DANIELA FERREIRA
PINTO DIAS RATO

Dentro de uma Célula Colaborativa: Calibração, Perceção e Medidas de Segurança

Inside a Collaborative Cell: Calibration, Perception and Safety Requirements

Universidade de Aveiro
2025

DANIELA FERREIRA
PINTO DIAS RATO

Dentro de uma Célula Colaborativa: Calibração, Perceção e Medidas de Segurança

Inside a Collaborative Cell: Calibration, Perception and Safety Requirements

Tese apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Doutor em Engenharia Mecânica, realizada sob a orientação científica do Doutor Miguel Armando Riem de Oliveira, Professor auxiliar do Departamento de Engenharia Mecânica da Universidade de Aveiro, do Doutor Vítor Manuel Ferreira dos Santos, Professor associado c/ agregação do Departamento de Engenharia Mecânica da Universidade de Aveiro, e do Doutor Angel Domingo Sappa, Senior Researcher do Computer Vision Center da Universitat Autònoma de Barcelona.

The author acknowledges the support of the Project Augmented Humanity [POCI-01-0247-FEDER-046103] and the CYTED Network: Ibero-American Thematic Network on ICT Applications for Smart Cities (REF-518RT0559).
This work was supported by FCT - Fundação para a Ciência e Tecnologia, I.P. by project reference 2021.04792.BD and DOI identifier https://doi.org/10.54499/2021.04792.BD.

Para a minha filha Amélia. Tornaste esta aventura ainda mais insana, mas muito muito melhor.

o júri / the jury

presidente / president

Doutor Armando José Formoso de Pinho
Professor Catedrático da Universidade de Aveiro

vogais / examiners committee

Doutor Marcelo Roberto Petry
Investigador Sénior da Universidade do Porto

Doutor Paulo Jorge Sequeira Gonçalves
Professor Coordenador do Instituto Politécnico de Castelo Branco

Doutora Verónica Maria Marques do Carreiro Silva Vasconcelos
Professora Adjunta do Instituto Superior de Engenharia de Coimbra

Paulo Miguel de Jesus Dias
Professor Auxiliar com Agregação da Universidade de Aveiro

Doutor Miguel Armando Riem de Oliveira
Professor Auxiliar da Universidade de Aveiro (orientador)

agradecimentos / acknowledgements

Quero começar por agradecer à minha equipa de orientação. Ao Prof. Dr. Miguel Oliveira pelas inúmeras horas que passámos juntos a programar e por ter estado sempre disponível para participar ativamente no desenvolvimento do trabalho desta tese. Ao Prof. Dr. Vítor Santos, que me acompanha desde o mestrado, pelas palavras sempre sábias e os inúmeros conselhos ao longo destes anos. Ao Prof. Dr. Angel Sappa que, mesmo estando do outro lado do mundo, fez questão de ser uma voz ativa no desenvolvimento deste trabalho. Por fim, gostaria de agradecer ao Dr. Bogdan Raducanu por me ter acompanhado durante o período que passei em Barcelona no Computer Vision Center.

Gostaria também de agradecer aos meus pais, por me terem dado condições de chegar até aqui, por me terem apoiado incondicionalmente em todos os momentos e me terem incentivado sempre a ser uma pessoa e profissional melhor. São uma grande parte do motivo de ter chegado aqui. Quero também agradecer à minha irmã, Alice, por ter sido sempre um ombro amigo e estar sempre disponível para ouvir os meus desabafos e para celebrar as minhas vitórias.

Quero também deixar uma palavra de agradecimento à Beatriz por me ter feito sempre sentir validada ao longo deste percurso e pelas longas conversas e desabafos entre duas alunas de doutoramento.

Obrigada à minha filha Amélia, cujo sorriso me deu a força necessária para continuar.

Por fim, quero agradecer ao meu marido, João, o meu companheiro, não só nesta jornada, como no resto da vida. O caminho para chegar aqui não foi fácil. Dois doutorandos muitas vezes consumidos pelo trabalho, com uma gravidez e uma bebé pelo caminho. Conseguimos. A maior aventura ainda agora começou.

Palavras Chave

sensores, calibração, célula colaborativa, multi-modal, multi-sensorial, perceção, robótica, cobots

Resumo

O objetivo principal de uma célula colaborativa industrial é a criação de um espaço onde os humanos e robots possam trabalhar para um objetivo comum com eficiência e segurança. Nos dias que correm, a maioria das células colaborativas requer que operadores humanos façam gestos pouco naturais para comunicarem com o robot. Estes sistemas têm também interações limitadas entre ambos: muitas das vezes a reação a movimentos inesperados do humano é uma paragem abrupta. Esta tese propõe resolver estas limitações criando mecanismos de perceção avançada com base em múltiplos sensores de múltiplas modalidades. Múltiplos sensores para garantir que a célula está inteiramente coberta, mesmo quando há oclusões não previstas. Múltiplas modalidades que permitem trazer diferentes tipos de informação essencial sobre o interior da célula. O uso de uma perceção avançada e robusta vai garantir uma linguagem entre o humano e o robot, onde o sistema automático suporta o peso da comunicação, criando assim operações mais eficientes.

Keywords sensors, calibration, collaborative cell, multi-modal, multi-sensor, perception, robotics, cobots

Abstract The ultimate goal of a collaborative manufacturing cell envisions a space where humans and robots work on a standard task with efficiency and complete safety. Nowadays, most collaborative cells required human operators to perform unnatural gestures to communicate with the robot. Additionally, these systems display limited interaction between humans and robots: often, the reaction to unexpected motions from the human operator is solved by halting the robot's movement. This plan proposes to tackle these limitations by developing advanced perceptual mechanisms based on multiple sensors of multiple modalities. Multi-modal because different modalities bring different types of information that give rich information about inside the cell. Multi-sensor to ensure that the cell is fully covered, even when there are unforeseen occlusions. Using advanced and robust perception will guarantee a language between the human and the robot, where the automatic system supports the communication burden, creating an efficient operation.

acknowledgement of use of AI tools/reconhecimento do uso de ferramentas IA

Recognition of the use of generative Artificial Intelligence technologies and tools, software and other support tools.

I acknowledge the use of Grammarly (www.grammarly.com/) and ChatGPT 4.0 (Open AI, https://chat.openai.com) to proofread the English of complex sentences and the use of Overleaf (https://www.overleaf.com) for text writing and productivity.
I acknowledge the use of Jetbrains tools (www.jetbrains.com) for code writing.

scientific papers produced and published in support of this thesis/artigos científicos produzidos ou publicados com base nesta tese

Rato D., Oliveira M., Santos V., Sappa A., Raducanu B. (2024), Multi-View 2D to 3D Lifting Video-Based Optimization: A Robust Approach for Human Pose Estimation with Occluded Joint Prediction. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems, doi:10.1109/IROS58592.2024.10802200

Rato D., Oliveira M., Santos V., Gomes M., Sappa A. (2022), A sensor-to-pattern calibration framework for multi-modal industrial collaborative cells. In: Journal of Manufacturing Systems (Journal), doi: 10.1016/j.jmsy.2022.07.006

Oliveira M., Pedrosa E., Aguiar A., Rato D., Neves F., Dias P., Santos V. (2020), ATOM: A general calibration framework for multi-modal, multi-sensor systems. In: Expert Systems with Applications (Journal), doi: 10.1016/j.eswa.2022.118000

Contents

List of Figures

List of Tables

List of Code Snippets

Glossary

LiDAR	Light Detection And Ranging
ROS	Robot Operating System
OpenCV	Open Source Computer Vision Library
ICP	Iterative Closest Point
ATOM	Atomic Transformations Optimisation Method
SLAM	Simultaneous Localization and Mapping
HRI	Human-Robot Interaction
MPJPE	Mean Per Joint Position Error
FoV	Field-of-View
HPE	Human Pose Estimation
3DPCK	Percentage of Correct Keypoints
AUC	Area Under the Curve
LARCC	Laboratory for Automation and Robotics Collaborative Cell
fps	Frames per Second
RMSE	Root Mean Square Error

CHAPTER 1

Introduction

1.1 Background and Context

The integration of robots designed to collaborate with humans has expanded significantly across sectors such as manufacturing, logistics, healthcare, and services [1], [2]. In contrast to traditional industrial robots, typically confined to isolated workspaces, collaborative robotic systems are developed to operate in proximity to humans, either by virtue of compliant mechanics (e.g., torque or force control) or through advanced perception capabilities. This shift towards shared workspaces is part of a broader trend in robotics that prioritises flexibility, adaptability, and responsiveness to complex and dynamic human behaviour. Whether supporting surgical procedures or operating in public environments, such robots are increasingly required to function within settings marked by unpredictability and variation.

While the term collaborative robots (or cobots) is often used to describe robots with inherent mechanical safety features, this thesis adopts a broader interpretation. The emphasis here is on perception-based collaborative interactions, those that depend on rich, multi-modal sensor inputs to interpret their surroundings and human partners. This conceptual framing is deliberate: the contributions of the thesis focus not on mechanical compliance, but on active perception, which is fundamental to intelligent and adaptive robotic behaviour. By making this distinction explicit, the thesis avoids conflating its contributions with earlier generations of cobots and underscores its relevance to perception-driven systems.

Robust perception in collaborative interactions requires a high degree of situational awareness and environmental understanding. Robots must perceive their surroundings accurately, interpret human motion in real-time, and adjust their behaviour responsively [3]. This necessitates reliable sensor fusion and the ability to operate under partial observations or occlusions. Two aspects of perception are especially critical: accurate extrinsic calibration across heterogeneous sensors, and reliable 3D human pose estimation under realistic deployment conditions.

Modern robotic platforms often incorporate combinations of RGB cameras, RGB-D sensors, and LiDAR devices. These may be fixed in the environment, mounted on robot bodies, or

attached to articulated arms. Integrating their outputs requires precise extrinsic calibration, especially where sensor fields of view only partially overlap or when the robot is in motion [4], [5]. However, the diversity of sensor modalities, asynchronous data streams, and the absence of shared features between modalities pose significant calibration challenges. Current approaches are often tailored to specific sensor pairings and fail to generalise to complex, mixed-modality configurations [6].

In parallel, estimating the 3D pose of humans in collaborative environments remains a difficult problem. Although multi-view approaches have progressed rapidly in recent years, their robustness diminishes in the presence of occlusions, sparse viewpoints, or noisy detections [7], [8]. In real-world scenarios, people may be partially visible due to obstructions or body posture, and their movements are often unpredictable. As a result, perception pipelines must be both real-time and resilient, capable of delivering accurate full-body estimates from incomplete data and integrating those estimates into robotic decision-making frameworks.

1.2 Motivation and Problem Statement

Despite substantial progress in the fields of sensor calibration and human pose estimation, current solutions tend to address these challenges in isolation. This fragmentation limits the scalability, generality, and real-world applicability of robotic perception systems. There is a pressing need for end-to-end pipelines that unify these tasks and accommodate the constraints of collaborative environments.

Several research gaps remain at the intersection of sensor calibration and human motion estimation. On the calibration side, there is a lack of general-purpose tools for multi-modal sensor networks that simultaneously handle RGB, RGB-D, and LiDAR devices. Factory calibrations are often insufficient when sensors are deployed in arbitrary arrangements or subjected to dynamic motion. Hand-eye calibration, involving sensors mounted on robotic arms, poses additional challenges, especially when depth data or constrained trajectories are involved. Existing approaches generally lack the flexibility and robustness required for such scenarios.

Regarding 3D pose estimation, most methods still depend heavily on high-quality 2D joint detections and often underperform when faced with occlusions or poor visibility. Deep learning models, although effective in curated datasets, frequently fail to generalise to unstructured settings. Moreover, pose estimation is rarely developed with explicit consideration for robotic integration, real-time responsiveness, or safety monitoring. Temporal continuity, anatomical constraints, and motion priors are underutilised despite their potential benefits.

Both perception components, calibration and pose estimation, also suffer from poor modularity and reproducibility. Tools that are accurate, accessible, and interpretable are still scarce, yet they are essential for practical deployment in human-robot collaboration.

This thesis is motivated by the need to bridge this gap by developing a unified, deployable framework for perception in collaborative robotics. The central hypothesis is that accurate sensor calibration and robust human pose estimation must be treated as interdependent challenges in order to achieve effective robot behaviour in shared workspaces.

The first research objective is to propose an extension of the Atomic Transformations Optimisation Method (ATOM) calibration framework that supports RGB, RGB-D, and LiDAR devices in both static and dynamic configurations. The method introduces a sensor-to-pattern optimisation strategy, supports hand-eye scenarios, and integrates into the Robot Operating System (ROS). It is designed to cope with practical constraints such as partial sensor overlap and constrained motion.

The second objective is to design a video-based, multi-view 3D human pose estimation method. This method incorporates temporal optimisation, anatomical priors, and uncertainty modelling to infer full-body poses from noisy or sparse 2D observations. It targets scenarios with limited camera coverage and is engineered for real-time use in robotics.

Together, these contributions aim to enhance the accuracy, robustness, and real-world usability of perception systems in collaborative settings. By addressing core perceptual challenges in a unified way, the thesis contributes towards safer, more adaptive, and more intelligent robotic collaboration in sensor-rich, human-populated environments.

1.3 Research Questions and Hypothesis

This thesis investigates two interrelated but independently addressed challenges in collaborative robotics: (i) the extrinsic calibration of heterogeneous and dynamic sensor configurations, and (ii) the estimation of 3D human pose using multi-camera RGB setups. These challenges are approached with the goal of improving spatial perception in environments where humans and robots operate in close proximity, often sharing tasks or space.

1.3.1 Research Questions

The investigation is guided by the following research questions:

1. **RQ1:** How can calibration algorithms be designed to support heterogeneous, multi-modal sensor setups—including both fixed and mobile configurations, while reducing reliance on manual procedures and maintaining spatial accuracy?
2. **RQ2:** What optimisation strategy enables consistent and accurate extrinsic calibration in dynamic environments where sensor positions or orientations may vary over time?
3. **RQ3:** How can 3D human pose estimation systems based solely on RGB imagery be configured to offer sufficient accuracy, robustness to occlusion, and computational efficiency for integration into collaborative robotic environments?

1.3.2 Research Hypothesis

The central hypothesis underpinning this thesis is as follows:

> *Robust extrinsic calibration methods for heterogeneous and dynamic sensor configurations, alongside reliable multi-view human pose estimation strategies, provide essential perceptual foundations for improving spatial awareness and safety in collaborative robotic systems.*

This hypothesis reflects the view that while calibration and pose estimation are developed as independent subsystems within this work, both are instrumental in enabling reliable, responsive, and human-aware perception pipelines. The research does not seek to demonstrate a direct causal link between them, but rather to show how advances in each contribute toward the shared goal of enabling perception-driven human-robot collaboration.

1.4 Objectives

The overarching aim of this thesis is to contribute to the development of perceptually aware collaborative robotic cells, with a focus on two fundamental enablers: multi-modal sensor calibration and 3D human pose estimation. These capabilities are essential for ensuring safe, efficient, and intelligent human-robot interaction (HRI) in industrial and manufacturing environments. While collaborative robots (cobots) have become increasingly common in production settings, their deployment is still constrained by limited environmental perception, inflexible calibration procedures, and a lack of human-awareness capabilities.

To address these limitations, the thesis pursues two core research objectives, each targeting a critical gap in the current state of the art:

- **Objective 1:** *To investigate how the existing ATOM calibration framework can be applied and extended for use in robotic collaborative cell scenarios.*
  The ATOM framework, initially developed as a general calibration solution for multi-modal, multi-sensor systems, already supports various sensor types and configurations. This thesis builds upon ATOM to assess its applicability in real-world robotic cells and to extend its capabilities where needed, most notably to support the calibration of depth sensors. Rather than creating a new calibration system from scratch, the contribution lies in adapting and enhancing ATOM to address domain-specific requirements, including modularity, support for RGB-D, and integration within robot operating environments.
- **Objective 2:** *To design and implement a 3D human pose estimation framework tailored to collaborative robotics.*
  Reliable detection and tracking of human posture is essential for safe and effective operation in human-robot shared environments. Although human pose estimation (HPE) has advanced rapidly, most systems have been developed outside the robotics context, often trained on unrealistic datasets or under assumptions that do not hold in constrained, cluttered, or occlusion-prone industrial scenarios.
  This thesis develops a multi-camera RGB-based pose estimation solution, optimised for robotic integration. It explores trade-offs in camera placement, algorithmic performance, and robustness to occlusion. The system is evaluated on both technical accuracy and usability, with the goal of enabling human-aware robot behaviours such as anticipatory motion planning or adaptive safety zones.

The outcomes of these two objectives aim to support the broader vision of integrating robust perception systems into collaborative robotic platforms that must operate safely, flexibly, and efficiently alongside humans.

1.5 Overview of Methodology and Evaluation Strategy

This thesis adopts a comprehensive methodology combining algorithm design, system development, and empirical validation to advance two critical components of perceptually aware collaborative robotic systems: extrinsic calibration of multi-modal, multi-sensor setups, and 3D human pose estimation using multiple RGB cameras. The approach is based on the requirements of real-world robotic applications, prioritising robustness, automation, and modularity.

The first core strand involves the development of an automatic, optimisation-based calibration pipeline designed to handle heterogeneous sensor systems, including fixed sensors and those mounted on robotic arms. The calibration process is formulated as a non-linear optimisation problem that minimises spatial and reprojection errors across sensor pairs, using a movable calibration target. It supports RGB, RGB-D, and LiDAR modalities, and includes procedures for hand-eye calibration. Implemented within the ROS framework, the system is designed for scalability and integration into collaborative robotic environments.

To evaluate the calibration solution, the thesis employs both synthetic and real-world setups. Controlled experiments are conducted in a laboratory robotic cell, using known geometric configurations and high-precision ground truth provided by simulation. Performance is measured using reprojection error across different modality combinations in both simulated and real-world data. Simulation-based evaluations allow controlled variation of parameters such as baseline distances, field of view overlap, and occlusion levels. Comparative baselines include established calibration tools such as Kalibr [9], against which the developed method is benchmarked for accuracy.

The second strand of the thesis focuses on the design of a 3D human pose estimation pipeline based on synchronised RGB video from multiple cameras. The system leverages existing deep-learning-based 2D keypoint detectors as a front-end and performs triangulation through a calibrated multi-camera setup to recover 3D joint positions. The framework is designed to be lightweight and compatible with real-time robotic systems, incorporating temporal filtering and spatial reasoning to maintain stability in the presence of occlusions, viewpoint variation, and rapid motion. Camera extrinsics are derived from the proposed calibration pipeline, ensuring full spatial consistency between the perception and control layers of the robotic system.

The Human Pose Estimation (HPE) component is evaluated using public datasets such as MPI-INF-3DHP [10], which provide initial comparisons for 3D joint localisation error and detection robustness. Evaluation metrics include 3D joint localisation error and percentage of correct keypoints (3DPCK). The system's real-time performance is assessed in end-to-end integration with a collaborative robot, focusing on response times and consistency in triggering safety or behavioural responses based on human motion.

By separating the evaluation of the calibration and HPE components, the thesis provides a transparent assessment of each system's capabilities and limitations. This dual-layered strategy ensures that each technical contribution is validated both in isolation and in the

context of its integration into a perceptually aware collaborative robotic cell.

1.6 Contributions

This thesis addresses fundamental challenges in multi-sensor calibration and 3D human pose estimation in collaborative robotics. It proposes novel methodological and system-level contributions aimed at improving accuracy, automation, and robustness in human-robot interaction contexts. The research spans from the design of calibration pipelines for heterogeneous sensor setups to the development of real-time, multi-view pose estimation systems suitable for dynamic and safety-critical environments.

The contributions can be grouped into two main domains:

1. **Multi-modal Sensor Calibration:** The thesis introduces an extrinsic calibration framework that supports multiple sensor types (RGB, RGB-D, LiDAR), including static and dynamic configurations. It offers an adaptable method for calibrating hand-eye and inter-sensor transforms within a unified system, implemented in ROS and validated in real-world robotic cells.
2. **3D Human Pose Estimation:** A multi-camera RGB-based pose estimation system is presented, enabling markerless, real-time inference of human skeletal pose with enhanced spatial consistency. The proposed approach is designed to function reliably under occlusions, varying lighting conditions, and changing viewpoints.

The main publications that result from this research:

- Rato D., Oliveira M., Santos V., Gomes M., Sappa A. (2022), A sensor-to-pattern calibration framework for multi-modal industrial collaborative cells. In: Journal of Manufacturing Systems (Journal), doi: 10.1016/j.jmsy.2022.07.006
- Rato D., Oliveira M., Santos V., Sappa A., Raducanu B. (2024), Multi-View 2D to 3D Lifting Video-Based Optimization: A Robust Approach for Human Pose Estimation with Occluded Joint Prediction. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems, doi:10.1109/IROS58592.2024.10802200
- Rato D., Oliveira M., Santos V., Sappa A., New Methodology to Calibrate Depth Sensors in Multi-Modal Dynamic Setups. Submitted: IEEE Access

Supplementary publications developed in parallel with or in support of this research:

- Oliveira M., Pedrosa E., Aguiar A., Rato D., Neves F., Dias P., Santos V. (2020), ATOM: A general calibration framework for multi-modal, multi-sensor systems. In: Expert Systems with Applications (Journal), doi: 10.1016/j.eswa.2022.118000
- Santos V., Dias P., Oliveira M., Rato D. (2022), Multimodal Sensor Calibration Approaches in the ATLASCAR Project. In: ICT Applications for Smart Cities (Book Chapter), doi: 10.1007/978-3-031-06307-7_7

1.7 Document Structure

The document is organised as follows:

- **Chapter 2** presents a comprehensive literature review that lays the theoretical and methodological foundations for the thesis. It is structured around two key pillars of perception in collaborative robotic systems: extrinsic calibration of multi-modal sensor setups and 3D human pose estimation from multi-camera RGB systems. The chapter begins by analysing the challenges of calibrating multi-modal sensors. It reviews target-based and targetless calibration strategies, including recent advances in motion-based, mirror-based, and learning-driven approaches, while highlighting their trade-offs in terms of accuracy, scalability, and deployability. The review then turns to hand-eye calibration techniques, examining both classical formulations and recent efforts to address RGB-D integration and modular robot architectures. In the second part, the chapter surveys the state of the art in 3D human pose estimation, comparing geometric triangulation, volumetric fusion, and learning-based models, including transformers and diffusion networks, while discussing their robustness, accuracy, and relevance to real-world, industrial applications. Together, these insights contextualise the research gaps addressed by this thesis and inform the design of the proposed calibration and perception framework.
- **Chapter 3** presents a sensor-to-pattern calibration framework tailored for complex, multi-modal collaborative robotic cells, addressing the challenge of achieving accurate extrinsic calibration in environments where numerous RGB, depth, and LiDAR sensors often lack overlapping fields of view. Departing from conventional sensor-to-sensor calibration techniques, the chapter introduces a unified optimisation-based approach that aligns each sensor to a common calibration pattern, enabling robust data fusion across modalities. It details the theoretical foundations of this method, the implementation of automatic and manual labelling mechanisms for different sensor types, and the deployment of a realistic simulation environment using ROS and Gazebo. The chapter also evaluates the framework through both simulated and real-world experiments, demonstrating its superiority over existing tools like OpenCV, Kalibr, and ICP in terms of calibration accuracy and scalability.
- **Chapter 4** introduces an extended calibration methodology designed to accurately align depth sensors alongside RGB and LiDAR devices in dynamic, multi-modal robotic systems, including hand-eye configurations. Building upon the ATOM framework, this chapter expands its capabilities to support depth modality calibration through a sensor-to-pattern approach, which optimises both sensor and pattern poses simultaneously. The methodology is validated in both simulated and real-world environments, demonstrating its capacity to calibrate fixed and mobile sensors even under partial detections or limited field-of-view overlap. It includes the development of dedicated cost functions for depth sensors, along with semi-automated labelling tools tailored to the nature of range data. The chapter presents extensive experimental results, first with standalone RGB-D systems, improving on factory calibrations, and then with a complex multi-sensor setup featuring a robotic manipulator. Across all tests, the method exhibits sub-pixel to low-pixel accuracy and surpasses existing tools, particularly in scenarios involving

non-overlapping, cross-modality sensor configurations.

- **Chapter 5** presents a robust multi-view, video-based approach for 3D human pose estimation, designed to handle occlusions and noise in collaborative robotic environments. By combining 2D-to-3D lifting with temporal optimisation and skeletal constraints, the method predicts occluded joints with high accuracy using limited camera views. The chapter details the algorithm's design, real-time adaptation, and evaluation on benchmark datasets, demonstrating its effectiveness under varying conditions, including joint detection error, partial visibility, and reduced camera setups.
- **Chapter 6** revisits the research objectives and synthesises the core contributions of the thesis, particularly in the areas of sensor calibration and 3D human pose estimation for collaborative robotic systems. It reflects on the methodological and practical limitations encountered, outlines key implications for robotic perception, and proposes promising directions for future research and deployment. The chapter consolidates the insights developed throughout the thesis and positions them within the broader context of human-aware and adaptable robotics.

CHAPTER 2

Literature Review

2.1 Introduction: Context and Perceptual Challenges

Collaborative robotic systems, particularly those deployed in manufacturing cells, represent a fundamental evolution in industrial automation, shifting from isolated, rigid automation toward dynamic, human-centred work environments. In these systems, human operators and robots share physical workspaces, perform interdependent tasks, and adapt in real time to changing operational conditions. The core objective of collaborative cells is to combine the precision, repeatability, and strength of robots with the contextual understanding, flexibility, and problem-solving capabilities of human workers.

This paradigm is central to both Industry 4.0 and emerging Industry 5.0 frameworks, which envision highly flexible, reconfigurable, and semantically enriched production systems. In collaborative cells, the division of labour between human and robot is fluid rather than fixed, requiring situational awareness, adaptive task allocation, and robust safety mechanisms. Such environments are characterised by spatial complexity, non-deterministic workflows, and high variability in human behaviour and sensor visibility. These constraints create significant technical challenges for perception, control, and system-level coordination.

Among these challenges, precise spatial perception and robust human–machine interaction stand out as enabling capabilities. Robots must operate with full awareness of their surroundings, dynamically adapting to human motion, object changes, and evolving task contexts. This requires the integration of heterogeneous sensors, including RGB cameras, depth sensors (e.g., structured light or time-of-flight), LiDARs, and panoramic or stereo vision systems, to construct a consistent and up-to-date spatial model of the environment. At the same time, robots must detect and interpret human pose, gestures, and intention to ensure safe and fluent interaction.

Two core perceptual problems arise in this context:

- Extrinsic calibration, the estimation of the rigid-body transformations between the coordinate frames of each sensor, is essential for fusing data into a unified spatial reference frame;

- 3D human pose estimation, the task of accurately reconstructing the body configuration of human operators, is critical for predicting motion, enforcing safety boundaries, and enabling collaborative planning.

Achieving these capabilities is especially demanding in collaborative cells, where sensors may be fixed or mobile, overlapping or non-overlapping in field of view, and subject to variable lighting and occlusion. Standard calibration routines often fail in such environments, particularly when sensors cannot simultaneously observe the same calibration pattern or must remain fixed during operation. Likewise, many human pose estimation algorithms assume clean, unobstructed views or known camera parameters, which are not always available in real-world scenarios.

These two perceptual problems, extrinsic calibration and 3D human pose estimation, are not only challenges in their own right, but also foundational dependencies for many of the higher-level approaches currently shaping collaborative robotics. The remainder of this section examines several such contributions. While diverse in their methods and scope, these approaches are either designed to directly mitigate these perception-related limitations or are predicated on the assumption that such perceptual capabilities are already robustly in place. As such, each must be understood in relation to how it leverages, constrains, or advances the two critical problems outlined above.

While the research contributions reviewed in the following paragraphs span different technical layers—from control and planning to semantics, monitoring, and simulation—they are closely linked by their dependence on the two perceptual capabilities identified above. In particular, extrinsic calibration underpins the reliable fusion of heterogeneous sensor data, enabling coherent spatial reasoning and consistent perception across the system. Similarly, 3D human pose estimation is foundational for anticipating human actions, ensuring safety, and enabling fluid interaction. Some approaches aim to directly improve interaction by assuming these capabilities as a given (e.g., gesture-based control or semantic interpretation), while others focus on higher-level coordination that implicitly relies on accurate perception to function correctly. Recognising this dependency allows us to view these contributions not as disconnected innovations but as components within a layered architecture—each advancing the broader objective of seamless human–robot collaboration by either addressing or operationalising perception in complex environments.

Recent research shifts toward more advanced, digitally enabled collaborative cells that incorporate learning-based interaction, multi-modal perception, and semantic awareness. Baptista et al. [11] develop a laboratory-scale collaborative cell featuring deep learning modules for gesture recognition, contact classification, and human intention anticipation, coordinated through a ROS-based architecture. This approach enables robots to adapt their behaviour in response to both explicit commands (e.g., hand gestures) and implicit cues (e.g., hand-object interactions), thereby creating more natural and fluent interactions.

At the level of control and execution, real-time adaptation becomes a key design objective. Wei et al. [12] propose a convex optimisation framework for online trajectory generation in dual-robot cells, capable of dynamically replanning in response to human presence and shifting

task priorities. Tonola et al. [13] introduce an anytime-informed re-planning algorithm that balances human safety and task efficiency, with an emphasis on trajectory legibility to improve human trust and predictability.

Complementing low-level control are advances in semantic reasoning and ontological modelling. The SOHO ontology (Sharework Ontology for Human-Robot Collaboration) is introduced by Umbrico, Orlandini, and Cesta [14] as a domain-specific ontology designed to endow collaborative robots with context-aware reasoning capabilities in HRC settings. Built upon foundational ontologies, SOHO structures knowledge into layered contexts, environment, behaviour, and production, enabling the formal representation of both observable properties (e.g., posture, motion) and abstract constructs (e.g., intentions, skills, risk levels). This framework supports robots in interpreting human actions, inferring production goals, and dynamically adapting plans in response to changing workplace conditions. Later developments by Umbrico et al. [15] refine this model and demonstrate its applicability in real industrial scenarios, enhancing robot awareness and decision-making in collaborative manufacturing environments.

To support real-time adaptation, multi-modal monitoring systems also emerge. Argyrou et al. [16] develop a modular data fusion platform that integrates sensor data from robotic controllers, human tracking modules, and safety systems to generate a shared situational model of the collaborative workspace. Such systems allow not only reactive responses to human actions but also proactive anticipation of task state transitions.

Finally, the emergence of digital twin technologies enables simulation-driven optimisation of collaborative cell layouts and behaviour prior to deployment. Cella et al. [17] introduce a digital twin architecture that uses AI-guided simulation to generate optimal robot programs and task allocations, significantly reducing the design-to-deployment gap and improving system resilience.

These advances in modelling, planning, and semantic reasoning increasingly depend on robust perceptual capabilities to operate reliably in real-world environments. As collaborative cells become more adaptive and simulation-driven, the need for accurate, real-time understanding of both the workspace and human behaviour becomes even more critical. In this context, perception is not merely a supporting function but a core enabler of intelligent collaboration.

This chapter builds on these developments by offering a critical review of the state of the art in extrinsic calibration and 3D human pose estimation, two perceptual foundations of collaborative robotics. The first part surveys calibration techniques across sensor modalities and spatial configurations, including target-based, targetless, and learning-driven methods. The second part explores multi-camera 3D pose estimation approaches, comparing geometric, volumetric, and transformer-based architectures in terms of their accuracy, robustness, and deployability in collaborative cell settings. Together, these analyses lay the groundwork for the perceptual framework proposed in this thesis, aimed at enabling safe, scalable, and semantically aware collaboration between humans and robots.

2.2 Extrinsic Calibration

Extrinsic calibration is a foundational requirement in multi-sensor robotic systems. It consists in estimating the rigid-body transformation between the coordinate frames of different sensors, enabling all measurements to be expressed within a common spatial reference frame. In collaborative robotic environments, where perception must be both precise and real-time, this calibration is critical to ensure coherent sensor fusion. These systems often integrate RGB cameras, depth sensors (e.g., structured light, time-of-flight), LiDARs, panoramic cameras, radars, and IMUs, each being characterised by distinct fields of view, resolutions, and measurement principles.

Figure 2.1 illustrates the concept of extrinsic calibration using homogeneous coordinates. Let sensor A and sensor B observe a common calibration target P. The pose of each sensor relative to the target is expressed by a transformation matrix in (2.1), mapping coordinates from the target frame to the sensor frame:

$$\mathbf{T}_{AP} = \begin{bmatrix} \mathbf{R}_{AP} & \mathbf{t}_{AP} \\ \mathbf{0}^\top & 1 \end{bmatrix}, \quad \mathbf{T}_{BP} = \begin{bmatrix} \mathbf{R}_{BP} & \mathbf{t}_{BP} \\ \mathbf{0}^\top & 1 \end{bmatrix}, \tag{2.1}$$

where $\mathbf{R} \in \mathbb{R}^{3\times3}$ is a rotation matrix and $\mathbf{t} \in \mathbb{R}^{3\times1}$ is a translation vector. The extrinsic transformation between sensor A and sensor B, $\mathbf{T}_{AB}$, is computed as:

$$\mathbf{T}_{AB} = \mathbf{T}_{AP}\,\mathbf{T}_{BP}^{-1}, \tag{2.2}$$

allowing a point $\mathbf{X}_B \in \mathbb{R}^{4\times1}$ measured in the homogeneous coordinate frame of sensor B to be expressed in the frame of sensor A via:

$$\mathbf{X}_A = \mathbf{T}_{AB}\,\mathbf{X}_B. \tag{2.3}$$

This formulation is general, efficient, and central to applications involving multi-sensor fusion, 3D reconstruction, and spatial alignment in collaborative robotic systems.

In practical applications, the effectiveness of multi-sensor systems hinges on the precision of these extrinsic transformations. Whether in relatively straightforward cases such as RGB–depth fusion [18], or in more complex environments like intelligent vehicles [19], [20], smart camera networks [21], underwater stereo systems [22], robotic inspection platforms [23], or satellite-based image integration [24], extrinsic calibration ensures geometric consistency and enables accurate data association across heterogeneous streams.

While traditional methods rely on geometric patterns and controlled motion to perform calibration, recent efforts aim to automate the process and reduce dependency on structured targets. Learning-based strategies emerge to tackle calibration in dynamic and unstructured scenarios, with promising results. For instance, Yaqing and Huaming [25] introduce a deep learning framework combining depth and height supervision with an attention-based fusion module, achieving robust calibration without predefined markers. Such approaches represent a broader shift in calibration research, from static procedures toward adaptive, data-driven pipelines better suited to the demands of modern robotics.

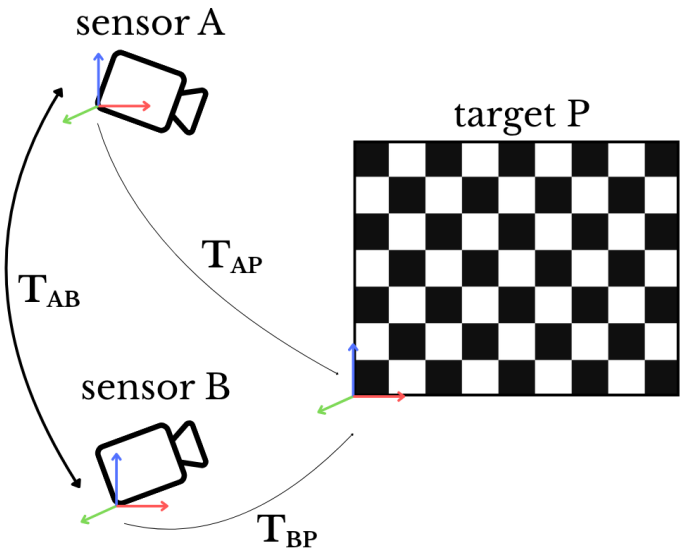

Figure 2.1: Extrinsic calibration between two sensors using a common reference target P. Each sensor estimates its pose relative to the target using a rigid transformation in homogeneous coordinates. The relative pose $\mathbf{T}_{AB}$ is obtained as $\mathbf{T}_{AB} = \mathbf{T}_{AP}\mathbf{T}_{BP}^{-1}$.

2.2.1 Sensor Combinations and Configurations

The majority of the literature focuses on pairwise calibration of sensor combinations, for example, RGB-RGB [26]–[30], RGB-depth [31]–[34], and RGB-LiDAR [35]–[42]. Traditional techniques, such as Zhang's method [8], prove particularly effective for RGB-RGB calibration in overlapping fields of view, typically leveraging checkerboard or Charuco patterns [43], [44] visible to multiple sensors. Extensions of these methods significantly improve accuracy in calibrating short-range depth sensors [33], [45], [46], achieving reprojection errors in the sub-millimetre to pixel range.

Figure 2.2 illustrates a typical setup in which multiple sensors share overlapping fields of view. In such configurations, a common calibration target can be simultaneously observed by all sensors, facilitating robust extrinsic calibration through standard pairwise or global optimisation methods.

However, the pairwise nature of these methods introduces limitations when applied to larger sensor networks. The number of required transformations increases combinatorially with the number of sensors, complicating calibration and amplifying cumulative errors. For instance, to calibrate sensors A, B, and C, one may choose direct or indirect paths (e.g., A–C, A–B–C, or B–A–C), with no guarantee of minimal error, particularly when dealing with heterogeneous modalities or non-symmetric objective functions. Moreover, the calibration order (A to B vs. B to A) influences outcomes, introducing ambiguity and error propagation across the network.

