# MocapGS: Robust Online 3D Mapping Using Motion Capture and Gaussian Splatting

Theodor Kapler[a,*], Markus Hillemann[a], Robert Langendörfer[a] and Markus Ulrich[a]

[a]Institute of Photogrammetry and Remote Sensing, Karlsruhe Institute of Technology, Englerstr. 7, Karlsruhe, 76131, Baden-Württemberg, Germany

ARTICLE INFO

*Keywords*:
Online 3D Reconstruction
Motion Capture
Gaussian Splatting
Time Synchronization
Camera Calibration
Hand-eye Calibration

ABSTRACT

Image-based online 3D mapping allows scenes to be reconstructed incrementally, with new images continuously integrated to provide direct feedback on the reconstructed scene. However, many online reconstruction methods, typically SLAM-applications, rely on image-based camera pose estimation, which is prone to drift and loop-closure errors. To address this issue, we present MocapGS, an online 3D reconstruction method that directly integrates accurate camera poses obtained from a Motion Capture system directly into a 3D Gaussian Splatting-based reconstruction pipeline. By decoupling pose estimation from scene reconstruction, MocapGS leverages externally provided, metrically accurate camera poses to improve robustness and global consistency. To enable this integration, a complete workflow is developed, including temporal synchronization between the camera and the Motion Capture system, camera calibration, hand-eye calibration, and online reconstruction from sequential image-pose pairs. The proposed method is evaluated on multiple datasets and compared to methods estimating poses from image data. Experimental results show that MocapGS yields more stable camera trajectories and improved reconstruction quality compared to such methods, particularly with respect to global consistency. The results demonstrate that Motion Capture can serve as a reliable primary source of camera poses for online 3D reconstruction, highlighting the potential of Motion Capture-driven reconstruction pipelines.

## 1. Introduction

Image-based 3D reconstruction has become a fundamental technology in computer vision, photogrammetry, and robotics, with applications ranging from industrial inspection [12] over autonomous driving [1] to digital heritage preservation [4]. As cameras are inexpensive, flexible, and widely available, image-based approaches remain particularly attractive compared to active sensing modalities.

Classical 3D reconstruction pipelines are predominantly batch-based. Methods such as Structure from Motion (SfM) [9] estimate camera poses, intrinsic parameters, and sparse geometry jointly from a complete set of images, followed by a densification step using Multi-view Stereo (MVS) [10] or learning-based representations, like 3D Gaussian Splatting (3DGS) [5]. While these approaches can achieve good reconstruction quality, they fundamentally require all images to be available in advance. As a consequence, they are unsuitable for scenarios where image data is acquired sequentially and where immediate feedback is desired.

Incremental reconstruction approaches address these issues. Simultaneous Localization and Mapping (SLAM)-methods [2] estimate camera poses and scene structure on-the-fly, enabling online operation and continuous map expansion as new images arrive. More recently, learning-based scene representations such as 3DGS have been adapted to online image acquisition [7]. These methods make it possible to visualize a scene in real-time while it is being reconstructed, providing immediate feedback and enabling adaptive data acquisition.

Despite these advantages, online approaches typically estimate camera poses sequentially from image data alone. As a result, they are prone to drift and loop-closure errors, which can accumulate over time and lead to globally inconsistent reconstructions.

In parallel, Motion Capture (Mocap) systems offer an alternative means of camera pose estimation. By tracking a rigid body with a mounted camera, Mocap systems can provide camera poses with high accuracy and global consistency, independent of visual scene content. However, in the context of 3D reconstruction, such systems are rarely used as the primary source of camera pose retrieval. Instead, they are typically employed for benchmarking, validation, or ground truth generation, leaving their potential for direct integration into online reconstruction pipelines largely unexplored.

This thesis aims to fill this gap by introducing *MocapGS*, a method that integrates accurate camera poses obtained from a Mocap system directly into an online 3D reconstruction pipeline based on 3DGS. To enable this integration, we propose a complete workflow consisting of:

1. Temporal Synchronization,
2. Camera Calibration,
3. Hand-eye Calibration,
4. Online 3D Reconstruction, and
5. 3D Mesh Extraction.

