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

2.1.2 Osteoarthritis and its Treatment

2.1.3 3D Bioprinting Techniques

2.2 Dense 3D Reconstruction

This chapter presents the theoretical background relevant to this work, with a focus on dense 3D reconstruction from monocular sequences. First, related work is reviewed, covering the evolution from traditional geometry-based methods such as Structure-from-Motion (SfM) and Multi-View Stereo (MVS) to recent feed-forward transformer-based reconstruction approaches. Then, the reconstruction model adopted in this work, Depth Anything 3 (DA3), is introduced in detail, providing the foundation for the subsequent methodology.

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.

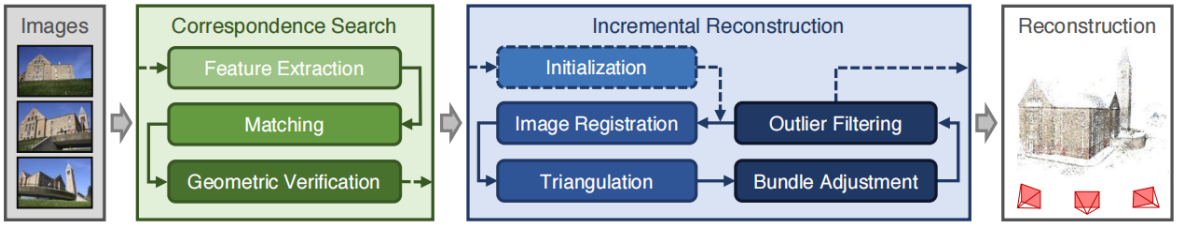

Figure 2.1: Structure-from-Motion pipeline

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, 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]

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 interactions 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 [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]

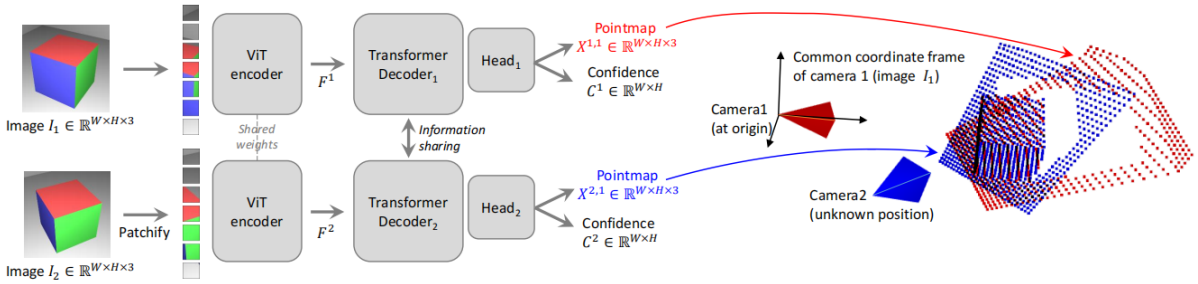

Figure 2.2: DUSt3R framework overview

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

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.