Modern sensor configurations, especially in collaborative cells, further complicate the

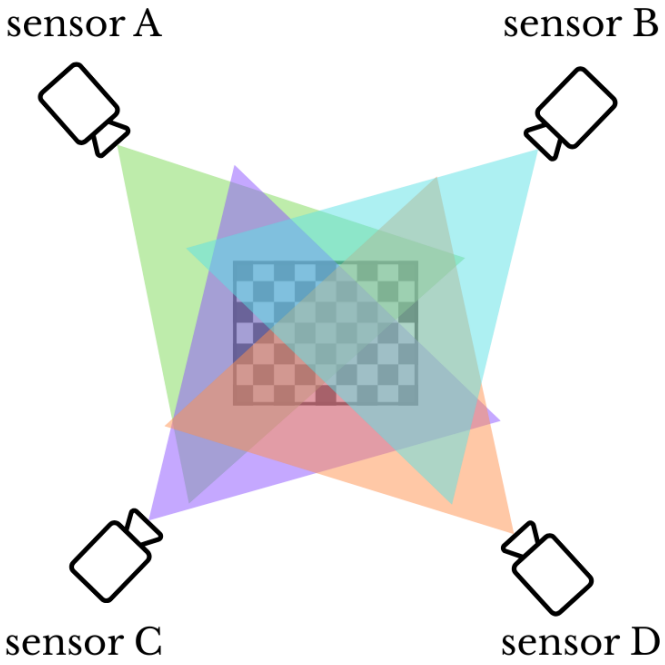

Figure 2.2: Example of overlapping fields of view (FoVs) among four sensors observing a common calibration pattern. This setup supports standard pairwise or global calibration using shared visual features.

process. Sensors are often mounted around the workspace in non-overlapping arrangements to avoid occlusion and to ensure safety coverage. In these scenarios, no single calibration pattern is visible to all sensors simultaneously, requiring alternative calibration strategies. Figure 2.3 shows such a case where each sensor observes only a portion of the environment, with minimal or no shared field of view. This layout reflects real-world setups in collaborative workcells, where static infrastructure and safety constraints preclude frequent repositioning.

While methods like that of Raposo et al. [47] address these challenges through mirror-based calibration, they still depend on pairwise registration and require physical movement of the sensor, conditions often incompatible with fixed installations in collaborative cells.

In response to these limitations, Oliveira et al. [48] propose ATOM, a general calibration framework capable of handling multi-sensor, multi-modal configurations without relying on rigid pairwise connections. ATOM reformulates calibration as a global optimisation problem over a transformation graph, where calibration constraints apply across sensor-to-sensor, sensor-in-motion, and sensor-to-frame configurations. Its generality stems from operating on indivisible, atomic transformations, thereby allowing the calibration of complex topologies within a unified framework. Notably, ATOM also integrates seamlessly with ROS and is validated across diverse robotic platforms, from autonomous vehicles to agricultural robots and collaborative arms [49], [50].

Recent contributions further advance calibration accuracy by leveraging novel geometric feature extraction techniques. Hua et al. [51] exploit edge correspondences from both LiDAR and camera data, enabling calibration using everyday objects like books or boxes. Kim et

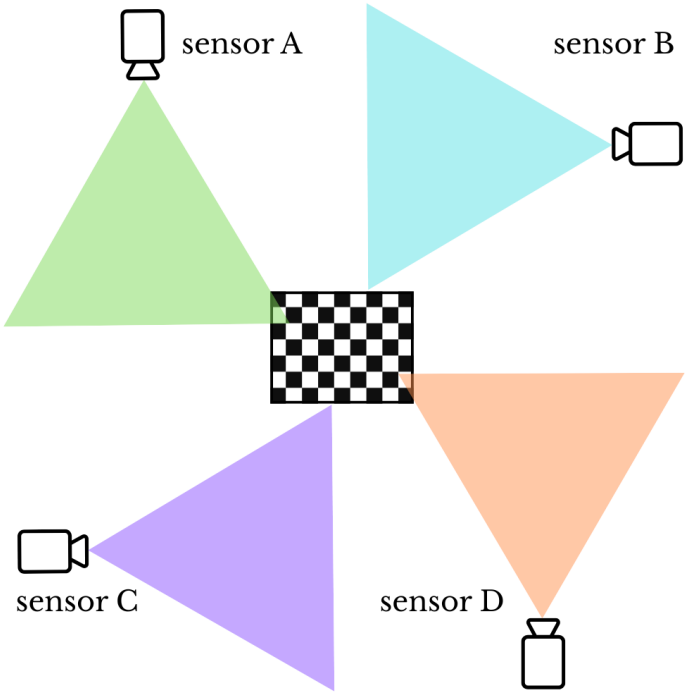

Figure 2.3: Example of non-overlapping fields of view (FoVs) in a collaborative robotic setup. Sensors are arranged to cover different zones of the workspace, preventing the use of conventional target-based calibration with full visibility.

al. [52] employ circular feature geometry combined with constrained optimisation, improving robustness under sensor noise. These approaches reflect a broader shift toward more flexible, accurate, and modular calibration pipelines.

These challenges become even more pronounced in RGB-D systems, motivating a closer examination of dedicated depth calibration techniques, as discussed in the next section.

2.2.2 RGB-RGB Calibration Methods

Calibration between RGB cameras, often referred to as stereo or multi-camera calibration, is a well-established field in computer vision. The goal is to estimate the relative pose between cameras such that measurements captured from each can be expressed in a common reference frame. Classical approaches rely on structured targets observed simultaneously across overlapping fields of view.

Zhang's planar method [8] remains the most widely used approach, employing multiple views of a checkerboard to jointly estimate intrinsic and extrinsic parameters. Subsequent works refine this pipeline to improve accuracy and robustness. For instance, Su et al. [26] and Dinh et al. [29] propose efficient stereo calibration routines, while Ling et al. [27] and Mueller et al. [28] address long-baseline calibration and robustness under geometric distortions.

Charuco boards, a hybrid of checkerboard grids and ArUco markers, gain popularity for improving marker detection in adverse conditions such as low lighting or oblique viewing angles. Studies by Romero-Ramirez et al. [44] and Hu et al. [43] highlight their performance advantages in real-world setups. Garrido-Jurado et al. [53] further contribute by optimising

ArUco marker dictionaries to improve detection rates and minimise false positives.

Recent research extends calibration to dynamic scenarios. Wu et al. [30] introduce a method that leverages temporal consistency in a real-time stereo pipeline. Others like Mueller et al. [28] explore multi-camera networks with partial overlaps and distributed topologies, relevant for large-scale environments such as warehouses or assembly cells.

In summary, RGB-RGB calibration methods are mature, reliable, and well-supported by toolkits. Current research focuses on increasing automation, supporting partial overlaps, and improving calibration in the absence of structured targets.

2.2.3 RGB-LiDAR Calibration Methods

Calibrating RGB cameras with LiDAR sensors poses distinct challenges due to the heterogeneity in sensing modalities: image-based projection versus spatial sampling. A robust transformation between these sensors is essential to merge rich texture information from images with accurate spatial geometry from LiDAR.

Classical approaches use shared planar calibration targets. Vasconcelos et al. [35] and Wang et al. [40] mount checkerboards or custom boards within both the camera's and LiDAR's Field-of-View (FoV). The calibration is typically achieved via plane fitting and pose estimation relative to the board. Guindel et al. [39] adapt these methods for vehicular platforms, using board detection for on-the-fly calibration. Other works, such as Yang et al. [41], employ triangulated targets or volumetric references to improve spatial alignment.

To reduce dependency on explicit targets, Zhou et al. [38] propose targetless calibration based on mutual information between image gradients and LiDAR reflectivity. This principle sees further extensions in recent years: Zhu et al. [54] introduce a reflectivity edge-matching algorithm; Wang et al. [55] propose a ray-traced mesh alignment scheme; and Yaqing and Huaming [25] combine depth, height, and learned features in a deep attention-based network.

Mirror-based methods are especially relevant for fixed installations where shared FoV is unavailable. Raposo et al. [47] and Huang et al. [37] use planar mirrors to simulate overlap between camera and LiDAR sensors. These configurations allow for robust calibration without moving the sensors, making them ideal for static environments.

Practical concerns such as mounting, environment reflectivity, and occlusions also influence method selection. While target-based methods typically yield sub-centimetre accuracy, targetless or motion-based techniques (e.g., Cattaneo et al. [56], Lin et al. [57]) prove more suitable for flexible or large-scale deployments. However, they often require more data and complex optimisation.

Overall, RGB-LiDAR calibration methods now span from pattern-dependent pipelines to targetless, learning-based frameworks. As these systems increasingly appear in dynamic and collaborative contexts, the trade-off between accuracy, autonomy, and scalability becomes central to design decisions.

2.2.4 RGB-D Calibration Methods

Extrinsic calibration of RGB-D systems is a well-established challenge in robotics and computer vision. It involves estimating the rigid-body transformation between the RGB

and depth components of an RGB-D sensor—a fundamental step for data fusion in multi-sensor setups [58]. Accurate calibration of these devices is especially crucial in fields such as autonomous robotics [59]–[61] and computer vision applications like object recognition or scene reconstruction [62], [63], where depth information enhances geometric understanding. Consequently, developing robust and precise calibration methodologies for RGB-D cameras becomes a significant research focus [64].

The calibration process typically starts with the identification of corresponding key points in both the RGB and depth images. These key points, usually extracted from artificial patterns, must then be associated across modalities, forming pairs that serve to compute the spatial transformation between the two sensor coordinate systems. This transformation allows projection of 3D points from one sensor's frame into the other, and the accuracy of the calibration is typically assessed by measuring the reprojection error between the estimated and observed points.

Most state-of-the-art methods rely on visual calibration patterns to ensure that key point detection and association are performed with high precision and reliability. Common approaches use chessboards or Charuco markers, which are well-supported by off-the-shelf detectors [43], [44], [53], [65]. While such patterns are easily detectable in RGB images, they often lack distinctive structure in the corresponding depth data, making direct matching between modalities more difficult.

To address this, some researchers develop custom calibration targets specifically designed to enhance feature detectability in depth maps. Park et al. [45], for instance, introduce a 3D Charuco tower, which enables calibration of multiple RGB-D cameras with limited overlap by providing features visible from different viewpoints. Chaochuan et al. [46] design a tower-like structure composed of circular markers, as shown in Fig. 2.4, which are detected via Hough transforms and refined using an adaptive cuckoo search algorithm. Their approach proves especially suitable for setups with restricted fields of view, as commonly found in stationary sensor arrangements.

Figure 2.4: Example of custom target [46].

In contrast to these target-based strategies, Horn et al. [66] propose a targetless calibration method based on dual optimisation. Their approach uses sensor ego-motion to match

corresponding features across frames without relying on artificial patterns. The authors distinguish between a fast optimiser, suitable for real-time applications, and a global optimiser, designed for more accurate offline calibration. Their results, obtained on both simulated data and the KITTI dataset [67], demonstrate translation errors of approximately 1 cm and rotation errors near 1°.

Other methods aim to estimate both extrinsic and intrinsic parameters simultaneously. Chen et al. [33] use heteroscedastic Gaussian process models to account for measurement uncertainty and propose a calibration procedure based on fixed chessboards augmented with fiducial markers. Their setup includes a motion capture system to track camera movement around the target, and results show reprojection errors of approximately 2 pixels. Chaochuan et al. [46] report reprojection errors of 2–4 pixels in their multi-camera setup, while Park et al. [45] achieve reconstruction errors as low as 3.5 millimetres.

Despite advances in pattern detection and optimisation strategies, challenges remain. Visual calibration patterns are often difficult to detect in cluttered or poorly lit environments, and precise pattern placement may not always be feasible. Patternless approaches like that of Horn et al. [66] offer more flexibility but may suffer from reduced accuracy or increased computational requirements.

Ultimately, the choice of calibration method depends on the specific application context: target-based methods tend to offer higher accuracy under controlled conditions, while targetless approaches prove better suited for in situ deployment in dynamic or constrained environments. The following section extends this discussion to hand-eye calibration, a crucial component in robotic systems that involve moving sensors or manipulators.

2.2.5 Hand-Eye Calibration Methods

Hand-eye calibration constitutes another essential aspect of multi-sensor systems. Rooted in the classical $AX = XB$ formulation [68], [69], contemporary approaches include closed-form and iterative solutions [70], [71]. Applications span robotic grasping, perception, and navigation, as shown in the works of Fu et al. [72], Liang et al. [73], and Pan et al. [74]. Techniques often involve decoupling camera calibration from robotic kinematics to improve system modularity [75], [76].

Two configurations dominate: eye-in-hand, which calibrates the transformation between the camera and the end-effector; and eye-to-base, which relates the sensor to the robot base. The $AX = XB$ problem also extends to $AX = ZB$ [77], allowing greater generalisation, as illustrated in Fig. 2.5.

While extensive work focuses on RGB sensors [78]–[80], increasing attention now turns to RGB-D calibration. Jiang et al. [70] propose both closed-form and reprojection-based methods, achieving rotation errors around 0.3 rad and translation errors near 3 mm in real-world scenarios.

A significant contribution in this domain is the work by Pedrosa et al. [49], which generalises hand–eye calibration by modelling it as a factor graph composed of atomic transformations. This formulation enables the flexible combination of heterogeneous constraints (e.g., kinematic,

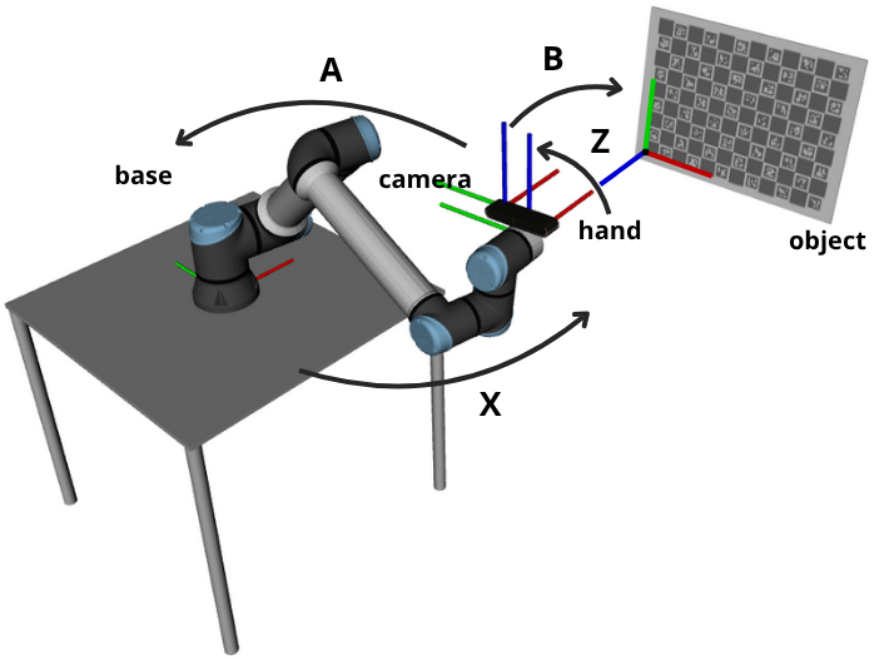

Figure 2.5: Representation of the $AX = XB$ hand-eye calibration philosophy.

visual, temporal) in both eye-in-hand and eye-to-base configurations, and demonstrates improved convergence and robustness under real-world noise conditions.

Importantly, most RGB-D hand-eye calibration methods treat RGB and depth units as inseparable, relying on factory calibration [72]–[74]. Some approaches decouple RGB and depth alignment entirely from hand-eye estimation [71], [75]. In contrast, our framework allows independent calibration of RGB and depth components and supports simultaneous calibration of mobile and fixed sensors from different modalities within a unified scheme.

While hand-eye calibration addresses dynamic scenarios involving mobile sensors, recent research also explores broader trends such as automatic and targetless methods applicable across various modalities.

2.2.6 Comparative Overview of Calibration Methods

Extrinsic calibration methods span a wide spectrum of sensor combinations, spatial arrangements, and calibration targets. To provide a more integrative understanding of the sensor calibration literature, Table 2.1 presents a comparative overview of representative extrinsic calibration methods across different sensor modalities. The table compiles key information on the number and type of sensors involved, calibration targets, FoV overlap requirements, and reported accuracy levels.

Table 2.1: Comparison of selected extrinsic calibration methods based on number of sensors, modalities, type of target, field of view (FoV) constraints, and accuracy.

Ref.	# Sensors	Modalities	Target	FoV	Accuracy
[26]	2	RGB-RGB	Checkerboard	Overlap	∼0.17 px
[27]	2	RGB-RGB	Checkerboard/Targetless	Overlap	0.1-10 cm / <1º
[29]	3	RGB-RGB	Checkerboard	Overlap	∼0.4 px
[31]	2	RGB-D	Spheres	Overlap	<2.5 px
[32]	2	RGB-D	Custom board	Partial	few cm
[33]	2	RGB-D	Checkerboard	Overlap	<4 px
[34]	n	RGB-D	Checkerboard	Partial	<15 px
[45]	4+	RGB-D	Charuco tower	Overlap	3.5 mm
[46]	4	RGB-D	Circular markers	Limited	3 mm
[9]	n	RGB-IMU	Checkerboard	Overlap	<2 mm / <0.5º
[35]	2	RGB-2D LiDAR	Checkerboard plane	Overlap	3% / 1º
[37]	2	RGB-LiDAR	Checkerboard	Overlap	∼4%
[38]	2	RGB-LiDAR	Checkerboard	Partial	12%/∼1.5°
[39]	2	RGB-LiDAR	Custom planar board	Partial	∼1 cm / 0.005 rad
[40]	2	RGB-LiDAR	Checkerboard	Overlap	∼0.8 px
[41]	3	RGB-LiDAR	Triangular target	Overlap	∼2 %
[42]	2	RGB-LiDAR	Checkerboard	Overlap	9 mm / ∼7 px
[25]	2	RGB-LiDAR	Targetless	Partial	∼4 cm / ∼ 0.5º
[47]	2	RGB-LiDAR	Mirror	Non-overlapping	4% / <3º
[51]	2	RGB-LiDAR	Checkerboard/Object	Overlap	∼1 cm/ ∼ 0.1 rad
[54]	2	RGB-LiDAR	Targetless	Overlap	∼7 cm/ ∼ 2.25º
[57]	2	RGB-LiDAR	Targetless	Overlap	0.75 m / 15º
[55]	2	RGB-LiDAR	Targetless	Overlap	2.43 cm / 0.25º
[56]	2	RGB-LiDAR	Targetless	Overlap	∼7 cm/ ∼ 0.23º
[52]	2	RGB-LiDAR	Costum target	Overlap	∼5 mm / <1°
[50]	3	RGB-LiDAR	Checkerboard	Partial	0.9-6 px
[66]	2	RGB-LiDAR-RGB-D	Targetless	Partial	∼1 cm / 1°
[81]	3	RGB-IR-LiDAR	Checkerboard/Costum Target	Overlap	NA
[82]	5	RGB-2D LiDAR-3D LiDAR	Checkerboard	Partial	0.5-6 px

Traditional RGB-RGB calibration methods form the backbone of early extrinsic calibration efforts, with techniques like those of Su et al. [26], Ling et al. [27], and Dinh et al. [29] relying on checkerboard targets in overlapping fields of view. These methods consistently report sub-pixel accuracy, confirming their maturity and reliability in structured setups. However, they remain intrinsically limited to scenarios with full visibility between cameras and are not easily extensible to larger sensor arrays or heterogeneous modalities.

RGB-D calibration introduces further challenges due to the noise characteristics and lower resolution of depth sensing. Works such as Liu et al. [31], Basso et al. [32], and Kwon et al. [34] explore diverse target designs including spheres, checkerboards, and planar patterns, often with partial field-of-view overlap. These methods typically achieve millimetre- to pixel-level accuracy, with Park et al. [45] demonstrating a notable improvement through a multi-layered Charuco tower, reaching sub-4 mm precision. Similarly, Chaochuan et al. [46] utilise circular markers and metaheuristic optimisation to deal with limited FoV scenarios. While such designs improve flexibility, they still depend on carefully engineered targets and spatial arrangements.

The calibration of RGB-LiDAR systems adds another layer of complexity due to the sparse and asynchronous nature of LiDAR data. A large body of work in this domain, such as Vasconcelos et al. [35], Guindel et al. [39], Wang et al. [40], and Verma et al. [42], proposes target-based strategies using planar boards, checkerboards, or triangular panels. Reported accuracy generally ranges between 1–2 cm and up to 1.5° rotation error, with variations

depending on point cloud density, target geometry, and calibration metric. Oliveira et al. [82] expand this line of work by proposing a practical calibration method involving both 2D and 3D LiDARs in addition to RGB sensors. Their system is validated in real-world agricultural settings, highlighting the method's robustness across multi-modal configurations with partial field-of-view overlap. Pinto de Aguiar et al. [50] present a method for calibrating fixed RGB and LiDAR sensors using a checkerboard observed from multiple static positions, achieving reprojection errors between 0.9 and 6 pixels.

In contrast, several recent works shift away from reliance on calibration targets. Targetless or minimally structured methods, such as those by Yaqing et al. [25], Zhu et al. [54], Lin et al. [57], Wang et al. [55], and Cattaneo et al. [56], introduce techniques based on edge alignment, surface geometry, or mesh constraints. These methods prove especially useful in unstructured or cluttered environments, although their accuracy is often lower, with translation errors ranging from 2 to 7 cm and rotation errors up to 2.5°. Still, their independence from specialised markers and greater scalability make them attractive for dynamic or large-scale deployments.

Calibration approaches in highly multi-modal setups, such as those incorporating IMUs [9], thermal sensors [81], or hybrid RGB-LiDAR-RGB-D arrangements [66], highlight the growing importance of integrated perception. These configurations are common in collaborative cells, where different sensing technologies cover complementary fields of view and operate under distinct measurement models. Furgale et al. [9] present an optimisation-based pipeline that jointly estimates spatial and temporal calibration parameters, achieving high precision across RGB-IMU combinations. Horn et al. [66] use ego-motion as a supervisory signal, bypassing the need for targets entirely and reporting errors below 1 cm and 1°.

Special mention should be made of efforts that address non-overlapping field-of-view configurations, a critical issue in collaborative robotic settings. Raposo et al. [47] offer a mirror-based method to simulate shared FoV across fixed RGB-D and LiDAR units, while Kim et al. [52] propose a custom target design embedded within SLAM optimisation routines to calibrate RGB-LiDAR pairs with high accuracy (<5 mm). These solutions illustrate creative ways to deal with occlusions and visibility gaps that are typical in industrial workcells.

Across the surveyed literature, accuracy remains highly dependent on the modality and methodology used. Target-based RGB-RGB and RGB-D setups often reach sub-pixel or millimetre-level performance. LiDAR-based methods typically yield translational errors between 1–3 cm and angular errors around 0.5–1.5°, although some approaches push these limits further under controlled conditions. Targetless methods generally sacrifice some precision in favour of flexibility, with higher error margins but greater applicability in unconstrained or online scenarios.

This comparative overview underscores both the diversity and the fragmentation of the calibration landscape. Classical methods provide excellent accuracy in controlled conditions but lack scalability. Learning-based and targetless techniques offer adaptability and automation but are not yet as precise or generalisable. The growing heterogeneity of sensor configurations, especially in collaborative cells, calls for unified frameworks that can flexibly accommodate

varying FoVs, sensor types, and spatial constraints without sacrificing accuracy or requiring extensive manual intervention.

2.2.7 Critical Analysis

The literature on extrinsic calibration evolves considerably in response to the increasing complexity of multi-sensor robotic systems, particularly within collaborative environments. Early efforts, such as Zhang's classic checkerboard-based method for stereo calibration [8], pave the way for highly accurate, pairwise calibration techniques under controlled and overlapping fields of view. These methods, including RGB-RGB approaches like those by Su et al. [26] and Ling and Shen [27], continue to offer reliable results in structured environments but scale poorly as sensor networks grow or diversify.

Recent research improves calibration accuracy and robustness through enhanced visual targets and detection pipelines. For example, Chaochuan et al. [46] propose a tower-based circular marker setup for constrained FoVs, while Kim et al. [52] introduce an optimised calibration procedure using circular features and constrained motion paths. The work of Park et al. [45] further demonstrates how multi-camera RGB-D calibration can achieve sub-millimetre accuracy through iterative refinement with a 3D Charuco tower. Meanwhile, Zhang et al. [65], Chen et al. [83], and Garrido-Jurado et al. [53] significantly enhance the reliability of marker detection under adverse conditions. These advances increase the usability of target-based calibration in industrial environments, but they still require visual access to shared patterns, spatial proximity, and sometimes manual setup.

In contrast, recent developments in targetless or semi-structured calibration methods seek to improve flexibility. Horn et al. [66] propose a dual-optimisation strategy that leverages sensor egomotion rather than fixed targets, while Raposo et al. [47] and Hua et al. [51] use mirrored surfaces or scene edges to align sensors with disjoint views. Similarly, motion and mesh-based methods [54]–[56] offer alignment based on dynamic feature correspondences or ray-traced mesh models. These techniques prove particularly relevant for non-overlapping and safety-critical configurations, such as those found in collaborative robotic cells. However, they introduce new trade-offs: greater computational burden, sensitivity to scene geometry or motion, and less predictable accuracy under real-world conditions.

Efforts to calibrate multi-sensor systems with RGB, depth, and LiDAR data also progress. Works by Vasconcelos et al. [35] and Rehder et al. [36] use planar targets and plane-fitting for RGB-LiDAR alignment, while Zhou et al. [38] exploit edge correspondences to reduce dependency on patterns. Nevertheless, most approaches still remain pairwise and assume overlap between sensors. When scaling to networks with multiple static and mobile sensors, including IMUs or thermal cameras, issues of error propagation, calibration path dependency, and toolchain heterogeneity emerge, limiting applicability in complex setups.

Another persistent gap concerns the calibration of sensors mounted on robotic arms (eye-in-hand) in parallel with fixed infrastructure sensors (eye-to-base). While some methods address hand-eye calibration specifically [70], [71], and others support static multi-view alignment, few frameworks prove capable of simultaneously calibrating heterogeneous sensors in mixed

configurations.

Spatiotemporal synchronisation, though not always treated as part of extrinsic calibration, plays a crucial role in systems with asynchronous sampling (e.g., RGB, LiDAR, IMU). Kalibr [9] and its extensions [36] enable joint optimisation of spatial and temporal offsets, but assume motion and may not suit fixed sensors. Liu et al. [58] go further with a modular calibration pipeline that addresses both synchronisation and alignment in a generalisable way, yet still fall short in real-time static deployments or setups requiring joint hand-eye and static calibration.

From a deployment perspective, the lack of general-purpose, reproducible toolkits remains a barrier. While ROS-based calibration packages exist, they are often limited to specific sensor pairs, require controlled motion sequences, or are cumbersome to configure. This is compounded by the uneven availability of benchmarking datasets. KITTI [67], YUTO MMS [84], and similar driving-focused benchmarks provide strong baselines for mobile robotic calibration, but there are few datasets that reflect the spatial constraints, occlusions, and static configurations of collaborative industrial cells.

A further limitation lies in the scarcity of calibration methods tailored to the real-world constraints of collaborative cells. Unlike mobile platforms or controlled environments, collaborative workspaces often require calibration of fixed, non-overlapping, and heterogeneous sensors, without disrupting the workspace or relying on motion. Few published approaches explicitly support these constraints while maintaining sub-centimetre accuracy. Notable exceptions include the ATOM framework [48], [50], [80], which enables the calibration of static RGB and LiDAR sensors through a modular, sequential strategy.

Learning-based calibration is an emerging but still maturing area. Deep learning models like those proposed by Yaqing and Huaming [25] demonstrate strong potential in generalisation and robustness, especially under occlusion or partial visibility. However, these approaches are not yet integrated into standard calibration pipelines and often require pretraining on synthetic data or specific scene assumptions.

In this context, the contribution of this thesis is both novel and operationally relevant. It proposes a multi-modal and multi-sensor calibration framework capable of estimating the extrinsic transformations between any number of RGB, depth, and LiDAR sensors, whether fixed or mounted on a robotic manipulator, with a single calibration target. The method supports mixed configurations (e.g., static infrastructure and mobile arms), partial or non-overlapping fields of view, and does not rely on predefined motion paths or elaborate marker setups. By integrating hand-eye calibration into a unified, scalable pipeline, this approach bridges a critical gap between the precision of traditional calibration and the flexibility demanded by real-world collaborative robotics. In doing so, it contributes to the development of perceptual infrastructures that are accurate, adaptable, and robust under industrial constraints.

2.3 3D Human Pose Estimation from Multi-Camera Systems

Human pose estimation refers to the task of inferring the configuration of the human body from sensor data. A distinction is typically made between two-dimensional (2D) and

three-dimensional (3D) pose estimation. While 2D estimation localises joints within the image plane, 3D estimation seeks to reconstruct the spatial configuration of body joints in world coordinates. The latter proves inherently more challenging due to depth ambiguities, occlusions, and the need for accurate geometric modelling.

In collaborative robotics and Human-Robot Interaction (HRI), 3D pose estimation is critical for ensuring safety, predicting human trajectories, and enabling responsive, context-aware behaviours. By accurately localising body joints, robots can avoid collisions, respect safety zones, and adapt their actions to human co-workers in shared environments. These applications demand robustness under occlusions, temporal consistency, and real-time performance.

2.3.1 RGB Multi-View 3D Human Pose Estimation Methods

Multi-camera RGB systems allow for redundant visual information across multiple viewpoints, facilitating robust 3D human pose estimation. Geometry-based triangulation is a classical approach that reconstructs 3D joint positions from 2D detections using calibrated camera parameters. While conceptually simple, it is highly sensitive to detection noise and missing joints.

To address these limitations, volumetric methods aggregate multi-view 2D detections into 3D heatmaps or voxel occupancy grids. These probabilistic fusion techniques, as explored by Zhao et al. [85], improve robustness in occluded scenarios but require large memory and computational resources.

A common option is to represent the human body through a skeleton of keypoints, typically capturing 14 to 25 major joints. This skeleton-based approach offers a simplified, lightweight representation that can be estimated efficiently and is often sufficient for applications such as gesture recognition, trajectory prediction, or real-time safety monitoring in robotics. Due to their reduced dimensionality and simpler structure, skeleton models are well suited to tasks where computational efficiency and fast inference are prioritised over anatomical detail.

In contrast, parametric body models such as SMPL [86] and SMPL-X [87] provide a more expressive and anatomically grounded mesh representation of the body surface. These models incorporate shape and pose parameters to generate full-body 3D meshes and, in the case of SMPL-X, also include articulated hands and facial expressions. While offering a richer depiction of human posture and emotion, such models require more processing time, additional parameters, and often tighter alignment with input data. The difference in representational complexity is illustrated in Figure 2.6, where the compactness of the skeleton-based model contrasts with the dense surface mesh of SMPL and SMPL-X. Several approaches fit SMPL to 2D keypoints [88], silhouettes [89], or contact constraints [90], using either optimisation [91] or learned regressors [92].

For this reason, skeleton-based approaches remain dominant in real-time systems, especially in human–robot interaction contexts, where fast decision-making is required and full mesh reconstruction is often unnecessary. On the other hand, SMPL-based representations are preferred in applications involving body modelling, biomechanics, or avatar generation, where surface accuracy and realism are critical.

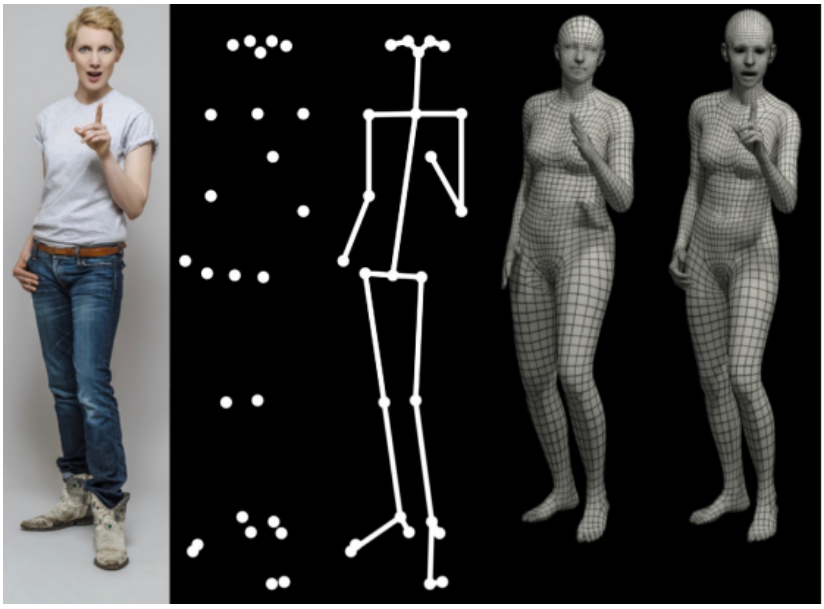

Figure 2.6: Comparison of pose representations. From left to right: RGB image, major joints, skeleton, SMPL, SMPL-X. Skeleton-based models are compact and computationally efficient, while SMPL(-X) models offer high fidelity and expressive detail at a higher computational cost. Image adapted from Pavlakos et al. [87].

Learning-based and transformer-driven architectures now dominate the field. Works such as PoseFormer [93] and PoseFormerV2 [94] model temporal dependencies across frames using attention or frequency-domain processing. Other architectures, like GLA-GCN [95] and MotionAGFormer [96], combine graph reasoning with temporal modelling to enhance spatial fidelity and temporal consistency. Self-supervised methods like EpipolarPose [97] and Bouazizi et al. [98] avoid the need for paired 2D-3D annotations by exploiting multi-view consistency. Meanwhile, diffusion-based methods like DiffuPose [99] and ZeDO [100] represent a shift toward generative pose estimation.

Evaluation typically relies on quantitative metrics such as Mean Per Joint Position Error (MPJPE) and its aligned variant, Procrustes-aligned MPJPE (P-MPJPE). MPJPE measures the average Euclidean distance (in millimetres) between predicted and ground truth joint positions across all frames and joints, calculated after aligning the skeletons by a rigid root translation. It is widely used in the Human3.6M benchmark [101], where it becomes the de facto standard for supervised pose estimation.

However, MPJPE is sensitive to scale, orientation, and body size, and it does not account for global pose shifts or camera misalignment. To address this, P-MPJPE applies a rigid-body Procrustes alignment before computing the joint distances, effectively removing errors due to global translation, rotation, and scale. This metric captures the structural correctness of the predicted pose independent of absolute position, making it particularly useful for comparing shape-preserving methods such as SMPL-based models.

Despite their widespread use, both MPJPE and P-MPJPE have limitations in downstream robotic applications, where temporal stability, semantic consistency (e.g., correct left-right labeling), and response latency may be more critical than small positional deviations. As

such, researchers have proposed complementary metrics like Percentage of Correct Keypoints (3DPCK) and Area Under the Curve (AUC), especially in mesh-based estimation. Yet, no consensus exists on standard metrics for real-world deployment, particularly in safety-critical or interactive scenarios.

A comparative overview of selected 3D pose estimation methods is presented in Table 2.2, highlighting their type, core innovations, performance in terms of MPJPE, and common evaluation datasets. This table provides a useful reference point for understanding trade-offs across model families.

Table 2.2: Comparison of selected 3D human pose estimation methods based on architecture type, methodological approach, and reported MPJPE in millimetres on benchmark datasets: Human3.6M (H36M) [101], MPI-INF-3DHP (3DPH) [10], and HumanEva (HE) [102]. MPJPE values are rounded to the nearest millimetre.

Method	**Type**	**Key Characteristics**	**MPJPE (mm)**		
			H36M	3DPH	HE
ZeDO [100]	Diffusion + optimisation	Zero-shot inference from 2D keypoints	42	55	-
SMPLify [88]	SMPL-based	Regression from 2D keypoints to mesh	69	-	62
MotionAGFormer [96]	Transformer-GCN	Lightweight spatio-temporal fusion	17	-	16
GLA-GCN [95]	GCN + attention	Local joint refinement with global reasoning	21	28	9
EpipolarPose [97]	Self-supervised	Triangulation-based training without GT labels	61	109	-
Martinez et al. [103]	Lifting (MLP)	Simple 2D-to-3D baseline, no temporal modelling	63	-	25
Pavllo et al. [104]	Temporal CNN	Semi-supervised training with back-projection	47	-	-
DiffuPose [99]	Diffusion	Denoising diffusion model from monocular RGB	49	-	-
Mono-3DHP [10]	CNN regression	Monocular in-the-wild pose using improved supervision	55	41	-
CameraPose [105]	Weak supervision	Lifting from 2D using weak supervision with camera priors	39	79	-
HoloPose [106]	Holistic reconstruction	End-to-end model for mesh and pose from monocular RGB	60	-	-
MotionBERT [107]	Transformer	Unified spatio-temporal representation learning	38	-	-
Bouazizi et al. [108]	Temporal + pseudo-labels	Temporal refinement using pseudo-3D labels from 2D detections	40	93	-
Bouazizi et al. [98]	Self-supervised	Self-supervised 3D pose using multi-view geometry	62	-	59
PoseFormerV2 [94]	Transformer	Frequency-domain transformer using DCT on 2D keypoints	45	28	-

The methods summarised in Table 2.2 illustrate the diversity of architectures and training strategies employed in 3D human pose estimation. Transformer-based models such as PoseFormerV2 and MotionBERT achieve state-of-the-art accuracy on Human3.6M [101] while maintaining temporal consistency and robustness in longer sequences. Diffusion-based methods like ZeDO and DiffuPose stand out for their ability to generalise to in-the-wild or cross-domain settings without requiring explicit 3D supervision. SMPL-based regressors (e.g., SMPLify) provide detailed mesh outputs but depend heavily on accurate 2D keypoints and are less robust to occlusions. Self-supervised models like EpipolarPose and CameraPose reduce dependence on annotated 3D data, offering promising solutions for deployment in unconstrained environments. Hybrid models that combine GCNs and transformers, such as GLA-GCN and MotionAGFormer, provide a good trade-off between local joint reasoning and global scene understanding. Overall, the field moves toward architectures that integrate multi-view, temporal, and semantic information in an efficient and scalable manner.