The main contributions of this work are threefold. First, we present the proposed workflow for integrating Mocap-based pose estimation into an online 3DGS reconstruction pipeline. Second, we evaluate the accuracy of the temporal

*Corresponding author
theodor.kapler@student.kit.edu (T. Kapler);
markus.hillemann@kit.edu (M. Hillemann);
robert.langendoerfer@kit.edu (R. Langendörfer);
markus.ulrich@kit.edu (M. Ulrich)
ORCID(s): 0000-0002-8906-0450 (M. Hillemann);
0000-0001-8457-5554 (M. Ulrich)

synchronization and calibration procedures. Third, we assess the reconstruction quality and compare our approach against standard 3DGS as well as an online method that estimates camera poses solely from image data.

We find that by decoupling pose estimation from scene optimization, MocapGS combines the online visualization and incremental processing capabilities of modern 3DGS-based methods with the robustness and global consistency of Mocap-based camera tracking.

## 2. Related Work

In this section, we summarize the related work regarding 3D reconstruction methods, as well as camera tracking with Mocap.

### 2.1. 3D Reconstruction Using Structure from Motion

Classical 3D reconstruction pipelines are predominantly batch-based and follow a two-stage processing scheme. In a first stage, SfM is applied to a complete set of images to jointly estimate camera poses, intrinsics, and a sparse 3D point cloud via bundle adjustment.

In a second stage, dense reconstruction is performed using the estimated poses. Traditional photogrammetric methods rely on MVS to densify the sparse reconstruction. Neural Radiance Fields (NeRFs) [8], a more recent learning-based approach, replaces explicit point-based reconstruction with a continuous scene representation, enabling high-quality novel view synthesis. 3DGS adopts a more explicit scene representation based on learned 3D Gaussian ellipsoids, enabling very fast rendering.

While highly accurate, these pipelines share a key limitation: SfM is always required as a preprocessing step and depends on the availability of the full image set. Consequently, such methods are non-incremental and do not support real-time reconstruction or interactive scene exploration.

### 2.2. Online 3D Reconstruction

Online 3D reconstruction methods aim to process image data sequentially as it becomes available, without prior knowledge of the full dataset. These approaches typically combine pose estimation and mapping in a single, incremental pipeline.

Classical point-based methods such as SLAM estimate camera poses on-the-fly while incrementally building a map of the environment. More recent learning-based methods extend this idea to neural or hybrid scene representations. Approaches such as MonoGS [6], WildGS-SLAM [14], or On-the-fly NVS [7] enable incremental reconstruction and real-time rendering while estimating camera poses directly from incoming images.

A major advantage of online methods is the ability to visualize the scene in real time and to actively densify under-reconstructed regions. However, pose estimation is sequential and therefore depends on previously estimated poses. As a result, these methods are susceptible to drift and loop-closure errors, which can lead to globally inconsistent reconstructions.

### 2.3. Camera Tracking with Motion Capture

Mocap systems provide an alternative source of camera pose information. By rigidly mounting a camera to a set of tracked markers, the pose of the camera can be recovered in real-time with high accuracy after hand-eye calibration. Such systems yield globally consistent poses and are not affected by drift or loop-closure failures.

In the literature, Mocap is most commonly used for validation or ground truth generation rather than as the primary source of camera poses for reconstruction [3, 13, 11].

This work is situated at the intersection of online 3D reconstruction and Mocap-based camera tracking. While online 3D reconstruction methods provide real-time feedback and incremental processing, Mocap offers accurate and globally consistent poses. The combination of these complementary properties motivates the approach presented in this work.

## 3. Methodology

The following section presents an overview of the methodology employed in this work. It systematically outlines the proposed workflow for enabling online 3D reconstruction with a synchronized and calibrated camera-Mocap system. Figure 1 outlines our proposed workflow, of which every single step is described in detail in the following.