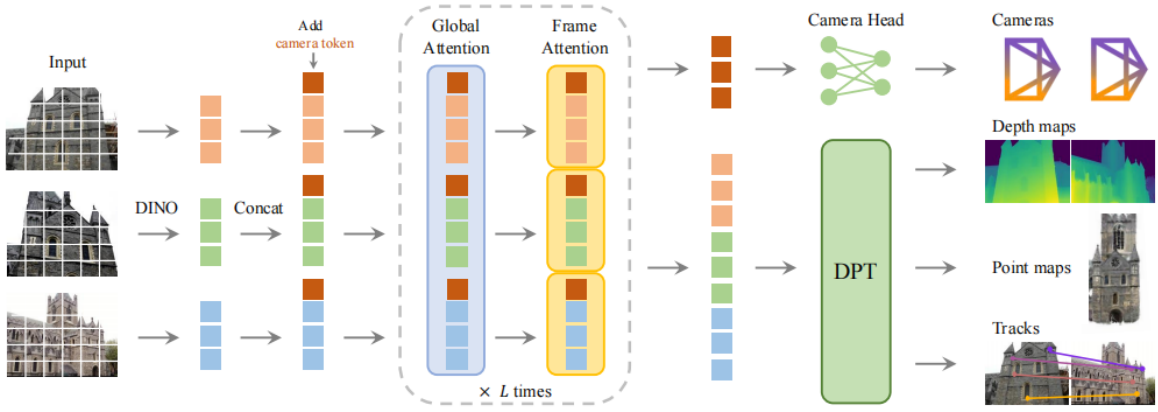

Figure 2.3: VGGT framework overview

2.2.2 State of the Art: DA3

While VGGT represents a significant advancement in unified feed-forward multi-view reconstruction, it still leaves room for improvement in terms of efficiency, scalability, and generalization. The VGGT framework is characterized by a large and complex architecture, with model sizes reaching up to one billion parameters, which makes it unsuitable for lightweight deployment scenarios. Although VGGT supports arbitrary-length input sequences, this flexibility comes at the cost of substantial memory consumption and computational overhead, especially when processing long image sequences. Moreover, despite being trained on large-scale and diverse datasets covering a wide range of common visual scenes, VGGT may still fail when confronted with domains that differ significantly from its training distribution, such as endoscopic imagery. These limitations highlight the need for a more compact and robust reconstruction model that preserves the strengths of VGGT while addressing its practical shortcomings. To this end, Depth Anything 3 (DA3) is introduced as an improved variant built upon the VGGT paradigm. Building upon this insight, DA3 proposes a unified feed-forward model that reconstructs the visual space from arbitrary visual inputs, including multi-view image collections and video streams. Given any number of input views, with or without known camera poses, DA3 jointly predicts dense depth maps and ray maps that can be directly fused into geometrically consistent 3D point clouds.

Compared to VGGT, DA3 significantly simplifies the overall framework by relying on a single DINO-based transformer backbone and enabling cross-view reasoning through token rearrangement rather than explicit multi-branch architectures. This design reduces the model size to a configurable range between 0.11B and 0.35B parameters, making DA3 substantially more suitable for lightweight and resource-constrained deployment while maintaining strong reconstruction performance.

In addition, DA3 leverages a powerful depth teacher model during training, allowing it to learn robust geometric priors from large-scale data. This training strategy improves generalization and enables the model to remain effective even in previously unseen domains. Finally, DA3 introduces a depth-ray representation as a minimal yet sufficient geometric abstraction, which explicitly encodes both scene structure and camera motion while avoiding redundant or constrained prediction targets.

The following sections provide a detailed description of the DA3 depth-ray formulation, its architecture, and the associated training paradigm.

The following sections provide a detailed description of the DA3 depth-ray formulation, its architecture, and the associated training paradigm.

Depth-ray Representation

At the core of DA3 lies the depth–ray representation, which explicitly encodes both scene geometry and camera motion without directly predicting constrained rotation matrices. For each pixel $p = (u, v, 1)^\top$, the model predicts a depth value $D(u, v)$ together with a corresponding ray

$$r = (t, d), \quad d = RK^{-1}p,$$

where $t \in \mathbb{R}^3$ denotes the camera center and $d \in \mathbb{R}^3$ is the unnormalized ray direction in world coordinates.

The predicted ray map is denoted as

$$M \in \mathbb{R}^{H \times W \times 6},$$

where the first three channels $M(:, :, :3)$ store per-pixel ray origins and the last three channels $M(:, :, 3:)$ store the corresponding ray directions. Given the depth and ray predictions, a 3D point can be recovered by

$$P(u, v) = t + D(u, v) \cdot d.$$

Beyond 3D point reconstruction, the depth–ray representation also enables direct recovery of camera parameters. The camera center t_c is estimated by averaging the per-pixel ray origins:

$$t_c = \frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} M(h, w, :3).$$

To recover the camera rotation R and intrinsic matrix K, DA3 formulates the problem as estimating a homography. An identity canonical camera is first defined with intrinsics $K_I = I$. For a pixel p, the corresponding ray direction in this canonical camera is

$$d_I = K_I^{-1} p = p.$$

The transformation from the canonical ray d_I to the ray direction d_{cam} in the target camera coordinate system can be written as

$$d_{\text{cam}} = KRd_I,$$

which establishes a homography relationship

$$H = KR.$$

The homography H is then estimated by minimizing the geometric error between transformed canonical rays and the predicted ray directions:

$$H^* = \arg\min_{\|H\|=1} \sum_{h=1}^{H} \sum_{w=1}^{W} \|Hp_{h,w} \times M(h, w, 3:)\|.$$