2.3.2 Multi-modal 3D Human Pose Estimation Methods

Some authors in 3D human pose estimation turn toward multi-modal architectures to overcome the limitations of single-sensor systems. In contrast to monocular or RGB-only methods, multi-modal approaches aim to exploit complementary strengths of sensors such as

LiDAR, RGB, depth, IMU, radar, and thermal to improve pose reconstruction under challenging conditions—occlusion, sparse observations, adverse lighting, and cluttered environments. These approaches prove particularly valuable in collaborative robotic systems and autonomous platforms, where perception must be robust, real-time, and adaptable to diverse operational contexts.

A key motivation for multi-modal estimation is the inherent sparsity and noise of LiDAR point clouds, which make accurate pose recovery difficult when relying solely on geometric priors. Early LiDAR-only systems (e.g., LiDARCap [109]) introduce dense temporal modelling and SMPL optimisation to recover human motion, but at the cost of complexity and limited scalability. Recent work like DAPT (Density-Aware Pose Transformer) [110] proposes an optimisation-free method that achieves robust single-frame LiDAR-based estimation by pre-training on a synthetically augmented dataset. This method integrates joint anchor representations and multi-density exchange modules, achieving state-of-the-art performance on benchmarks like Waymo [111] and SLOPER4D [112].

While LiDAR-only approaches become increasingly viable, many state-of-the-art systems pursue multi-modal fusion to boost performance, especially in outdoor or cluttered indoor environments. For instance, HPERL [113] integrates RGB and LiDAR data to extract complementary semantic and geometric features. Similarly, LPFormer [114] employs a transformer architecture with multi-task learning over LiDAR data, jointly performing segmentation and pose estimation. The inclusion of semantic cues improves disambiguation in dense scenes and offers improved generalisation across domains.

A growing number of works address weakly supervised multi-modal learning, particularly in automotive and large-scale outdoor scenarios. Bauer et al. [115] and Zheng et al. [116] propose frameworks that learn 3D poses from sparse LiDAR and RGB images, guided by 2D annotations. These systems benefit from the geometric constraints of LiDAR and the texture and semantic richness of RGB images while minimising the need for expensive 3D ground truth. Similarly, HUM3DIL [117] introduces a semi-supervised fusion approach that integrates RGB and LiDAR with 2D annotations, tailored for autonomous driving contexts. These methods report competitive accuracy with limited supervision, making them well suited for industrial domains where fully annotated datasets are rare.

Beyond RGB and LiDAR, additional modalities emerge. LidPose [118] demonstrates that even sparse LiDAR scans can yield reliable 3D pose estimations when the scanning pattern is optimised, using circular scanning and density-aware training. Meanwhile, SPiKE [119] exploits sequential point cloud frames to perform efficient spatio-temporal inference, underlining the role of motion continuity even without image data.

Some authors combine other types of modalities. The mRI dataset [120] integrates mmWave radar, RGB-D, and inertial sensors, providing a challenging testbed for robust pose estimation under occlusion and motion. It serves as the basis for cross-modal learning, particularly relevant for wearable applications and safety-critical collaborative robotics. These systems excel in scenes where line-of-sight is frequently blocked—e.g., around heavy machinery or behind obstacles—making them highly suitable for safety monitoring and real-time decision

support in manufacturing environments.

Multi-modal pose estimation methods typically rely on early or late fusion strategies, depending on the task and sensor alignment. Early fusion combines raw data streams, often requiring tight spatiotemporal synchronisation and calibration. Late fusion operates on intermediate representations (e.g., keypoints, feature maps), offering greater flexibility and robustness but often requiring careful domain adaptation between modalities.

Despite their promise, multi-modal approaches pose challenges in calibration, synchronisation, and deployment scalability. Achieving real-time performance while fusing heterogeneous data streams from distributed sensors remains non-trivial. Nonetheless, their potential to enable reliable, occlusion-robust, and generalisable human pose estimation positions them as a cornerstone for the next generation of collaborative robotic systems.

2.3.3 Datasets and Benchmarks

Benchmark datasets are instrumental in advancing research in 3D HPE, offering standardised platforms for training, validation, and comparison across models and tasks. Most widely used datasets are designed for general-purpose human activity understanding and are collected under controlled laboratory conditions. While these datasets drive significant methodological progress, they present certain limitations when applied to industrial and collaborative robotic contexts, particularly regarding sensor configurations, environmental variability, and occlusion patterns.

Human3.6M [101] remains the most widely adopted dataset for 3D HPE evaluation. It includes over 3.6 million 3D pose annotations derived from marker-based motion capture, paired with synchronized video from four calibrated RGB cameras. Subjects perform a fixed set of actions in a studio environment, offering consistent ground truth but limited scene diversity. It supports both supervised and weakly supervised learning setups, and it is the de facto benchmark for methods trained in indoor, single-person, full-body scenarios.

MPI-INF-3DHP [10] is introduced to address some limitations of Human3.6M, notably the lack of diversity in clothing, viewpoints, and environments. It includes both indoor and outdoor sequences, with 3D annotations acquired from markerless systems and manual correction. The dataset supports generalisation studies and is especially useful for assessing model robustness in semi-constrained environments with varied lighting and attire.

CMU Panoptic Studio [121] is a large-scale multi-person dataset featuring over 500 synchronised RGB cameras arranged spherically around a central capture area. It enables high-resolution tracking of group interactions and complex occlusion scenarios, making it a valuable resource for multi-view learning and multi-person pose estimation. However, its scale and complexity limit its use to a smaller set of research groups with high computational resources.

HumanEva [102] is among the first datasets to provide synchronized video and 3D pose data, enabling probabilistic and optimisation-based evaluation in early 3D HPE research. While smaller in scale and resolution compared to newer datasets, it remains relevant for benchmarking monocular and stereo reconstruction algorithms due to its simplicity and high

annotation fidelity.

3DPW (3D Poses in the Wild) [122] represents a significant step toward in-the-wild evaluation, capturing natural human motion in outdoor settings using a portable motion capture setup (IMUs + video). It provides 3D mesh annotations aligned with the SMPL model and supports evaluation of methods in unconstrained environments. However, its annotations are less precise than optical marker-based setups and it remains underused in industrial robotics research.

Despite their contributions, most of these datasets fall short of reflecting the complexities of industrial or collaborative robotic environments. Specifically, they rarely incorporate environmental occlusion, sensor noise, reflective surfaces, non-standard camera placements, or partial observability, conditions commonly found in manufacturing cells or HRI scenarios. Furthermore, calibration data is often limited or not aligned with the sensor setups used in industrial settings.

Ultimately, there is a pressing need for more representative datasets that align with the specific spatial, perceptual, and temporal constraints of collaborative robotics. Such datasets would ideally feature non-overlapping camera networks, dynamic lighting, industrial artefacts, real operator behaviour, and accurate 3D ground truth under occlusion. The lack of such resources remains a bottleneck in the development of deployable human pose estimation solutions for real-world collaborative systems.

2.3.4 Applications

3D human pose estimation evolves into a core capability for a wide variety of systems requiring real-time, spatial understanding of human motion. Its capacity to accurately reconstruct joint locations in three-dimensional space supports an expanding set of applications across fields such as robotics, autonomous systems, healthcare, and sports. While much of the research focuses on improving the precision and generalisation of pose estimation models, their practical utility is increasingly demonstrated in domain-specific deployments.

In clinical and rehabilitative contexts, 3D human pose estimation gains traction as a tool for non-invasive assessment of musculoskeletal health. Koleini et al. [123] introduce BioPose, a monocular video-based method that integrates biomechanical constraints into a pose estimation pipeline. By enforcing anatomical plausibility and joint angle limits, BioPose enables the reconstruction of physically consistent motions suitable for use in musculoskeletal assessment and virtual rehabilitation. The system achieves accurate reconstruction of gait and therapeutic exercises from standard camera input, making it well-suited for at-home or outpatient clinical use where marker-based motion capture is impractical.

Expanding into applied biotechnology, Ge and Mariano [124] propose a multi-feature fusion framework that leverages 3D HPE for wearable diagnostics and human performance monitoring. Their approach integrates skeletal pose estimation with physiological signal analysis in commercial health platforms, facilitating applications such as real-time posture tracking, fatigue monitoring, and behavioural biometrics. This integration exemplifies a growing convergence between human pose estimation and digital health technologies, where

accurate motion analysis becomes a proxy for broader physical or cognitive state inference.

In the sports domain, 3D HPE emerges as a central tool for performance assessment, skill acquisition, and injury prevention. Park et al. [125] develop GolfPoseNet, a pose estimation network specialised for the analysis of golf swings. By tailoring the architecture to domain-specific kinematics, the system reconstructs the full-body motion of athletes in outdoor or indoor settings, enabling frame-by-frame feedback for coaches and players. Similarly, Liu et al. [126] introduce STIGANet, a low-latency pose estimation model that fuses dynamic graph convolutional networks with attention mechanisms. Designed for deployment in high-speed sports scenarios, STIGANet supports real-time tracking and action segmentation across diverse motion types.

Zhang et al. [127] propose STAPFormer, a transformer-based architecture aimed at sports and health applications where biomechanical accuracy is critical. It integrates temporal context with skeletal dynamics to enhance anatomical consistency and temporal stability in pose sequences. Also addressing posture correction, Yuan and Zhou [128] present GTA-Net, an IoT-integrated system for adolescent sports training. Their model enables real-time assessment of spinal alignment and postural deviations, with applications in injury prevention and youth athlete development. These systems reflect a broader trend toward specialised 3D HPE models that prioritise domain knowledge, real-time responsiveness, and usability in non-laboratory settings.

3D HPE also becomes increasingly relevant in perception systems for autonomous vehicles. In complex traffic scenarios, understanding the body posture and motion intent of pedestrians and cyclists proves vital for safe navigation and planning. The HUM3DIL framework introduced by Zanfir et al. [117] addresses this challenge by employing a semi-supervised, multi-modal pipeline that fuses stereo, LiDAR, and image data to estimate pedestrian pose in diverse street scenes. The system learns from both labelled and unlabelled data, enabling robust generalisation under variable lighting, occlusion, and movement dynamics.

In a related line of work, Bauer et al. [115] present a weakly supervised multi-modal pose estimation framework for autonomous driving, combining RGB and LiDAR inputs with 2D annotations. Their system operates in urban environments, providing accurate 3D pose estimates for external human actors such as pedestrians, scooter riders, and construction workers. These applications demonstrate how 3D HPE contributes not only to safety and intent prediction but also to broader goals of human-aware autonomy.

Within the robotics domain, 3D human pose estimation serves as a perceptual foundation for enabling physical and semantic interaction between robots and humans in shared environments. In collaborative manufacturing cells, pose estimation is used to monitor human proximity, predict motion intent, and adapt robot trajectories in real time, allowing for the enforcement of dynamic safety zones and fluid task-sharing. For example, Peppas et al. [129] integrate a multi-modal 3D HPE system into a collaborative robotic setup, enabling real-time ergonomic analysis and adaptive robot responses to worker motion. Similarly, Fürst et al. [113] demonstrate the effectiveness of combining LiDAR and RGB input to estimate human pose for safety-aware robotic co-working.

Beyond reactive safety measures, 3D HPE is increasingly used for semantic interaction and intent recognition. Systems capable of detecting hand gestures, tool-use postures, or whole-body motion can anticipate the user's next action and proactively adjust the robot's behaviour. This is further explored by Baptista et al. [11], who incorporate pose-based gesture recognition into a ROS-based collaborative cell to improve fluency in human–robot cooperation. In more advanced learning-from-demonstration frameworks, pose estimation enables the capture of human motion trajectories without markers or wearable sensors, which can then be translated into executable robotic actions—a paradigm increasingly employed in task teaching and motion programming.

As robotic systems continue to evolve toward greater autonomy and shared agency, the integration of robust, real-time 3D pose estimation becomes indispensable. Not only does it enhance physical safety and task efficiency, but it also contributes to the robot's ability to understand, anticipate, and interact meaningfully with human collaborators, aligning with broader objectives in Industry 5.0 and human-centric automation [15], [130].

Taken together, the reviewed applications demonstrate that 3D human pose estimation moves beyond the confines of computer vision research and into real-world systems that require reliable, real-time human motion understanding. Whether tracking gait in clinical rehabilitation, guiding swing mechanics in sports, enabling pedestrian awareness in autonomous driving, or supporting seamless interaction in collaborative robotics, 3D HPE now plays a vital role in human-centred system design. These advances are driven not only by algorithmic improvements but also by closer integration with domain-specific constraints and goals. As the next chapters of this thesis will show, these capabilities are essential in developing perceptually aware, safety-compliant, and semantically rich robotic systems designed for industrial and collaborative environments.

2.3.5 Critical Analysis

The field of 3D human pose estimation makes substantial strides, yet a number of conceptual, methodological, and practical challenges continue to constrain its deployment in real-world scenarios, particularly in robotics and human-machine collaboration.

One of the most persistent limitations across existing methods is their reliance on precise camera calibration. Multi-view triangulation, volumetric fusion, and SMPL-based fitting typically assume known intrinsics and extrinsics, which is seldom the case in ad hoc or dynamic camera arrangements. Even methods trained on large-scale benchmark datasets such as Human3.6M [101] or MPI-INF-3DHP [10] often lack robustness to shifts in viewpoint geometry, sensor fidelity, or background complexity. This poses a barrier to their use in non-laboratory settings, such as industrial floors, warehouses, or mobile robotic platforms, where scene parameters may change frequently or remain partially unknown.

Generalisation to out-of-distribution contexts remains a significant obstacle, especially for learning-based and monocular approaches. While transformer-based architectures (e.g., PoseFormerV2 [94]) and diffusion models (e.g., ZeDO [100]) show promising generalisation across datasets, they are still predominantly evaluated in curated environments. Models

often degrade when confronted with unseen lighting, occlusions, clothing variation, or motion profiles not represented in the training data. This indicates that current benchmarks may insufficiently reflect the variance encountered in operational settings, and that generalisation remains more an aspiration than a standard feature of state-of-the-art models.

Optimisation-based methods, particularly those relying on parametric models like SMPL [88], [89], offer interpretable and anatomically grounded outputs. They are well suited to applications requiring detailed mesh representations or physical simulation. However, these approaches are sensitive to 2D detection quality and often require extensive computation. Moreover, iterative fitting procedures may fail to converge under occlusion or when joints are misdetected, which is common in cluttered environments.

Temporal models such as those proposed by Pavllo et al. [104] or MotionAGFormer [96] introduce valuable tools for smoothing pose estimates and leveraging continuity in movement. However, these gains in temporal coherence frequently come at the cost of higher latency or dependence on future frames, making them less suitable for causal inference in reactive robotic systems. Moreover, while many models implicitly capture motion regularities, they often lack mechanisms for reasoning about human intent, interaction contexts, or safety-critical states.

Self-supervised learning [97], [98] and weakly supervised techniques [105] mark a critical step toward reducing data dependence and improving adaptability. These approaches are especially relevant for scenarios where collecting 3D ground truth is impractical. However, current self-supervised models tend to rely on restrictive assumptions (e.g., multiview consistency, geometric priors) and are not yet as competitive in accuracy as fully supervised counterparts.

The emergence of multi-modal approaches offers a promising direction for increasing robustness and contextual awareness. Systems that fuse RGB, LiDAR, radar, IMU, and depth data, such as DAPT [110], LPFormer [114], and HUM3DIL [117], demonstrate improved resilience to occlusion and sensor degradation. However, these pipelines face non-trivial integration challenges, including sensor synchronisation, calibration, and computational load. The trade-offs between early and late fusion remain an open question, and the deployment of such architectures in real-time collaborative robotics is still limited by hardware constraints and processing requirements.

A further point of concern lies in the fragmented nature of evaluation protocols. While MPJPE remains the most common metric, it does not account for semantic plausibility, temporal stability, or real-world usability. Metrics such as 3DPCK, AUC, and PVE (Per Vertex Error) offer partial alternatives, but no consensus has emerged on how to benchmark robustness under occlusion, generalisation to unseen domains, or resilience to calibration noise. Without unified standards, it becomes difficult to assess the practical suitability of competing approaches for safety-critical applications.

From an application perspective, recent research begins to address more specific and functionally relevant challenges. In healthcare, systems like BioPose [123] incorporate anatomical priors to produce physically meaningful reconstructions, making them more suitable for gait analysis or rehabilitation. In sports, task-specific models such as GolfPoseNet [125] or STI-GANet [126] demonstrate the value of domain adaptation, where training objectives are closely

aligned with target use cases. Similarly, autonomous driving platforms like those developed by Bauer et al. [115] and Zanfir et al. [117] show how weak supervision and multi-modal integration can improve safety in outdoor, cluttered environments. However, the transfer of these insights into collaborative robotics remains uneven. While some systems, such as those proposed by Peppas et al. [129] and Fürst et al. [113], demonstrate effective integration of pose data for human-aware control, many research contributions still remain detached from robot execution pipelines or ergonomic assessment tools.

Additionally, the dataset landscape is not yet aligned with real-world demands. While Human3.6M [101] and MPI-INF-3DHP [10] remain standard references, they do not represent the spatial, temporal, and contextual variability found in industrial or collaborative environments. Most datasets lack occlusion, variable lighting, or realistic sensor configurations, and few provide synchronised multi-modal data or labelled interaction scenarios. As a result, the evaluation of systems intended for deployment in collaborative robotic workcells, digital health, or mobile autonomy remains fragmentary.

Against this backdrop, our work proposes a methodologically grounded approach to 3D human pose estimation that explicitly addresses some of the most pressing constraints for deployment in collaborative robotics. By operating in a multi-camera RGB setup, we demonstrate that it is possible to achieve robust pose estimation in dynamic, semi-structured environments. Our method integrates a modular calibration procedure and uncertainty-aware triangulation pipeline, implemented in ROS, that supports real-time performance while tolerating minor camera misalignment. In contrast to models reliant on tightly controlled parameters or future frame dependencies, our approach is designed for causal inference and continuous operation in open-ended human-machine interaction scenarios.

This effort speaks directly to ongoing concerns about generalisability, calibration dependency, and practical integration. It also contributes to the growing need for systems that can operate beyond idealised conditions and benchmark environments. Looking ahead, we argue that the most promising research trajectories lie at the intersection of representation learning, sensor fusion, and runtime uncertainty estimation. Lightweight transformer architectures with frequency-domain compression [94] or hybrid spatial-temporal attention [95] are likely to support greater scalability. Generative techniques such as diffusion models [100] offer pathways to handle occlusion and data scarcity, but their translation into robotics will hinge on the development of methods that enable causal decision-making under uncertainty. In summary, our work contributes to this agenda by showing how calibration-aware, real-time HPE systems can be made more resilient and suitable for human-centred, collaborative applications.

2.4 Conclusion

This chapter reviews the state of the art in extrinsic calibration and 3D human pose estimation within the context of multi-sensor collaborative robotic systems. These two domains, while often treated independently, are fundamentally linked in the design and operation of human-centred manufacturing environments. As collaborative cells evolve toward

greater autonomy, safety, and semantic awareness, they increasingly rely on precise spatial alignment of heterogeneous sensor data and robust, real-time perception of human activity.

In the first part of the chapter, we examine extrinsic calibration techniques across a range of modalities, including RGB, depth, LiDAR, and radar. We highlight the technical challenges posed by non-overlapping fields of view, static sensor configurations, and modality-specific noise profiles. While traditional target-based methods offer high accuracy in controlled conditions, their scalability and applicability to complex, fixed sensor networks remain limited. Recent advances in targetless calibration, geometric feature extraction, and learning-based approaches represent a shift toward more flexible and automated solutions, but questions of generalisation, repeatability, and deployment readiness persist, especially in industrial contexts that demand minimal intervention and high precision.

The second part of the chapter focuses on 3D human pose estimation using multi-camera RGB systems. We survey a wide range of methods, from classical geometric triangulation to volumetric fusion, mesh-based regression, and recent transformer and diffusion architectures. These methods demonstrate remarkable accuracy on benchmark datasets, yet their deployment in collaborative robotics faces practical constraints such as occlusions, latency, calibration sensitivity, and limited generalisation to unstructured environments. Emerging approaches in self-supervised learning, frequency-domain modelling, and generative inference show promise in addressing these gaps, though their robustness under real-world conditions remains an active area of research.

The literature reveals a clear trend: both extrinsic calibration and pose estimation transition from static, highly constrained pipelines toward more integrated, adaptive, and semantically aware perception systems. However, this transition remains incomplete. Current methods often rely on assumptions that do not hold in collaborative industrial environments, such as perfect synchronisation, unobstructed views, or homogeneous sensor layouts. Moreover, there is a lack of evaluation protocols and datasets that reflect the practical demands of human–robot interaction in complex settings.

This thesis builds on these insights by proposing novel methodologies that address the key limitations identified in this chapter. Specifically, it advances a unified calibration framework that accommodates fixed and moving sensors across modalities and a pose estimation pipeline that exploits temporal and individual-specific priors to enhance robustness under occlusion. In doing so, it aims to bridge the gap between theoretical accuracy and practical applicability, contributing to the development of safe, perceptually capable, and context-aware collaborative robotic systems.

CHAPTER 3

A sensor-to-pattern calibration framework for multi-modal industrial collaborative cells

3.1 Introduction

According to the European Commission, Industry 5.0 aims to strengthen the contribution of the industry to society by thinking beyond efficiency and productivity, aiming for the development of technology towards the improvement of the worker's quality of life, while also respecting the planet [131], [132]. With this expansion of Industry 5.0, many new technologies are arriving to facilitate industrial and manufacturing jobs for humans by removing heavy burdens such as lifting heavy weights and repetitive movements. For that matter, collaboration with robots has become a highly researched topic because it can combine the expertise of humans with the workload of a robot [133], [134].

A collaborative cell is a three-dimensional space in which a collaborative robot and humans can safely coexist and carry out shared tasks. Safety requirements in such setups are standardised [135], imposing strict constraints on robot motion during human–robot interaction. These constraints primarily concern limitations on speed and torque when the robot operates in proximity to humans. To meet these demanding requirements, a robust perception framework is essential, particularly in the areas surrounding the robot. Achieving this requires a dense network of strategically placed sensors throughout the cell, capable of mitigating occlusions caused by the movements of people, objects, and the robot itself. Beyond a multi-sensor configuration, employing a multi-modal sensor system introduces complementary information that further enhances safety. For instance, range data from LiDARs and RGB-D cameras enables volumetric monitoring, while RGB images from standard cameras facilitate object detection. These data streams can also be fused for human pose estimation, ensuring continuous awareness of human positions within the cell and thereby safeguarding their

security.

Despite its significant advantages, data fusion remains a complex challenge, particularly when integrating information from multiple multi-modal sensors to create a unified intelligent system within the collaborative cell. As outlined by Baltrušaitis et al. [136], the challenges of multi-modality can be categorised into five key areas: representation, translation, alignment, fusion, and co-learning. Representation concerns the structure of multi-modal data, leveraging its complementarity and redundancy. Translation involves mapping data from one modality to another. Alignment refers to identifying direct relationships between elements or sub-elements across different modalities. Fusion entails combining multiple modalities to support prediction tasks. Finally, co-learning addresses the transfer of knowledge between modalities.

It follows, then, that determining the alignment between sensors is a crucial step in fusing multi-modal information, which brings us to the problem of extrinsic calibration: the process of establishing the transformation between a set of sensors. In the context of a collaborative cell, two major concerns emerge. First, as previously noted, the strict safety requirements demand highly accurate calibration, which is not trivial with such complex systems. Second, the large number of sensors and their differing modalities generate vast volumes of data, making simultaneous processing particularly challenging. Moreover, the heterogeneous nature of the sensor types adds complexity to data integration, such as the difficulty in selecting calibration targets that are detectable across all modalities.

Although existing methods do address the extrinsic calibration problem, they are often inadequate when applied to collaborative cells, which pose additional challenges. These include multi-sensor, multi-modal configurations; the need for high-precision calibration; and the sheer density of data generated. Additionally, collaborative cells can be physically large, and the sensors' FoV may not fully overlap. This alone rules out commonly used techniques, such as the *Open Source Computer Vision Library (OpenCV)* calibration tool, which assume overlapping FoVs between cameras.

To address these limitations, we propose a calibration framework based on optimising sensor-to-pattern transformations. This approach is capable of calibrating complex robotic systems involving RGB, depth, and LiDAR modalities. Unlike conventional sensor-to-sensor methodologies, our sensor-to-pattern strategy simplifies the calibration process and, crucially, accommodates sensors with non-overlapping FoVs.

The contributions of this chapter are:

- The development of a calibration framework able to calibrate complex, multi-modal and multi-sensor setups
- A solution to calibrate sensors with non-overlapping FoVs
- A calibration framework able to calibrate RGB, LiDAR and depth modalities

The remainder of this chapter describes and demonstrates the concept of the methodology undertaken to tackle the extrinsic calibration problem in a generic way. Section 3.2 describes the process of data acquisition, labelling and calibration itself. Section 3.3 presents the results obtained in our collaborative cell with a robotic system of 7 sensors with three different

modalities and a comparative study with other approaches in the literature. Finally, section 3.4 summarises the contributions of this chapter.

3.2 Methodology

The standard procedure for calibrating multi-sensor systems is to use a calibration pattern, which is positioned in such a way that it is accurately detected by all sensors. Then, the classic approach is to formulate the calibration as an optimisation procedure that minimises a set of errors. These errors are computed by an objective function, which is designed to translate the quality of alignment between the sensors, given the transformation between those sensors. The issue is that the errors are computed as a function of pairs of sensors. One example is the usage of the reprojection error (e) for calibrating multiple camera systems, expressed in equation (3.1):

$$e = \underset{\{^{s_i}\hat{\mathbf{T}}^{s_j}\}}{\arg\min} \sum_{\mathbb{S}} \sum_{\mathcal{I}} e\Big(^{s_i}\hat{\mathbf{T}}^{s_j}, d_{s_i}, d_{s_j}, \lambda_{s_i}, \lambda_{s_j}\Big), \tag{3.1}$$

where $^{s_i}\hat{\mathbf{T}}^{s_j}$ is the estimated transformation between sensors s_i and s_j, $\mathbb{S}$ represents the set of pair-wise combinations of the sensors in the system, $\mathcal{I}$ represent the set of images used to calibrate the system, d denotes the detections of the pattern by a sensor, and finally λ represents the intrinsic parameters of the sensor.

However, in the case at hand, a very complex system, the usage of errors derived from pairwise combinations of sensors (represented by $\mathbb{S}$ in (3.1)) is not scalable, since one must develop different mechanisms, i.e., different versions of the objective function e, for each pairwise combination of modalities. Moreover, in a collaborative cell, there will be many pairs of sensors that do not overlap. In these cases, it would not be possible to compute the errors using this problem formulation.

The differentiating aspect of our calibration methodology w.r.t. to others is that it uses a sensor-to-pattern approach instead of the classic sensor-to-sensor error estimation. Since each sensor views the pattern as a function of its intrinsic properties, pose, and pose of the pattern, instead of using the pairwise transformation error to optimise the transformations between sensors, we estimate the error by defining a function that uses the transformation between each sensor and the calibration pattern, as expressed in equation (3.2).

$$e = \underset{\{^{s_i}\hat{\mathbf{T}}^{w}\},\{^{p}\hat{\mathbf{T}}^{w}_{c}\}}{\arg\min} \sum_{\mathcal{S}} \sum_{\mathcal{C}} e\Big(^{s_i}\hat{\mathbf{T}}^{w}, {}^{p}\hat{\mathbf{T}}^{w}_{c}, d_{s_i}, \lambda_{s_i}\Big), \tag{3.2}$$

where $^{s_i}\hat{\mathbf{T}}^{w}$ is the estimated transformation between sensor s_i and the world coordinate frame w, $^{p}\hat{\mathbf{T}}^{w}_{c}$ is estimated transformation between the pattern p and w, which varies according to each collection c. Since our calibration framework tackles multi-modal systems, we refer to a moment in time where data from all sensors in the system is collected as a *collection* c and not an image i, to account for the fact that it may contain data from other modalities. To calibrate, we use a set of collections $\mathcal{C}$ to which we refer to as a dataset. In this case, the

overall error is computed by summing up the contributions of the sensors in the set of sensors $\mathcal{S}$, as opposed to the set of pairwise combinations of the sensors $\mathbb{S}$.

One advantage of our methodology is that only one mechanism per modality must be designed, as opposed to one mechanism per pairwise combination of modalities. Another advantage is that it is possible to estimate an error for each sensor, provided that it views the calibration pattern. This approach is much better suited to tackle calibration systems with several non-overlapping FoVs. Because we now use the transformation between the sensors and the pattern to estimate the errors, the pose of the pattern must also be included as a parameter to be optimised ${}^{w}\hat{\mathbf{T}}_{c}^{p}$. As such, our calibration system estimates not only the pose of the sensors but also the pose of the calibration pattern. This may sound counterintuitive, but in fact, by enlarging the optimisation problem, we reduce its complexity.

In order to ensure robustness, the optimisation should consider errors from multiple viewpoints, i.e., the sensors should observe the pattern from different viewpoints. This is a standard requirement of any calibration procedure. For example, for calibrating a stereo system, several images of both cameras are used.

In each collection, we store not only the raw sensor data but also the labels that describe where the pattern was detected in the sensor data. Naturally, the detection of the pattern may fail in some cases. This may be caused by issues in the detection algorithm, such as sensitivity to illumination, or simply because the sensor does not view the pattern in that collection. When, in a collection, there is at least one sensor that does not detect the pattern, we refer to it as an incomplete collection. It is also possible that only a portion of the pattern is identified, which occurs primarily due to a partial view of the calibration pattern in RGB images. We refer to these cases as partial detections. Since a collaborative cell is a large tridimensional space and the goal is to monitor its complete volume, it is common to have small or non-existent overlapping FoVs between different sensors. It is often very difficult to find a position of the calibration pattern which is viewed by all sensors simultaneously. For that reason, the number of incomplete collections and partial detections is larger than usual in a collaborative cell system. A sensor-to-pattern paradigm is clearly much more adequate to tackle such complex multi-modal and multi-sensor systems.

The next sections describe the configuration of a calibration procedure and the automatic labeling and manual annotation mechanisms which are available. Finally, we detail the objective functions for each of the three presented modalities.

3.2.1 Setup and Data Acquisition

Since the goal is to calibrate complex robotic systems, a prior step is required for configuring the calibration. This step defines which sensors are to be calibrated. The coordinate frames in the system are hierarchically organised in a topologically tree-like structure called the transformation tree. The transformation tree of the collaborative cell used in this work is shown in Figure 3.1.

The calibration of a sensor requires the definition of which specific transformations that is to be changed during the optimisation, in order to assess if the error is minimised. This

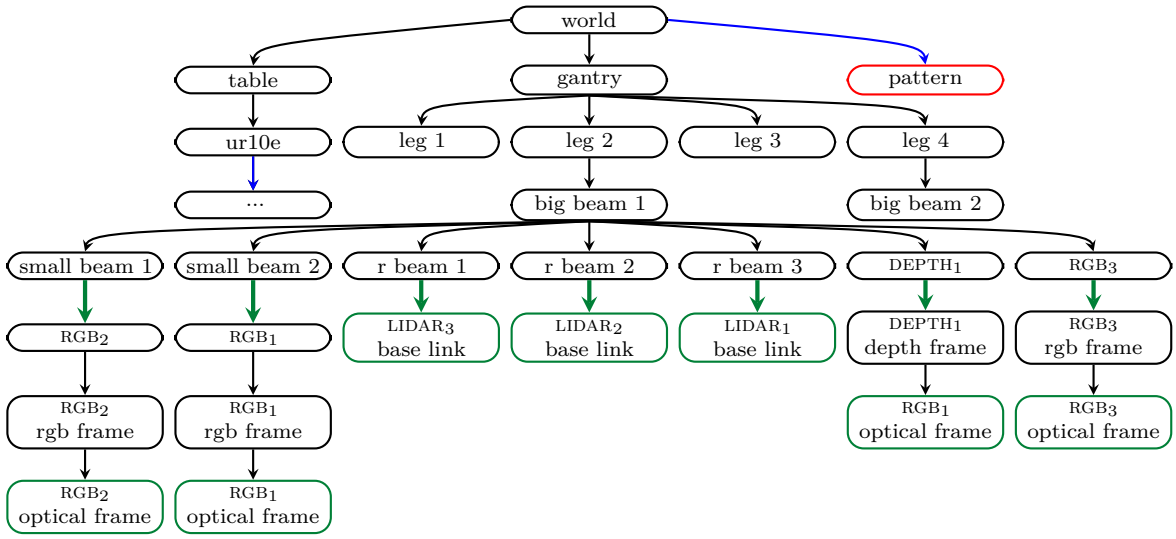

Figure 3.1: Example of a transformation tree that represents the chain of transformations between coordinate systems of the collaborative cell. Blue arrows signal that transformations are dynamic, green arrows denote that the transformation will be optimised, frames are highlighted in green when sensors output data in that coordinate frame, the red node represents the calibration pattern link, which is both dynamic and is to be calibrated.

transformation must belong to the chain of transformations that go from the common reference frame (world) to the sensor's coordinate frame. For example, in Figure 3.1, the sensor RGB_2 is mounted on the *small beam 1* coordinate frame, which is in turn assembled in the *big beam* coordinate frame. The idea is to define one transformation along the world-to-sensor chain to be estimated. These selected transformations are highlighted with green arrows in the figure. The other transformations remain unaltered during the calibration. In the example of RGB_2, the selected transformation was between the *small beam 1* and RGB_3 coordinate frames. Note that the selected transformation does not necessarily need to include the sensor's coordinate frame, as in the example of the sensor RGB_2.

The calibration procedure computes the overall sensor-to-world transformation, i.e., ${}^{s_i}\hat{\mathbf{T}}^{w}$ in (3.2), from the chain of transformations for that respective sensor, where one selected transformation is changing during optimisation and the others are static.

The process of acquiring data consists of moving a calibration pattern in front of the sensors in a way that the pattern is viewed by all sensors at some moments of the acquisition. The acquisition does not require that the pattern be visible to all sensors at the same time. The dimensions of the calibration pattern must also be specified during the calibration setup.

3.2.2 Simulation Setup

To support the development and validation of the perception and calibration strategies explored in this thesis, a complete simulation environment replicating the Laboratory for Automation and Robotics Collaborative Cell (LARCC) collaborative cell was constructed using the Gazebo simulator and the ROS. This environment integrates the physical structure

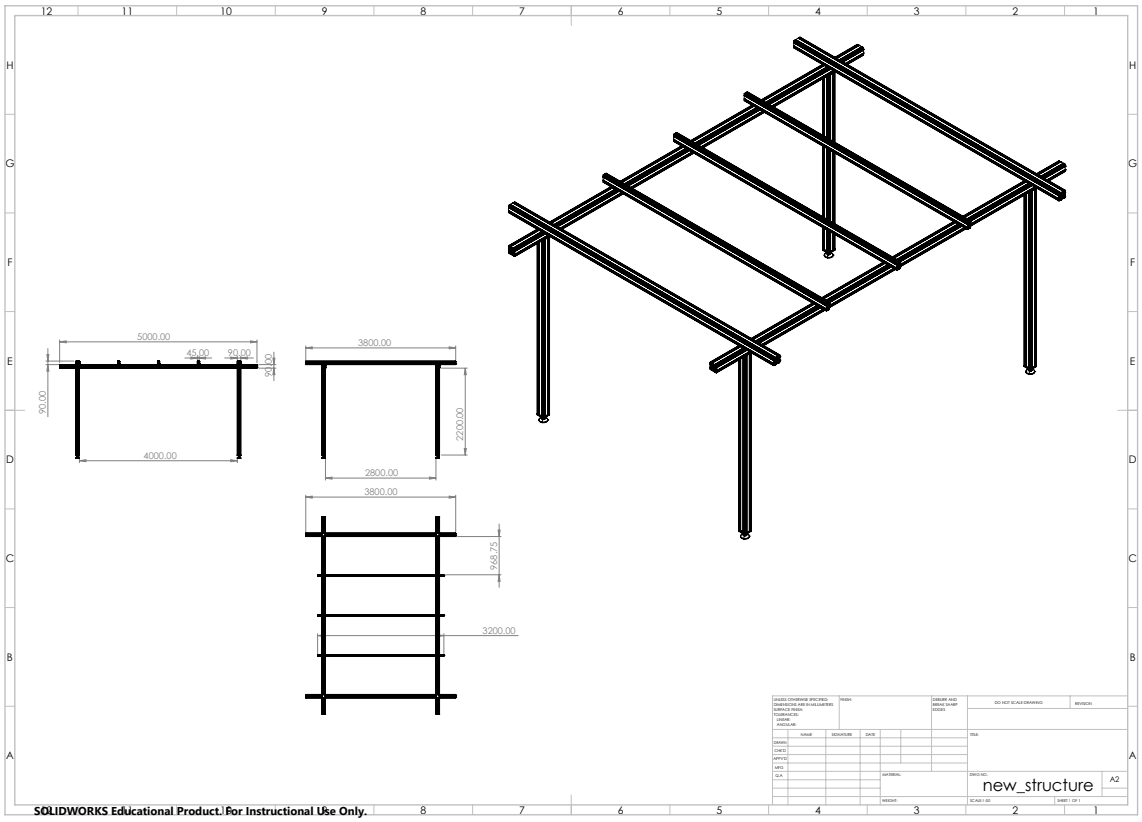

Figure 3.2: Details of the collaborative cell's main structure.

of the cell, the sensing equipment, and the robotic manipulator, allowing for realistic testing of algorithms under controlled yet representative conditions.

The simulation process began with detailed measurements of the physical structure. Based on these, a 3D model of the collaborative cell was developed, including elements such as supporting beams, structural feet, and table surfaces. These components were modeled using CAD tools and exported in formats compatible with ROS and Gazebo, ensuring geometric fidelity. The physical structure defines the internal space of the cell and serves as the foundation for sensor placement. A visual representation of the design is shown in Fig. 3.2.

