ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DEPARTMENT of Industrial Engineering - DIN
Second Cycle Degree
Mechanical Engineering for Sustainability

Testing of AI-driven gesture recognition techniques to support human-centered design of industrial systems

Supervisor:
Prof.
Margherita Peruzzini
Co-Supervisor:
Simone Borghi
Rocco De Ciantis

Defended by:
Emad Barandan

Graduation Session / July / 2026
Academic Year 2025/2026

Contents

Abstract

Work-related musculoskeletal disorders (WMSDs) affecting the hand and upper limb remain among the most prevalent occupational health problems in manufacturing and assembly. Their prevention depends on reliable ergonomic assessment of manual tasks — a process that is still largely carried out through manual observation by trained experts. Vision-based Hand Pose Estimation (HPE) offers a promising alternative, as it can potentially provide continuous, objective information about hand posture and movement without constraining the operator. However, state-of-the-art HPE models are trained on datasets featuring bare hands in controlled environments and consistently degrade in performance when transferred to real industrial conditions, where safety gloves, tool-driven occlusions, and cluttered workstations alter the visual appearance of the hand.

This thesis addresses the gap between the potential of vision-based hand tracking and its practical reliability in assembly environments. The work proposes a multi-modal acquisition and alignment strategy that combines ZED-2 stereo RGB-D cameras with Quantum Manus instrumented gloves. A Kabsch-based spatio-temporal alignment algorithm was developed to synchronize the two data streams and map the glove kinematics into the camera coordinate frame, making it possible to generate reliable visual supervision for gloved hands even under occlusion. This pipeline was applied to 18 assembly trials of a Bonfiglioli gearbox reducer task, covering both bare-hand and gloved-hand conditions, producing a structured and traceable dataset of more than 35,000 hand-centred samples for training and evaluating the proposed ErgoHand model.

A benchmark of representative off-the-shelf pipelines — MediaPipe, OpenPose, and WiLoR — was conducted under a binary hand-presence formulation. The results confirm a substantial industrial domain gap: MediaPipe and OpenPose achieved recall values below 0.07, while WiLoR improved recall to 0.39 but still missed more than half of the hands that were actually present. Under these conditions, reliable ergonomic assessment methods such as OCRA and HAL cannot be applied without critical data loss. The ErgoHand pipeline, trained on the generated industrial dataset, is designed to overcome these limitations; its full quantitative evaluation is currently in progress.

The contribution of this thesis lies in establishing the data infrastructure, alignment methodology, and domain-specific learning strategy that are necessary prerequisites for reliable hand monitoring in human-centered indus-

trial systems, in line with the sustainability and human-centricity pillars of Industry 5.0.

Keywords: Hand Pose Estimation, Industrial Ergonomics, Safety Gloves, Multi-modal Data Acquisition, Deep Learning, Industry 5.0, ErgoHand

Acknowledgements

"Listen to the reed, how it tells a tale of separations."
— Jalal ad-Din Rumi *(Mowlana)*

This thesis would not have been possible without the support of many people. First and foremost, I thank my supervisor, Prof. Margherita Peruzzini, for her constant guidance and ability to keep the broader purpose of this work in focus. Her perspective on human-centered design shaped this thesis in a fundamental way.

A special and heartfelt thank you goes to Simone Borghi, who was far more than a co-supervisor. From the earliest conceptual discussions to the most technical decisions, he was consistently present, generous with his knowledge, and genuinely invested in this research. This thesis would not have taken the shape it has without him. I am also grateful to Rocco De Ciantis for his technical support throughout the process.

I would like to thank my colleagues in the research group, and all the participants who volunteered their time for the recording sessions — their patience was essential to building this dataset.

This has been a particularly difficult year. Completing this thesis far from home, while carrying the weight of what my country and my people have been going through, was not easy. Many lives were lost. I dedicate part of this work to them — to those who could not pursue their own.

Finally, I want to thank my family for their unconditional support across thousands of kilometres — the most constant source of strength from the beginning to the very end.

Emad Barandan
Forlì, 2026

Glossary of Terms

This glossary summarizes the main technical terms, acronyms, models, datasets, sensors, and evaluation concepts used throughout the thesis.

2D Hand Pose Estimation Estimation of hand landmarks on the image plane, expressed as pixel coordinates of anatomical joints and fingertips.

3D Hand Pose Estimation Estimation of hand landmarks in three-dimensional space, providing spatial coordinates suitable for ergonomic interpretation and motion analysis.

3D Keypoints Regression Learning task in which a model predicts the 3D coordinates of the 21 hand joints from an input image or cropped hand region.

AdamW Optimization algorithm that decouples weight decay from the gradient update, improving regularization compared to standard Adam.