This optimization corresponds to a standard least-squares problem and can be efficiently solved using the Direct Linear Transform (DLT) algorithm. Once the optimal homography H^* is obtained, the camera intrinsics K and rotation R are uniquely recovered via RQ decomposition, exploiting the upper-triangular structure of K and the orthonormality of R.

By jointly modeling depth and rays, this representation avoids explicit pose constraints while ensuring geometric consistency across views, and provides a compact yet expressive abstraction for feed-forward 3D reconstruction.

Architecture

As illustrated in Fig. 2.4 shows an overview of the DA3 framework. Architecturally, DA3 adopts a deliberately simple design centered around a single pretrained Vision Transformer (ViT) backbone. For an input set

$$\mathcal{I} = \{I_i\}_{i=1}^{N_v}, I_i \in \mathbb{R}^{H \times W \times 3}$$

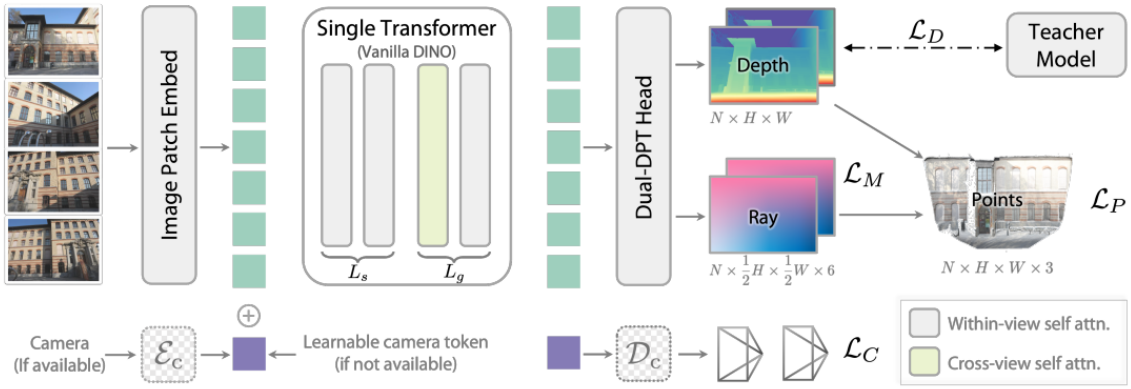

Figure 2.4: DA3 framework overview

the transformer serves as a shared feature extractor across all views, enabling the model to naturally handle variable input cardinalities.

Cross-view geometric reasoning is achieved through an input-adaptive self-attention mechanism. Instead of introducing additional transformer stacks or specialized modules, DA3 rearranges token sequences in selected layers, allowing standard self-attention to operate either within individual views or across all views jointly.

Dense geometric predictions in DA3 are produced by a novel dual Dense Prediction Transformer (Dual-DPT) head, which jointly regresses dense depth maps and ray maps from a shared set of visual features. As illustrated in Fig. 2.5, the Dual-DPT head first processes backbone features through a common set of reassembly modules, ensuring that both prediction tasks operate on a unified and compact intermediate representation.

After this shared processing stage, the features are routed into two parallel branches, corresponding to depth prediction and ray prediction, respectively. Each branch applies its own set of fusion layers to aggregate contextual information relevant to the target geometric quantity. Finally, separate output layers generate the depth map and ray map predictions. By sharing the majority of the feature processing pipeline and diverging only at the final fusion stage, the Dual-DPT design promotes strong interaction between depth and ray estimation while avoiding redundant intermediate representations.

In addition to dense geometric outputs, DA3 optionally predicts camera parameters through a lightweight camera head operating on dedicated camera tokens, incurring minimal computational overhead. This joint prediction framework further reinforces geometric consistency without compromising the overall efficiency of the model.