Each element of the system was described using modular Xacro files. Individual files were created for the structural components, the UR10 robotic arm, and each of the sensors, including RGB, RGB-D, and LiDAR devices. This modular structure facilitated both reusability and maintenance. A macro Xacro file was then developed to define the spatial configuration of all components, establishing the relative positions and orientations of each within a shared reference frame. This ensured consistent deployment of the simulation setup and simplified further development. The specifications used for simulating each sensor, such as resolution, FoV, and update rate, are summarised in Table 3.1, which reflects both manufacturer data and the parameters applied within the simulation environment.

Following the structural modelling, a set of adapted Gazebo simulation packages was integrated. Although simulation support and ROS drivers were available for the selected

Table 3.1: Sensor configurations used in the simulation environment.

Sensor	Resolution (px)	FoV (°)	FPS	Range (m)
Orbbec Astra	640 × 480	63.1 (H) × 49.4 (V) (RGB) 58.4 (H) × 45.5 (V) (Depth)	30	0.6-8
Orbbec Astra Pro	1280 × 720 (RGB) 640 × 480 (Depth)	66.1 (H) × 40.2 (V) (RGB) 58.4 (H) × 45.5 (V) (Depth)	30	0.6-8
Orbbec Astra Mini	640 × 480	63.1 (H) × 49.4 (V) (RGB) 58.4 (H) × 45.5 (V) (Depth)	30	0.35-1
Kinect v1 (Xbox 360)	640 × 480	62 (H) × 48.6 (V) (RGB) 57 (H) × 43 (V) (Depth)	30	0.8-4
Velodyne VLP-16	N/A	360 (H) × 30 (V)	20	200

Note: Specifications based on manufacturer datasheets and default simulation plugin parameters. Field of view (FoV) values refer to horizontal (H) and vertical (V) angles.

sensors (Orbbec Astra[1], Velodyne VLP-16[2], Xbox Kinect[3]) and the UR10 arm[4], several adaptations were necessary. These included aligning coordinate frames, refining mesh files, and tuning plugin parameters to ensure accurate behaviour in the simulated environment. The resulting setup enabled the simulation of RGB and depth images, LiDAR point clouds, and full robotic arm motion, thus supporting both perception and manipulation scenarios.

A key addition to the simulation was a movable calibration target based on a ChArUco pattern, which was incorporated to support extrinsic calibration experiments. The pattern could be repositioned in six degrees of freedom via an interactive marker, allowing for controlled simulation of various calibration conditions. This included scenarios with incomplete pattern visibility, limited sensor overlap, or non-ideal viewing angles. By manually adjusting the pattern's pose, it was possible to simulate realistic edge cases and evaluate the robustness of the calibration procedures. The synthetic sensor data, capturing the target across multiple modalities, was recorded into `rosbag` files. While intrinsic parameters were initially defined by the simulation, they could also be optimised during calibration, allowing flexible testing of both extrinsic and intrinsic refinement strategies. An example of the simulated pattern across modalities is illustrated in Fig. 3.3.

To streamline deployment and testing, a structured set of ROS launch files was developed. Each sensor and the robotic arm had their own dedicated launch file, allowing for modular development and debugging. A unified bring-up launch file was also created to initialise the entire simulation environment in a single command, simplifying full-system experiments and reducing setup time.

This simulated environment provided a crucial platform for early-stage evaluation of system performance. It enabled comprehensive testing of calibration procedures, sensor configurations, and perception pipelines prior to any physical deployment in the LARCC collaborative cell, ensuring greater robustness and efficiency in subsequent experimental phases.

[1] `github.com/orbbec/ros_astra_camera.git`
[2] `github.com/ros-drivers/velodyne`
[3] `github.com/ros-drivers/freenect_stack.git`
[4] `github.com/UniversalRobots/Universal_Robots_ROS_Driver`

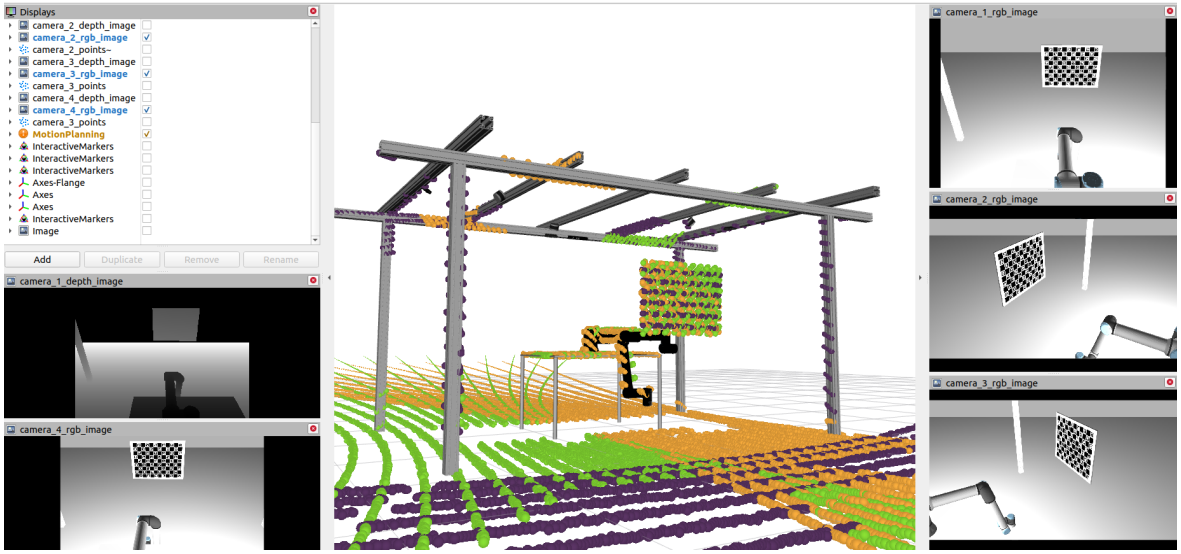

Figure 3.3: Representation of the simulated pattern across different sensors and modalities. The simulation shows 4 RGB cameras, 1 depth camera, and 3 3D LiDARs, with the green, purple, and orange point clouds.

3.2.3 Automatic and Manual Labelling

When saving a collection, the sensor data is labelled automatically. That means that information on where the pattern is identified in the data of each sensor is generated automatically. Naturally, the format of these pattern labels differs from modality to modality. For range data, a label is defined as the position of the outer edges of the physical chessboard. For RGB data, a label consists of the pixel coordinates of the inside corners of the pattern.

RGB

The RGB automatic labeling uses the *OpenCV ArUco Marker Detection* toolbox. We have configured it so that at least 25% of the total number of corners must be identified to assume a valid pattern detection. This automatic labeling of the RGB data using *ChArUcO* patterns is very accurate and efficient. For this reason, we have found no need to develop interactive tools to correct the automatic labels and produce manual annotations. Figure 3.4 shows an example of a labeled RGB image.

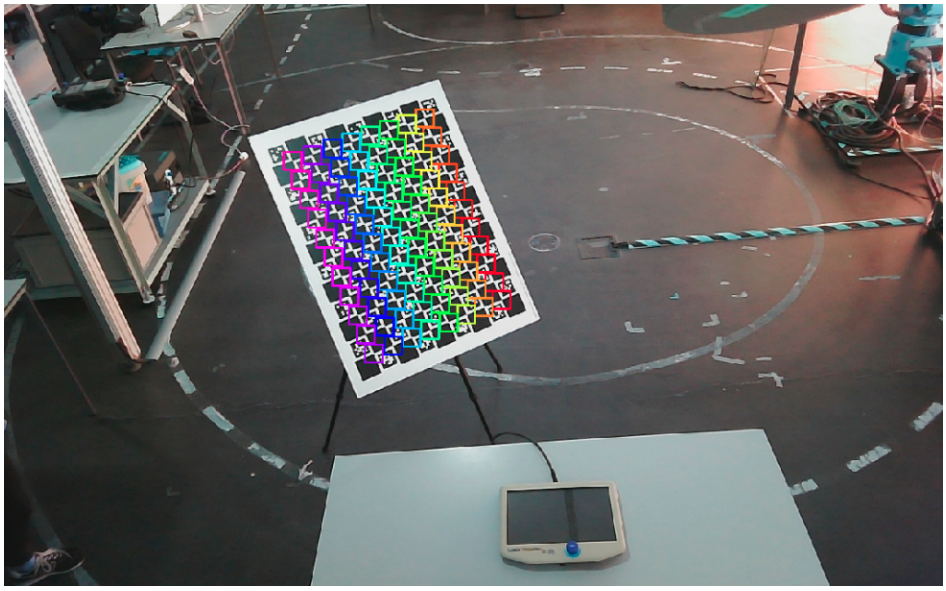

Figure 3.4: Example of a labeled image in a RGB camera.

3D LiDAR

The label representation for 3D LiDAR data consists of a list of points that belong to the sensor raw data and are identified as intersecting the pattern. In addition, we define two separate classes: points that lie on the pattern plane, and a subset of the former that are located on the boundaries of the pattern.

The procedure is semi-automatic, since it requires the user to define a seed point close to the pattern. This is done using an interactive marker in *Rviz*. That seed is used as the center of a sphere of predefined radius that selects only a small set of points where the pattern should be located. Then, the support plane of the pattern is searched using a *RanSaC* algorithm. Finally, the points that belong to the pattern are obtained as those that are close enough to the support plane, i.e., the *RanSaC* inliers. The boundary points are then found by collecting, for each vertical LiDARlayer, the left and rightmost inliers. Figure 3.5 shows an example of a labeled point cloud, where black points represent the physical limits of the calibration pattern, and green points represent the points inside the calibration pattern.

It must be noted that, if needed, the LiDAR labeling can be reviewed and corrected manually by selecting points in the point cloud and assigning them the proper labels.

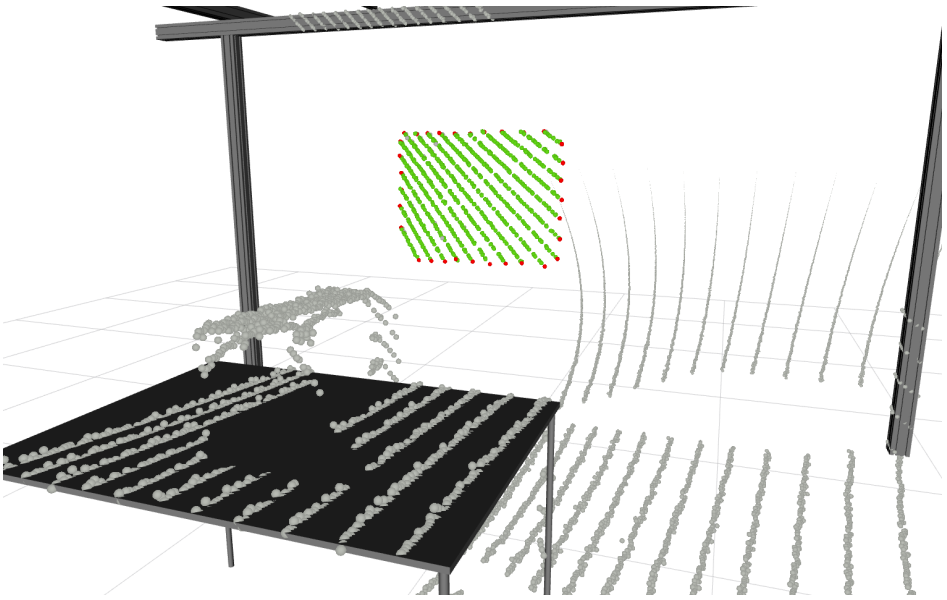

Figure 3.5: Example of a labeled point cloud from a 3D LiDAR. Gray points are raw data, points annotated as belonging to the pattern are highlighted in green, and points annotated as belonging to the boundaries of the pattern are annotated in red.

Depth Camera

The depth camera is labeled using a propagation mechanism, starting from an initial seed point. Similarly to the LiDAR labeling mechanism, the seed point is manually given in the first frame, and from then on it is automatically tracked from frame to frame, using the centre of mass of the detected pattern in the previous frame. As we are labeling adjacent frames, we can assume that, from frame to frame, the movement of the calibration pattern is small enough so that the centre of mass of the previous frame is still inside the area of the pattern in the subsequent frame. Note that it can be redefined manually if, for some reason, the detection of the pattern is not working correctly, for example, when the pattern leaves the FoV of the camera and returns moments later. The propagation algorithm starts from the initial seed point and uses a tetra-directional flood fill technique to propagate through the area of the calibration pattern. The labels of depth images are separated into two categories: boundary points and inside points. Figure 3.6 shows an example of a labeled depth image.

The automatic procedure detailed above does not work accurately in all frames. This is due to the nature of depth images and to the proximity of other objects to the pattern. For this reason, we have also developed a dataset reviewer in which incorrectly labeled images can be manually annotated by defining a polygon around the pattern. Then, the previously mentioned propagation algorithm is executed with this polygon acting as a propagation constraint, which results in accurately defined labels for depth data.

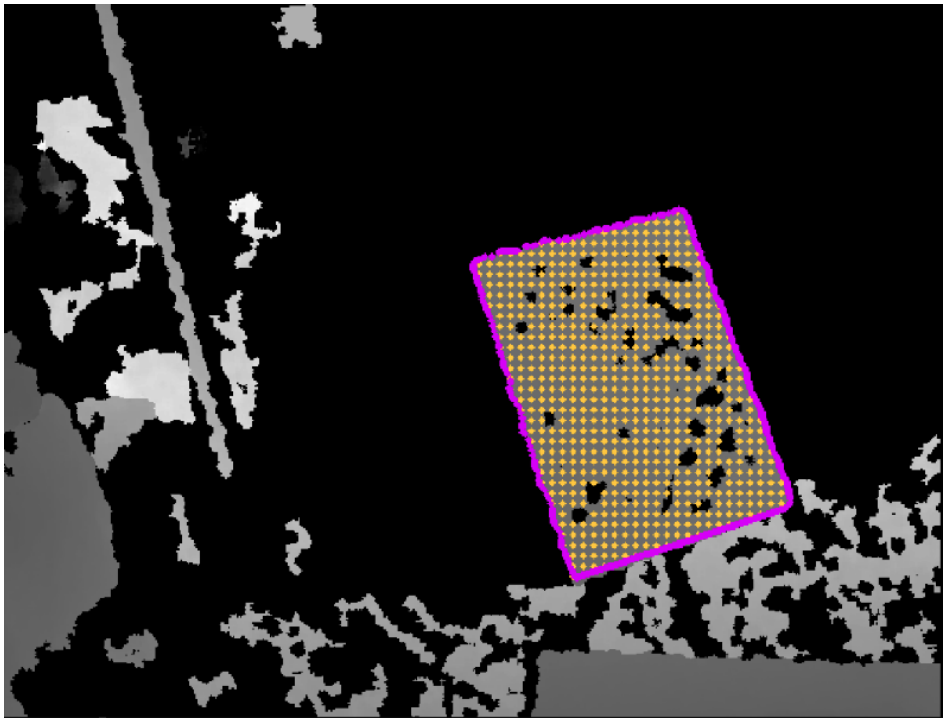

Figure 3.6: Example of a labeled depth map. Yellow points signal the subsampled LiDAR points annotated as belonging to the pattern, while purple points denote the annotation of the boundaries of the pattern.

3.2.4 Calibration

The calibration process begins by reading a previously recorded dataset, from which it extracts and organises labels for each sensor and each data collection into a dedicated data structure. These labels, along with the estimated positions of a calibration pattern (typically a chessboard), are used to compute modality-specific errors. These errors form the basis of an objective function, which is then optimised to estimate the rigid-body transformations between sensors and the calibration pattern.

A distinguishing feature of our approach is that both the sensors and the calibration pattern are allowed six degrees of freedom (translation and rotation). Granting this freedom to the pattern often leads to more accurate calibration results, as it reflects the true variability in its positioning during acquisition. However, this approach introduces a limitation: while the sensors are calibrated relative to one another and to the calibration pattern, they are not necessarily aligned with the fixed mechanical structures of the robotic system.

To address this, we propose an optional anchoring strategy. If the position of at least one sensor is known with sufficient confidence, it can be fixed in space during the optimisation process. This anchored sensor serves as a reference, and all other transformations (including those of the calibration pattern) are optimised relative to it. This constraint ensures that the final calibrated setup remains consistent with the physical configuration of the system.

The optimisation is performed using a non-linear least squares method, which minimises the defined objective function over the set of parameters. Each sensor is represented by six extrinsic parameters: three for translation (x, y, z) and three for rotation (r_1, r_2, r_3); while, in

cases where intrinsic calibration is also performed, the intrinsic parameters of the cameras are simultaneously estimated.

RGB Sensors

For RGB cameras, the calibration error is computed by comparing the detected 2D image coordinates of the calibration pattern's corners with their projected positions based on the estimated 3D transformations. The objective function used is presented in Equation (3.3), where the error $e_{[c,s,d]}$ for a given collection c, sensor s, and detection d, is defined as the squared norm of the difference between the detected image label $x_{[c,s,d]}$ and the projection of the corresponding 3D pattern point:

$$e_{[c,s,d]} = \left\| x_{[c,s,d]} - \mathcal{P}\left([{}^{p}_{c}T^{s} \times x_d]_{xyz}, k_s, u_s \right) \right\|^2, \tag{3.3}$$

here, $\mathcal{P}(\cdot)$ denotes the perspective projection function using the intrinsic parameters k_s and distortion coefficients u_s of sensor s, and ${}^{p}_{c}T^{s}$ is the estimated transformation from the calibration pattern to the sensor frame for collection c.

3D LiDAR

For 3D LiDAR sensors, the objective function is decomposed into two complementary components: orthogonal error and longitudinal error, capturing the alignment between the observed point cloud and the known geometry of the pattern.

The orthogonal error, $e_{o_{[c,s,d]}}$, measures how far the labelled points deviate from the pattern's surface along the pattern's local Z-axis. It is computed by transforming the labelled point x_i from the sensor frame to the pattern frame and extracting its Z component:

$$e_{o_{[c,s,d]}} = [({}^{s}T^{p}_{c})^{-1} \times x_i]_z, \tag{3.4}$$

the longitudinal error, $e_{l_{[c,s,d]}}$, evaluates the planimetric alignment in the XY-plane. It is defined as the squared distance from each detected boundary point, projected into the pattern frame, to its closest ground truth point on the known pattern geometry:

$$e_{l_{[c,s,d]}} = \min_{q \in \mathcal{Q}} \left(\left\| [x_q - ({}^{s}T^{p}_{c})^{-1} \times x_b]_{xy} \right\|^2 \right), \tag{3.5}$$

here, x_q represents a ground truth point on the pattern, and x_b is a point on the boundary of the detected label.

Depth Cameras

The calibration strategy for depth cameras mirrors that of LiDARs, as both types of sensors provide range data. However, a key distinction lies in the form of the raw data: while LiDAR produces 3D point clouds directly, depth cameras provide depth images in the 2D image domain. Each pixel encodes the distance from the camera to a scene point, which can be converted into 3D coordinates using the camera's intrinsic parameters. Equation (3.6) formalises this conversion process:

$$\mathcal{F}(x_d) = \begin{cases} Z = image(x_{pix}, y_{pix}) \\ X = \dfrac{x_{pix} - c_x}{f_x} \times Z \\ Y = \dfrac{y_{pix} - c_y}{f_y} \times Z \end{cases} , x_d = (x_{pix}, y_{pix}), \tag{3.6}$$

where (f_x, f_y) denote the focal lengths, and (c_x, c_y) represent the optical centre of the camera. The result (X, Y, Z) is the 3D coordinate of the point labelled by pixel x_d.

As with LiDAR, the orthogonal error is computed as the Z component of the transformed 3D label into the pattern's frame:

$$e_{o_{[c,s,d]}} = \left[({}^{s}T_{c}^{p})^{-1} \times \mathcal{F}(x_i) \right]_z . \tag{3.7}$$

The longitudinal error measures how closely the converted label points lie in the XY-plane of the known pattern geometry:

$$e_{l_{[c,s,d]}} = \min_{q \in \mathcal{Q}} \left(\left\| \left[x_q - ({}^{s}T_{c}^{p})^{-1} \times \mathcal{F}(x_b) \right]_{xy} \right\|^2 \right), \tag{3.8}$$

in this formulation, x_q is a known ground truth corner or edge point on the pattern, and $\mathcal{F}(x_b)$ is the 3D coordinate of a boundary label, converted from its pixel location.

It is important to note that this error function considers only the boundary points of the label. However, given prior knowledge of the pattern's dimensions, it is possible to synthesise a full set of ground truth edge points to increase precision. Each labelled point is projected from the sensor's frame into the pattern's local coordinate system and compared to its closest ground truth counterpart. Since only the longitudinal misalignment is of interest, the error is computed in the XY-plane only.

3.3 Tests and Results

As discussed in previous sections, our approach enables the simultaneous calibration of all the sensors in the system. However, other approaches cannot carry out this global optimisation since they operate with pairs of sensors. Because of this, the assessment of the calibration accuracy is conducted in a pairwise configuration, so that it may be applied both to our methodology (despite the fact that it calibrates the complete system) and also to other approaches.

Tests and results are divided and detailed in the following parts: Collaborative Cell Setup Calibration; RGB to RGB Evaluation; LiDAR to LiDAR Evaluation; LiDAR to RGB Evaluation; LiDAR to Depth Evaluation and Depth to RGB Evaluation.

3.3.1 Collaborative Cell Setup Calibration

As mentioned previously, a collaborative cell is a space where collaborative robots and humans can safely work together. The ultimate goal would be that the robot and human could participate in tasks with a common goal to achieve a more efficient work.

In our particular case, we have built a collaborative cell with 4 m × 2.8 m and 2.29 m high. Figure 3.7 shows the collaborative cell in simulated and real environment. In terms of sensors, the cell includes three LiDARs, one RGB-D camera and three RGB cameras. From now on, these sensors will be referred to as LIDAR_1, LIDAR_2 and LIDAR_3, DEPTH_1 and RGB_1, RGB_2 and RGB_3.

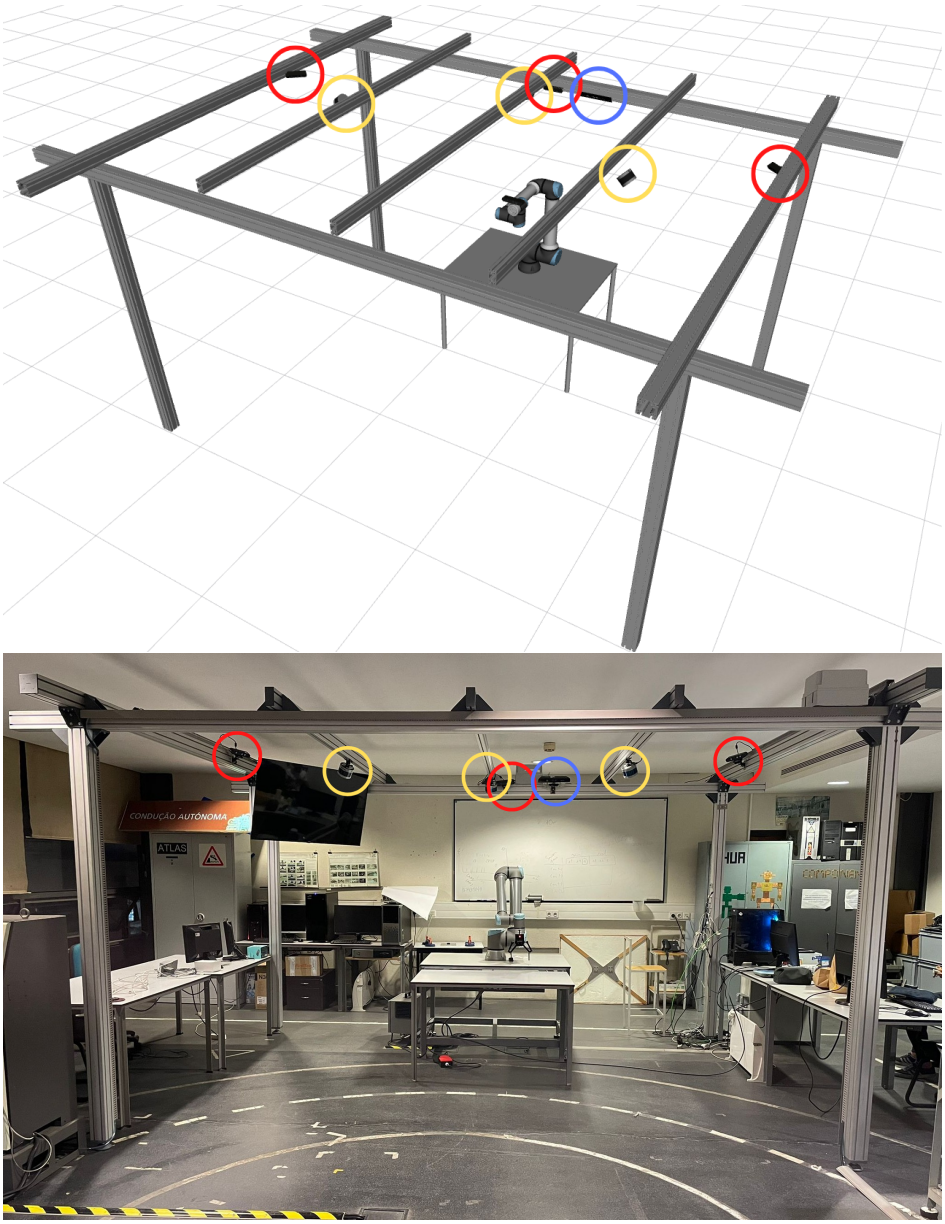

Figure 3.7: Simulated and real representation of the collaborative cell that serves as a case study. The cell contains a gantry where several RGB, depth, and LiDAR sensors are mounted. In the middle of the volume there is table and a robotic manipulator which will interact with human operators. Red circles represent RGB cameras, blue circles represent depth cameras and yellow circles represent 3D LiDAR.

ATOM allows to see the FoV of the different cameras within the configured system, according to their intrinsic parameters. Figure 3.8 shows a representation of the FoVs of the sensors and the coverage of the cell by the LiDAR point clouds. In the image, the gray point

clouds come from the LIDAR$_1$ (right), the green ones from the LIDAR$_2$ (centre) and the yellow ones from the LIDAR$_3$ (left). The purple frustum represents the FoVs of the sensor DEPTH$_1$. The RGB FoVs are represented by the light orange, gray and green for RGB$_1$ (left), RGB$_3$ (centre) and RGB$_2$ (right) respectively.

A video[5] has been made available that includes a demonstration of the complete calibration procedure for this collaborative cell.

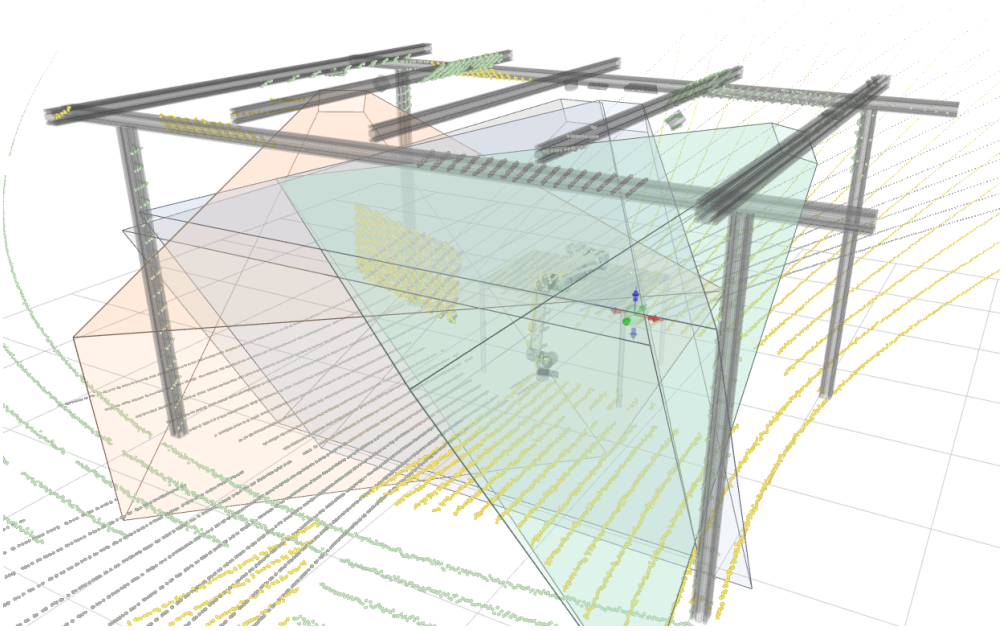

Figure 3.8: Fields of view (FoV) of the cameras mounted on the collaborative cell. The point clouds produced by the LiDAR are also shown.

Table 3.2 shows the details of the used train and test datasets for the results presented in this section. The train dataset is the dataset that is used for calibration and where the transformations between sensors are estimated. The test dataset is a non-calibrated dataset, with the sensors in the same position as the train dataset, where the results will be evaluated with the transformations obtained during the calibration of the train dataset.

Table 3.2: Descriptions of the datasets used in the experiments, where RGB partials mean the number of partial calibration pattern detections in the RGB sensors and complete denotes the number of collections where the calibration pattern was detected by all seven sensors.

Type of data	Dataset	# collections	# RGB partials	# complete
simulation	train dataset	23	35	5
	test dataset	17	26	4
real	train dataset	29	61	6
	test dataset	14	29	4

The evaluation of the calibration is conducted in a pairwise manner. The results will be presented both in the simulated system and using real data. Note that to calibrate the

[5] `https://youtu.be/KFPUTGR4rBw`

simulated system, we induced an initial estimate random error of 0.1 m and 0.1 rad to the initial position of the sensors to reinforce the validity of the method in simulation.

3.3.2 RGB to RGB Evaluation

The RGB to RGB sensor evaluation is computed by projecting the labels of the source sensor, using the calculated transformation matrix, to the target sensor image and calculating the reprojection errors.

Table 3.3 shows the root mean square errors for both simulation and real data calibration. In the calibration using simulated data, we obtained sub-pixel accuracy with an average of half a pixel. As expected, the accuracy in real data is lower, with an error of around 1.2 pixels. The reason for this could be that real data is less controlled and has more sources of error than simulation, such as, for example, illumination, reflectivity, and background noise that might influence the accuracy of detection of the calibration pattern. This table also presents the results using the same data for the *OpenCV Calibration Tool*, a very popular computer vision library used for stereo camera calibration. As discussed in Chapter 2, most calibration algorithms use a pairwise methodology, which is the case with *OpenCV Calibration Tool*. This means that, to calibrate the entire system, it would require a sequential pairwise calibration by calibrating all the possible combinations of two sensors. Note that for some camera pairs it was not possible to calibrate using *OpenCV*. This is because these pairs contain cameras with a very small overlapping FoV. Moreover, *OpenCV* uses a pattern detection that requires that the pattern be fully visible in the image in order to be detected. Because of this, there were no collections in which both cameras in the pair were able to detect the pattern. Since *OpenCV* is a sensor-to-sensor approach, it cannot operate in these circumstances.

We can also conclude that, even calibrating seven different sensors simultaneously, the proposed approach still managed to obtain better RGB pairwise results when compared to *OpenCV*.

We have also compared our approach with *Kalibr* [9][36], which is a more recent multi-camera intrinsic and extrinsic calibration tool. We were only able to use *Kalibr* in a pairwise configuration. This calibration framework, unlike *OpenCV Calibration Tool*, is already a multi-sensor method based on optimisation. Nonetheless, this method is not multi-modal and is only able to calibrate RGB cameras. Results are also presented in Table 3.3.

Table 3.3: Pairwise root mean square errors for the RGB to RGB evaluation in pixels.

	Our Framework		**OpenCV**		**Kalibr**	
Sensor Pair	Sim.	Real	Sim.	Real	Sim.	Real
RGB_1-RGB_2	0.684	1.536	(1)	(1)	(2)	1.010
RGB_1-RGB_3	0.463	1.085	0.675	1.828	1.290	0.906
RGB_2-RGB_3	0.541	1.113	0.578	(1)	(2)	0.743
average	0.563	1.245	0.627	1.828	1.290	0.825

(1) OpenCV error: No complete detections of the chessboard.

(2) Kalibr error: Cameras are not connected through mutual observations.

Kalibr also shows the same problem as OpenCV, since in some situations it is not able to calibrate when cameras have minimal overlapping FoVs. In the real-world scenario, *Kalibr* was able to perform calibration in all three camera pairs with sub-pixel performance. Although the performance is slightly better in comparison to our method, it should be noted that this calibration framework is only able to calibrate cameras. In contrast, our system calibrates several modalities all at the same time.

As we can see by looking at Figure 3.8, RGB_1 (orange frustum) and RGB_2 (green frustum) have minimal overlap and very different orientations. This makes it difficult to position the calibration pattern in such a way that it is visible by both sensors. For that reason, that camera pair is the most difficult to detect. Unlike OpenCV, *Kalibr* can calibrate this pair in the real data scenario because it uses a different chessboard detector, which does not require that the pattern is fully visible in the image. Even so, of the 29 available collections in the training real dataset, *Kalibr* was only able to use 8 for calibration.

Regarding the RGB_2-RGB_3 pair, *OpenCV* was not able to calibrate with the real data because the detections of camera RGB_2 are all partial, which is a problem that *Kalibr* does not have. Regarding simulation, camera RGB_2 has more complete detections using our pattern detection algorithm. However, the *Kalibr* detection algorithm was not able to produce detections of the pattern for both images in the same collection, and for that reason, it was not able to calibrate.

3.3.3 LiDAR to LiDAR Evaluation

The evaluation between LiDAR pairs is conducted by transforming the points of the source LiDAR into the coordinate system of the target LiDAR. Then, for each point in the target LiDAR, we obtain the closest transformed point in the source LiDAR and compute the distance between both.

Table 3.4 shows the calibration errors for the LiDAR-LiDAR pairs. Considering that the calibration pattern is at a distance of 2-2.5 m from each LiDAR, the maximum distance between LiDAR points measuring the pattern is around 100 mm. When transforming the labelled points of the source LiDAR to the target LiDAR coordinate system, the labels could end up in such a way that the scan of the two LiDAR have a significant displacement between them caused by the low sensor resolution.

Table 3.4: Pairwise root mean square errors for the LiDAR to LiDAR sensors evaluation in mm.

	Ours		**ICP**							
			Initial				Aligned			
			Average		Best		Average		Best	
Sensor Pair	Sim.	Real	Sim.	Real	Sim.	Real	Sim.	Real	Sim.	Real
LIDAR_1-LIDAR_2	26.472	77.504	76.541	294.977	28.088	207.025	30.398	68.524	127.468	75.617
LIDAR_1-LIDAR_3	33.049	69.234	248.526	84.633	43.230	265.582	33.033	67.312	32.367	70.103
LIDAR_2-LIDAR_3	39.401	13.946	423.132	140.611	38.212	15.900	38.361	24.105	115.198	183.892
average	32.974	53.561	249.400	173.407	36.510	162.836	33.931	53.314	91.678	109.871

Considering this low sensor resolution it is natural that the error values in Table 3.4 are in the magnitude of a few tenths of millimetres. Table 3.4 also shows the calibration results for the same datasets using the Iterative Closest Point (ICP) algorithm. ICP is a very common iterative solution for the alignment of two sets of 3D data. The difference between Initial and Aligned ICPs indications in the table is that Initial has the same initial estimate as our framework, and Aligned has a better initial estimate created by manually aligning the point clouds. The ICP algorithm is executed for the pair of point clouds in each collection, which means that, as *OpenCV Calibration Tool*, it is also a pairwise algorithm and requires some form of sequential pairwise calibration to calibrate all of the sensors in the system. Thus, there is an estimated transformation for each collection. The difference between the Best and Average ICPs is that the Average uses the average transformation estimated for all collections, while the Best makes use of the estimated transformation which had the least amount of estimated ICP error.

When comparing the ICP results with our framework, we can conclude that the only one that comes close is the ICP Aligned Average. Nevertheless, the ICP is apairwise method, while our method obtained similar results while calibrating all sensors simultaneously.

3.3.4 LiDAR to RGB Evaluation

The LiDAR to RGB camera error metric is assessed by projecting the LiDAR labelled points to the RGB image using the transformation between sensors estimated during calibration. However, the labels of the RGB data correspond to the inside corners of the chessboard or *ChArUcO*. In contrast, the LiDAR data labels correspond to the physical limits of the chessboard. Therefore, the RGB images for each collection need to be manually labelled in the test dataset to identify the physical limits of the pattern. Those labels can then be compared to the LiDARlabels using the reprojection error.

Table 3.5 shows the reprojection errors obtained from pairwise evaluations of the calibration of both the simulated and the real data. As explained before, the LiDAR have a low resolution so it is expected that these errors have higher magnitude when compared with RGB-to-RGB evaluation errors. In this evaluation, we can see that results are on average around 1 or 2 pixels of reprojection error. The difference between simulated and real results is approximately 0.5 pixels. This shows that the evaluation is consistent: as expected, the real data is less accurate. Also, there is no significant difference between the several pairs of sensors. Our explanation is that since the LiDAR has a lateral 360° FoV, there is complete overlap between all camera-LiDAR pairs.

Figure 3.9 shows the projection of the three LiDAR point clouds into an RGB frame after calibration. The point clouds are colored according to the distance to each sensor. As such, changes in an object in the image should align with changes in the colour of the point clouds. As we can see, point clouds align almost perfectly with the shape of the chessboard in the image, which demonstrates that the calibration was successful.

Table 3.5: Pairwise root mean square errors for the LiDAR to RGB sensor evaluations in pixels.

