Aachener Verfahrenstechnik | RWTH AACHEN UNIVERSITY

**Diese Arbeit wurde vorgelegt am Lehrstuhl für Chemische Verfahrenstechnik**
The present work was submitted to Chair of Chemical Process Engineering

# Dichte 3D-Echtzeit-Rekonstruktion für die endoskopische Chirurgie mittels monokularer Sequenzen durch ein Transformer-basiertes Feed-Forward neuronales Netz

# Real-Time Dense 3D Reconstruction for Endoscopic Surgery using Monocular Sequences via a Transformer-Based Feed-Forward Neural Network

Masterarbeit
Master Thesis

von / presented by
Guan, Zhouyi (406259)

Betreuer*in / Supervisor
Yoo, Sang-Whon, M. Sc.

Prof. Dr.-Ing. Matthias Wessling

Aachen, DATE

# Abstract

Osteoarthritis remains a global socioeconomic burden characterized by the irreversible degradation of Articular Cartilage (AC). Within the framework of the DWI Leibniz-Institute for Interactive Materials, research is directed toward developing fully automatic in vivo 3D bioprinting to overcome the limitations of traditional ex vivo scaffolding. Conventional dense 3D reconstruction techniques in endoscopic scenarios have struggled with weak structural textures, leading to insufficient robustness and precision. Furthermore, existing algorithmic frameworks have remained computationally heavy with prohibitive processing times, and their performance has been frequently compromised by dynamic surgical instruments or flowing tissues that obstruct the surgical field. This study aims to design a high-precision, real-time dense SLAM framework capable of maintaining robust 3D reconstruction even in the presence of intraoperative occlusions and dynamic disturbances. A Transformer-based feed-forward neural network was developed to perform high-fidelity dense 3D reconstruction. Besides a dedicated motion head was integrated and trained to identify and decouple dynamic surgical instruments and flowing tissues from the scenes. Furthermore, a real-time SLAM framework was constructed utilizing $SL(4)$ manifold representation. /TODO: Add quantitative results/

# Assignment

# Artificial Intelligence (AI) Statement

This thesis has been partially rewritten using an artificial intelligence (AI) model to improve readability. All ideas, considerations, and content are original and created by the author. Therefore, AI systems are not concerned with intellectual property. The AI usage rules stipulated by the Faculty of Mechanical Engineering of RWTH Aachen University on July 29, 2025, have been fully satisfied.

# Contents

# 1 Introduction

Osteoarthritis (OA) stands as a premier cause of global disability, characterized by the irreversible degradation of Articular Cartilage (AC). Due to the avascular nature of AC, its intrinsic repair capacity is remarkably limited; even minor focal lesions often fail to heal, eventually leading to chronic pain and impaired mobility. While traditional ex vivo scaffolding techniques have been widely researched, recent advancements in medical robotics and materials science are paving the way for direct, in vivo 3D bioprinting. As envisioned within the major initiative at the DWI Leibniz-Institute for Interactive Materials, a fully automatic robotic process for in situ cartilage repair offers a transformative alternative by minimizing contamination risks and bypassing the time-consuming nature of conventional methods.[1]

The success of in vivo bioprinting is fundamentally predicated on the precise 3D scanning and geometry estimation of the target lesion. Traditionally, pre-operative Magnetic Resonance Imaging (MRI) has been the gold standard; however, it suffers from systematic underestimation of lesion thickness and limited out-of-plane resolution. While photogrammetry presents a promising non-contact alternative, conventional dense 3D reconstruction algorithms often falter in the challenging endoscopic environment. These techniques are frequently hampered by weak textural features, prohibitive computational overhead, and the presence of dynamic occlusions—such as surgical instruments and flowing tissues—which compromise both the accuracy and the robustness required for real-time clinical intervention.[1]

Building upon the foundational work at DWI, this thesis aims to bridge the gap between static photogrammetry and real-time surgical navigation. The primary objective is to develop a high-precision, real-time dense 3D reconstruction and SLAM framework. By leveraging Transformer-based architectures and advanced manifold optimization, this work seeks to provide a robust spatial mapping solution that can effectively handle intraoperative occlusions and provide the necessary geometric intelligence for autonomous robotic bioprinting.