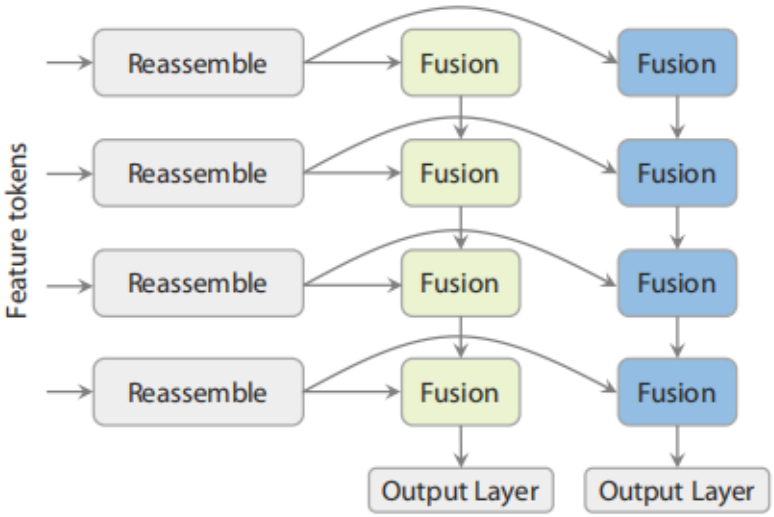

Figure 2.5: DA3 Dual-DPT Head

Training

To train a generalist geometry model across diverse data sources, DA3 adopts a teacher–student learning paradigm. A teacher model is first trained on large-scale synthetic data to generate dense and reliable pseudo-depth supervision, which is then combined with sparse or noisy real-world depth measurements. This strategy enables effective knowledge transfer while preserving geometric fidelity across diverse domains.

Formally, the DA3 model F_θ maps an input image set $\mathcal{I}$ to a set of geometric predictions, including a depth map $\hat{D}$, a ray map $\hat{R}$, and an optional camera pose $\hat{c}$:

$$F_\theta(\mathcal{I}) \rightarrow \{\hat{D}, \hat{R}, \hat{c}\}.$$

The camera pose prediction is optional and primarily included for practical convenience.

Prior to loss computation, all ground-truth geometric signals are normalized by a common scale factor to ensure consistent magnitude across different modalities. This scale is defined as the mean ℓ_2 norm of the valid reprojected 3D point maps P, which stabilizes optimization when jointly supervising depth, rays, and reconstructed points.

The overall training objective is defined as a weighted sum of several loss terms:

$$L = L_D(\hat{D}, D) + L_M(\hat{R}, M) + L_P(\hat{D} \odot d + t, P) + \beta L_C(\hat{c}, c) + \alpha L_{\text{grad}}(\hat{D}, D),$$

where L_D supervises depth prediction, L_M enforces consistency of ray maps, L_P penalizes geometric errors in reconstructed 3D points, and L_C is an optional camera pose loss.

The depth supervision term is defined as

$$L_D(\hat{D}, D; D_c) = \frac{1}{|\Omega|} \sum_{p \in \Omega} m_p (D_{c,p} |\hat{D}_p - D_p| - \lambda_c \log D_{c,p}),$$

where $D_{c,p}$ denotes the confidence associated with the ground-truth depth D_p, m_p is a validity mask, and Ω denotes the set of valid pixels. All loss terms are based on the ℓ_1 norm.

To further regularize the depth predictions, a gradient loss is introduced:

$$L_{\text{grad}}(\hat{D}, D) = \|\nabla_x \hat{D} - \nabla_x D\|_1 + \|\nabla_y \hat{D} - \nabla_y D\|_1,$$

where ∇_x and ∇_y denote horizontal and vertical finite difference operators. This term preserves sharp depth discontinuities while encouraging smoothness in planar regions. In practice, the weighting factors are set to $\alpha = 1$ and $\beta = 1$.

This training strategy enables DA3 to effectively leverage heterogeneous datasets while maintaining consistent geometric supervision, resulting in a scalable and robust feed-forward reconstruction model.

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. 6).

[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. 6).

[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 pp. 6, 7).

List of Figures

List of Tables

A Appendix