Source Sensor	Target Sensor	Simulation	Real Data
LIDAR_1	RGB_1	1.726	2.516
LIDAR_2	RGB_1	1.801	2.582
LIDAR_3	RGB_1	2.692	3.297
LIDAR_1	RGB_2	3.502	3.837
LIDAR_2	RGB_2	3.659	3.173
LIDAR_3	RGB_2	2.854	2.754
LIDAR_1	RGB_3	2.892	3.767
LIDAR_2	RGB_3	1.763	2.768
LIDAR_3	RGB_3	2.352	3.189
Average		2.582	3.098

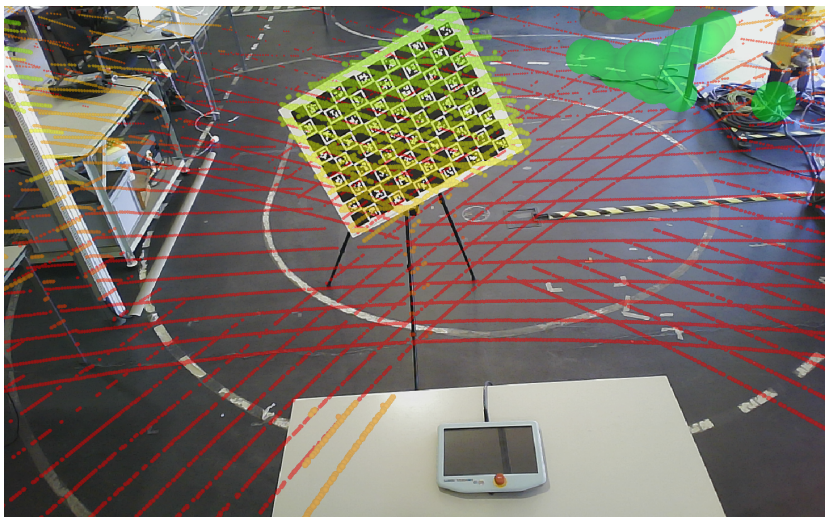

Figure 3.9: Projection of point clouds from all LiDAR to the image of camera RGB_3 after calibration. The point clouds are colored according to the distance to each sensor. As such, changes in an object in the image should align with changes in colour of the point clouds.

3.3.5 LiDAR to Depth Evaluation

Similarly to the LiDAR-RGB evaluation, the LiDAR-depth evaluation consists of projecting the LiDARpoints to the depth image. The difference is that the depth labels are also the physical limits of the chessboard, so we can directly compare the points without needing additional manual labeling.

Table 3.6 shows the results of the calibration error for simulated and real data. Once again, simulated and real results have a small sub-pixel difference, which shows consistency. Table 3.6 also shows calibration results using the ICP technique, where the different variants are the same as the ones in the LiDAR-LiDAR evaluation. None of these techniques obtain calibration results as good as the ones obtained with our methodology.

Figure 3.10 shows the projection of the point clouds from three LiDAR into the depth map. As we can see, the calibrated point clouds align well with the pattern and other features in the image, like the table at the bottom, and the structure of the cell on the left side.

Table 3.6: Pairwise root mean square errors for the LiDAR to depth sensors evaluation in pixels.

	Ours		**ICP**							
			Initial				Aligned			
			Average		Best		Average		Best	
Sensor Pair	Sim.	Real	Sim.	Real	Sim.	Real	Sim.	Real	Sim.	Real
LIDAR_1-DEPTH_1	1.281	1.791	12.591	166.459	1.575	84.895	5.563	9.170	2.094	8.339
LIDAR_2-DEPTH_1	1.054	1.608	16.478	30.654	45.970	4.883	2.477	4.392	1.915	6.114
LIDAR_3-DEPTH_1	1.584	2.058	34.246	148.168	5.951	144.052	3.025	2.055	3.493	2.050
average	1.306	1.819	21.105	115.094	17.832	77.943	2.751	5.206	2.501	5.501

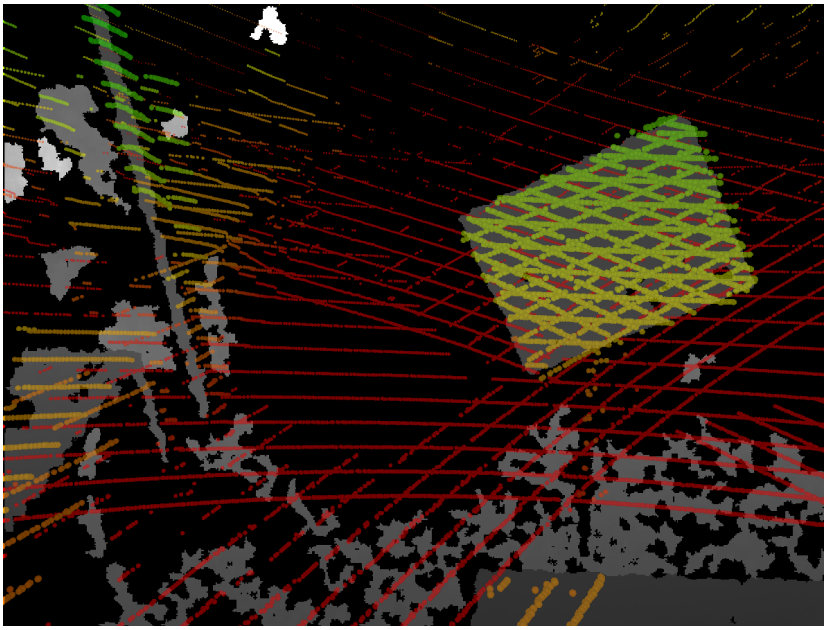

Figure 3.10: Projection of point clouds in the DEPTH_1 sensor depth map after calibration. The point clouds are colored according to the distance to each sensor. As such, changes in an object in the image should align with changes in the colour of the point clouds.

3.3.6 Depth to RGB Evaluation

On the depth-to-RGB pairwise evaluation, we project the depth labels to the RGB image using the transformations obtained during calibration. Again, there is a difference between the nature of the labels, so we use the annotations of the RGB images that were already made for the LiDAR-RGB evaluation to compare the physical pattern limits.

Table 3.7 shows the calibration errors of the depth-RGB pairs. The average errors are around 3 pixels, which are clearly above those for the LiDAR-to-RGB evaluation. We believe this is because the depth estimation is as precise in the depth sensors when compared to LiDAR.

Table 3.7: Pairwise root mean square errors for depth-to-RGB sensors evaluation in pixels.

Source Sensor	Target Sensor	Simulation	Real Data
DEPTH_1	RGB_1	3.328	3.990
DEPTH_1	RGB_2	3.212	4.553
DEPTH_1	RGB_3	3.642	3.584
Average		3.394	4.042

3.4 Final Considerations

This method solves the problem of the calibration of complex, multi-sensor, and multi-modal systems. To do so, we created a calibration framework based on a sensor-to-pattern paradigm, which has clear advantages over sensor-to-sensor calibrations, which are the basis for most of the current calibration approaches. Our approach provides several improvements w.r.t. the state-of-the-art, such as:

- a solution to calibrate any number of sensors and several modalities;
- a solution for systems with non-overlapping FoVs.
- the ability to accurately calibrate RGB cameras with partial detections;
- the simultaneous calibration of any number of sensors;

Furthermore, we provide a complete calibration framework with seamless integration with the ROS ecosystem, available at `https://github.com/lardemua/atom`.

Results show that our framework is able to achieve similar, or even better performance when compared with other state-of-the-art pairwise calibration methods, while calibrating all sensors from three different modalities simultaneously.

One shortcoming of our approach is the inability to calibrate the sensors with the structure of the robotic system. For example, in the case of the collaborative cell used in the experiments, it was necessary to manually calibrate one sensor w.r.t. the gantry structure. Then, this sensor is fixed, and the calibration moves all other sensors w.r.t. the fixed one. A better, automatic procedure for solving this problem would be an interesting addition.

As discussed throughout the chapter, collaborative cells are highly complex systems that render current calibration approaches unusable. Furthermore, collaborative cells have several additional challenges, such as the minimal overlapping FoVs between sensors. Our approach is able to tackle all these challenges, as the method was able to carry out a successful calibration of a highly complex collaborative cell.

CHAPTER 4

New Methodology to Calibrate Depth Sensors in Multi-Modal Dynamic Setups

4.1 Introduction

The importance of calibration becomes increasingly evident as sensor technology advances and multi-modal learning gains prominence in diverse fields like remote sensing, robotics [137], Simultaneous Localization and Mapping (SLAM) [138], scene classification [62], and autonomous driving [84]. Data fusion can occur at different levels, including early fusion (data-level and feature-level) and late fusion (decision-level) [139]. While late fusion performs classification independently for each sensor, early fusion relies heavily on accurate calibration to seamlessly integrate data from various sources by aligning them within a common reference system. This multi-sensor calibration process, both intrinsic and extrinsic, is essential for transforming data from distinct sources into a unified coordinate framework, enabling effective and robust multi-modal sensor fusion [139].

RGB-D cameras, which combine RGB and depth sensors, are widely used in applications requiring both types of data. Although these cameras come with factory calibration, it often fails to meet the precision demands of certain applications, particularly those requiring extrinsic calibration between sensors. When multiple RGB-D sensors are involved, determining the extrinsic parameters between various sensor pairs becomes critical. In industrial environments, such as collaborative robotic cells, deploying multiple sensors ensures complete coverage for human safety and efficient operation. These sensors may include several RGB-D cameras or devices of varying modalities, making cross-calibration between them a necessity.

A specific and increasingly common calibration scenario is the hand-eye configuration, where cameras are mounted on the end-effector of a robotic manipulator. Calibration in this setup involves determining the transformation between the camera and the end-effector.

While existing approaches focus predominantly on calibrating RGB sensors in hand-eye setups, incorporating depth sensors into this configuration is a promising innovation. Such an approach enables applications like 3D object reconstruction through manipulator-controlled multi-view imaging. To the best of our knowledge, state-of-the-art hand-eye calibration methods largely neglect the depth modality. This chapter addresses this gap by introducing a novel methodology that includes depth sensors in hand-eye calibration, expanding the potential applications of such systems.

In large-scale industrial environments where hand-eye configurations coexist with fixed sensors, extrinsic calibration between mobile and stationary sensors adds complexity. This scenario requires cross-calibration between sensors of different modalities, including those mounted on robotic manipulators and those fixed to the environment.

Despite ongoing research, current calibration methods face several limitations. Most approaches treat RGB-D cameras as inseparable sensor units and offer little support for the flexible, independent calibration of RGB and depth components, particularly in dynamic contexts such as hand-eye setups. Furthermore, there is a lack of unified frameworks that enable accurate extrinsic calibration across heterogeneous sensor types, such as RGB, depth, and LiDAR, especially when combining fixed and mobile sensors.

This work tackles the challenge of calibrating complex multi-modal sensor systems in dynamic robotic environments, with a particular focus on configurations that combine mobile (hand-eye) and static sensors. To address these challenges, we detail a methodological extension of the ATOM framework [48], [80], [140] to support depth sensors via a dedicated cost function, enabling independent calibration of RGB, depth, and LiDAR modalities while maintaining global consistency. Our approach leverages a sensor-to-pattern strategy based on a transformation tree of coordinate frames and is validated in both simulated and real-world scenarios.

The contributions of this chapter can be summarised as follows:

- A detailed and extensible methodology for integrating depth sensors into existing calibration workflows, specifically in robotic systems involving hand-eye configurations;
- Comprehensive experimental validation of the proposed calibration pipeline across mobile and fixed multi-modal sensors—including RGB, depth, and LiDAR—highlighting robustness in both simulated and real-world industrial scenarios;
- Extension of the ATOM framework [48], [80], [140] to support the calibration of depth sensors, broadening its application to multi-modal sensor systems.

The subsequent sections provide a detailed account of the methodology and results. Section 4.2 explains the calibration process, emphasising the depth modality and its unique challenges. Section 4.3 presents experimental outcomes, beginning with standalone RGB-D calibration and culminating in a multi-modal system featuring hand-eye setups and fixed sensors, including LiDAR, RGB, and depth. Finally, Section 4.4 summarises the chapter's contributions.

4.2 Methodology

The process of determining the extrinsic calibration of a system, specifically the transformations between various coordinate systems, typically necessitates the presence of a common feature, usually a calibration pattern, within the scene. This feature facilitates the estimation of the relative positions of each sensor. In our approach, we refer to this feature as a "calibration pattern", which can take the form of either a chessboard [65] or charuco [43], [44], [53]. The calibration pattern, depending on the sensor modality, provides visual or physical features that enable the calibration process.

The methodology outlined in this manuscript extends a previously developed calibration tool, known as ATOM [48], to include the depth modality within its spectrum of sensors for calibration. The calibration of a single RGB hand-eye was already approached previously with the ATOM framework [80]. However, this extension demonstrates that the framework can calibrate systems that combine both fixed and dynamic sensors. Furthermore, the addition of the depth modality aligns with ATOM's sensor-to-pattern calibration approach, deviating from the more conventional sensor-to-sensor calibration methods. This sensor-to-pattern approach sees calibration as a single optimisation process for all sensors, ensuring that the position of the pattern and the sensors are optimised so that the detections of the pattern match its actual position.

In our method, a *detection* occurs when a sensor identifies the calibration pattern at a given instant. When multiple sensors simultaneously detect the same instance of the pattern, this forms a *collection*, which also includes the corresponding (uncalibrated) transformations of the system. A series of such collections, gathered from different sensor poses or pattern configurations, constitutes a *dataset*. This structure is central to our calibration framework, as it enables joint optimisation of sensor-to-pattern transformations across heterogeneous and potentially asynchronous sensors. Figure 4.1 illustrates the definitions of a detection, collection and dataset.

ATOM's approach to extrinsic calibration deviates from the norm by utilising a sensor-to-pattern framework instead of the conventional sensor-to-sensor approach. Equation (4.1) clarifies the underlying philosophy of the sensor-to-pattern approach. It expresses that the cost function error, F, is a function of the transformation between each sensor and the world, denoted as ${}^{s_i}\hat{T}^{w}$, the detection of the pattern for that sensor, d_{S_i}, and its intrinsic parameters, K_{S_i}. The error is also contingent on the transformation between the world and the pattern, represented as ${}^{w}\hat{T}_{c}^{p}$ for each collection. The error term $e(\cdot)$ is modality-specific: RGB sensors compute a reprojection error in image space, while range sensors (LiDAR and depth) use 3D geometric criteria. Specifically, the depth error includes the orthogonal component e_o (Equation (4.2)), which measures misalignment along the pattern's normal axis, and the longitudinal component e_l (Equation (4.3)), which evaluates the distance between detected boundaries and the known pattern contour in the XY plane. This unified error formulation enables a joint optimisation across sensor modalities.

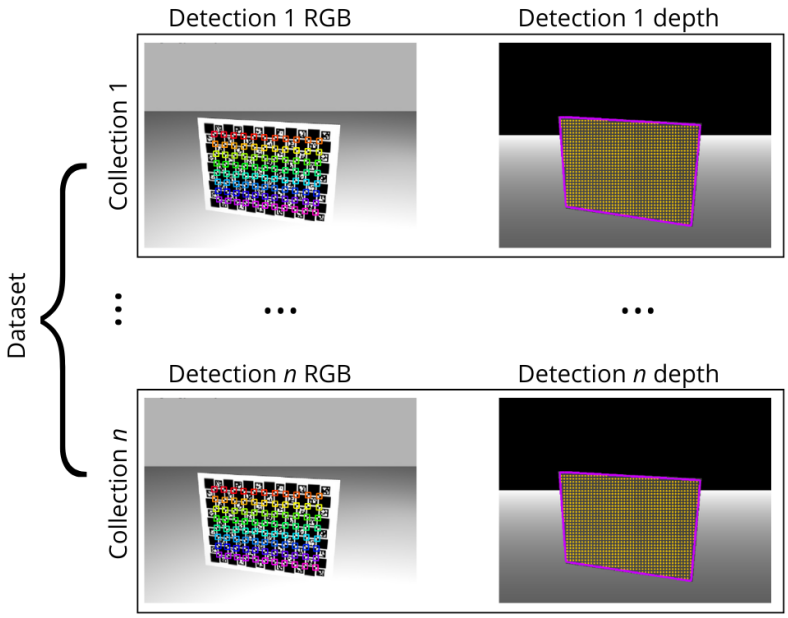

Figure 4.1: Illustration of the definitions of a detection, collection and dataset.

$$F = \underset{{}^{s_i}\hat{T}^w}{\arg\min} \left(\sum_{\mathcal{C}} \sum_{\mathcal{S}} e\left({}^{w}\hat{T}_c^p, {}^{s_i}\hat{T}^w, d_{s_i}, K_{s_i} \right) \right) \tag{4.1}$$

The cost function employed by the algorithm for solving the optimisation problem is modality-specific and depends on how the sensor perceives the calibration pattern. Range sensors (such as LiDAR and depth sensors) detect the physical boundaries of the pattern, while RGB sensors identify the black and white patterns on the calibration board. For detailed information on the cost functions for RGB and LiDAR sensors, please refer to [48].

Incorporating the depth modality into the ATOM framework was motivated by the growing demand for RGB-D camera calibration. A distinct cost function was developed to accommodate depth sensors. Depth sensors operate by measuring the distance to objects in the scene by projecting specific infrared light patterns, classifying them as range sensors. They are incapable of interpreting the printed pattern on the calibration board, but they can detect the physical boundaries of the pattern in the image or depth cloud, akin to LiDARs. The cost function for depth sensors comprises two main components: orthogonal (e_o) and longitudinal (e_l) errors. The orthogonal error, represented by equation (4.2), measures the distance between the pattern at its current pose and the points detected as pattern points in the depth data. In theory, this means that the detected pattern points for that collection and sensor, $X_{[c,s]}$, need to be transformed from the camera coordinate frame to the pattern coordinate frame using the estimated transformation between the sensor and the pattern for that particular collection, ${}^{s}\hat{T}_c^p$. The distance that equals the orthogonal error is the variation of the Z coordinate, considering that the pattern is in the XY plane in its own defined coordinate frame, as is represented in Fig. 4.2. This means that it is the Z coordinate from the detected pattern points in the depth image transformed to the pattern's coordinate system.

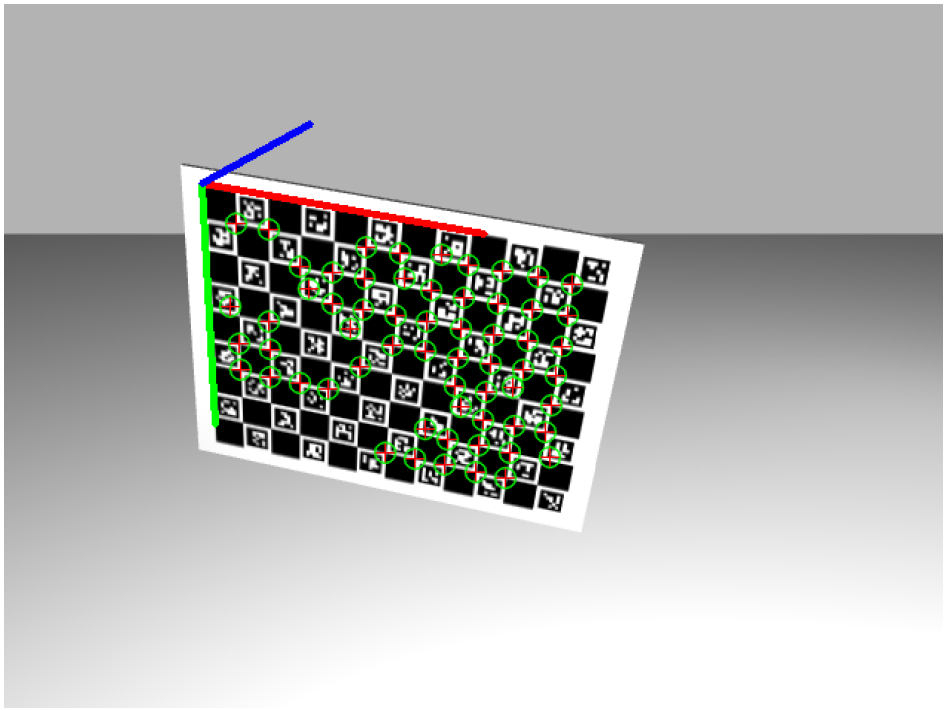

Figure 4.2: Calibration pattern's coordinate frame, where circles correspond to the detected corners of the pattern.

$$e_{o_{[c,s]}} = \left[\left({}^{s}\hat{T}_{c}^{p} \right)^{-1} \cdot X_{[c,s]} \right]_{z} \tag{4.2}$$

The longitudinal error, expressed by equation (4.3) is calculated using the points that belong to the physical limits of the calibration pattern. The boundary points detected in the depth image, $x_{[c,s,b]}$, are once again projected to the pattern's coordinate frame using the estimated transformation between the sensor and the pattern in that particular collection, ${}^{s}T_{c}^{p}$. Then, we calculate the Euclidean distance (in the XY plane) between that point and all the points in the pattern's border that are sampled using the lines that define the pattern's body that are defined as $\mathcal{Q}$. This allows us to find, for each labeled boundary point, the closest point from the set of sampled points, P, which will be considered the error. The objective is for all the boundary points to have corresponding pattern points and for all their Euclidean distances to be the closest possible, which would mean that the labeled data matches the pattern's position.

$$e_{l_{[c,s,b]}} = \min_{q \in \mathcal{Q}} \left(\left\| \left[q - \left({}^{s}\hat{T}_{c}^{p} \right)^{-1} \cdot X_{[c,s,b]} \right]_{xy} \right\|_{F}^{2} \right) \tag{4.3}$$

4.2.1 Labelling Depth Data

Depth images require labelling to establish the pattern's location within each collection. The pattern is identified through the selection of a seed point within the image, which serves as the starting point for a propagation algorithm that tries to find its boundaries. Initially, this seed point is positioned at the centre of the image. However, if it is not within the

boundaries of the pattern, the user has the option to choose a point within the pattern as a new seed point for propagation.

Once the seed point is accurately defined, the algorithm calculates the seed point for the subsequent frame by determining the centre of the detected pattern's shape in the current frame. This approach is valid under the assumption of continuous pattern movement, which ensures that the seed point will remain within the boundaries of the pattern in the next image. With a well-defined seed point, a flood-fill algorithm is applied to determine the rest of the shape of the pattern.

Post-processing of the image is carried out to fill holes and enhance the definition of the pattern's shape. The pattern's boundaries are identified using a "find contours" algorithm. In Fig. 4.3, we can observe a 3D representation of points in space with respect to the camera's coordinate system on the left, along with an overlay of the pattern aligned with the detected depth boundaries on the right.

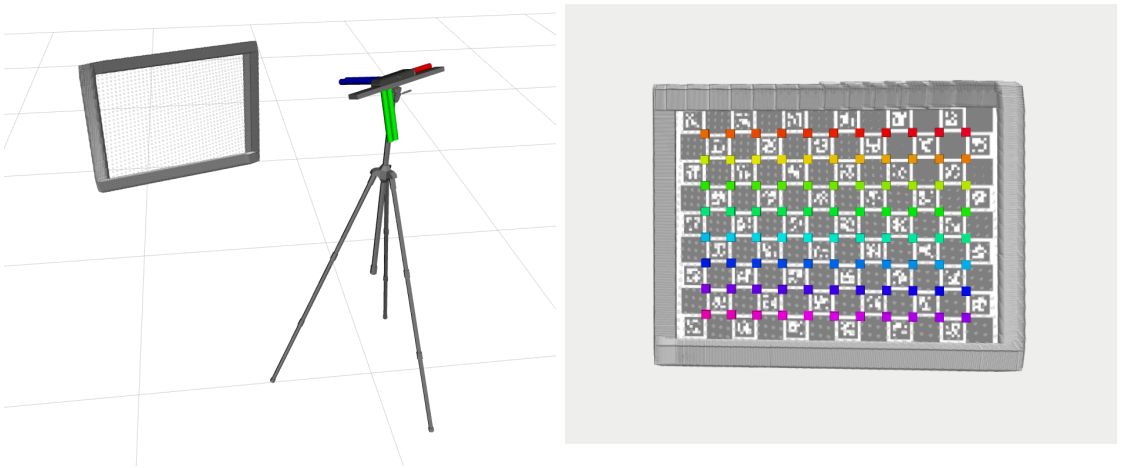

Figure 4.3: Representation of the detection of the 3D boundary points of the calibration pattern by depth sensors.

4.3 Tests and Results

In this section, we present results for two systems, both in simulated and real environments. The first system is composed of a single RGB-D camera. The hardware used in this setup was an Asus Xtion Pro. The second system is a complex system including an RGB-D hand-eye robot and three additional fixed sensors (RGB, depth, and LiDAR). The hardware used in this setup was an Orbecc Astra as the RGB camera, a Microsoft Kinect1 as the depth camera, an Orbecc Astra Mini as the hand-eye RGB-D and a Velodyne VLP-16 as the 3D LiDAR.

Digital twins are built in ROS to match the real-world systems. The use of simulated data allows for controlled data and a known ground truth, which can be used to evaluate the performance of the framework in different environments and induce different types of errors.

Although the framework uses a sensor-to-pattern, all-in-one approach, evaluations are pairwise. The general approach when evaluating a pair of sensors is to project the pattern

detections from one sensor to another, using the transformation estimated during calibration, and obtain the error between the projected and detected data. The evaluation uses the physical limits of the pattern, except in RGB pairs, where the projected points are the Charuco or chessboard pattern detection corners. When evaluating combinations that involve RGB and other modalities, the physical borders in the RGB images of the dataset are annotated by hand to enable a fair comparison.

4.3.1 RGB-D Calibration

RGB-D cameras are composed of RGB and depth sensors. In this experiment, we will carry out a simple calibration of an RGB-D system and compare it with the factory calibration and other state-of-the-art RGB-D calibration methods.

Fig. 4.4 shows the transformation tree of our RGB-D system, that is, the topology of coordinate frames that compose the system. When configuring ATOM's framework, the user needs to define a transformation to optimise for each sensor. The transformation does not necessarily need to be related to the link from which the sensor data is produced. For example, in this case, the frame ID for the RGB and depth data are the optical frames but, for calibration, it is more interesting to calibrate the transformations from the world link to the RGB and depth links, respectively. In more complex systems, the range of options for optimisation is even wider and it is up to the user to define which transformation they want to calibrate for each sensor.

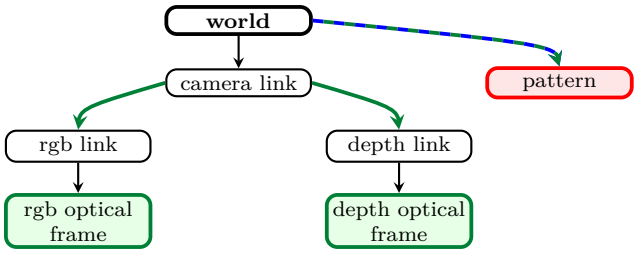

Figure 4.4: Transformation tree for an RGB-D system. Green arrows represent the transformation that will be estimated during calibration. Blue arrows represent dynamic transformations. Green nodes represent the sensor links from which data is output.

Fig. 4.5 shows an example of data labelling for RGB and depth sensors for simulated and real data. As mentioned in previous sections, RGB sensors detect the printed ArUcO codes on the pattern, while depth sensors detect the limits of the board. To allow comparison between calibration metrics, the physical limits of the board are manually labelled for the RGB images for each collection, since the automated proposal only detects the ArUcOs in the images.

Table 4.1 summarises the key characteristics of the datasets employed in the RGB-D calibration experiments. The train dataset refers to the data used to compute the calibration parameters. In contrast, the test dataset, although uncalibrated, shares the same sensor configuration and is used to evaluate the accuracy of the estimated transformation. Evaluation is performed in a pairwise manner by projecting the annotated boundary points, X_b^{ss}, from the source sensor (ss) onto the target sensor (ts). These boundary points are obtained using

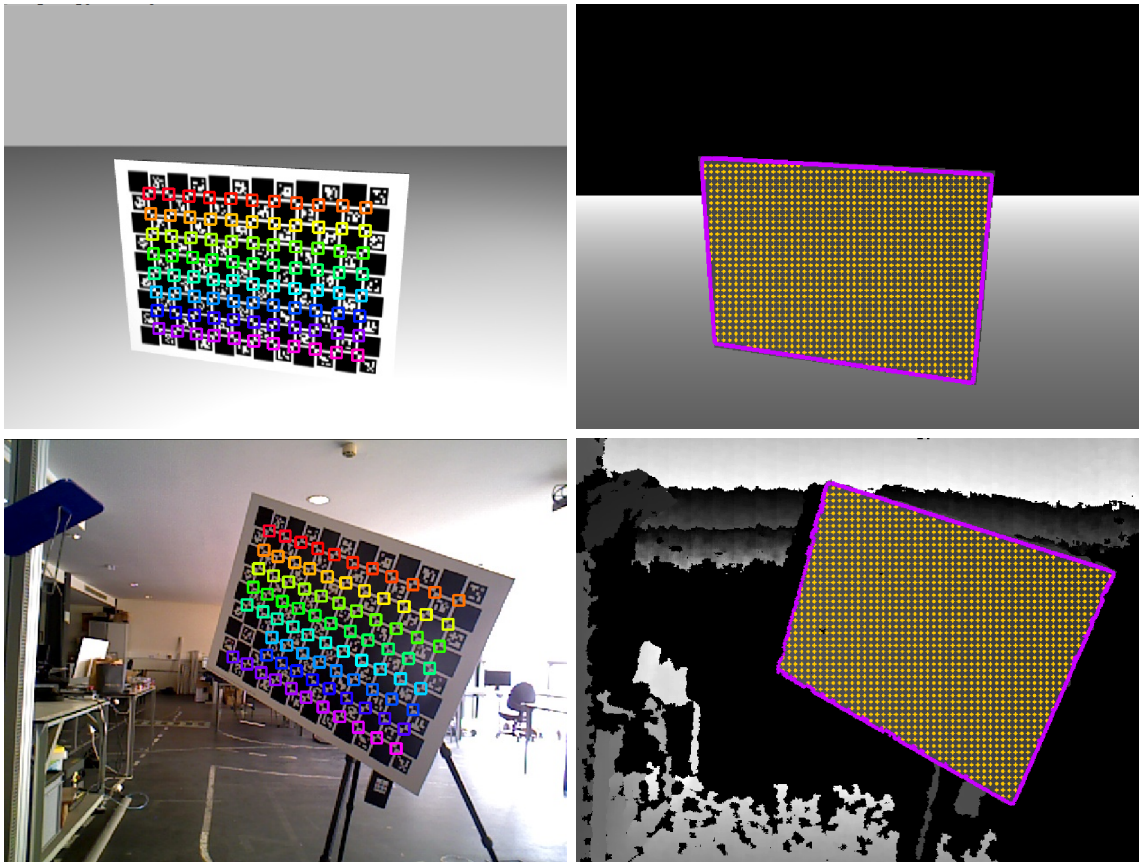

Figure 4.5: Example of detection for RGB and depth in an RGB-D system for simulation (2 top images) and real data (2 bottom images).

a custom annotation script that sequentially displays the RGB images from each collection, allowing for manual annotation of the board's physical limits, as shown in Fig. 4.6.

The projected points are then compared against the corresponding boundary labels, X_b^{ts}, of the target sensor. The projection error is computed using equation (4.4).

$$e_{rms} = \sum_{s \in X_b^{ss}} \left(\min_{t \in X_b^{ts}} \left(\left\| t - {}^{ss}T^{ts} \cdot s \right\|^2 \right) \right) \tag{4.4}$$

The real-world experiments were conducted in an indoor laboratory setting with controlled ambient lighting, although some variability in lighting occurred due to natural daylight. The Asus Xtion Pro RGB-D sensor was mounted on a fixed tripod, and the calibration pattern was also mounted on a custom-made structure that allows us to manually move the pattern across the scene to obtain different poses. Pattern distances from the sensor varied between 1 and 3 m. For the real dataset, a total of 16 collections were used for training and 9 for testing. Each collection corresponded to a distinct pose of the calibration pattern, ensuring variation in orientation and distance. Despite efforts to maximise coverage, some RGB frames only partially captured the pattern, as indicated in Table 4.1.

First, we tested our calibration methodology in a simulated environment. In this context, it is possible to know the exact transformations between the sensors, i.e, the ground truth pose, which allows us to evaluate exactly how well the algorithm is performing. In real scenarios,

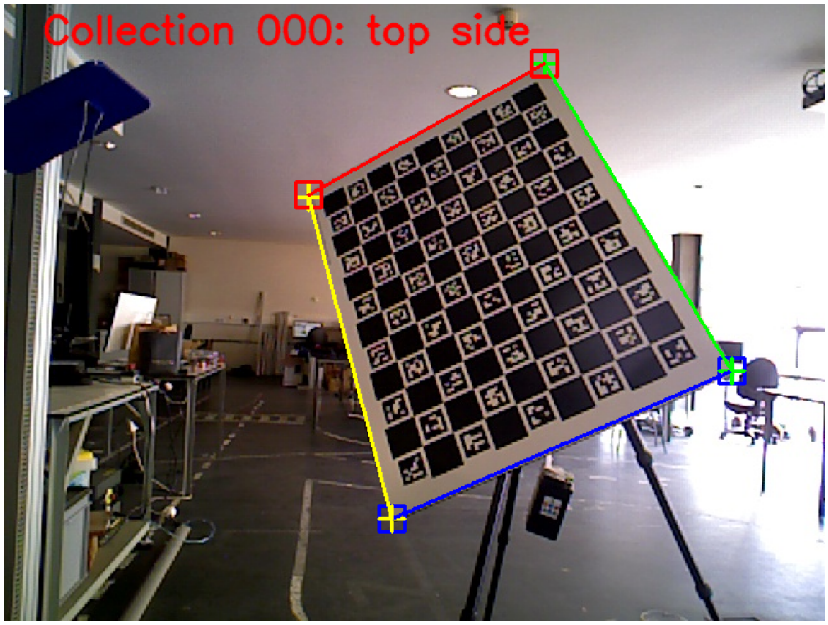

Figure 4.6: Manual annotation of the physical limits of the calibration pattern in RGB images.

Table 4.1: Descriptions of the datasets used in this experiment, where RGB partials mean the number of partial calibration pattern detections in the RGB sensors. All the collections in the datasets were complete, meaning that all the collections had detections for both sensors.

Type of data	Dataset	# collections	# RGB partials
simulation	train dataset	15	6
	test dataset	11	7
real asus	train dataset	16	9
	test dataset	9	7

we cannot use the transformations given by the simulation as an initial guess pose, because these are already in the correct position. To counteract this, we induce random noise to the initial guess pose for calibration in both translation and rotation. In this case, we induced an error of 0.1 m and 0.1 rad to the data to allow for a fair comparison. Table 4.2 shows the displacement between the ground truth links and the calibrated depth link. In this system, the RGB camera was fixed and only the pattern and the depth link moved around it. For that reason, this evaluation is only for the depth sensor. As we can see, the calibration error is low, below 1 cm, which demonstrates that the system was accurately calibrated.

Table 4.2: Ground truth evaluation of calibrated transformations for simulated results.

From	To	t (m)	R (rad)	X (m)	Y (m)	Z(m)	ϕ (rad)	θ (rad)	ψ (rad)
camera link	*depth link*	0.0090	0.0030	0.0007	0.0037	0.0082	0.0003	0.0026	0.0015

Table 4.3 displays the projection errors resulting from the calibration of simulated and real data. Our evaluation includes outcomes from the calibration using the 3D information generated by the depth sensor. Following the philosophy of the OpenNI Library, we also

calibrate the IR data as an RGB camera, which is viable because the depth sensors use IR projections to obtain the 3D data, and therefore IR and depth have the same coordinate frame. Results are also presented for calibration using IR and RGB data with the OpenCV [141] and Kalibr [9][36] libraries. It is worth noting that the IR projector needs to be covered by paper to diffuse IR illumination for the calibration pattern to be detected.

However, this calibration approach faces practical challenges. Due to insufficient bandwidth, RGB-D cameras cannot simultaneously stream RGB and IR data, leading us to fix the calibration pattern and the sensor and record separate data for the RGB and IR sensors for each pattern pose. Subsequently, we move the calibration pattern multiple times to capture the pattern in several poses. Merging these recordings and tampering with the timestamps to simulate simultaneous RGB and IR detections is necessary to run the calibration processes. This process is tiresome, unintuitive and becomes unfeasible in a more complex system, which ATOM also supports. It was carried out to enable the comparison with other approaches.

Considering the resolution of both simulated and real cameras as 640×480, percentage errors for both coordinates are calculated. For simulated data using RGB and depth pairs (3D), we observe errors of 0.14 % for the x direction and 0.23 % for the y direction. In real data, errors amount to 0.23 % for the x direction and 0.32 % for the y direction. Furthermore, we assessed errors in the factory calibration of the Asus Xtion using real data from the driver of the camera and obtained calibration errors of 8 pixels. Factory errors amount to 0.1 % in the x direction and 0.8 % in the y directions. This error is significantly high, especially when dealing with applications that require high precision such as data fusion. This demonstrates that our methodology is effective in improving factory calibration for RGB-D cameras.

Table 4.3: Pairwise errors for depth-to-RGB sensors evaluation in pixels.

Method	**Simulation**			**Real Data**		
	e_{rms}	x	y	e_{rms}	x	y
Factory	-	-	-	8.38	4.64	5.07
OpenCV w/ intrinsics [141]	(1)	(1)	(1)	0.30	0.15	0.20
OpenCV w/o intrinsics [141]	(1)	(1)	(1)	1.41	0.95	0.66
Kalibr w/ intrinsics [9][36]	(1)	(1)	(1)	0.66	0.31	0.39
ATOM IR w/ intrinsics	(1)	(1)	(1)	0.31	0.16	0.21
ATOM IR w/o intrinsics	(1)	(1)	(1)	1.40	0.94	0.66
ATOM 3D	1.79	0.92	1.11	2.90	1.45	1.52

(1) Simulation doesn't produce IR data.

In a comparative analysis with other algorithms, ATOM demonstrates performance on par with OpenCV's RGB-to-RGB extrinsic calibration. Optimisation of intrinsic parameters, which ATOM also supports, significantly enhances calibration performance for both algorithms. When calibrating the IR-RGB pair with Kalibr [9][36], subpixel accuracy is achieved, but with twice the error compared to the ATOM calibration. Kalibr's calibration also optimises intrinsic parameters.