AMP (Automatic Mixed Precision) Training strategy that combines different numerical precisions to reduce GPU memory usage and accelerate computation while maintaining numerical stability.

Bare Hand Experimental condition in which the operator performs the assembly task without protective gloves.

Bounding Box Rectangular region surrounding a detected hand in an image, used to crop hand regions before downstream classification or regression.

Clock Drift Gradual mismatch between the internal clocks of two acquisition systems, potentially affecting synchronization between ZED and Manus recordings during longer trials.

CNN (Convolutional Neural Network) Deep-learning architecture that learns spatial features from images through convolutional filters.

Cross-Entropy Loss Loss function for classification tasks that measures the difference between predicted class probabilities and target labels.

Data Augmentation Technique that increases model robustness by applying transformations such as rotations, scaling, noise, or synthetic occlusions to training samples.

Depth Map Image-like representation in which each pixel stores distance from the camera, provided by RGB-D systems such as the ZED-2 camera.

DoF (Degrees of Freedom) Number of independent parameters required to describe the configuration of a system; human hands have many DoF, making pose estimation challenging.

Domain Gap Performance drop observed when a model designed for one environment is applied to another, particularly between controlled datasets and real industrial scenes.

Early Stopping Training strategy that halts optimization when validation performance ceases to improve, reducing overfitting.

ErgoHand Proposed learning pipeline designed to improve hand recognition in industrial scenes characterized by PPE gloves, clutter, and tool-driven occlusions.

False Negative Case in which a hand is present but the model predicts its absence.

Fine-Tuning Procedure in which a pretrained model is adapted to a new task by updating its parameters on domain-specific data.

FNR (False Negative Rate) Proportion of present hands missed by a model; critical in ergonomic monitoring because missed detections interrupt temporal continuity.

GCN (Graph Convolutional Network) Neural network that applies convolution-like operations on graph-structured data, such as hand skeletons represented as joints and connections.

Glove-Hand Classification Binary classification task that distinguishes whether a detected hand is bare or covered by a PPE glove.

GNN (Graph Neural Network) Neural-network family designed to process graph-structured data, used here to classify hand-related information from 3D keypoint graphs.

Ground Truth Reference label or measurement used to train and evaluate a model, generated by aligning Manus glove data with ZED camera recordings.

HAL (Hand Activity Level) Ergonomic assessment index defined by the American Conference of Governmental Industrial Hygienists (ACGIH) to quantify exposure to repetitive hand-intensive tasks; used alongside normalised peak force to define threshold limit values for the upper limb.

Hand Detection Task of identifying whether a hand is present in an image and locating it with a bounding box.

Hand Keypoints Anatomical landmarks of the hand — wrist, finger joints, and fingertips — commonly represented as a 21-point skeleton.

Hand Presence Recognition Binary classification task that determines whether a hand is visible in a given frame or cropped image region.

HCI (Human–Computer Interaction) Research area concerned with human–digital system interaction, in which hand tracking is a common input modality.

HPE (Hand Pose Estimation) Computer-vision task that estimates the spatial configuration of a hand from sensory data such as RGB images or RGB-D streams.

Industry 4.0 Industrial paradigm based on automation, connectivity, and cyber-physical systems.

Industry 5.0 Human-centric industrial paradigm that places worker well-being, resilience, and sustainability alongside technological advancement.

Kabsch Algorithm Point-set alignment method that estimates the optimal rigid transformation between corresponding 3D landmarks, used here to align Manus and ZED coordinate frames.

Keypoint Snapping Failure mode in which a hand-pose model places landmarks on nearby objects or edges instead of the true anatomical hand points.

Label Smoothing Regularization technique that softens target class labels during training to prevent overconfident predictions.

Landmark Specific point describing an anatomical or geometric feature, such as a wrist point, fingertip, or finger joint.

MAE (Mean Absolute Error) Regression metric that computes the average absolute difference between predicted and reference values.

Manus Gloves Wearable sensing system that acquires dense 3D hand kinematics, used to provide supervision when vision-based tracking is degraded by gloves or occlusion.

MCP Joint (Metacarpophalangeal Joint) Base joint of a finger where it connects to the palm; one of the anatomical references in the 21-joint hand representation.

MediaPipe Hands Real-time hand-tracking framework evaluated as an off-the-shelf baseline.

Multimodal Dataset Dataset created from more than one sensing modality; here, ZED RGB-D camera data and Manus glove kinematics.

OCRA (Occupational Repetitive Action) Standardised ergonomic index, recommended by ISO 11228-3, for assessing exposure to repetitive upper limb movements over a work shift; requires knowledge of hand action frequency, posture, force, and recovery time.

Occlusion Condition in which a hand or part of a hand is hidden by another body part, a tool, a workpiece, or the workstation environment.

