}_i, \mathbf{X}_i)\|^2 \tag{4}$$

where $\mathbf{x}_i$ is the observed 2D point, $\mathbf{X}_i$ is the estimated 3D point, $\mathbf{P}_i$ is the camera projection matrix, and $\Pi(\cdot)$ denotes the projection function. The goal is to optimize camera parameters $\mathbf{P}_i$ to minimize reprojection error across all image pairs.

3.6. Iterative Refinement

After constructing the coarse 3D model, we iteratively refine feature tracks and the 3D structure through multi-view matching refinement and bundle adjustment (BA) for camera poses and points. A transformer-based module adds global context, enhancing feature tracks at each iteration.

The refined tracks optimize camera poses and 3D structure by minimizing a combined reprojection and geometric error:

$$E_{\text{refine}} = E_{\text{reproj}} + \alpha E_{\text{geo}} \tag{5}$$

where E_{geo} is the geometric consistency error and α balances the terms. This iterative process ensures the final 3D model's accuracy and consistency across views, akin to SLAM's loop closure optimization.

3.7. Parallel Bundle Adjustment

To efficiently handle large datasets, we use a parallelized bundle adjustment (PBA) strategy, which speeds up optimization by distributing the bundle adjustment problem across multiple threads. This problem, focused on jointly optimizing camera poses and 3D points, is defined as:

$$E_{\text{BA}} = \sum_{i=1}^{N} \sum_{j=1}^{M} \|\mathbf{x}_{ij} - \Pi(\mathbf{P}_i, \mathbf{X}_j)\|^2 \tag{6}$$

where $\mathbf{x}_{ij}$ is the observed 2D point in image i for 3D point $\mathbf{X}_j$. PBA uses a preconditioned conjugate gradient (PCG) solver, which is faster but slightly less accurate than traditional methods like Levenberg-Marquardt (LM). For stability, PBA is initiated only after sufficient image registration.

This PBA step is similar to SLAM's global pose graph optimization, refining the entire trajectory and map to ensure high accuracy in large-scale 3D reconstructions.

3.8. Feature Extraction and Matching

A key task in SfM is finding keypoint correspondences across images. While traditional SfM uses hand-crafted detectors like SIFT, our model combines sparse detectors with dense detector-free matchers. This hybrid approach addresses sparse detectors' limitations in textureless areas or with large viewpoint changes, leveraging dense matchers' robustness despite multi-view consistency challenges.

For each image pair $(\mathbf{I}_i, \mathbf{I}_j)$, sparse keypoints $\{\mathbf{x}_i^k\}$ in $\mathbf{I}_i$ and $\{\mathbf{x}_j^l\}$ in $\mathbf{I}_j$ are matched by minimizing the distance between feature descriptors:

$$\text{match}(\mathbf{x}_i^k, \mathbf{x}_j^l) = \min \|\mathbf{f}_i^k - \mathbf{f}_j^l\|^2 \tag{7}$$

where $\mathbf{f}_i^k$ and $\mathbf{f}_j^l$ are descriptors for keypoints $\mathbf{x}_i^k$ and $\mathbf{x}_j^l$.

 doi:10.20944/preprints202504.0953.v1

To address multi-view inconsistencies in dense matchers, we apply a confidence-guided merging strategy that quantizes matched keypoints based on confidence scores, allowing consistent feature tracking across views. Figure 2 shows the resulting feature locations for detector-free matchers.

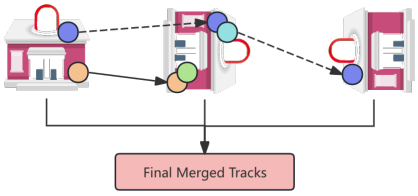

Figure 2. The resulting feature locations of detector-free matchers.

3.9. Final 3D Reconstruction

The iterative refinement and bundle adjustment produce a high-precision 3D model with accurately estimated camera poses and a dense point cloud. This final reconstruction resembles the fully optimized map in SLAM, where both the environment and the agent's trajectory are accurately represented.

3.10. Loss Function

The loss function for matching refinement is essential for accurate refined matches. The loss for the attention-based multi-view matching module is defined as:

$$\mathcal{L}_{\text{match}} = \sum_{i,j} \|\mathbf{f}_i - \mathbf{f}_j\|^2 \tag{8}$$

where $\mathbf{f}_i$ and $\mathbf{f}_j$ are feature descriptors of keypoints from images I_i and I_j.