When evaluating the pairwise errors in pixels, our methodology also outputs images with the depth boundaries projected to the corresponding RGB from the same collection, as seen

in Fig. 4.7. These images allow for a better visual understanding of the calibration accuracy.

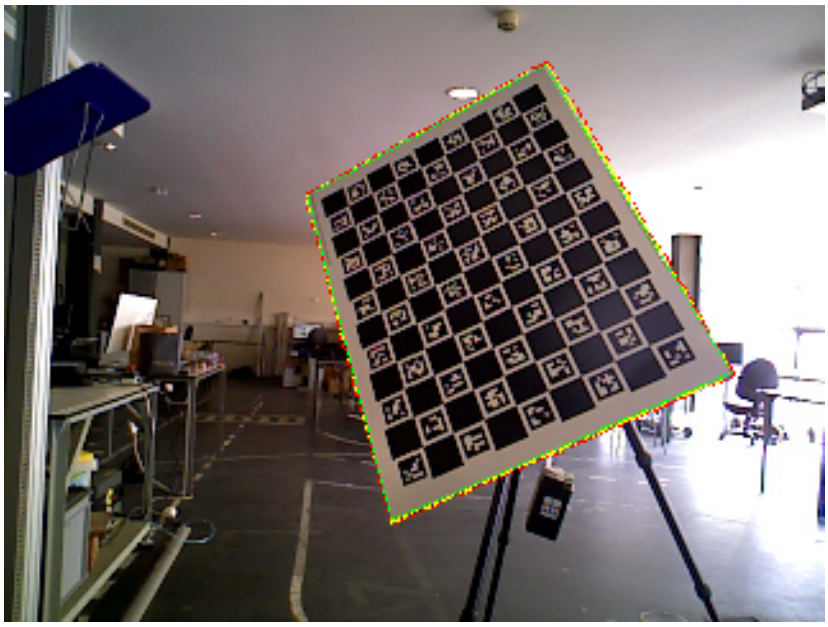

Figure 4.7: Depth-to-RGB evaluation image output, where green points are the physical limits of the pattern annotated in the RGB image, red points are the limits of the physical board detected in the depth images, and yellow lines represent the distances between corresponding points in depth and RGB.

To conclude, we would like to point out that, when using ATOM to calibrate an RGB-D camera, we can use both the depth or the IR data to calibrate the *depth_link*. As expected, the calibration using the IR data achieves a higher performance than using depth data, because the detection of chessboards and Charuco in images is very accurate and 3D labelling of depth data is less accurate due to a lower quality of the original data that is more affected by external factors like illumination. Nevertheless, both methods highly improve factory calibration.

4.3.2 Hand-Eye with Fixed Sensors

This subsection details the calibration results achieved when calibrating a complex system that incorporates both mobile and fixed sensors. Fig. 4.8 presents an illustration of this system. It is constituted of an RGB-D camera fixed to the end-effector of a robotic manipulator, along with three stationary sensors: a LiDAR, an RGB sensor, and a depth sensor. During the calibration procedure, we manipulate both the calibration pattern and the robotic manipulator to acquire an array of collections that encompass various perspectives of the pattern and diverse poses of the hand-eye camera.

Fig. 4.9 shows the transformation for the above-mentioned system. In green, we can see the nodes that represent the frame ID from which data is output and the arrows that are associated with that sensor's calibration. In blue are the dynamic transformations of the system: the whole kinematic chain of the robotic manipulator, at the left of the figure, and the pattern, which moves from collection to collection, to guarantee that the dataset is diverse with several poses and points of view.

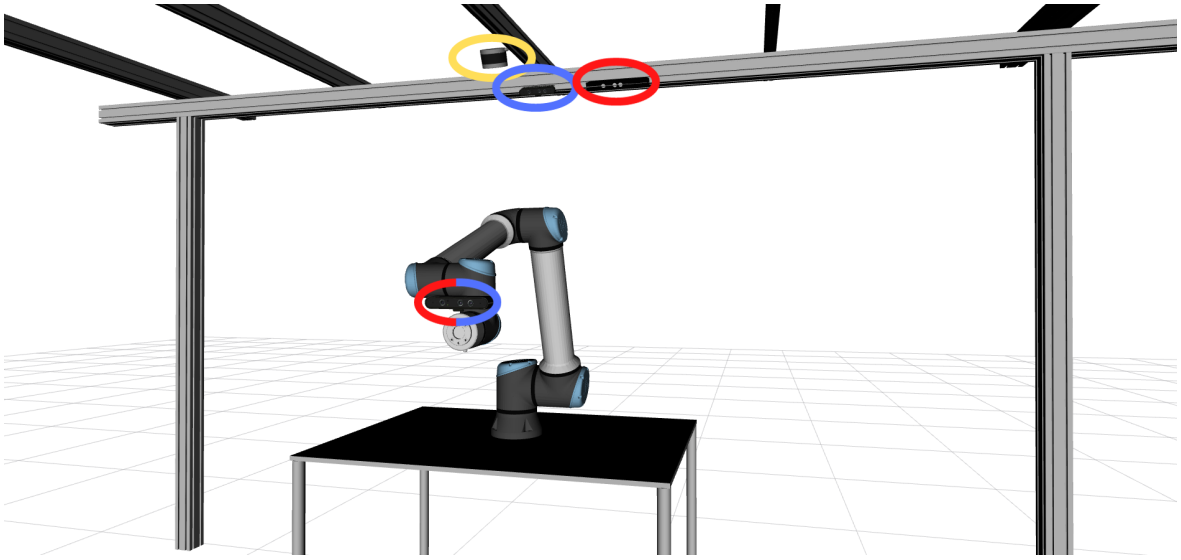

Figure 4.8: Illustration of the system to be calibrated, where red ellipses represent depth cameras, yellow ellipses represent LiDARs, and blue ellipses represent RGB cameras.

Table 4.4 summarises the datasets used for calibrating the robotic system, including results from both simulated and real-world data. This system's complexity leads to a significant number of incomplete collections, particularly in real-world scenarios. Several factors impact the data quality, such as lighting conditions, sensor interference (notably with depth cameras), and the distance between the cameras and the calibration pattern. These challenges contribute to the prevalence of partial detections in the RGB sensors. The distance between the cameras and the pattern further affects detection quality.

Table 4.4: Descriptions of the datasets used in this experiment. "RGB partials" indicates the number of partial detections of the calibration pattern by the RGB sensors, while "complete" refers to collections where the calibration pattern was detected by all seven sensors.

Data	Dataset	# collections	# RGB partials	# complete
simulation	train	22	4	22
	test	10	2	9
real	train	29	58	11
	test	13	26	5

The real-world setup was deployed in our laboratory, featuring artificial ceiling lights and minimal natural light interference, although some variability in lighting occurred due to natural daylight. The robotic manipulator was placed on a stable surface, and its end-effector-mounted RGB-D sensor was moved using automated trajectories. The calibration pattern was positioned at different locations within the field of view of the fixed sensors and robot arm. Pattern distances ranged from 1 m to 3 m. We acquired 29 training collections and 13 testing collections, each corresponding to a distinct scene or pattern pose. Variability in data quality occurred due to ambient light, depth sensor interference (especially between structured light systems), and sensor-to-pattern distance. This setup aimed to emulate realistic calibration

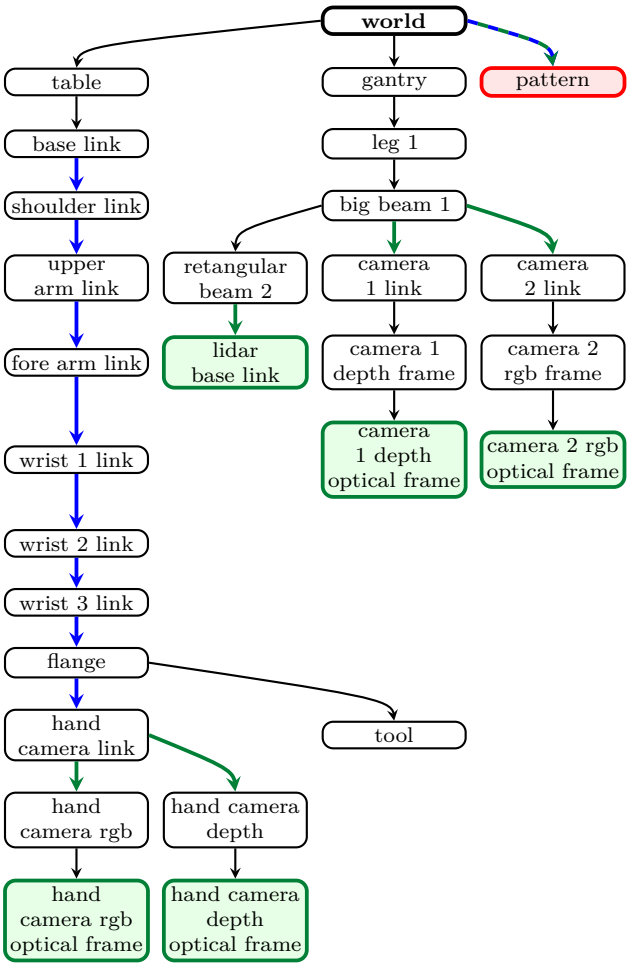

Figure 4.9: Transformation tree for the robotic system to be calibrated. Blue arrows represent dynamic transformations. Green arrows represent the transformation that will be calibrated during optimisation. The red node represents the calibration pattern and the green nodes represent the sensor links from which data is output.

challenges in heterogeneous multi-sensor systems. Table 4.4 provides a breakdown of complete and partial pattern detections.

Following the same line of thought as in the RGB-D calibration, first, we calibrated the simulated system to prove our concept with precise ground truth. To calibrate, we induced an error of 0.1 m and 0.1 rad to the initial guess transformations for the links in the system. Table 4.5 provides an assessment of the calibrated transformations by comparing the obtained transformations to the ground truth values for simulated results. It outlines the displacement between the ground truth and the obtained transformations for all the links to be calibrated, highlighted in green in Fig. 4.9, providing data on translation, rotation, and displacement along the X, Y, and Z axes, as well as Euler angles in the form of ϕ, θ, and ψ. These values serve as a reference for evaluating the accuracy of the calibration process, allowing for a comparison of the calibrated transformations with the ground truth. Overall, the low values for both translation and rotation discrepancies, as well as for the displacement and rotational angles, demonstrate the high quality of the calibration. The calibrated transformations closely

match the ground truth values, ensuring precise alignment of the sensors in the simulated environment. This level of accuracy is crucial for applications where sensor data needs to be combined or where precise sensor positioning is required, such as in robotics and data fusion.

Table 4.5: Ground truth evaluation of calibrated transformations for simulated results.

From	To	t (m)	R (rad)	X (m)	Y (m)	Z (m)	ϕ (rad)	θ (rad)	ψ (rad)
big beam 1	*camera 1 link*	0.0085	0.0061	0.0006	0.0047	0.0066	0.0001	0.0040	0.0046
hand camera link	*hand camera depth frame*	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
hand camera link	*hand camera rgb frame*	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
rectangular beam	*lidar 2*	0.0109	0.0090	0.0102	0.0037	0.0010	0.0048	0.0052	0.0056
big beam 1	*camera 2 link*	0.0101	0.0065	0.0021	0.0098	0.0013	0.0022	0.0032	0.0052

Table 4.6 displays pairwise Root Mean Square Error (RMSE) of various sensor pairs in pixels, offering valuable insights into the accuracy of extrinsic sensor calibration. The sensor pairs are a combination of RGB, depth and LiDAR sensors, both in a simulated and real data context. Notably, the RMSE values in the simulated data scenario range from 0.70 to 2.28 pixels, showcasing the precision of calibration between different sensors. In the real data setting, the LARCC values are slightly higher, varying from 3.45 to 6.53 pixels. These values reflect the discrepancies in sensor alignment in both simulated and real-world environments. The "average" row provides a useful summary of the overall calibration accuracy for the entire set of sensor pairs. This table serves as a critical reference for evaluating the quality of sensor calibration, which is crucial in various applications, such as robotics and computer vision.

Table 4.6: Pairwise root mean square errors of sensor pairs in pixels.

Sensor Pair	Target Sensor	Simulation	Real Data
camera 2 rgb	*hand camera rgb*	0.70	3.72
lidar	*camera 2 rgb*	1.78	5.57
lidar	*hand camera rgb*	2.09	5.35
lidar	*camera 1 depth*	1.87	4.08
lidar	*hand camera depth*	1.93	5.34
camera 1 depth	*camera 2 rgb*	2.28	4.22
camera 1 depth	*hand camera rgb*	2.21	3.45
hand camera depth	*camera 1 depth*	1.52	3.58
hand camera depth	*camera 2 rgb*	1.58	6.53
hand camera depth	*hand camera rgb*	1.69	4.40
average		1.77	4.67

The results show a clear variation in calibration performance across different sensor pair types, which can be attributed to their sensing modalities and the accuracy of pattern detection. In both simulated and real environments, RGB-to-RGB calibration pairs, such as *camera 2 rgb* to *hand camera rgb*, yield the lowest LARCC values (e.g., 0.70 pixels in simulation and 3.72 pixels in real data). This performance is primarily due to the high precision of 2D pattern detection in RGB images, where Charuco markers can be located with sub-pixel accuracy. In contrast, sensor pairs involving depth or LiDAR, such as *lidar* to *hand camera depth* or

camera 1 depth to *camera 2 rgb*, exhibit higher LARCC values. These sensors often suffer from lower spatial resolution and higher sensitivity to ambient conditions like lighting and surface reflectance, leading to noisier pattern boundaries.

Another factor contributing to performance variability is the difference in detection strategies across modalities. While RGB sensors rely on image-based marker detection, depth sensors and LiDAR require geometric interpretation of surface discontinuities or board contours, which are often less reliably segmented. Furthermore, the temporal and spatial alignment of captures in real-world setups may introduce additional minor inconsistencies, especially when bandwidth or hardware constraints force sequential rather than synchronous acquisition. Overall, the ATOM framework achieves strong performance in heterogeneous setups, particularly when using IR data for depth calibration or when combining sensors with more precise detection capabilities. Nevertheless, results highlight that achieving sub-pixel calibration with modalities like LiDAR-depth pairs remains a challenge due to inherent sensing limitations.

4.4 Final Considerations

In this chapter, we have presented a comprehensive methodology for the extrinsic calibration of a diverse set of sensors, including RGB, depth, and LiDAR sensors, both in simulation and real-world scenarios. The results indicate that the proposed approach can effectively and accurately calibrate these sensors, providing essential data for various applications in robotics and computer vision. We have demonstrated that our methodology substantially improves RGB-D factory calibration. The calibration results show low root mean square errors for both systems, confirming the reliability of the methodology. We have also demonstrated that ATOM is capable of calibrating a complex combination of dynamic and static sensors.

Overall, this study provides a robust foundation for the calibration of sensors in a range of scenarios and paves the way for advancements in the field of sensor fusion and perception in robotics and computer vision.

CHAPTER 5

Multi-View 2D to 3D Lifting Video-Based Optimisation: A Robust Approach for Human Pose Estimation with Occluded Joint Prediction

5.1 Introduction

3D human pose estimation is essential for human-robot collaboration, as it equips robots with the ability to comprehend and respond to human movements in a three-dimensional space. This technology facilitates seamless and intuitive interactions by enabling robots to interpret gestures, body language, and spatial relationships. Accurate pose estimation ensures precise coordination between humans and robots, enhancing safety and efficiency in shared workspaces [142], [143]. It allows robots to adapt their actions based on human poses, allowing smoother collaboration in diverse applications such as manufacturing, healthcare, and assistive robotics [142], [144]. Thus, it forms a fundamental bridge for effective communication and cooperation between humans and robots, promoting a productive collaborative environment.

However, despite the advances in 3D human pose estimation, current methods often struggle to handle occluded joints effectively. Occlusions occur when certain body parts are temporarily hidden from view, challenging the ability of the algorithm to predict the complete pose accurately. Occlusions can be caused by either an object in front of the human or by the human itself (self-occlusions) when part of the body occludes certain joints. In the context of human-robot collaboration, predicting occlusions becomes crucial. When robots cannot accurately perceive occluded joints, it may lead to misinterpretations of human actions, potentially resulting in errors or accidents. For instance, if a robot fails to recognise that an

arm of the person is temporarily hidden behind an object, it might misunderstand the intended action, impacting the collaborative task. Therefore, developing robust algorithms that can predict and account for occluded joints is paramount for enhancing the reliability and safety of human-robot collaboration scenarios. It ensures that the robot can adapt appropriately even when parts of the human body are temporarily occluded, contributing to a more effective and secure collaborative environment.

To address the limitations identified in Chapter2, particularly the challenges of occlusion, reliance on single-frame inference, and lack of individual-specific modelling, this manuscript introduces a novel approach to 3D human pose estimation, summarised in Fig.5.1. Building upon the insights from recent literature, we employ a 2D-to-3D lifting optimisation technique that leverages temporal information from multiple video frames. Unlike traditional single-frame methods discussed in the literature review, our approach integrates temporal consistency to enhance robustness against joint occlusions. By analysing consecutive frames, the algorithm improves the prediction of occluded joints and maintains spatial coherence. Furthermore, we incorporate subject-specific skeletal information to tailor the 3D reconstruction, addressing another gap highlighted in Chapter 2 regarding the generalisation limits of current models. This dual strategy, temporal integration and skeleton-specific modelling, enables a more accurate and resilient estimation of human poses, especially under occlusion and variation in morphology. We evaluate our framework on a representative 3D human pose estimation dataset, the MPI-INF-3DHP Dataset [10], and present comparative results with other state-of-the-art methods. A video[1] has been made available that includes a brief explanation of the methodology and some qualitative results. The contributions of this chapter can be summarised as follows:

- to propose a multi-camera video-based 3D human pose estimation algorithm;
- to predict accurately the position of occluded 3D joints;
- to compare with other 3D human pose estimation state-of-the-art approaches.

[1] `https://youtu.be/EiUbGgs2Wsk`

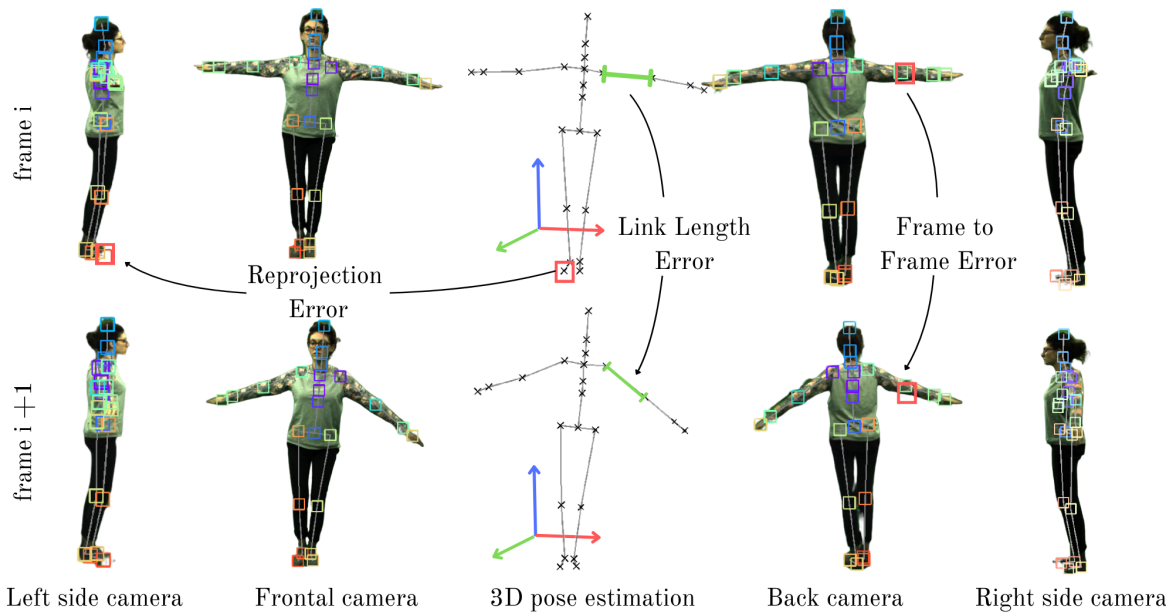

Figure 5.1: Schematic representation of the proposed approach. The main framework is divided into three key components: the reprojection component, the link length component, and the frame-to-frame component. The reprojection component aims to minimise the distance between the projection of the 3D joints and their 2D detections. The link length component aims to uniformise the tridimensional link length in all the frames. The frame-to-frame component helps predict the position of occluded joints using the position of the same joint in adjacent frames.

5.2 Methodology

The proposed video-based optimisation approach uses the least-squares method to determine the 3D position of each joint in a predefined skeleton. This approach uses 2D to 3D lifting, meaning that it assumes that the 2D poses are known in certain images (in pixels) and also requires the extrinsic parameters of the cameras in the system.

Besides the information from the 2D keypoints, our approach uses temporal information that helps detect occluded joints while trying to predict the movement of the occluded joint by extrapolating from frames where that joint was previously seen. It integrates knowledge from the anatomical configuration of the human skeleton by aiming to homogenise the estimated three-dimensional length of each skeletal link across all frames. More precisely, the objective is to ensure uniformity in the three-dimensional length of each link across the entire sequence of frames.

The optimisation problem is solved using a nonlinear least squares method. This algorithm aims to find the parameter vector θ that minimises the sum of squared residuals $Q(\theta) = \sum_{i=1}^{n} r_i^2$, where $r_i = y_i - f(x_i, \theta)$ represents the difference between the observed data y_i and the model prediction $f(x_i, \theta)$, where x_i represents the input features or predictors used in the model to make predictions. The Jacobian matrix $\mathbf{J}$ is a key component, containing partial derivatives of the residuals with respect to the parameters: $J_{ij} = \frac{\partial r_i}{\partial \theta_j}$. The nonlinear least squares

method iteratively updates the parameter estimates using the linearised system of equations represented by eq. 5.1

$$\theta_{k+1} = \theta_k - (\mathbf{J}^\intercal \mathbf{J})^{-1} \mathbf{J}^\intercal r \ , \tag{5.1}$$

where $\mathbf{J}^\intercal$ is the transpose of the Jacobian matrix, and r is the vector of residuals. This update rule adjusts the parameter estimates in the direction that reduces the sum of squared residuals, and the process is repeated iteratively until convergence is achieved. The final θ represents the optimal parameter that provides the best fit of the nonlinear model to the observed data. The optimisation is handled as a sparse problem because the parameters do not influence all of the residuals, and it is solved with the trust region reflective algorithm [145], which is a suitable method for large bounded sparse problems.

5.2.1 Objective Function

We aim to estimate the 3D coordinates, X, Y, Z, for each joint j and each frame f, by minimising an objective function composed of three main error components: the *reprojection error* (e_{rp}), the *link length error* (e_{ll}), and the *frame-to-frame temporal error* (e_{ff}). These terms ensure, respectively, consistency between projected and observed 2D points, anatomical plausibility of the estimated skeleton, and temporal smoothness across frames. The objective function is formally defined as:

$$f_{obj} = \underset{(X,Y,Z)_{j,f}}{\arg\min} \sum_j e_{ll} + \sum_j \sum_f \left(e_{rp} + e_{ff}\right). \tag{5.2}$$

The error components are detailed as follows.

Reprojection Error

In multi-camera systems for 3D human pose estimation, the reprojection function plays a fundamental role in assessing and optimising the consistency between reconstructed 3D joint positions and their corresponding 2D observations across multiple views. The basic principle is that a 3D point in the world, when projected onto a calibrated camera's image plane, should ideally coincide with the observed 2D detection of that point in the image. Any deviation between the projected 3D point and the actual 2D detection (obtained using either classical or learning-based computer vision procedures that use the image as input) can be interpreted as an error, often referred to as reprojection error.

This error provides a direct measure of how well the estimated 3D pose explains the set of 2D detections. The reprojection function itself is defined by the camera projection model, which encompasses both intrinsic parameters (e.g., focal length, principal point, distortion coefficients) and extrinsic parameters (e.g., rotation and translation with respect to a global reference frame). Together, these parameters define the mapping from a 3D point in world coordinates to a 2D point in pixel coordinates on the image plane.

The reprojection error is defined as:

$$e_{rp} = \left\| \text{proj}\Big((X, Y, Z)_{j,f}, \lambda_i\Big) - d_{j,f,i} \right\| \cdot c_{j,f,i} \ , \tag{5.3}$$

where $(X, Y, Z)_{j,f}$, for each joint j, and each frame f, projected to the frame in question for each camera image i, with the intrinsic and extrinsic parameters λ_i, and the coordinates 2D detection of that joint, $d_{j,f,i}$ in pixel. The confidence value for each joint in each frame and camera $c_{j,f,i}$ is also used as a multiplying factor for the reprojection residuals, highlighting joints that have high confidence detection values. The confidence value is provided by the 2D detector.

Link Length Error

An essential physical constraint in 3D human pose estimation is the assumption of constant bone lengths across time. Human limbs can bend at joints, but the skeletal link between any two adjacent joints (e.g., shoulder to elbow, hip to knee) has a fixed anatomical length for a given individual. To ensure anatomical plausibility and temporal consistency of the estimated poses, this constraint is enforced via a link length residual term in the objective function.

The link length residual measures the deviation of the estimated 3D link lengths from their average value over time. It penalises temporal fluctuations in bone length, which may arise due to noise in 2D detections, camera calibration inaccuracies, or imperfections in the reconstruction algorithm. By minimising this residual, the optimisation process is encouraged to produce temporally stable and physically consistent 3D skeletons.

Formally, the residual for a given skeletal link l is defined as the standard deviation of its estimated length over all frames:

$$e_{ll} = \sqrt{\frac{\sum_f (l_{j,f} - \bar{l})^2}{F}} \ , \tag{5.4}$$

where:

- $l_{j,f}$ is the Euclidean distance (i.e., the 3D length) between two joints defining link l in frame f,
- $\bar{l}$ is the average length of that link over all frames,
- F is the total number of frames.

This term acts as a regulariser, ensuring that for each skeletal link, its length remains approximately constant across time. While it does not enforce a specific anatomical value, it promotes internal consistency within the sequence, which is especially important in scenarios with occlusions or ambiguous joint detections. This improves the robustness of the 3D reconstruction by anchoring it to a biologically plausible structural prior.

Frame to Frame Error

Another component of the optimisation process in multi-frame 3D human pose estimation is the frame-to-frame residual, which enforces temporal smoothness by penalising abrupt changes in joint positions across consecutive frames. This term is particularly important for improving estimation quality in the presence of occlusions, missing detections, or highly noisy observations.

The frame-to-frame residual is defined as:

$$e_{ff} = \begin{cases} \|(X,Y,Z)_{j,f} - (X,Y,Z)_{j,f-1}\| & \text{if } j \text{ occluded} \\ 0 & \text{otherwise} \end{cases} \tag{5.5}$$

where

- $(X,Y,Z)_{j,f}$ denotes the 3D coordinates of joint j in frame f,
- $(X,Y,Z)_{j,f-1}$ denotes the 3D coordinates of the same joint in the $f-1$ frame.

The activation function is only activated when the joint is *occluded* or the inter-frame displacement exceeds a predefined threshold (in this case, 150 mm). The rationale behind this formulation is based on the assumption that, at sufficiently high frame rates, human joints do not exhibit large displacements between consecutive frames. As such, sudden or erratic movements in the reconstruction are likely indicative of detection failures or occlusions. By introducing a residual that penalises large inter-frame differences for such cases, the optimisation process is guided towards producing temporally coherent and kinematically plausible trajectories.

The threshold value of 150 mm aligns with the evaluation metric used in the 3DPCK (3D Percentage of Correct Keypoints) benchmark [10], where a keypoint is considered correctly localised if it lies within 150 mm of the ground truth. This threshold ensures that the frame-to-frame constraint does not interfere with normal joint movement, preserving natural motion while only activating when discontinuities (that could otherwise only be explained by extremely fast movement, which is not plausible) arise.

By incorporating e_{ff} into the overall cost function, the model benefits from a temporal regularisation that complements spatial constraints (e.g., reprojection error and link length consistency). This leads to smoother, more realistic motion sequences, and provides a mechanism for estimating occluded joints based on previous valid positions, thus improving the robustness and reliability of the final 3D pose reconstruction.

5.3 Inference Speed Improvement

To enable the deployment of the proposed multi-view 2D-to-3D lifting method in time-sensitive robotic applications, a real-time adaptation of the original batch-based optimisation algorithm was developed. This was motivated by the need for continuous and low-latency pose estimation in collaborative scenarios, where robots must interpret and respond to human motion instantaneously to ensure both functionality and safety.

The original method, which performs a global optimisation over an entire sequence of video frames, was restructured into a sliding-window framework. In this variant, the optimisation is performed on a fixed-length buffer of the most recent frames (typically five frames, as a compromise in the speed–accuracy trade-off) updated incrementally as new frames are acquired. Rather than recomputing the pose from scratch at each timestep, the previously optimised 3D pose is used as the initial guess for the current frame. This recursive approach significantly reduces convergence time while ensuring temporal coherence.

All three components of the original objective function were retained: the reprojection residual, which ensures alignment between the estimated 3D joints and the detected 2D

keypoints; the link length residual, which enforces physical consistency across frames; and the frame-to-frame residual, which acts as a temporal prior to improve robustness in cases of occlusion or uncertain detections. In the real-time implementation, the frame-to-frame residual takes on added importance, as it allows the algorithm to infer joint trajectories even when direct visual data is temporarily unavailable.

One of the primary challenges during this adaptation was achieving sufficient time efficiency. In early tests, optimising five frames without the link length and temporal smoothness terms required approximately 2.5 seconds per frame. Through progressive refinements, including algorithmic simplification and more effective initialisation, the full version, with all residuals active, now achieves processing times ranging from 150 to 750 milliseconds per frame, depending on the frame's complexity and visibility conditions. While this range does not yet support real-time performance at video frame rates (e.g., 30 Frames per Second (fps)), it represents a significant improvement and brings the method within reach of interactive, frame-by-frame applications in robotics.

The adapted system was integrated into a ROS-based architecture, enabling direct communication between the pose estimation module and other components of a robotic system. Each camera provides synchronised 2D keypoints, which are fed into the optimiser, and the resulting 3D skeleton is immediately available for downstream tasks such as robot trajectory planning, proximity detection, or adaptive behaviour modeling.

In summary, the real-time variant of the algorithm preserves the strengths of the original approach while addressing its practical limitations. By shifting to a recursive, windowed optimisation strategy and incorporating all three residual components, the system maintains robustness to occlusions and noise while moving towards operational viability in real-world robotic environments

5.4 ROS Integration

To enable real-time operation of the proposed 3D HPE methodology within robotic environments, the full pipeline was adapted to the ROS framework. This integration allows for modular, synchronised processing of image data from multiple cameras, the dissemination of keypoint and pose information, and visual feedback through standard ROS visualisation tools.

The integration begins with the creation of a dedicated ROS node that subscribes to camera image topics, specified dynamically via arguments, and processes these streams using the MediaPipe [146] library to detect 2D keypoints. The node continuously publishes the results in a customised message format for each camera, making the data accessible to the rest of the system.

To represent the detected keypoints in 2D, a custom message `keypoint2D.msg` was defined as follows:

Each keypoint contains 2D image coordinates and a confidence score. These are grouped into a `person2D.msg` message, defined as:

```
float32 x
float32 y
float32 score
```

Code 1: Structure of keypoint2D.msg.

```
Header header
keypoint2D[] keypoints
```

Code 2: Structure of person2D.msg.

This message is published on a dedicated `/camera_id/skeleton` topic for each camera. A second node subscribes to these topics and aggregates the information in a global dictionary indexed by camera ID. This architecture enables flexible scaling to arbitrary numbers of views.

The robot model is instantiated via a *xacro* file, ensuring proper structural configuration within the ROS ecosystem. Intrinsic parameters for each camera (including focal lengths, distortion coefficients, and image dimensions) are read from `CameraInfo` topics and stored in a dictionary for subsequent use. Extrinsic parameters, which were previously obtained with the ATOM calibration framework, are retrieved from the `tf` transformation tree, enabling accurate spatial localisation of each camera within the system.

For 3D reconstruction, a separate set of custom messages is used to represent poses in space. The fundamental unit is the `keypoint3D.msg`:

```
float32 x
float32 y
float32 z
float32 score
```

Code 3: Structure of keypoint3D.msg.

These are grouped into a `person3D.msg`, published as follows:

```
Header header
keypoint3D[] keypoints
```

Code 4: Structure of person3D.msg.

The estimated 3D skeletons are published in real-time, and visualised through `visualization_msgs/MarkerArray` objects in *RViz*. Each marker encodes the spatial structure of the human body using a fixed skeleton topology.

One of the principal challenges lies in ensuring temporal synchronisation across asynchronous camera streams. As individuals move through the scene, 2D detections may intermittently fail, either due to occlusion or subjects exiting a camera's field of view. To address this, a mechanism was implemented to assign frame indices to incoming detections and verify whether a sufficient number of valid 2D keypoints (from at least two cameras) are available for triangulation and optimisation within that timestamp. This is a non-trivial aspect of the system, as effective frame index synchronisation remains a limiting factor under partial visibility and frame drops.

Once the necessary conditions are met, a series of verification steps are performed before initiating the optimisation routine. Residuals are constructed to model reprojection errors, and a non-linear optimisation is executed to estimate 3D poses. When successful, the results are published using the aforementioned `person3D` messages and visualised in *RViz* using the skeletal marker structure.

Crucially, this ROS-based integration supports seamless deployment of the HPE framework within the LARCC collaborative cell described in previous chapters. It enables real-time monitoring and pose estimation within human-robot interaction scenarios, aligning with the operational requirements of collaborative workspaces. However, due to the absence of a ground truth reference in the deployed environment, the system currently cannot be used to conduct formal accuracy evaluations. While qualitative assessments of pose plausibility can be made, quantitative benchmarking remains limited, representing a key area for future development, potentially through the inclusion of motion capture or depth-based ground truth systems.

In summary, the ROS integration facilitates real-time, modular, and extensible deployment of the HPE methodology within collaborative robotics contexts, while also highlighting important limitations in synchronisation and accuracy assessment that must be addressed in future work.

5.5 Tests and Results

This section presents the results and experiments developed to prove the accuracy of our method. We present comparative results and three separate experiments where we evaluate the impact of the initial 2D detection noise, the number of occluded joints, and the number of cameras.

5.5.1 Dataset and Metrics

For all the tests and evaluations presented in this manuscript, we use the MPI-INF-3DHP dataset [10], a widely adopted benchmark for 3D human pose estimation. This dataset features both indoor and outdoor recordings of diverse actions performed by multiple subjects, captured with markerless motion capture systems. It includes data from 8 synchronised cameras positioned at different viewpoints, along with corresponding 2D and 3D ground-truth annotations, enabling rigorous evaluation under varying perspectives and conditions. The 2D keypoints used in our experiments are projected from four camera views (cameras 0, 4, 5, and 8) of the MPI-INF-3DHP dataset, chosen to represent diverse viewpoints, as illustrated in Fig. 5.3. As the skeleton model, we adopt the MPI-INF-3DHP skeleton structure, consisting of 23 joints, illustrated in Fig. 5.2. This configuration ensures consistency with the dataset's annotation scheme and facilitates reliable benchmarking of our method.

To assess the performance of the algorithm, we use established evaluation metrics commonly utilised in recent 3D human pose estimation articles: MPJPE (Mean Per Joint Position Error)

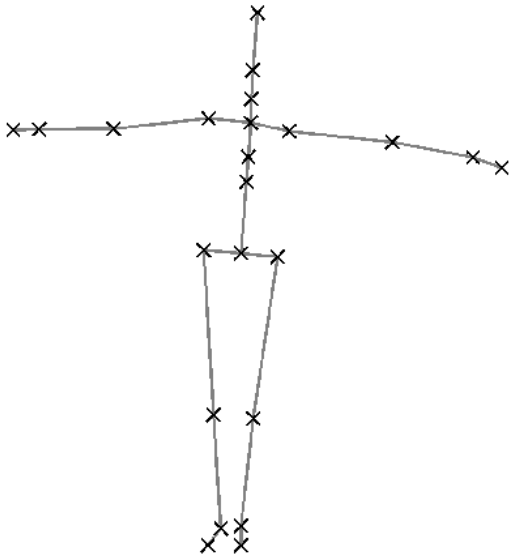

Figure 5.2: Representation of the skeleton used in the experiments, where X represent joints.

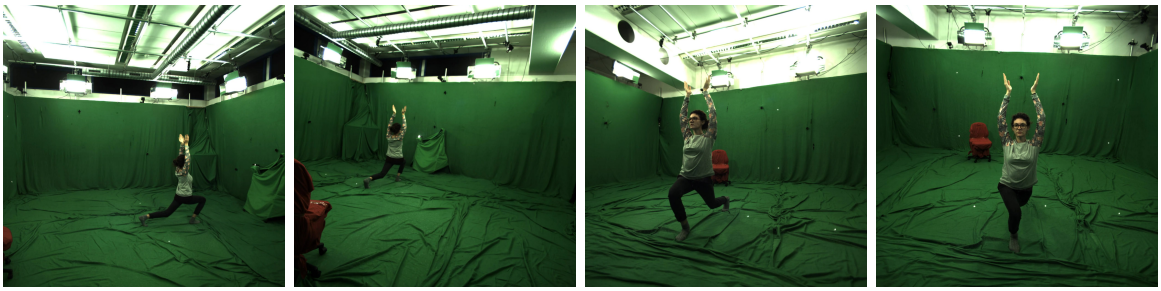

Figure 5.3: Example of image set used in calibration from cameras 0, 4, 5 and 8 of the MPI-INF-3DHP skeleton [10].

and 3DPCK (3D Percentage of Correct Keypoints). The MPJPE is represented by eq. (5.6),

$$\text{MPJPE} = \frac{\sum_{f=0}^{F} \sum_{j=1}^{J} \|P_{j,f} - P_{gt}\|}{F \cdot J} \ , \tag{5.6}$$

where $P = (X, Y, Z)$. The MPJPE quantifies the average error per joint position, determined by the Euclidean distance between the ground truth (X_{gt}, Y_{gt}, Z_{gt}) and the estimated joint positions $(X_{j,f}, Y_{j,f}, Z_{j,f})$ for each joint j and frame f, divided by the total number of frames F and the total number of joints, J.