### 3.1. Definition of a Rigid Body

To enable camera pose tracking with the Mocap system, we attach a camera rigidly to a fixed set of Mocap markers. These markers define a rigid body that can be tracked by the Mocap system. To maintain the analogy to robotics, in the following, we refer to the rigid body as *tool*. The pose of this tool coordinate system (TCS) being tracked by the Mocap system differs from the pose of the camera coordinate system (CCS), that we are interested in, by an unknown pose ${}^{\mathrm{C}}\mathbf{T}_{\mathrm{T}}$. The determination of this

### 3.2. Temporal Synchronization
### 3.3. Camera Calibration
### 3.4. Hand-eye Calibration
### 3.5. Online 3D Reconstruction
### 3.6. 3D Mesh Extraction
## 4. Experiments
## 5. Discussion
## 6. Conclusion
## References

[1] Behley, J., Stachniss, C., 2018. Efficient Surfel-Based SLAM using 3D Laser Range Data in Urban Environments, in: Robotics: Science and Systems XIV, Robotics: Science and Systems Foundation. URL: http://www.roboticsproceedings.org/rss14/p16.pdf, doi:10.15607/RSS.2018.XIV.016.

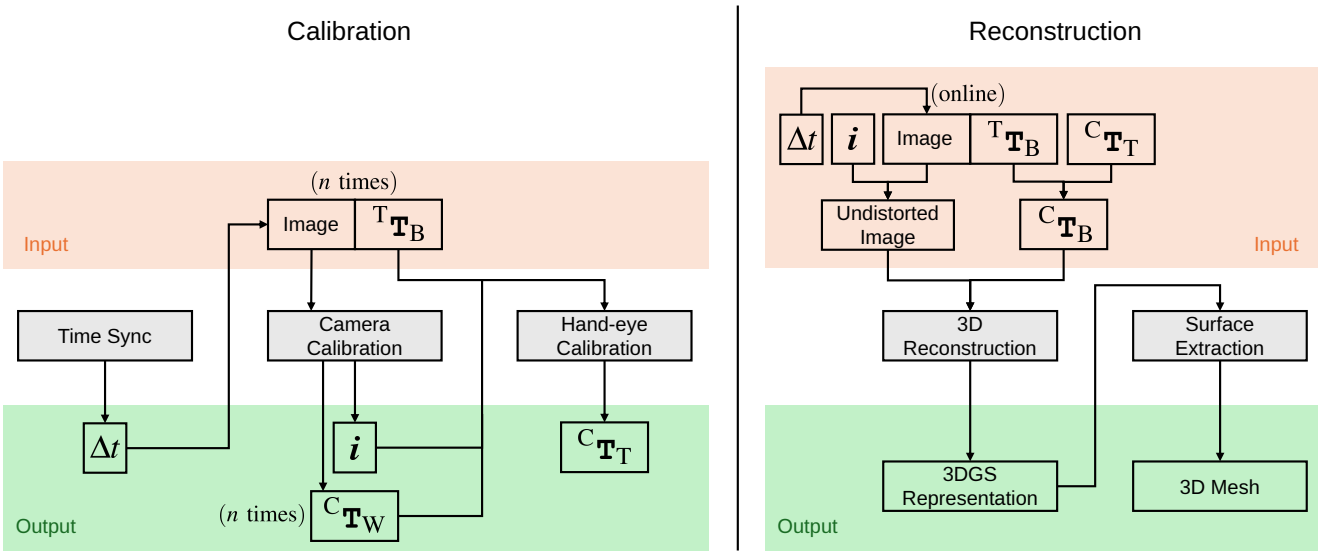


**Figure 1:** XXX


[2] Cadena, C., Carlone, L., Carrillo, H., Latif, Y., Scaramuzza, D., Neira, J., Reid, I., Leonard, J.J., 2016. Past, Present, and Future of Simultaneous Localization and Mapping: Toward the Robust-Perception Age. IEEE Transactions on Robotics 32, 1309–1332. URL: https://ieeexplore.ieee.org/abstract/document/7747236, doi:10.1109/TRO.2016.2624754.