The proposed pipeline introduces a Transformer-based feed-forward neural network designed for

high-fidelity geometry estimation, capturing global dependencies that traditional methods often miss. To ensure clinical reliability, a dedicated motion masking head is integrated to identify and decouple dynamic surgical disturbances from the static cartilage surface. Furthermore, the framework achieves high-efficiency real-time performance by utilizing an $SL(4)$ (Special Linear Group) manifold representation for SLAM, allowing for robust tracking under complex projective transformations inherent in monocular endoscopy.

The remainder of this thesis is structured as follows:

Section 2.1 discusses the medical background of knee AC and OA, emphasizing state-of-the-art treatment and 3D bioprinting techniques.

Section 2.2 introduces the theoretical foundations of dense 3D reconstruction technologies.

Section 2.3 provides an overview of SLAM technologies and manifold optimization.

Section 3.1 details the architectural design and training process of the proposed Transformer-based feed-forward neural network.

Section 3.2 outlines the construction of the real-time dense SLAM framework built upon the aforementioned model.

Chapter 4 analyzes the experimental results obtained from the pipeline, using well plates and simulated environments as precursors.

Chapter 5 concludes the thesis with a summary of findings and an outlook on potential optimizations for clinical application.

# 2 Theoretical Background

## 2.1 Medical Background

In this chapter, the focus is on the theoretical background of the medical topics covered in the scope of this work, which is the basis for the development of the processing pipeline. Firstly, AC will be introduced, followed by a description of OA and its common treatment methods. The chapter ends with the depiction of state-of-the-art methods for clinically treating OA.[1]

### 2.1.1 Articular Cartilage

The human body is composed of various cells that not only form organs but also define different types of organic tissue. The fundamental four tissue types are nervous, muscular, epithelial, and connective tissue. Among these, connective tissue is the most widespread and diverse, encompassing fat tissue, blood, fibrous tissue, bone marrow, and cartilage. These connective tissues stand out due to being highly specialized and rich in matrix, which in turn ensures cohesion, mechanical support, and protection of the body's organs and structures.

### 2.1.2 Osteoarthritis and its Treatment

### 2.1.3 3D Bioprinting Techniques

## 2.2 Dense 3D Reconstruction

This chapter will first introduce the traditional geometry-based 3D reconstruction methods, including Structure-from-Motion (SfM) and Multi-View Stereo (MVS). Then, it will present the recent feed-forward Transformer-based 3D reconstruction methods, including DUSt3R and VGGT.

### 2.2.1 Related Work

#### Traditional Geometry-based 3D Reconstruction

The field of 3D reconstruction has been traditionally dominated by pipeline-based approaches, namely Structure-from-Motion (SfM)[2] and Multi-View Stereo (MVS)[3].

**Structure from Motion** is a classic computer vision problem [45, 77, 80] that involves estimating camera parameters and reconstructing sparse point clouds from a set of images of a static scene captured from different viewpoints. The traditional SfM pipeline [2, 36, 70, 94, 103, 134] consists of multiple stages, including image matching, triangulation, and bundle adjustment. COLMAP [94] is the most popular framework based on the traditional pipeline.[4]

Figure2.1 illustrates a fundamental SfM framework.The pipeline starts from multi-view image inputs, establishes inter-image correspondences through feature matching and geometric verification, and then incrementally estimates camera poses and 3D scene structure. The reconstruction is continuously refined through triangulation and bundle adjustment, resulting in a globally consistent 3D representation.

**Multi-view Stereo (MVS)** aims to densely reconstruct the geometry of a scene from multiple overlapping images, typically assuming known camera parameters which are often estimated via SfM [4]. The fundamental principle underlying MVS is the Photo Consistency Assumption,