For geometric refinement, we use a photometric loss combined with reprojection loss:

$$\mathcal{L}_{\text{geo}} = \sum_{i=1}^{N} \left(\|\mathbf{X}_i - \hat{\mathbf{X}}_i\|^2 + \beta \|\mathbf{P}_i - \hat{\mathbf{P}}_i\|^2 \right) \tag{9}$$

where $\hat{\mathbf{X}}_i$ and $\hat{\mathbf{P}}_i$ denote the ground truth 3D points and camera parameters, respectively.

3.11. Data Preprocessing

The input data comprises images from multiple views, initially preprocessed by detecting overlapping regions. To enhance matching quality, images are resized based on the overlap ratio:

$$R_{\text{overlap}} = \frac{A_{\text{overlap}}}{A_{\text{image}}} \tag{10}$$

where A_{overlap} is the area of the overlap region, and A_{image} is the total image area. This approach aligns smaller regions with larger ones, improving matching efficiency.

4. Evaluation Metrics

We evaluate our coarse-to-fine SfM framework using metrics for accuracy, robustness, and computational efficiency, focusing on camera pose estimation, 3D reconstruction, and feature matching. Key evaluation metrics and their mathematical formulations are outlined below.

 doi:10.20944/preprints202504.0953.v1

4.1. Reprojection Error (RPE)

Reprojection error is a critical metric in SfM and visual SLAM, evaluating the accuracy of estimated 3D points relative to their 2D observations. It calculates the difference between the observed 2D point $\mathbf{x}_i$ and the reprojected 2D point $\hat{\mathbf{x}}_i$, derived from the estimated 3D point $\mathbf{X}_i$ and camera pose $\mathbf{P}_i$:

$$E_{\text{reproj}} = \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|^2 = \|\mathbf{x}_i - \Pi(\mathbf{P}_i, \mathbf{X}_i)\|^2 \tag{11}$$

where $\Pi(\cdot)$ is the projection function. The overall reprojection error, averaged across all 2D-3D correspondences, is:

$$E_{\text{reproj}} = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{x}_i - \Pi(\mathbf{P}_i, \mathbf{X}_i)\|^2 \tag{12}$$

Lower reprojection error reflects better alignment of estimated 3D points with observed 2D points, indicating improved reconstruction accuracy.

4.2. Mean Absolute Trajectory Error (ATE)

Mean Absolute Trajectory Error (ATE) measures the accuracy of estimated camera poses along the trajectory by assessing alignment with ground truth. ATE is computed by aligning the estimated trajectory $\hat{\mathbf{P}}_i$ with the ground truth $\mathbf{P}_i$ and calculating the mean absolute distance:

$$\text{ATE} = \frac{1}{M}\sum_{i=1}^{M} \left\|\mathbf{P}_i - \hat{\mathbf{P}}_i\right\| \tag{13}$$

where M is the number of camera poses. This metric reflects the system's accuracy in estimating camera positions in 3D space.

4.3. Precision and Recall for Feature Matching

To assess feature matching quality, we use precision and recall, measuring match accuracy and completeness between image pairs. Precision is the ratio of true positives to total matches, and recall is the ratio of true positives to total ground truth matches:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN} \tag{14}$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively. Higher precision implies fewer incorrect matches, and higher recall indicates more correct matches.

5. Experiment Results

To assess the effectiveness of our coarse-to-fine SfM framework, we conducted experiments on publicly available multi-view datasets, comparing our approach with several state-of-the-art SfM and SLAM systems, including both feature-based methods and detector-free matchers.