The 3DPCK is represented by eq. (5.7),

$$\text{3DPCK} = \frac{J_{correct}}{J} \times 100 \ , \tag{5.7}$$

where $J_{correct}$ is the number of correct joints and J is the total number of joints. The 3DPCK measures the percentage of correctly identified 3D keypoints. A detection is a true positive if the Euclidean distance between the estimated joint position and its corresponding ground truth falls within a specified threshold. In alignment with standard practices from other state-of-the-art approaches, we applied a threshold value of 150 mm like suggested in [10].

Table 5.1: Comparative analysis with other 2D to 3D lifting state-of-the-art methodologies on the MPI-INF-3DHP [10] dataset.

Methodology	**Optimization**	**Multi-view**	**Video**	**MPJPE ↓**
Kocabas et al. [97]		✓		109.0
Bouazizi et al. [108]		✓	✓	93.0
Pavvlo et al. [104]			✓	86.6
Bouazizi et al. [98]		✓		65.9
Jiang et al. [100]	✓			55.2
Zhao et al. [94]			✓	27.8
Yu et al. [95]			✓	27.8
Ours (20px)	✓	✓	✓	36.4
Ours (10px)				18.1

These metrics collectively provide a comprehensive evaluation of the accuracy of the algorithm in estimating 3D human poses.

5.5.2 Comparative Analysis

Table 5.1 presents a comparative analysis of the proposed methodology with other state-of-the-art algorithms on the MPI-INF-3DHP dataset [10]. The evaluated algorithms include Bouazizi et al. [108], Bouazizi et al. [98], Kocabas et al. [97], Jiang et al. [100], Yu et al. [95], and Pavvlo et al. [104] as benchmark references.

We restrict our comparison to state-of-the-art 2D-to-3D lifting approaches, as these most closely align with the core assumptions and methodological scope of our proposal, namely that 2D keypoints are available as input and the goal is to reconstruct their 3D counterparts. To assess the robustness of our approach under realistic detection uncertainty, we simulate two levels of error, 10 and 20 pixels, in the 2D input keypoints by adding Gaussian noise to the ground-truth 2D joint locations. This choice is motivated by the observation that although modern 2D keypoint detectors typically achieve sub-10-pixel precision, performance can degrade in more complex settings, particularly under occlusion or motion blur. Including the 20-pixel condition allows us to stress-test the model and highlight its behaviour under more challenging, yet still plausible, circumstances. These controlled perturbations offer a consistent and reproducible means to evaluate how different lifting methods respond to degraded input accuracy, thereby facilitating a fair and focused comparison.

Our algorithm outperforms all other methodologies in MPJPE values. The MPJPE stands at 18.06 for a 10-pixel error scenario and increases to 36.40 for a more challenging 20-pixel error scenario. These results prove the robustness and accuracy of the proposed approach, demonstrating its efficacy in achieving accurate 3D human pose estimation under conditions with varying degrees of 2D pixel errors.

Our approach exhibits state-of-the-art performance for several key reasons, setting it apart from existing methodologies. One factor is the use of a video-based approach. Our algorithm improves the simple reprojection function by taking into account how human poses change between frames by using the frame-to-frame approach. Utilising information from multiple

frames enables our algorithm to gather context and refine estimations by considering the consistency of pose configurations across frames. This not only improves the robustness but also enhances the ability to handle dynamic and occluded complex movements.

Furthermore, our approach relies on the unique characteristics of each human skeleton. The link length component of the objective function (see section 5.2.1) will estimate different link lengths for each human. By tailoring the optimisation process to the individual characteristics of skeletons, our algorithm achieves a higher degree of precision. This customised optimisation contributes significantly to mitigating errors and enhances the overall accuracy of 3D pose estimations by guaranteeing that the link length does not change in different frames.

The proposed approach, using 2D detection with a 10-pixel error, gives the best results overall. It is followed by Yu et al. [95] and Zhao et al. [94]. These good results may be related to the fact that all the mentioned approaches are video-based, which allows to improve 3D poses by leveraging information from all the frames.

5.5.3 Impact of 2D Joint Detection Error

The following experiment evaluates the impact of the 2D joint detection pixel error on the 3D joint results. The test set consisted of 500 frames (roughly 20 s) from the MPI-INF-3DHP [10]. To evaluate the impact of the quality of 2D keypoint detection on the outcome, we used the 2D ground-truth keypoint values as input. We added a systematic absolute pixel error to all the keypoints in a random direction. The added error varied from 0 to 100 pixels. The test dataset did not include any occluded joints, and the optimisation used 4 of the 8 available cameras.

Fig. 5.4 shows a plot of the evolution of the indicators MJPJE and 3DPCK indicators with the increase of the 2D joint detection error. It presents an analytical view of the correlation between 2D joint detection errors and the subsequent impact on the detection of human 3D poses. The MJPJE, illustrated by the ascending orange curve, shows a gradual increase in millimetres as 2D joint errors in pixels rise. This positive correlation underscores the sensitivity of 3D pose predictions to inaccuracies in 2D joint localisation. The trend suggests that as the precision of 2D joint detection diminishes, the accuracy of predicting the spatial positions of joints in the 3D space becomes compromised. The 3DPCK, represented by the descending blue curve, reflects the percentage of accurately estimated 3D keypoints in relation to increasing 2D joint errors. The decline in 3DPCK underscores a more pronounced sensitivity to higher 2D detection errors.

Nevertheless, the algorithm demonstrates robust performance up to a 2D joint detection error of 25 pixels. Within this range, both the MJPJE and 3DPCK show favourable behaviours. The MJPJE remains relatively low, up to 50 mm, indicating an accurate prediction of 3D joint positions, while the 3DPCK remains consistently high, demonstrating a high percentage of correctly estimated keypoints.

Up to the 25-pixel threshold, the algorithm effectively compensates for minor inaccuracies in 2D joint detection, showcasing resilience to moderate 2D detection errors. Beyond 25 pixels, however, the performance trends diverge, with both MJPJE and 3DPCK responding more sensitively to increasing 2D joint errors.

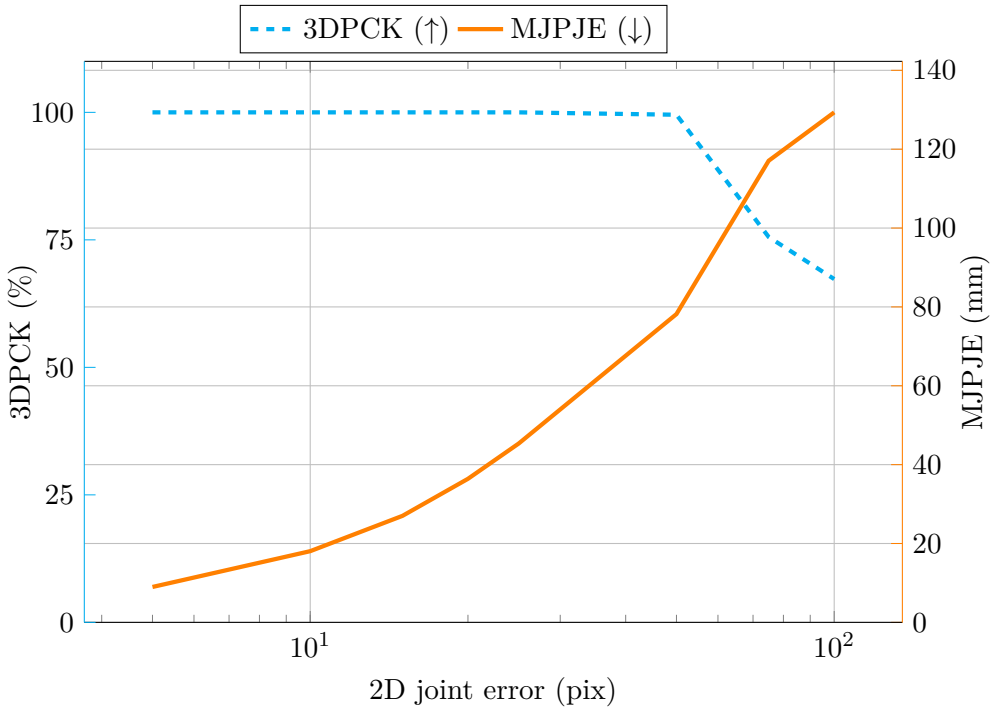

Figure 5.4: Impact of 2D joint detection error in detection of human 3D poses. A detailed explanation of the indicators can be found in section 5.5.1.

5.5.4 Impact of Occlusions

This subsection aims to determine the robustness of the algorithm to occlusions. For this, we designed two different experiments: one where we randomly occluded an increasing number of 2D joints that served as input for optimisation; in the second experiment, we occluded the same joint in all cameras for a period of time and evaluated how precisely the position of that joint was being predicted.

Random Occlusions

This experiment aims to assess the influence of 2D joint occlusions on the prediction of 3D joint values. The test set contains 500 frames from the MPI-INF-3DHP dataset [10]. To simulate occlusions, we systematically remove keypoints from the ground-truth 2D keypoints. The quantity of deleted keypoints per frame and point-of-view ranged from 0 to 15. Additionally, an absolute error of 10 pixels was added to each ground-truth keypoint value. This controlled variation in occlusion and error scenarios allows for a comprehensive evaluation of the robustness of the algorithm under realistic conditions.

Fig. 5.5 shows the results obtained from the experience mentioned earlier. Comparative analyses evaluate the efficacy of a simple reprojection function, optimised through the least squares method (plots with the tag $_{reproj}$), against our proposal.

In the MJPJE plot, our algorithm (denoted as "MJPJE") consistently outperforms the reprojection function ("MJPJE$_{reproj}$"), particularly with higher occurrences of occluded joints. The stability in MJPJE values for our methodology signifies very good precision in joint

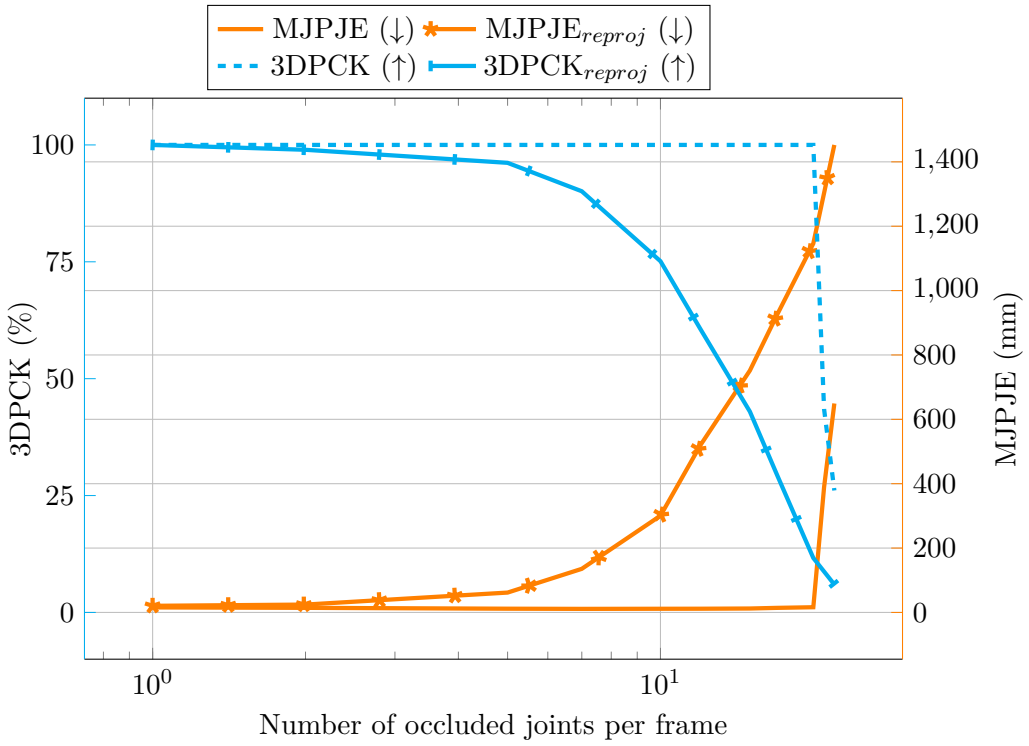

Figure 5.5: Impact of occluded 2D joints in the detection of human 3D poses, where simple lines represent the performance of our proposal and marked lines represent the performance of an optimisation using only the reprojection of 3D coordinates to 2D images as the objective function.

position estimation, even in highly occluded scenarios, with the performance of the algorithm maintaining resilience.

In the 3DPCK plot, our algorithm (denoted as "3DPCK") shows higher correctness percentages, even with a substantial number of occluded joints. In contrast, the reprojection approach ("3DPCK$_{reproj}$") evidences a decline in correctness with escalating occluded joints, highlighting the efficacy of our algorithm in sustaining keypoint accuracy under challenging occlusive scenarios and significantly improving the performance of the reprojection function.

In conclusion, the figure successfully demonstrates that our proposal greatly improves the reprojection function, particularly in scenarios where joints are occluded. The observed stability in performance highlights the robustness and potential of our solution, enhancing its utility for precise 3D human pose estimation within intricate real-world scenarios.

Consistent Occlusions

This experiment evaluates the impact of occluded joints in all points of view for a period of time in 3D joint poses. For this, the dataset used, for each 10 normal frames, had 5 frames where the left elbow was occluded in all cameras. The dataset also had 10 pixels of 2D joint error in every joint. The obtained MJPJE was 18.08 mm and 100 % 3DPCK. Regarding the left elbow, the 3D joint error was 20.30 mm, which is slightly above average but demonstrates that the position of that joint was well predicted by the proposed approach.

5.5.5 Impact of Number of Cameras

This experiment intends to determine the impact of the number of cameras used to optimise the quality of the 3D human poses. The test set contains 500 frames from MPI-INF-3DHP [10]. We chose a test set with frames with 5 randomly occluded joints and 10 pixels of error.

Table 5.2 shows the results obtained when calibrating the same dataset with a varying number of cameras. We can conclude that even with the constraint of optimising only 2 cameras, the performance remains robust. The MPJPE is 47.4 mm, indicating good precision in estimating joint positions. The 3DPCK is registered at 96.2 %, a very good accuracy considering the limited number of cameras. As the number of cameras increases to 3 and 4, the precision further improves, as is evident in the reduced MPJPE values (13.8 and 11.6, respectively) and the 100 % 3DPCK accuracy. This analysis emphasises the resilience of the algorithm, demonstrating good performance even in scenarios where only two cameras are utilised.

Table 5.2: Impact of the number of cameras in the MPJPE (mm) and 3DPCK (%) in the MPI-INF-3DHP dataset.

# cameras	MPJPE ↓	3DPCK ↑
2	47.4	96.2
3	13.8	100
4	11.6	100

From a practical standpoint, these results have significant implications for real-world applications. The ability to achieve accurate 3D human pose estimation with only two cameras makes the system more feasible and cost-effective for deployment in real-world scenarios, where the number of available cameras might be limited. The method's robustness in low-camera scenarios enhances its versatility and potential for broader adoption across different fields and use cases. Additionally, the point of view of the cameras also influences the quality of detection, as optimal camera placement can further enhance the accuracy and reliability of the system.

5.5.6 Experimental Results with Synthetic ROS Integration

To evaluate the performance of the proposed 3D human pose estimation pipeline within the ROS framework, a synthetic testing procedure was developed based on the MPI-INF-3DHP dataset [10]. Given that the ROS-based system is primarily designed for real-time applications and lacks built-in mechanisms for offline benchmarking, a controlled dataset replay strategy was implemented to emulate real-time operation while enabling consistent input and replication.

A dedicated Python script was developed to convert a subset of the MPI-INF-3DHP dataset into a `rosbag` file. This conversion involved assigning artificial yet synchronised timestamps to ensure cross-camera temporal alignment. The RGB images from four camera perspectives (`camera_0`, `camera_4`, `camera_5`, and `camera_8`) were saved as `sensor_msgs/Image` messages, along with corresponding intrinsic calibration data via `CameraInfo` topics. In parallel, the

associated 2D poses, extracted using the MediaPipe [146] library, were written into a custom `person2D` message and published on individual skeleton topics for each camera.

This setup allowed the entire ROS integration to function as if it were receiving real-time data, including skeleton detection, inter-camera synchronisation, 3D triangulation, optimisation, and RViz-based visualisation. The approach effectively bridges dataset-based testing with real-time ROS architecture, enabling reproducible and systematic testing of the pipeline components.

Figures 5.6 and 5.7 illustrate the reconstructed 3D skeletons corresponding to two sample poses. The top panels show the synchronised views from the four cameras, while the bottom panels depict the 3D skeleton rendered in RViz. In both examples, the reconstructed skeleton appears spatially coherent and dynamically plausible, demonstrating the integration's potential for use in collaborative robotics contexts.

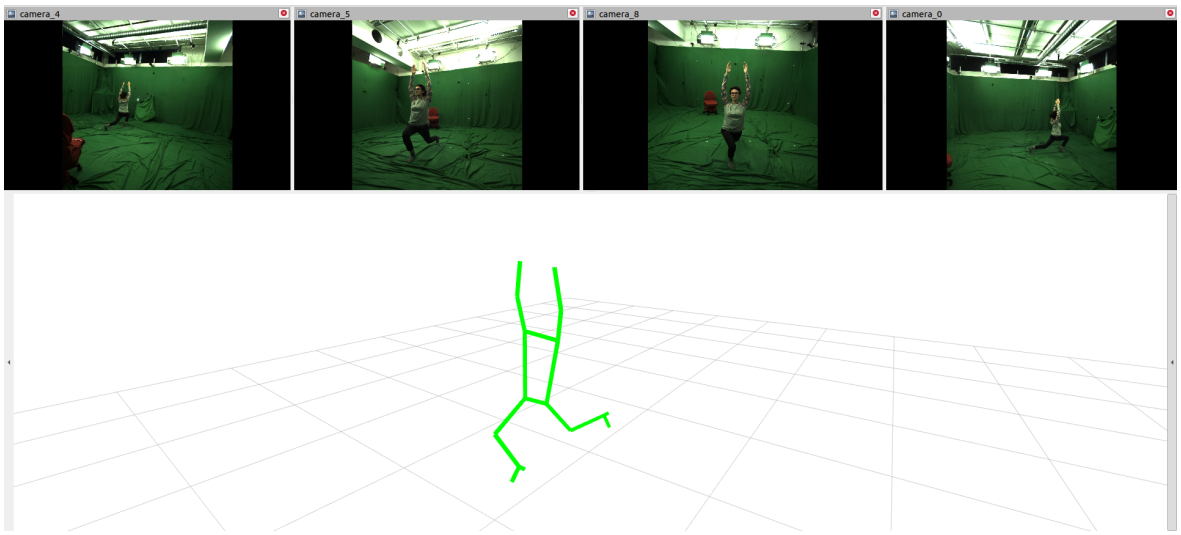

Figure 5.6: RViz visualisation of the reconstruction of a dynamic warrior-like pose using synchronised images from four cameras. The bottom panel shows the estimated 3D skeleton in RViz, generated from MediaPipe 2D keypoints processed through the ROS pipeline.

While this setup does not incorporate the ground-truth annotations available in the original MPI dataset, it provides a valuable and realistic context for evaluating the coherence and stability of the reconstructed 3D poses. The assessment conducted is therefore primarily qualitative, focusing on the visual plausibility and consistency of the skeleton across views and time. Future developments may explore complementing this analysis with offline comparisons or integrated overlays, further enhancing the validation process.

This synthetic evaluation nevertheless provides a valuable intermediary step between offline model testing and real-time deployment in collaborative robotics scenarios, such as the LARCC cell described in earlier chapters.

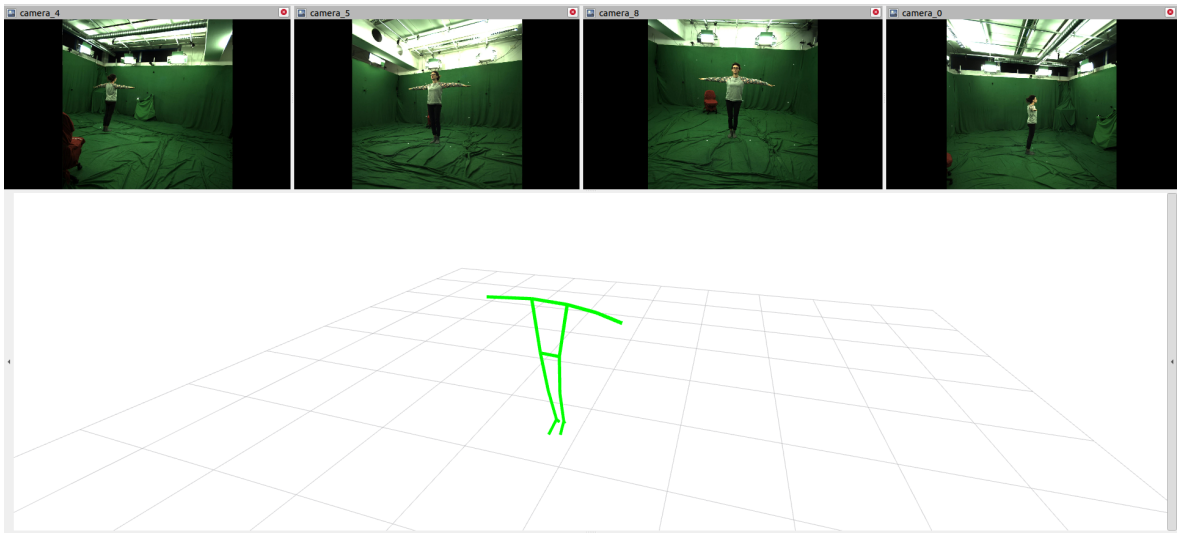

Figure 5.7: RViz visualisation of the reconstruction of a T-pose from the MPI dataset using the synthetic ROS rosbag. The estimated 3D skeleton (bottom panel) confirms consistent pose inference from multi-view inputs.

5.6 Final Considerations

This chapter presented a comprehensive investigation into a 3D human pose estimation framework designed for multi-camera environments, with particular emphasis on its deployment in collaborative robotics contexts. Building on accurate 2D keypoint detection and multi-view triangulation, the proposed methodology was developed to be modular, robust to noise, and suitable for real-time integration within robotic systems.

Extensive quantitative evaluation was carried out to assess the performance of the system under varied conditions, including different levels of 2D joint detection error, partial occlusions, and reductions in the number of available camera views. The results demonstrated that the algorithm maintains high reconstruction accuracy even when confronted with significant noise or occluded joints, highlighting its robustness and reliability. The method consistently delivered stable and plausible 3D skeletons across a wide range of human poses, confirming its applicability to real-world scenarios.

A key contribution of this work lies in the integration of the HPE pipeline within the ROS ecosystem. The system supports synchronised multi-camera processing, real-time publishing of 2D and 3D keypoints via custom ROS messages, and visualisation in RViz. This architecture enables deployment in live collaborative environments and facilitates interaction with robotic agents. The use of synthetic ROS bag files, constructed from a publicly available dataset with artificially synchronised timestamps, enabled controlled testing of the complete pipeline in realistic conditions.

Taken together, the proposed approach offers a reliable and extensible framework for human pose estimation in robotics. It is particularly well-suited to applications requiring accurate spatial perception and human-aware decision-making, such as those found in collaborative work cells. The experimental evidence provided in this chapter underlines the method's effectiveness

and positions it as a solid foundation for future research on real-time, multi-person tracking, dynamic scene understanding, and human-robot cooperation.

CHAPTER 6

Conclusions and Final Remarks

6.1 Overview

This thesis addressed two fundamental challenges in collaborative robotics: (i) the accurate extrinsic calibration of heterogeneous sensor systems, and (ii) the robust estimation of 3D human pose using multi-camera RGB configurations. Motivated by the growing demand for safer, more perceptually aware robotic systems in industrial and semi-structured environments, the research combined geometric modelling, system integration, and experimental validation to develop and evaluate novel contributions in both areas. Through the work presented in Chapters 3 to 5, the thesis advances the state of the art in perception-driven collaboration between humans and robots.

The thesis also revisited and addressed the research questions introduced in Chapter 1, demonstrating through its findings that both core objectives were effectively accomplished. The proposed calibration framework confirmed the hypothesis that accurate, scalable extrinsic calibration is feasible even under dynamic, multi-modal conditions. Likewise, the human pose estimation pipeline validated the hypothesis that view redundancy and system calibration significantly enhance 3D localisation in collaborative settings. These contributions support the central thesis that perception systems must be not only technically precise but also operationally adaptable in order to meet the demands of modern human–robot interaction.

6.2 Summary of Contributions

Following the problem definition and motivation established in Chapter 1, this thesis has addressed two fundamental perception challenges in collaborative robotics: the extrinsic calibration of heterogeneous sensor setups and the estimation of 3D human pose using multi-camera RGB systems. These capabilities were identified in Chapter 2 as critical enablers of spatial awareness, safety, and interaction fluency in human–robot collaboration. The literature review highlighted persistent gaps in current methodologies, particularly regarding the lack of scalable calibration procedures for static and mobile sensors across RGB, RGB-D, and LiDAR

modalities [45], [66], and the difficulty of achieving reliable pose estimation under occlusion, clutter, or partial calibration [11], [98], [117].

In response to these limitations, Chapter 3 presented an original calibration method extending the ATOM framework [48] to support depth cameras and to integrate hand–eye calibration into a unified, optimisation-based pipeline. The proposed system is designed for ROS-based robotic platforms and supports flexible configurations involving fixed, mobile, or robot-mounted sensors. It enables the estimation of extrinsic transformations in heterogeneous and partially overlapping fields of view, addressing key scalability and modularity concerns identified in the literature review [46], [47], [52].

Chapter 4 established a rigorous validation protocol, combining synthetic datasets with controlled real-world robotic deployments. The evaluation focused on translational and rotational accuracy, robustness to motion, and repeatability across multiple sensor types. Benchmarking against widely adopted toolkits, such as Kalibr [9] and the ROS camera calibration suite, demonstrated that the proposed approach achieves sub-centimetre precision and superior stability under motion, a scenario where many traditional methods degrade [38], [66]. Notably, the method proved effective in calibrating robot-mounted LiDAR and RGB-D units without requiring strict synchronisation or shared calibration targets, overcoming common deployment constraints in industrial settings.

Building on this calibrated perception backbone, Chapter 5 investigated 3D human pose estimation in multi-camera RGB environments, with particular emphasis on its application to collaborative workspaces. A comparative analysis of pose reconstruction pipelines was performed, assessing triangulation strategies, robustness to occlusion, and runtime performance. The findings confirm that calibrated multi-view systems significantly outperform monocular approaches in terms of localisation stability and joint accuracy, especially in cluttered or partially occluded environments, as discussed in the literature [96], [114], [126]. Furthermore, the chapter highlighted the trade-offs between frame rate, processing load, and spatial consistency, providing design guidelines for configuring perception systems in real-world collaborative robotic cells.

Together, the contributions of this thesis demonstrate that scalable, multi-modal calibration and robust 3D pose estimation are not only feasible, but also practically integrable in dynamic robotic systems. The proposed methods bridge theoretical gaps identified in the state of the art while delivering operational advances in accuracy, flexibility, and deployability. By addressing core challenges in perception, this work contributes to the broader goal of enabling safe, adaptive, and human-aware collaboration between robots and their human co-workers.

6.3 Discussion

This thesis contributes to two core perception tasks in collaborative robotics: extrinsic calibration of heterogeneous sensor configurations and multi-camera 3D human pose estimation. This discussion revisits the implications of the work in light of ongoing technological trends, the broader academic discourse, and future industrial demands.

6.3.1 Calibration in Dynamic and Multi-Modal Contexts

The calibration framework developed in this thesis addresses an increasingly common scenario in robotics: dynamic, multi-modal sensor arrangements that do not conform to static or tightly integrated hardware setups. As collaborative and modular robotic platforms proliferate, so too does the complexity of spatial configuration, making traditional, manual calibration tools insufficient or impractical.

What distinguishes the proposed method from existing solutions is not only its accuracy but also its attention to operational feasibility, integrating hand-eye calibration, supporting RGB-D and LiDAR sensors, and aligning with ROS workflows. By building on and extending ATOM, the work acknowledges the importance of leveraging existing structures while addressing their limitations. In doing so, the research contributes to an emerging design philosophy in robotics: that perception systems must be both technically rigorous and adaptable to fluid, often unpredictable, working conditions.

These contributions are detailed in two of the thesis's main publications. The journal article published in *Journal of Manufacturing Systems* [140] presents the core calibration framework and its validation in an industrial collaborative cell. A complementary article, submitted to *IEEE Access*, further refines this framework by introducing new procedures tailored to the calibration of depth sensors under motion. Together, these works advance the state of the art in sensor fusion and contribute to the broader push for perception pipelines that can support flexible, semantically aware robotic systems.

While the method has shown strong performance in controlled and semi-structured environments, further testing is necessary in more adversarial conditions (e.g., outdoor scenes, non-rigid mounts, or autonomous vehicle contexts). Nonetheless, the results clearly demonstrate that high-accuracy calibration is achievable even when the assumptions of static geometry or fixed sensor poses are relaxed.

6.3.2 Using Pose Estimation to Enable Human-Aware Collaboration

The second contribution of this thesis lies in analysing how human pose estimation methods perform in multi-camera settings designed for robotic collaboration. Unlike conventional vision applications, robotic systems require not only accurate human detection, but spatially coherent, continuous tracking that can be trusted by the control system. The evaluation in Chapter 5 goes beyond accuracy metrics to address issues of occlusion, visibility, and triangulation geometry, factors critical for industrial application but often neglected in vision-only literature.

This approach positions human pose estimation not as an isolated computer vision problem, but as a perceptual foundation for context-aware robotics. By highlighting the advantages of view redundancy and calibration fidelity, the thesis supports a paradigm in which multiple, well-placed cameras offer not just more data, but better decisions. In safety-critical domains, where failure to detect a human limb can result in injury, this level of precision is not optional.

Moreover, the experimental findings reinforce the interplay between hardware design (e.g., camera placement and number) and algorithmic strategy (e.g., keypoint detection and fusion).

These results offer practical guidance to engineers designing collaborative workspaces and underline the importance of calibrating perception systems as holistic ecosystems rather than isolated components. This contribution has also been formalised and validated in a peer-reviewed publication at the 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), which presents the developed multi-view lifting pipeline with occluded joint prediction as a robust solution for pose estimation in cluttered and collaborative environments [147].

6.3.3 Integration and Real-World Deployment Potential

A strength of this thesis is its commitment to bridging the gap between conceptual advances and practical deployment. The open-source implementation of the calibration and estimation tools, and their compatibility with standard robotic platforms, means they can be readily tested, adapted, and scaled. This practical orientation strengthens the research's impact and speaks to the demands of applied robotics research, where reproducibility and integration are as valued as innovation.

The work also responds to the growing consensus that modern robotic systems must be designed for resilience, reconfigurability, and human co-presence. As such, it contributes to a broader trend in mechanical engineering: one that favours modularity, sensor fusion, and semantic understanding over hardwired, pre-specified behaviours.

In sum, the thesis demonstrates that perception in collaborative robotics is no longer a question of isolated performance gains in calibration or detection. Rather, it is a systems-level challenge—one where reliability, flexibility, and integration define the quality and safety of human-robot interaction.

6.4 Limitations and Open Challenges

Despite the contributions presented in this thesis, several limitations and open challenges remain, highlighting areas for further refinement and future investigation. These constraints affect both the calibration and pose estimation components and reflect broader gaps identified in the literature on perceptual systems for collaborative robotics.

On the calibration side, while the proposed framework achieves accurate results across both static and hand-eye configurations, it assumes sufficient environmental structure and visibility for robust visual feature extraction. In environments with poor texture, reflective surfaces, or limited field of view, conditions common in industrial or cluttered settings, the reliability of calibration may deteriorate. This limitation is consistent with findings by Horn et al. [66] and Hua et al. [51], who note that traditional calibration methods often struggle in unstructured or low-feature scenes. Although targetless or geometric feature-based approaches mitigate this issue to some extent, they may introduce higher uncertainty or require longer observation windows.

A second limitation concerns the temporal rigidity of the current calibration method. The pipeline operates in an offline mode and does not yet support online or incremental recalibration, which is a critical requirement in reconfigurable or long-duration deployments.

As pointed out by Furgale et al. [9] and discussed in Chapter 2, online calibration remains an unsolved problem for many robotic platforms. In dynamic settings, where sensors may shift due to thermal drift, vibration, or workspace rearrangement, the lack of adaptive recalibration could lead to spatial inconsistencies and degraded performance in downstream perception tasks.

Regarding 3D human pose estimation, the system developed in this thesis is limited to static, pre-calibrated multi-camera setups and assumes the presence of a single user within the workspace. While this design simplifies geometric reasoning and enables accurate triangulation, it does not address scenarios involving multiple overlapping individuals, occlusions, or rapid motion. Multi-user tracking, in particular, introduces significant complexity in terms of identity maintenance, temporal coherence, and robustness to visual ambiguity [113], [117]. These challenges are especially relevant for industrial settings, where operators may move unpredictably or interact with the robot simultaneously.

Real-time performance is another critical constraint. Although the pipeline achieves acceptable processing rates for moderate frame sizes, it has not yet been optimised for embedded or resource-constrained platforms. Many industrial deployments require tight latency budgets and deterministic behaviour, especially in safety-critical applications. Models such as STIGANet [126] and MotionAGFormer [96] attempt to address this through architectural efficiency and temporal reasoning, but trade-offs remain between speed, accuracy, and generalisability. The lack of formal evaluation of latency, memory usage, or jitter under deployment conditions represents a limitation of the current work.

Additionally, the current implementation assumes full camera calibration is available and remains valid throughout operation. As noted in the literature [38], [47], pose estimation pipelines are often sensitive to small calibration errors, especially in triangulation-based systems. In practice, even minor misalignments may introduce significant degradation in 3D joint accuracy. This highlights the need for tighter coupling between calibration and estimation modules, as well as the development of error-aware or self-correcting pose estimation pipelines.

Finally, the system has been evaluated in controlled environments with standard illumination and limited background clutter. Its robustness under varying lighting conditions, sensor noise, or environmental occlusions remains to be systematically studied. As observed by Bauer et al. [115] and Ye et al. [114], real-world deployments often involve dynamic changes that cannot be fully captured in standard lab-based validations.

In summary, while the methods presented in this thesis address several key challenges in sensor calibration and 3D human pose estimation for collaborative robotics, important limitations persist. Addressing these open challenges will be essential for transitioning these capabilities from prototype systems to robust, deployable perception modules in real-world industrial environments.

6.5 Future Research Directions

Following the limitations and open challenges identified throughout this thesis, several promising directions emerge for future research. One clear extension involves the development

of online or adaptive calibration strategies. While the proposed method achieves accurate extrinsic estimation for both simulated and real-world deployments, it currently assumes a static configuration during the calibration phase. This limits applicability in long-term or reconfigurable setups, where sensor positioning may change due to drift, mechanical vibration, or workspace reconfiguration. As highlighted by Furgale et al. [9], continuous or opportunistic calibration is essential for maintaining perception accuracy over time. Adaptive approaches that integrate calibration into runtime estimation, potentially through joint optimisation frameworks or factor graph methods, could enable robust operation in dynamic environments.

A second line of inquiry concerns the integration of richer multi-modal sensing. Although this thesis supports RGB, RGB-D, and LiDAR configurations, perception systems could benefit from the fusion of inertial, thermal, or radar data. Prior work has demonstrated that combining inertial cues with vision improves robustness under motion and lighting variation [9], while others have explored thermal–RGB fusion in constrained domains [130]. Extending the current framework to handle loosely or asynchronously coupled sensors would pose challenges in synchronisation and calibration but could unlock new application domains in poorly lit, cluttered, or safety-critical environments.

Third, there is a need to strengthen the coupling between perception and control. Although this thesis focuses on the perceptual layer, real-world systems require integration of calibrated sensing into control loops that govern robot behaviour in response to human motion. Recent developments in adaptive planning and semantic reasoning, such as those described by Baptista et al. [11] and Argyrou et al. [16], rely on accurate spatial understanding to implement anticipatory behaviours. Future research should explore how pose estimates and uncertainty metrics can inform predictive or reactive control strategies, closing the loop between sensing and action in collaborative cells.

Lastly, broader benchmarking under diverse real-world conditions is essential to assess generalisability. Although this thesis validates its methods on representative use cases, further experimentation is needed across varying lighting, body types, workspace geometries, and occlusion levels. As emphasised by authors such as Bauer et al. [115] and Fürst et al. [113], most current datasets fail to reflect the complexity of industrial human–robot interaction. Creating realistic benchmarks with annotated 3D pose and ground truth calibration, especially under partial visibility and asynchronous sensing, would greatly advance reproducibility and deployment-readiness.

In summary, future developments should aim for perception systems that are not only accurate, but also adaptive, multi-modal, and tightly integrated into robotic control architectures. The contributions of this thesis provide a foundation upon which these capabilities can be developed and validated.

6.6 Final Remarks

In conclusion, this thesis provides a cohesive set of methodological, experimental, and implementation contributions to the field of collaborative robotics. The work demonstrates that accurate sensor calibration and robust human pose estimation are not isolated challenges,

but mutually reinforcing components of a perceptually intelligent robotic system. By offering a novel calibration framework, validating it in dynamic multi-sensor environments, and applying it to improve spatial awareness in human tracking, the thesis lays a strong foundation for future research and development. The tools and insights presented here are expected to support safer, more reliable, and more capable human-robot collaboration across a range of industrial and research settings.