[3] Endres, F., Hess, J., Engelhard, N., Sturm, J., Cremers, D., Burgard, W., 2012. An evaluation of the RGB-D SLAM system, in: 2012 IEEE International Conference on Robotics and Automation, IEEE, St Paul, MN, USA. pp. 1691–1696. URL: http://ieeexplore.ieee.org/document/6225199/, doi:10.1109/ICRA.2012.6225199.

[4] Gomes, L., Regina Pereira Bellon, O., Silva, L., 2014. 3D reconstruction methods for digital preservation of cultural heritage: A survey. Pattern Recognition Letters 50, 3–14. URL: https://www.sciencedirect.com/science/article/pii/S0167865514001032, doi:10.1016/j.patrec.2014.03.023.

[5] Kerbl, B., Kopanas, G., Leimkuehler, T., Drettakis, G., 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. ACM Transactions on Graphics 42, 1–14. URL: https://dl.acm.org/doi/10.1145/3592433, doi:10.1145/3592433.

[6] Matsuki, H., Murai, R., Kelly, P.H.J., Davison, A.J., 2024. Gaussian Splatting SLAM, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Seattle, WA, USA. pp. 18039–18048. URL: https://ieeexplore.ieee.org/document/10657715/, doi:10.1109/CVPR52733.2024.01708.

[7] Meuleman, A., Shah, I., Lanvin, A., Kerbl, B., Drettakis, G., 2025. On-the-fly Reconstruction for Large-Scale Novel View Synthesis from Unposed Images. ACM Transactions on Graphics 44, 1–14. URL: https://dl.acm.org/doi/10.1145/3730913, doi:10.1145/3730913.

[8] Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R., 2021. NeRF: representing scenes as neural radiance fields for view synthesis. Commun. ACM 65, 99–106. URL: https://dl.acm.org/doi/10.1145/3503250, doi:10.1145/3503250.

[9] Schönberger, J.L., Frahm, J.M., 2016. Structure-from-Motion Revisited, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4104–4113. URL: https://ieeexplore.ieee.org/document/7780814, doi:10.1109/CVPR.2016.445.

[10] Schönberger, J.L., Zheng, E., Frahm, J.M., Pollefeys, M., 2016. Pixelwise View Selection for Unstructured Multi-View Stereo, in: Leibe, B., Matas, J., Sebe, N., Welling, M. (Eds.), Computer Vision – ECCV 2016, Springer International Publishing, Cham. pp. 501–518. doi:10.1007/978-3-319-46487-9_31.

[11] Shu, Z., Bei, S., Dai, J., Li, L., Chen, Z., Wang, J., 2026. MoCap2GT: A High-Precision Ground Truth Estimator for SLAM Benchmarking Based on Motion Capture and IMU Fusion. IEEE Robotics and Automation Letters 11, 1538–1545. URL: https://ieeexplore.ieee.org/document/11297812, doi:10.1109/LRA.2025.3643287.

[12] Steger, C., Ulrich, M., Wiedemann, C. (Eds.), 2018. Machine vision algorithms and applications. 2nd, completely revised and enlarged edition ed., Wiley-VCH, Weinheim.

[13] Sturm, J., Engelhard, N., Endres, F., Burgard, W., Cremers, D., 2012. A benchmark for the evaluation of RGB-D SLAM systems, in: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, Vilamoura-Algarve, Portugal. pp. 573–580. URL: http://ieeexplore.ieee.org/document/6385773/, doi:10.1109/IROS.2012.6385773.

[14] Zheng, J., Zhu, Z., Bieri, V., Pollefeys, M., Peng, S., Armeni, I., 2025. WildGS-SLAM: Monocular Gaussian Splatting SLAM in Dynamic Environments, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11461–11471.