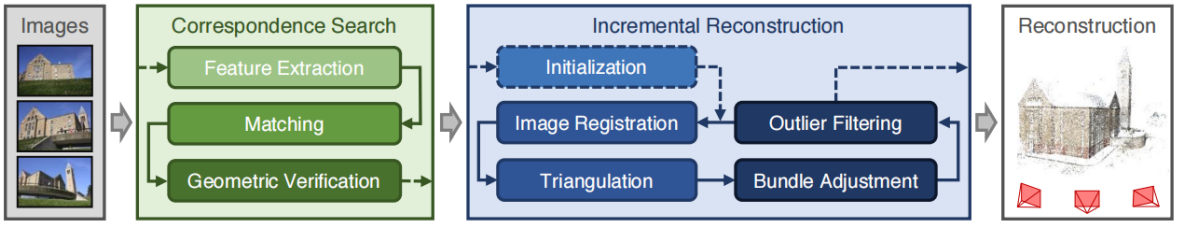


Figure 2.1: Structure-from-Motion pipeline

which posits that a 3D point in space should exhibit a similar appearance (color or intensity) when projected onto different image planes, provided the surface is approximately Lambertian. To efficiently identify corresponding points across views, MVS leverages Epipolar Geometry. Instead of searching the entire image domain, pairwise matching is constrained to epipolar lines determined by the relative camera poses, thereby significantly reducing computational complexity. Based on these geometric constraints, the algorithm estimates a depth map for each reference image by aggregating information from neighboring views. Finally, in the Fusion stage, these individual depth maps are integrated into a globally consistent point cloud or mesh, filtering out outliers to produce a high-fidelity dense estimation of the scene geometry.

Early learning methods injected robustness at the component level: learned detectors [20], descriptors for matching [22], and differentiable optimization layers that expose pose/depth updates to gradient flow [31, 33, 62]. On the dense side, cost-volume networks [106, 114] for MVS replaced hand-crafted regularization with 3D CNNs, improving depth accuracy especially at large baselines and thin structures compared with classical PatchMatch.[5] They did reduce engineering complexity and demonstrated the feasibility of learned joint depth pose estimation, but they often struggled with scalability, generalization, and handling arbitrary input cardinalities.[5]

### 2.2.2 Feed-forward Transformer-based 3D Reconstruction

Traditional Geometry-based methods remain strong on well-textured scenes, but their modularity and brittle correspondences complicate robustness under low texture, specularities, or large viewpoint changes.

The emergence of feed-forward 3D reconstruction methods is closely tied to the introduction of attention mechanisms. Unlike convolutional operations that rely on local receptive fields, attention enables explicit modeling of long-range dependencies by computing pairwise inter-

actions between elements in a sequence. In the context of multi-view geometry, this property is particularly appealing, as 3D reconstruction inherently requires reasoning over global spatial relationships and correspondences across views. By allowing features from different images to directly attend to each other, attention mechanisms provide a natural alternative to explicit feature matching and geometric verification used in traditional pipelines.

Leveraging attention-based architectures, recent approaches formulate 3D reconstruction as a feed-forward prediction problem, where geometric quantities such as depth, camera motion, or point maps are directly inferred from image features. This paradigm eliminates the need for iterative optimization and decoupled processing stages, enabling end-to-end learning of visual geometry from data.

### DUSt3R: A Pioneering Feed-forward Transformer-based Method

A turning point came with DUSt3R,a radically novel paradigm for Dense and Unconstrained Stereo 3D Reconstruction of arbitrary image[6], which leveraged transformers to directly predict point map between two views and compute both depth and relative pose in a purely feed-forward manner. This work laid the foundation for subsequent transformer-based methods aiming to unify multi-view geometry estimation at scale.[5]

As illustrated in Fig. 2.2, given two input images $I_1$ and $I_2$, both images are first processed by a shared image feature extractor to obtain dense visual representations. The feature extractor can be implemented using either convolutional neural networks or transformer-based architectures; in DUSt3R, a Vision Transformer (ViT) is adopted as the encoder, whose architectural details will be discussed in the next chapter. The extracted feature tokens from the two views are then passed to two transformer decoders, which iteratively exchange information through cross-attention. This mechanism allows features from one view to attend to and reason about features from the other view, enabling implicit correspondence discovery without explicit feature matching.