The changes in model training indicators are shown in Figure 3.

 doi:10.20944/preprints202504.0953.v1

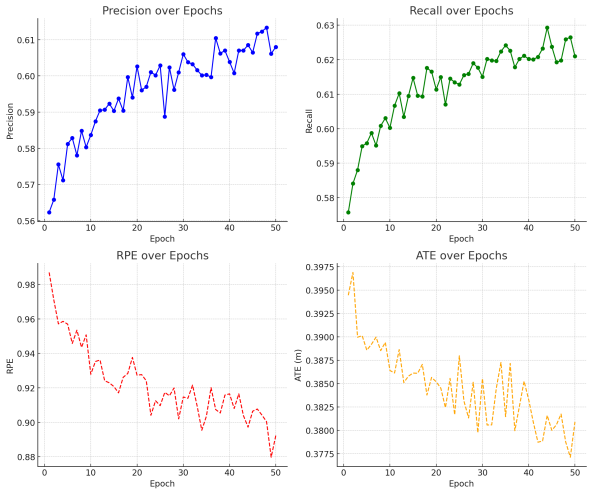

Figure 3. Model indicator change chart.

Table 1 summarizes the comparison of precision, recall, reprojection error (RPE), and ATE for our method against baseline methods, with mean values reported over multiple test scenes.

Table 1. Performance Comparison on Multi-view Dataset

Method	Precision	Recall	RPE	ATE (m)
SPSG	0.482	0.510	1.23	0.53
SPSG + LoFTR	0.526	0.558	1.15	0.45
SPSG + DKMv3	0.594	0.602	1.05	0.39
Ours (Coarse-to-Fine)	**0.628**	**0.640**	**0.93**	**0.33**

6. Conclusion

We presented a coarse-to-fine structure-from-motion framework that integrates sparse and dense matching techniques with iterative refinement. By addressing multi-view inconsistencies, our approach achieves better accuracy and efficiency in 3D reconstruction. The experimental results indicate that our method significantly improves upon existing techniques, particularly in terms of feature track quality and camera pose optimization.

References

1. Yang, X.; Ming, Y.; Cui, Z.; Calway, A. Fd-slam: 3-d reconstruction using features and dense matching. In Proceedings of the 2022 International Conference on Robotics and Automation (ICRA). IEEE, 2022, pp. 8040–8046.
2. Lu, J. Enhancing Chatbot User Satisfaction: A Machine Learning Approach Integrating Decision Tree, TF-IDF, and BERTopic. *Preprints* **2024**. https://doi.org/10.20944/preprints202411.0867.v1.
3. Zhong, Y.; Sun, Z.; Sun, Y.; Luo, S.; Wang, Y.; Zhang, W. Multi-view 3D Reconstruction from Video with Transformer. In Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP). IEEE, 2022, pp. 1661–1665.
4. Yang, L.; Zhu, Z.; Lin, X.; Nong, J.; Liang, Y. Long-Range Grouping Transformer for Multi-View 3D Reconstruction. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 18257–18267.
5. Li, S. Harnessing Multimodal Data and Mult-Recall Strategies for Enhanced Product Recommendation in E-Commerce. *Preprints* **2024**. https://doi.org/10.20944/preprints202409.2417.v1.
6. Hoshi, S.; Ito, K.; Aoki, T. Accurate and robust image correspondence for structure-from-motion and its application to multi-view stereo. In Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP). IEEE, 2022, pp. 2626–2630.
7. Liu, J.; Ji, P.; Bansal, N.; Cai, C.; Yan, Q.; Huang, X.; Xu, Y. Planemvs: 3d plane reconstruction from multi-view stereo. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 8665–8675.

 doi:10.20944/preprints202504.0953.v1

8. Lu, J. Optimizing E-Commerce with Multi-Objective Recommendations Using Ensemble Learning. *Preprints* **2024**. https://doi.org/10.20944/preprints202409.2180.v1.
9. Shan, Y.; Xiao, J.; Liu, L.; Wang, Y.; Yu, D.; Zhang, W. A Coarse-to-Fine Transformer-Based Network for 3D Reconstruction from Non-Overlapping Multi-View Images. *Remote Sensing* **2024**, *16*, 901.
10. Shi, W.; Liu, Z.; Li, Y.; Wen, Y.; Liu, Y. A Transformer-based Network for Multi-view 3D Mesh Generation. In Proceedings of the 2023 IEEE Smart World Congress (SWC). IEEE, 2023, pp. 1–8.
11. Li, S.; Zhou, X.; Wu, Z.; Long, Y.; Shen, Y. Strategic Deductive Reasoning in Large Language Models: A Dual-Agent Approach. *Preprints* **2024**. https://doi.org/10.20944/preprints202409.1875.v1.