References

[1] V. Villani, F. Pini, F. Leali, and C. Secchi, «Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications», *Mechatronics*, vol. 55, pp. 248–266, 2018, ISSN: 0957-4158. DOI: 10.1016/j.mechatronics.2018.02.009.

[2] A. Bauer, D. Wollherr, and M. Buss, «Human-robot collaboration: A survey.», *International Journal of Humanoid Robotics*, vol. 5, pp. 47–66, Mar. 2008. DOI: 10.1142/S0219843608001303.

[3] A. Ajoudani, A. M. Zanchettin, S. Ivaldi, A. Albu-Schäffer, K. Kosuge, and O. Khatib, «Progress and prospects of the human-robot collaboration», *Autonomous Robots*, vol. 42, Jun. 2018. DOI: 10.1007/s10514-017-9677-2.

[4] R. Unnikrishnan and M. Hebert, «Fast extrinsic calibration of a laser rangefinder to a camera», Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMU-RI-TR-05-09, 2005.

[5] J. Levinson and S. Thrun, «Automatic online calibration of cameras and lasers», in *Proceedings of Robotics: Science and Systems*, Jun. 2013. DOI: 10.15607/RSS.2013.IX.029.

[6] Z. Taylor and J. Nieto, «A mutual information approach to automatic calibration of camera and lidar in natural environments», in *Australasian Conference on Robotics and Automation 2012*, Dec. 2012. DOI: 10.13140/2.1.4124.0321.

[7] K. Iskakov, E. Burkov, V. Lempitsky, and Y. Malkov, «Learnable triangulation of human pose», in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. DOI: 10.1109/ICCV.2019.00781.

[8] Z. Zhang, «A flexible new technique for camera calibration», *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330–1334, 2000. DOI: 10.1109/34.888718.

[9] P. Furgale, J. Rehder, and R. Siegwart, «Unified temporal and spatial calibration for multi-sensor systems», in *IEEE International Conference on Intelligent Robots and Systems*, 2013, pp. 1280–1286. DOI: 10.1109/IROS.2013.6696514.

[10] D. Mehta, H. Rhodin, D. Casas, *et al.*, «Monocular 3D human pose estimation in the wild using improved cnn supervision», in *Fifth International Conference on 3D Vision*, IEEE, 2017. DOI: 10.1109/3dv.2017.00064.

[11] J. Baptista, A. Castro, M. Gomes, *et al.*, «Human–robot collaborative manufacturing cell with learning-based interaction abilities», *Robotics*, vol. 13, no. 7, 2024. DOI: 10.3390/robotics13070107.

[12] L. Wei, Y. Wang, Y. Hu, T. Lam, and Y. Wei, «Online dual robot–human collaboration trajectory generation by convex optimization», *Robotics and Computer-Integrated Manufacturing*, vol. 91, 2025. DOI: 10.1016/j.rcim.2024.102850.

[13] C. Tonola, M. Faroni, N. Pedrocchi, and M. Beschi, «Anytime informed path re-planning and optimization for human-robot collaboration», in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, 2021, pp. 997–1002. DOI: 10.1109/RO-MAN50785.2021.9515422.

[14] A. Umbrico, A. Orlandini, and A. Cesta, «An ontology for human-robot collaboration», *Procedia CIRP*, vol. 93, pp. 1097–1102, 2020, 53rd CIRP Conference on Manufacturing Systems 2020, ISSN: 2212-8271. DOI: https://doi.org/10.1016/j.procir.2020.04.045.

[15] A. Umbrico, A. Cesta, and A. Orlandini, «Enhancing awareness of industrial robots in collaborative manufacturing», *Semantic Web*, vol. 15, no. 2, pp. 389–428, 2024. DOI: 10.3233/SW-233394.

[16] A. Argyrou, C. Giannoulis, A. Sardelis, P. Karagiannis, G. Michalos, and S. Makris, «A data fusion system for controlling the execution status in human-robot collaborative cells», vol. 76, 2018, pp. 193–198. DOI: 10.1016/j.procir.2018.01.012.

[17] C. Cella, A. Maria Zanchettin, and P. Rocco, «Digital technologies for the design of human-robot collaborative cells», in *2023 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering, MetroXRAINE 2023 - Proceedings*, 2023, pp. 438–443. DOI: 10.1109/MetroXRAINE58569.2023.10405765.

[18] T. Hanning, A. Lasaruk, and T. Tatschke, «Calibration and low-level data fusion algorithms for a parallel 2d/3d-camera», *Information Fusion*, vol. 12, no. 1, pp. 37–47, 2011, Special Issue on Intelligent Transportation Systems, ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2010.01.006.

[19] M. Tsogas, N. Floudas, P. Lytrivis, A. Amditis, and A. Polychronopoulos, «Combined lane and road attributes extraction by fusing data from digital map, laser scanner and camera», *Information Fusion*, vol. 12, no. 1, pp. 28–36, 2011, Special Issue on Intelligent Transportation Systems, ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2010.01.005.

[20] M. Oliveira, V. Santos, and A. D. Sappa, «Multimodal inverse perspective mapping», *Information Fusion*, vol. 24, pp. 108–121, 2015, ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2014.09.003.

[21] W. Jiuqing, C. Xu, B. Shaocong, and L. Li, «Distributed data association in smart camera network via dual decomposition», *Information Fusion*, vol. 39, pp. 120–138, 2018, ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2017.04.007.

[22] A. M. Pinto and A. C. Matos, «Maresye: A hybrid imaging system for underwater robotic applications», *Information Fusion*, vol. 55, pp. 16–29, 2020, ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2019.07.014.

[23] R. Arrais, M. Oliveira, C. Toscano, and G. Veiga, «A mobile robot based sensing approach for assessing spatial inconsistencies of a logistic system», *Journal of Manufacturing Systems*, vol. 43, pp. 129–138, 2017, ISSN: 0278-6125. DOI: 10.1016/j.jmsy.2017.02.016.

[24] B. Rasti and P. Ghamisi, «Remote sensing image classification using subspace sensor fusion», *Information Fusion*, vol. 64, pp. 121–130, 2020, ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2020.07.002.

[25] C. Yaqing and W. Huaming, «Robust extrinsic calibration for lidar-camera systems via depth and height complementary supervision network», *IEEE Access*, vol. 13, pp. 35 818–35 828, 2025. DOI: 10.1109/ACCESS.2025.3542279.

[26] R. Su, J. Zhong, Q. Li, S. Qi, H. Zhang, and T. Wang, «An automatic calibration system for binocular stereo imaging», in *2016 IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, 2016, pp. 896–900. DOI: 10.1109/IMCEC.2016.7867340.

[27] Y. Ling and S. Shen, «High-precision online markerless stereo extrinsic calibration», in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016, pp. 1771–1778. DOI: 10.1109/IROS.2016.7759283.

[28] G. Mueller and H. Wuensche, «Continuous stereo camera calibration in urban scenarios», in *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, 2017, pp. 1–6. DOI: 10.1109/ITSC.2017.8317675.

[29] V. Dinh, T. Nguyen, and J. Jeon, «Rectification using different types of cameras attached to a vehicle», *IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 815–826, 2019, ISSN: 1057-7149. DOI: 10.1109/TIP.2018.2870930.

[30] L. Wu and B. Zhu, «2015 ieee international conference on mechatronics and automation (icma)», in *Binocular stereovision camera calibration*, 2015, pp. 2638–2642. DOI: 10.1109/ICMA.2015.7237903.

[31] H. Liu, D. Qu, F. Xu, F. Zou, J. Song, and K. Jia, «Approach for accurate calibration of RGB-D cameras using spheres», *Optica Express*, vol. 28, no. 13, pp. 19 058–19 073, 2020. DOI: 10.1364/OE.392414.

[32] F. Basso, E. Menegatti, and A. Pretto, «Robust intrinsic and extrinsic calibration of rgb-d cameras», *IEEE Transactions on Robotics*, vol. 34, no. 5, pp. 1315–1332, 2018. DOI: 10.1109/TRO.2018.2853742.

[33] G. Chen, G. Cui, Z. Jin, F. Wu, and X. Chen, «Accurate intrinsic and extrinsic calibration of RGB-D cameras with GP-based depth correction», *IEEE Sensors Journal*, vol. 19, no. 7, pp. 2685–2694, 2019. DOI: 10.1109/JSEN.2018.2889805.

[34] Y. Kwon, J. Jang, and O. Choi, «2018 18th international conference on control, automation and systems (iccas)», in *Automatic sphere detection for extrinsic calibration of multiple RGBD cameras*, 2018, pp. 1451–1454.

[35] F. Vasconcelos, J. Barreto, and U. Nunes, «A minimal solution for the extrinsic calibration of a camera and a laser-rangefinder», *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2097–2107, 2012. DOI: 10.1109/TPAMI.2012.18.

[36] J. Rehder, R. Siegwart, and P. Furgale, «A general approach to spatiotemporal calibration in multisensor systems», *IEEE Transactions on Robotics*, vol. 32, no. 2, pp. 383–398, 2016. DOI: 10.1109/TRO.2016.2529645.

[37] L. Huang and M. Barth, «Ieee intelligent vehicles symposium», 2009, pp. 117–122. DOI: 10.1109/IVS.2009.5164263.

[38] L. Zhou, Z. Li, and M. Kaess, «Automatic extrinsic calibration of a camera and a 3d lidar using line and plane correspondences», in *IEEE International Conference on Intelligent Robots and Systems*, 2018, pp. 5562–5569. DOI: 10.1109/IROS.2018.8593660.

[39] C. Guindel, J. Beltrán, D. Martín, and F. García, «Automatic extrinsic calibration for LiDAR-stereo vehicle sensor setups», in *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, 2017, pp. 1–6. DOI: 10.1109/ITSC.2017.8317829.

[40] W. Wang, K. Sakurada, and N. Kawaguchi, «Reflectance intensity assisted automatic and accurate extrinsic calibration of 3D LiDAR and panoramic camera using a printed chessboard», *Remote Sensing*, vol. 9, no. 8, 2017. DOI: 10.3390/rs9080851.

[41] C. Yang, P. Curtis, and P. Payeur, «Calibration of an integrated robotic multimodal range scanner», *IEEE Transactions on Instrumentation and Measurement*, vol. 55, no. 4, pp. 1148–1159, 2006. DOI: 10.1109/TIM.2006.876410.

[42] S. Verma, J. Berrio, S. Worrall, and E. Nebot, «Automatic extrinsic calibration between a camera and a 3D LiDAR using 3D point and plane correspondences», in *2019 IEEE Intelligent Transportation Systems Conference, ITSC 2019*, 2019, pp. 3906–3912. DOI: 10.1109/ITSC.2019.8917108.

[43] D. Hu, D. DeTone, and T. Malisiewicz, «Deep charuco: Dark charuco marker pose estimation», in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. DOI: 10.1109/CVPR.2019.00863.

[44] F. J. Romero-Ramirez, R. Muñoz-Salinas, and R. Medina-Carnicer, «Speeded up detection of squared fiducial markers», *Image and Vision Computing*, vol. 76, pp. 38–47, 2018, ISSN: 0262-8856. DOI: 10.1016/j.imavis.2018.05.004.

[45] B.-S. Park, W. Kim, J.-K. Kim, D.-W. Kim, and Y.-H. Seo, «Iterative extrinsic calibration using virtual viewpoint for 3D reconstruction», *Signal Processing*, vol. 197, 2022. DOI: 10.1016/j.sigpro.2022.108535.

[46] J. Chaochuan, Y. Ting, W. Chuanjiang, F. Binghui, and H. Fugui, «An extrinsic calibration method for multiple RGB-D cameras in a limited field of view», *Measurement Science and Technology*, 2020. DOI: 10.1088/1361-6501/ab48b3.

[47] C. Raposo, J. Barreto, and U. Nunes, «Extrinsic calibration of multi-modal sensor arrangements with non-overlapping field-of-view», *Machine Vision and Applications*, vol. 28, no. 1-2, pp. 141–155, 2017. DOI: 10.1007/s00138-016-0815-1.

[48] M. Oliveira, E. Pedrosa, A. P. de Aguiar, *et al.*, «Atom: A general calibration framework for multi-modal, multi-sensor systems», *Expert Systems with Applications*, vol. 207, p. 118 000, 2022, ISSN: 0957-4174. DOI: 10.1016/j.eswa.2022.118000.

[49] E. Pedrosa, M. Oliveira, N. Lau, and V. Santos, «A general approach to hand–eye calibration through the optimization of atomic transformations», *IEEE Transactions on Robotics*, 1–15, 2021. DOI: 10.1109/TRO.2021.3062306.

[50] A. Aguiar, M. Oliveira, E. Pedrosa, and F. Santos, «A camera to LiDAR calibration approach through the optimization of atomic transformations», *Expert Systems with Applications*, 2021, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2021.114894.

[51] Y. Hua, Q. Liu, T. Jiang, J. Zhang, W. Xu, and Y. Tian, «Accurate lidar–camera calibration using feature edges», *Image and Vision Computing*, vol. 155, 2025. DOI: 10.1016/j.imavis.2024.105394.

[52] D. Kim, S. Shin, and H. Hwang, «Camera-lidar extrinsic calibration using constrained optimization with circle placement», *IEEE Robotics and Automation Letters*, vol. 10, no. 2, pp. 883–890, 2025. DOI: 10.1109/LRA.2024.3512253.

[53] S. Garrido-Jurado, R. Muñoz-Salinas, F. Madrid-Cuevas, and R. Medina-Carnicer, «Generation of fiducial marker dictionaries using mixed integer linear programming», *Pattern Recognition*, vol. 51, pp. 481–491, 2016, ISSN: 0031-3203. DOI: 10.1016/j.patcog.2015.09.023.

[54] W. Zhu, S. Shan, C. Tong, K. Zhang, and H. Wei, «Targetless automatic edge-to-edge extrinsic calibration of lidar-camera system based on pure laser reflectivity», *IEEE Transactions on Instrumentation and Measurement*, 2025. DOI: 10.1109/TIM.2025.3560719.

[55] S. Wang, F. Tang, C. Shi, and Y. Wu, «Targetless LiDAR-camera extrinsic calibration with mesh-based constraints», *IEEE Transactions on Instrumentation and Measurement*, vol. 74, 2025. DOI: 10.1109/TIM.2025.3551570.

[56] D. Cattaneo and A. Valada, «Cmrnext: Camera to lidar matching in the wild for localization and extrinsic calibration», *IEEE Transactions on Robotics*, vol. 41, pp. 1995–2013, 2025. DOI: 10.1109/TRO.2025.3546784.

[57] P.-T. Lin, Y.-S. Huang, W.-C. Lin, C.-C. Wang, and H.-Y. Lin, «Online LiDAR-camera extrinsic calibration using selected semantic features», *IEEE Open Journal of Intelligent Transportation Systems*, vol. 6, pp. 456–464, 2025. DOI: 10.1109/OJITS.2025.3555574.

[58] W. Liu, Z. Li, S. Sun, H. Du, and M. A. Sotelo, «A novel motion-based online temporal calibration method for multi-rate sensors fusion», *Information Fusion*, vol. 88, pp. 59–77, 2022, ISSN: 1566-2535. DOI: 10.1016/j.inffus.2022.07.004.

[59] X. Gao and T. Zhang, «Robust RGB-D simultaneous localization and mapping using planar point features», *Robotics and Autonomous Systems*, vol. 72, pp. 1–14, 2015, ISSN: 0921-8890. DOI: 10.1016/j.robot.2015.03.007.

[60] P. Kurtser and S. Lowry, «RGB-D datasets for robotic perception in site-specific agricultural operations—A survey», *Computers and Electronics in Agriculture*, vol. 212, p. 108 035, 2023, ISSN: 0168-1699. DOI: 10.1016/j.compag.2023.108035.

[61] M. Schwarz, A. Milan, A. S. Periyasamy, and S. Behnke, «RGB-D object detection and semantic segmentation for autonomous manipulation in clutter», *The International Journal of Robotics Research*, vol. 37, no. 4-5, pp. 437–451, 2018. DOI: 10.1177/0278364917713117.

[62] A. Mosella-Montoro and J. Ruiz-Hidalgo, «2D–3D Geometric Fusion network using Multi-Neighbourhood Graph Convolution for RGB-D indoor scene classification», *Information Fusion*, vol. 76, pp. 46–54, 2021, ISSN: 1566-2535. DOI: 10.1016/j.inffus.2021.05.002.

[63] L. Bo, X. Ren, and D. Fox, «Learning hierarchical sparse features for rgb-(d) object recognition», *The International Journal of Robotics Research*, vol. 33, no. 4, pp. 581–599, 2014. DOI: 10.1177/0278364913514283.

[64] A. N. Staranowicz, G. R. Brown, F. Morbidi, and G.-L. Mariottini, «Practical and accurate calibration of RGB-D cameras using spheres», *Computer Vision and Image Understanding*, vol. 137, pp. 102–114, 2015, ISSN: 1077-3142. DOI: 10.1016/j.cviu.2015.03.013.

[65] Y. Zhang, G. Li, X. Xie, and Z. Wang, «A new algorithm for accurate and automatic chessboard corner detection», in *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2017, pp. 1–4. DOI: 10.1109/ISCAS.2017.8050637.

[66] M. Horn, T. Wodtko, M. Buchholz, and K. Dietmayer, «Online extrinsic calibration based on per-sensor ego-motion using dual quaternions», *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 982–989, 2021. DOI: 10.1109/LRA.2021.3056352.

[67] A. Geiger, P. Lenz, and R. Urtasun, «Are we ready for autonomous driving? the KITTI vision benchmark suite», in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. DOI: 10.1109/CVPR.2012.6248074.

[68] R. Y. Tsai, «A new technique for fully autonomous and efficient 3D robotics hand/eye calibration», *IEEE Transactions on Robotics and Automation*, vol. 5, no. 3, pp. 345–358, 1989. DOI: 10.1109/70.34770.

[69] Y. Shiu and S. Ahmad, «Calibration of wrist-mounted robotic sensors by solving homogeneous transform equations of the form ax=xb», *IEEE Transactions on Robotics and Automation*, vol. 5, no. 1, pp. 16–29, 1989. DOI: 10.1109/70.88014.

[70] J. Jiang, X. Luo, S. Xu, Q. Luo, and M. Li, «Hand-Eye Calibration of EOD Robot by Solving the AXB = YCZD Problem», *IEEE Access*, vol. 10, pp. 3415–3429, 2022. DOI: 10.1109/ACCESS.2021.3136850.

[71] H. Tian, K. Song, J. Xu, S. Ma, and Y. Yan, «Antipodal-points-aware dual-decoding network for robotic visual grasp detection oriented to multi-object clutter scenes», *Expert Systems with Applications*, vol. 230, p. 120 545, 2023, ISSN: 0957-4174. DOI: 10.1016/j.eswa.2023.120545.

[72] H. Fu, D. Xu, and J. Wu, «Robotic arm intelligent grasping system for garbage recycling», *2021 China Automation Congress, CAC 2021*, pp. 6821–6826, 2021. DOI: 10.1109/CAC53003.2021.9728547.

[73] X. Liang and H. Cheng, «RGB-D camera based 3D object pose estimation and grasping», *9th IEEE International Conference on Cyber Technology in Automation*, pp. 1279–1284, 2019. DOI: 10.1109/CYBER46603.2019.9066550.

[74] Y. Pan, C. Chen, Z. Zhao, T. Hu, and J. Zhang, «Robot teaching system based on hand-robot contact state detection and motion intention recognition», *Robotics and Computer-Integrated Manufacturing*, vol. 81, p. 102 492, 2023, ISSN: 0736-5845. DOI: 10.1016/j.rcim.2022.102492.

[75] H. Cheng, H. Chen, and Y. Liu, «Object handling using autonomous industrial mobile manipulator», *2013 IEEE International Conference on Cyber Technology in Automation, Control and Intelligent Systems, IEEE-CYBER 2013*, pp. 36–41, 2013. DOI: 10.1109/CYBER.2013.6705416.

[76] D. Cao, W. Liu, S. Liu, *et al.*, «Simultaneous calibration of hand-eye and kinematics for industrial robot using line-structured light sensor», *Measurement*, vol. 221, p. 113 508, 2023, ISSN: 0263-2241. DOI: 10.1016/j.measurement.2023.113508.

[77] F. Dornaika and R. Horaud, «Simultaneous robot-world and hand-eye calibration», *Robotics and Automation, IEEE Transactions on*, vol. 14, pp. 617–622, Sep. 1998. DOI: 10.1109/70.704233.

[78] A. Tabb and K. M. Ahmad Yousef, «Solving the robot-world hand-eye(s) calibration problem with iterative methods», *Machine Vision and Applications*, vol. 28, no. 5-6, pp. 569–590, 2017. DOI: 10.1007/s00138-017-0841-7.

[79] Z. Zhao, «Simultaneous robot-world and hand-eye calibration by the alternative linear programming», *Pattern Recognition Letters*, vol. 127, pp. 174–180, 2019, ISSN: 0167-8655. DOI: 10.1016/j.patrec.2018.08.023.

[80] E. Pedrosa, M. Oliveira, N. Lau, and V. Santos, «A general approach to hand–eye calibration through the optimization of atomic transformations», *IEEE Transactions on Robotics*, vol. 37, no. 5, pp. 1619–1633, 2021. DOI: 10.1109/TRO.2021.3062306.

[81] A. Krishnan and S. Saripalli, «Cross-calibration of RGB and thermal cameras with a LIDAR for RGB-depth-thermal mapping», *Unmanned Systems*, vol. 5, no. 2, pp. 59–78, 2017. DOI: 10.1142/S2301385017500054.

[82] M. Oliveira, A. Castro, T. Madeira, P. Dias, and V. Santos, «A general approach to the extrinsic calibration of intelligent vehicles using ROS», *Fourth Iberian Robotics Conference*, 2019.

[83] X. Chen, S. Xiang, and J. Zhou, «A svm based extrinsic calibration method for RGB-D camera», *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, BMSB*, vol. 2021-August, 2021. DOI: 10.1109/BMSB53066.2021.9547097.

[84] Y. Zhang, S. Ahmadi, J. Kang, Z. Arjmandi, and G. Sohn, «Yuto mms: A comprehensive slam dataset for urban mobile mapping with tilted lidar and panoramic camera integration», *The International Journal of Robotics Research*, vol. 44, no. 1, pp. 3–21, 2025. DOI: 10.1177/02783649241261079.

[85] L. Zhao, X. Peng, Y. Tian, M. Kapadia, and D. N. Metaxas, «Semantic graph convolutional networks for 3D human pose regression», in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3420–3430. DOI: 10.1109/CVPR.2019.00354.

[86] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, «SMPL: A skinned multi-person linear model», *ACM Transactions Graphics (Proc. SIGGRAPH Asia)*, vol. 34, 248:1–248:16, 2015. DOI: 10.1145/2816795.2818013.

[87] G. Pavlakos, V. Choutas, N. Ghorbani, *et al.*, «Expressive body capture: 3D hands, face, and body from a single image», in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 10 975–10 985. DOI: 10.1109/CVPR.2019.01123.

[88] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, «Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image», in *European Conference on Computer Vision*, 2016. DOI: 10.1007/978-3-319-46454-1_34.

[89] V. Choutas, F. Bogo, J. Shen, and J. Valentin, «Learning to fit morphable models», in *17th European Conference on Computer Vision*, 2022. DOI: 10.1007/978-3-031-20068-7_10.

[90] L. Müller, A. A. A. Osman, S. Tang, C.-H. P. Huang, and M. J. Black, «On self-contact and human pose», in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. DOI: 10.1109/CVPR46437.2021.00986.

[91] J. Song, X. Chen, and O. Hilliges, «Human body model fitting by learned gradient descent», in *2020 European Conference on Computer Vision*. 2020, pp. 744–760, ISBN: 978-3-030-58564-8. DOI: 10.1007/978-3-030-58565-5_44.

[92] Y. Sun, Q. Bao, W. Liu, Y. Fu, M. J. Black, and T. Mei, «Monocular, one-stage, regression of multiple 3D people», in *2021 IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 159–11 168. DOI: 10.1109/ICCV48922.2021.01099.

[93] C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen, and Z. Ding, «3D human pose estimation with spatial and temporal transformers», in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 11 636–11 645. DOI: 10.1109/ICCV48922.2021.01145.

[94] Q. Zhao, C. Zheng, M. Liu, P. Wang, and C. Chen, «PoseFormerV2: Exploring frequency domain for efficient and robust 3d human pose estimation», in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 8877–8886. DOI: 10.1109/CVPR52729.2023.00857.

[95] B. X. Yu, Z. Zhang, Y. Liu, S.-h. Zhong, Y. Liu, and C. W. Chen, «Gla-gcn: Global-local adaptive graph convolutional network for 3d human pose estimation from monocular video», in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 8818–8829. DOI: 10.1109/ICCV51070.2023.00810.

[96] B. T. Soroush Mehraban Vida Adeli, «Motionagformer: Enhancing 3D human pose estimation with a transformer-gcnformer network», in *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024. DOI: 10.1109/WACV57701.2024.00677.

[97] M. Kocabas, S. Karagoz, and E. Akbas, «Self-supervised learning of 3D human pose using multi-view geometry», in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1077–1086. DOI: 10.1109/CVPR.2019.00117.

[98] A. Bouazizi, J. Wiederer, U. Kressel, and V. Belagiannis, «Self-supervised 3d human pose estimation with multiple-view geometry», in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition*, 2021, pp. 1–8. DOI: 10.1109/FG52635.2021.9667074.

[99] J. Choi, D. Shim, and H. J. Kim, «Diffupose: Monocular 3D human pose estimation via denoising diffusion probabilistic model», in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2023, pp. 3773–3780. DOI: 10.1109/IROS55552.2023.10342204.

[100] Z. Jiang, Z. Zhou, L. Li, W. Chai, C.-Y. Yang, and J.-N. Hwang, «Back to optimization: Diffusion-based zero-shot 3D human pose estimation», in *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024. DOI: 10.1109/WACV57701.2024.00603.

[101] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, «Human3.6m: Large scale datasets and predictive methods for 3D human sensing in natural environments», *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, pp. 1325–1339, 2014. DOI: 10.1109/TPAMI.2013.248.

[102] L. Sigal, A. O. Balan, and M. J. Black, «HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion», *International Journal of Computer Vision*, vol. 87, pp. 4–27, 2010. DOI: 10.1007/s11263-009-0273-6.

[103] J. Martinez, R. Hossain, J. Romero, and J. J. Little, «A simple yet effective baseline for 3D human pose estimation», in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2659–2668. DOI: 10.1109/ICCV.2017.288.

[104] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, «3D human pose estimation in video with temporal convolutions and semi-supervised training», in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. DOI: 10.1109/CVPR.2019.00794.

[105] C.-Y. Yang, J. Luo, L. Xia, *et al.*, «Camerapose: Weakly-supervised monocular 3D human pose estimation by leveraging in-the-wild 2d annotations», *2023 IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2923–2932, 2023. DOI: 10.1109/WACV56688.2023.00294.

[106] R. A. Güler and I. Kokkinos, «Holopose: Holistic 3D human reconstruction in-the-wild», in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 876–10 886. DOI: 10.1109/CVPR.2019.01114.

[107] W. Zhu, X. Ma, Z. Liu, L. Liu, W. Wu, and Y. Wang, «Motionbert: A unified perspective on learning human motion representations», in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. DOI: 10.1109/ICCV51070.2023.01385.

[108] A. Bouazizi, U. Kressel, and V. Belagiannis, «Learning temporal 3d human pose estimation with pseudo-labels», in *17th IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2021, pp. 1–8. DOI: 10.1109/AVSS52988.2021.9663755.

[109] J. Li, J. Zhang, Z. Wang, *et al.*, «Lidarcap: Long-range markerless 3d human motion capture with lidar point clouds», in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 20 470–20 480. DOI: 10.1109/CVPR52688.2022.01985.

[110] X. An, L. Zhao, C. Gong, J. Li, and J. Yang, «Pre-training a density-aware pose transformer for robust LiDAR-based 3D human pose estimation», in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, 2025, pp. 1755–1763. DOI: 10.1609/aaai.v39i2.32169.

[111] P. Sun, H. Kretzschmar, X. Dotiwalla, *et al.*, «Scalability in perception for autonomous driving: Waymo open dataset», in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2443–2451. DOI: 10.1109/CVPR42600.2020.00252.

[112] Y. Dai, Y. Lin, X. Lin, *et al.*, «Sloper4d: A scene-aware dataset for global 4d human pose estimation in urban environments», in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 682–692. DOI: 10.1109/CVPR52729.2023.00073.

[113] M. Fürst, S. Gupta, R. Schuster, O. Wasenmüller, and D. Stricker, «HPERL: 3D human pose estimation from RGB and LiDAR», in *Proceedings - International Conference on Pattern Recognition*, 2020, pp. 7321–7327. DOI: 10.1109/ICPR48806.2021.9412785.

[114] D. Ye, Y. Xie, W. Chen, Z. Zhou, L. Ge, and H. Foroosh, «Lpformer: Lidar pose estimation transformer with multi-task network», in *Proceedings - IEEE International Conference on Robotics and Automation*, 2024, pp. 16 432–16 438. DOI: 10.1109/ICRA57147.2024.10611405.

[115] P. Bauer, A. Bouazizi, U. Kressel, and F. Flohr, «Weakly supervised multi-modal 3d human body pose estimation for autonomous driving», in *IEEE Intelligent Vehicles Symposium, Proceedings*, vol. 2023-June, 2023. DOI: 10.1109/IV55152.2023.10186575.

[116] J. Zheng, X. Shi, A. Gorban, *et al.*, «Multi-modal 3D human pose estimation with 2D weak supervision in autonomous driving», in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, vol. 2022-June, 2022, pp. 4477–4486. DOI: 10.1109/CVPRW56347.2022.00494.

[117] A. Zanfir, M. Zanfir, A. Gorban, *et al.*, «HUM3DIL: Semi-supervised multi-modal 3D humanpose estimation for autonomous driving», in *6th Annual Conference on Robot Learning*, 2022. [Online]. Available: https://openreview.net/forum?id=jTh3rdEF3LH.

[118] L. Kovács, B. Bódis, and C. Benedek, «LidPose: Real-time 3D human pose estimation in sparse lidar point clouds with non-repetitive circular scanning pattern», *Sensors*, vol. 24, no. 11, 2024. DOI: 10.3390/s24113427.

[119] I. Ballester, O. Peterka, and M. Kampel, «SPiKE: 3D human pose from point cloud sequences», *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 15318 LNCS, pp. 470–486, 2025. DOI: 10.1007/978-3-031-78456-9_30.

[120] S. An, Y. Li, and U. Ogras, «MRI: Multi-modal 3D human pose estimation dataset using mmWave, RGB-D, and inertial sensors», in *Advances in Neural Information Processing Systems*, vol. 35, 2022. DOI: 10.48550/ARXIV.2210.08394.

[121] H. Joo, T. Simon, X. Li, *et al.*, «Panoptic studio: A massively multiview system for social interaction capture», *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[122] T. von Marcard, R. Henschel, M. Black, B. Rosenhahn, and G. Pons-Moll, «Recovering accurate 3D human pose in the wild using imus and a moving camera», in *15th European Conference on Computer Vision*, 2018. DOI: 10.1007/978-3-030-01249-6_37.

[123] F. Koleini, M. U. Saleem, P. Wang, H. Xue, A. Helmy, and A. Fenwick, «Biopose: Biomechanically-accurate 3d pose estimation from monocular videos», in *2025 IEEE Winter Conference on Applications of Computer Vision (WACV 2025)*, 2025, pp. 6330–6339. DOI: 10.1109/WACV61041.2025.00617.

[124] X. Ge and V. Mariano, «Commercial applications of 3d human pose estimation in biotechnology: A two-stage fusion with multi-feature integration approach», *Journal of Commercial Biotechnology*, vol. 29, no. 3, pp. 402–415, 2024. DOI: 10.5912/jcb1880.

[125] B. Park, J. Kim, S. Mun, Y. Choi, and H. Kim, «Golfposenet: Golf-specific 3d human pose estimation network», in *2025 International Conference on Electronics, Information, and Communication (ICEIC 2025)*, 2025. DOI: 10.1109/ICEIC64972.2025.10879645.

[126] Q. Liu, Z. Wang, H. Zhang, and C. Miao, «Stiganet: Integrating dgcns and attention mechanisms for real-time 3d pose estimation in sports», *Alexandria Engineering Journal*, vol. 121, pp. 236–247, 2025. DOI: 10.1016/j.aej.2025.02.058.

[127] Z. Zhang, Q. Peng, L. Zhang, Z. Zhang, and W. Huang, «Stapformer: A new 3d human pose estimation framework in sports and health», in *15th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB 2024)*, 2024. DOI: 10.1145/3698587.3701367.

[128] S. Yuan and L. Zhou, «Gta-net: An iot-integrated 3d human pose estimation system for real-time adolescent sports posture correction», *Alexandria Engineering Journal*, vol. 112, pp. 585–597, 2025. DOI: 10.1016/j.aej.2024.10.099.

[129] K. Peppas, K. Tsiolis, I. Mariolis, A. Topalidou-Kyniazopoulou, and D. Tzovaras, «Multi-modal 3D human pose estimation for human-robot collaborative applications», *Lecture Notes in Computer Science*, vol. 12644 LNCS, pp. 355–364, 2021. DOI: 10.1007/978-3-030-73973-7_34.

[130] D. Antonelli and G. Bruno, «Ontology-based framework to design a collaborative human-robotic workcell», *IFIP Advances in Information and Communication Technology*, vol. 506, pp. 167–174, 2017. DOI: 10.1007/978-3-319-65151-4_16.

[131] European Commission, Directorate-General for Research and Innovation, and J. Müller, *Enabling Technologies for Industry 5.0 : results of a workshop with Europe's technology leaders*. Publications Office, 2020. DOI: doi/10.2777/082634.

[132] European Commission, Directorate-General for Research and Innovation, M. Breque, L. De Nul, and A. Petridis, *Industry 5.0 : towards a sustainable, human-centric and resilient European industry*. Publications Office, 2021. DOI: doi/10.2777/308407.

[133] E. Coronado, T. Kiyokawa, G. A. G. Ricardez, I. G. Ramirez-Alpizar, G. Venture, and N. Yamanobe, «Evaluating quality in human-robot interaction: A systematic search and classification of performance and human-centered factors, measures and metrics towards an industry 5.0», *Journal of Manufacturing Systems*, vol. 63, pp. 392–410, 2022, ISSN: 0278-6125. DOI: 10.1016/j.jmsy.2022.04.007.

[134] A. C. Simões, A. Pinto, J. Santos, S. Pinheiro, and D. Romero, «Designing human-robot collaboration (hrc) workspaces in industrial settings: A systematic literature review», *Journal of Manufacturing Systems*, vol. 62, pp. 28–43, 2022, ISSN: 0278-6125. DOI: 10.1016/j.jmsy.2021.11.007.

[135] I. O. for Standardization, *Robotics — Safety requirements*. ISO 10218, 2025.

[136] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, «Multimodal Machine Learning: A Survey and Taxonomy», en, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. DOI: 10.1109/TPAMI.2018.279860. (visited on 11/04/2019).

[137] J. Yuan, J. Zhang, S. Ding, and X. Dong, «Cooperative localization for disconnected sensor networks and a mobile robot in friendly environments», *Information Fusion*, vol. 37, pp. 22–36, 2017, ISSN: 1566-2535. DOI: doi.org/10.1016/j.inffus.2017.01.001.

[138] H. Yang, J. Yuan, Y. Gao, X. Sun, and X. Zhang, «UPLP-SLAM: Unified point-line-plane feature fusion for RGB-D visual SLAM», *Information Fusion*, vol. 96, pp. 51–65, 2023, ISSN: 1566-2535. DOI: 10.1016/j.inffus.2023.03.006.

[139] Z. Qiu, J. Martínez-Sánchez, P. Arias-Sánchez, and R. Rashdi, «External multi-modal imaging sensor calibration for sensor fusion: A review», *Information Fusion*, vol. 97, p. 101 806, 2023, ISSN: 1566-2535. DOI: 10.1016/j.inffus.2023.101806.

[140] D. Rato, M. Oliveira, V. Santos, M. Gomes, and A. Sappa, «A sensor-to-pattern calibration framework for multi-modal industrial collaborative cells», *Journal of Manufacturing Systems*, vol. 64, pp. 497–507, 2022. DOI: 10.1016/j.jmsy.2022.07.006.

[141] G. Bradski, «The opencv library», *Dr. Dobb's Journal of Software Tools*, 2000.

[142] H. M. Clever, A. Kapusta, D. Park, Z. Erickson, Y. Chitalia, and C. C. Kemp, «3D human pose estimation on a configurable bed from a pressure image», in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2018, pp. 54–61. DOI: 10.1109/IROS.2018.8593545.

[143] S. Hu, C. Zheng, Z. Zhou, C. Chen, and G. Sukthankar, «Lamp: Leveraging language prompts for multi-person pose estimation», in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2023, pp. 3759–3766. DOI: 10.1109/IROS55552.2023.10341430.

[144] A. Casalino, S. Guzman, A. Maria Zanchettin, and P. Rocco, «Human pose estimation in presence of occlusion using depth camera sensors, in human-robot coexistence scenarios», in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2018, pp. 1–7. DOI: 10.1109/IROS.2018.8593816.

[145] M. A. Branch, T. F. Coleman, and Y. Li, «A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems», *SIAM Journal on Scientific Computing*, vol. 21, pp. 1–23, 1999. DOI: 10.1137/S1064827595289108.

[146] C. Lugaresi, J. Tang, H. Nash, *et al.*, «Mediapipe: A framework for perceiving and processing reality», in *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*, 2019.

[147] D. Rato, M. Oliveira, V. Santos, A. Sappa, and B. Raducanu, «Multi-view 2d to 3d lifting video-based optimization: A robust approach for human pose estimation with occluded joint prediction*», in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, pp. 11 508–11 514. DOI: 10.1109/IROS58592.2024.10802200.