Based on the fused multi-view features, two regression heads predict dense point maps for each input image, along with associated confidence maps. Importantly, both point maps are expressed in a common coordinate frame defined by the first camera view $I_1$, thereby establishing a shared geometric reference across the two images. From the predicted point maps, camera intrinsics such as focal lengths can be estimated using the Weiszfeld algorithm [7], and relative camera poses can be recovered using minimal solvers such as 3-point RANSAC

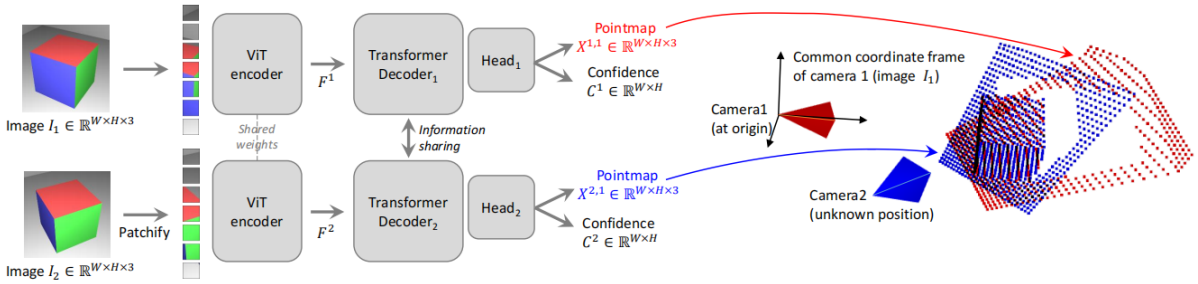


Figure 2.2: DUSt3R framework overview

[8, 9].

By directly regressing dense geometry in a feed-forward manner, DUSt3R eliminates the need for explicit feature matching, geometric verification, and iterative optimization, demonstrating the feasibility of learning-based geometric reasoning from image pairs. [10]

### Follow-up Works

MASt3R follows a similar design but also outputs descriptors that can be used to generate pairwise correspondences between the two frames. MASt3R-SFM [14] demonstrates global optimization of multiple images using MASt3R but computation scales quickly with the number of frames. [10]

To extend the idea of DUSt3R to multiple frames, Spann3R [67] leverages a learned memory module and Cut3R [69] uses a recurrent state model. Both can incrementally reconstruct a scene using multiple images, but are each limited to short sequences. Recently, Pow3R [27] extends the DUSt3R framework to optionally take in any estimates of any combination of camera intrinsics, poses, and depth (which may be sparse or dense) and demonstrates substantial improvement in scene reconstruction and pose estimation given the added inputs. Splatt3R [60] extends the DUSt3R idea to Gaussian Splatting [29] by directly outputting the Gaussian Splatting parameters given two views, and PreF3R [8] extends this to multiple views using a similar memory framework as Spann3R. Reloc3r [13] modifies the DUSt3R framework for directly outputs relative camera poses and uses motion averaging to recover absolute poses with respect to a map database.[10]

### Visual Geometry Grounded Transformer and Its Variants

While DUSt3R and its follow-up works demonstrate that feed-forward models can successfully infer dense geometry from image pairs, their pairwise formulation inherently limits scalability and global reasoning. Extending such methods to multi-view settings typically requires additional post-processing steps, such as pose graph optimization or incremental fusion, which reintroduce elements of traditional pipelines.

These limitations motivate the need for a unified framework that can jointly reason over an arbitrary number of views and directly model global geometric relationships in a feed-forward manner.

As a culmination of feed-forward reconstruction models inspired by DUSt3R, VGGT (Visual Geometry Grounded Transformer) extends pairwise geometric reasoning to arbitrary-length image sequences within a unified architecture. An overview of the VGGT framework is shown in Figure 2.3. Given a set of input images, VGGT first decomposes each image into a sequence of visual tokens using a pretrained image feature extractor. In the original formulation, self-supervised features obtained from DINO are employed to provide semantically rich and geometrically consistent representations. These per-image token sequences are then concatenated, and camera-specific tokens are appended to enable explicit modeling of camera parameters.

The core of VGGT consists of multiple transformer layers that alternate between global attention and frame-wise attention. Global attention allows tokens from different views to directly interact, facilitating cross-view information exchange and long-range geometric reasoning. In contrast, frame-wise attention focuses on refining features within each individual image. By alternating between these two attention mechanisms, VGGT jointly captures both intra-view structure and inter-view geometric relationships across the entire image set.

Based on the aggregated token representations, VGGT employs multiple prediction heads to regress diverse geometric quantities. A dedicated camera head estimates camera intrinsics and extrinsics, while dense prediction transformer (DPT) heads are used to generate dense depth maps, point maps, and feature tracks. Importantly, these outputs are expressed in a common coordinate frame, enabling consistent multi-view reconstruction without explicit correspondence estimation or incremental optimization.

By unifying multi-view feature aggregation and geometry prediction within a single feed-forward

transformer, VGGT represents a significant step toward holistic and scalable dense 3D reconstruction from unconstrained image collections.

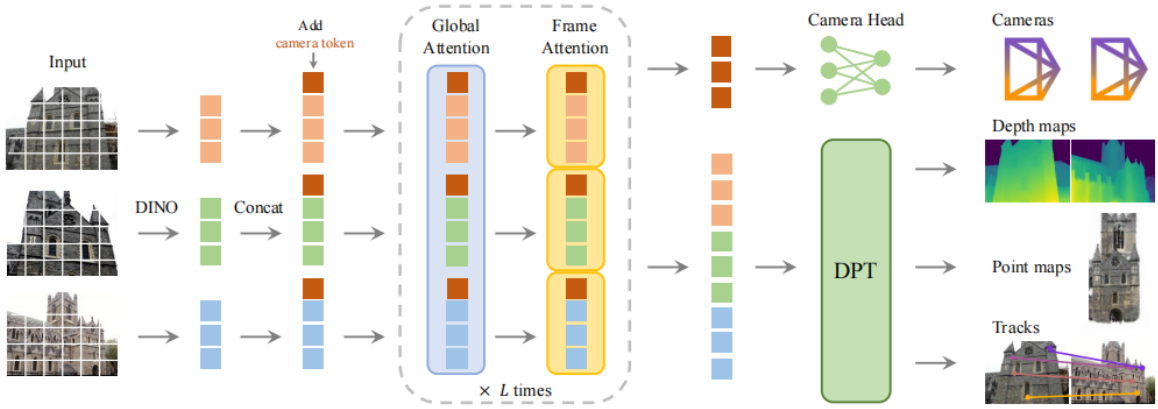


Figure 2.3: VGGT framework overview

### 2.2.3 State of the Art: Depth Anything 3

DINO

## 2.3 SLAM and Manifold Optimization

# 3 Materials and Methods

## 3.1 Neural Network Architecture and Training

## 3.2 Real-Time Dense SLAM Framework

# 4 Results and Discussion

# 5 Conclusion and Outlook

# Bibliography


[1] S.-W. Yoo. "Neural recognition and 3D-reconstruction pipeline for automated replacement of knee cartilage tissue". MA thesis. Aachen, Germany: Chair of Chemical Process Engineering, RWTH Aachen University, May 2024 (cit. on pp. 1, 3).

[2] J. L. Schönberger and J.-M. Frahm. "Structure-from-motion revisited". In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, 4104–4113 (cit. on p. 4).

[3] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm. "Pixelwise view selection for unstructured multi-view stereo". In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*. Springer. 2016, 501–518 (cit. on p. 4).

[4] J. Wang, M. Chen, N. Karaev, C. Rupprecht, A. Vedaldi, and D. Novotny. "VGGT: Visual Geometry Grounded Transformer". *arXiv preprint arXiv:2503.11651* (2025) (cit. on p. 4).

[5] H. Lin, S. Chen, J. H. Liew, D. Y. Chen, Z. Li, G. Shi, J. Feng, and B. Kang. "Depth Anything 3: Recovering the Visual Space from Any Views". *arXiv preprint arXiv:2511.10647* (2025) (cit. on pp. 5, 6).

[6] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud. "DUST3R: Geometric 3D vision made easy". In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, 20697–20709 (cit. on p. 6).

[7] E. Weiszfeld. "Sur le point pour lequel la somme des distances de n points donnés est minimum". *Tohoku Mathematical Journal, First Series* 43 (1937) 355–386. (Cit. on p. 6).

[8] D. Nistér. "An efficient solution to the five-point relative pose problem". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26.6 (2004) 756–770. (Cit. on p. 7).

[9] J. Matas and O. Chum. "Randomized RANSAC with T d, d test". *Image and vision computing* 22.10 (2004) 837–842. (Cit. on p. 7).

[10] D. Maggio, H. Lim, and L. Carlone. "VGGT-SLAM: Dense RGB SLAM Optimized on the SL(4) Manifold". *arXiv preprint arXiv:2505.12549* (2025) (cit. on p. 7).

# List of Figures

# List of Tables

# A Appendix