OpenPose Vision-based pose-estimation framework evaluated as an off-the-shelf baseline for hand keypoint detection under industrial conditions.

Out-of-the-Box Model Model applied without fine-tuning on the target industrial dataset, used to quantify baseline reliability limits.

Pairwise Agreement Analysis Evaluation approach that compares predictions from model pairs to identify agreements and disagreements in hand/no-hand decisions.

PPE (Personal Protective Equipment) Protective items worn by operators, such as safety gloves; central to this thesis because gloves significantly affect visual hand tracking.

Recall Proportion of actual positive cases correctly detected by a model.

ResNet-50 Convolutional neural-network backbone fine-tuned on the industrial dataset for classification and regression tasks.

RGB-D Data Data combining RGB color information with per-pixel depth, provided by the ZED-2 camera.

RMSE (Root Mean Squared Error) Regression metric computed as the square root of the average squared error between predictions and reference values.

RULA (Rapid Upper Limb Assessment) Observational ergonomic method for rapidly evaluating upper limb postural loading during work tasks; produces a risk score based on arm, wrist, neck, and trunk posture combined with force and activity type.

SOTA (State of the Art) Best-performing methods currently available for a given research problem or benchmark.

Spatial Alignment Procedure that maps 3D points from one coordinate frame into another, used to express Manus hand joints in the ZED camera frame.

SX and DX Labels distinguishing left (SX) and right (DX) hands in the dataset and experiments.

Temporal Synchronisation Procedure that associates samples from different sensors according to time, compensating for offsets between ZED frames and Manus measurements.

THGs (Technical Hand Grips) Set of semantic grasp classes used to categorize hand-grasp types during industrial manual tasks.

Tool-Driven Occlusion Occlusion caused specifically by tools or workpieces during manipulation; a primary failure mode for generic hand-tracking models in the studied task.

Train/Validation/Test Split Division of the dataset into subsets for learning model parameters, selecting configurations, and evaluating final generalization.

WiLoR Hand-pose estimation model evaluated as an off-the-shelf baseline; it improves recall compared with some alternatives but does not fully resolve industrial reliability issues.

WMSDs Work-related Musculoskeletal Disorders: a group of injuries and disorders affecting the muscles, tendons, ligaments, nerves, and joints caused or worsened by workplace activities.

YOLO Object-detection model family used for hand detection; a YOLOv8s model is fine-tuned on the hand-detection dataset.

ZED-2 Camera Stereo RGB-D camera used to capture industrial assembly recordings, depth information, and body-tracking outputs.

ZED–Manus Alignment Combined temporal and spatial registration procedure that links ZED camera frames with Manus glove kinematics to generate supervisory labels.

List of Figures

List of Tables

1 Introduction

Manufacturing systems are increasingly expected to combine productivity with worker well-being, environmental responsibility, and adaptability to changing operating conditions. This transition is central to Industry 5.0, where technology is not considered an end in itself, but a means to support human-centered, sustainable, and resilient industrial systems. In this context, ergonomic assessment has a strategic role because many assembly and manufacturing activities still depend on repetitive manual operations that expose workers to biomechanical risk, particularly at the hand and wrist level. The scope of this thesis is therefore not limited to the development of an AI model, but concerns the use of robust hand tracking as an enabling tool for manufacturing ergonomics and for the design of safer industrial systems.

1.1 Industry 5.0 and Human-centricity

The existing industrial landscape is undergoing a major paradigm change. It is moving away from the automation-centric principles of Industry 4.0 and towards the more comprehensive and collaborative framework of Industry 5.0. Industry 4.0 was mainly associated with digitalization and the establishment of smart factories using technologies such as IoT and cyber-physical systems. However, its dominant objective was often to maximize efficiency and productivity through automation [1]. Industry 5.0 extends this perspective by placing the human worker at the centre of industrial transformation and by promoting production systems that are technologically advanced, sustainable, resilient, and human-centric [2].

Industry 5.0 is commonly described through three interrelated pillars: human-centricity, sustainability, and resilience [3]. Human-centricity requires production systems that protect workers and adapt to their needs rather than forcing workers to adapt to technology [3, 4]. In this thesis, this pillar is addressed by developing a tool that can support the observation of hand-intensive work and help identify critical actions that may expose operators to ergonomic risk. Sustainability is supported when tasks and workstations are designed to reduce physical overload, prevent work-related injuries, and preserve the long-term health and availability of the workforce [4, 5]. Resilience is strengthened when industrial systems can rely on objective information about human activity in real production settings, allowing risks to be detected earlier and work processes to be redesigned more effectively [4,

5]. From this perspective, robust hand tracking contributes to all three pillars by providing reliable information on how manual work is actually performed during industrial tasks [5].

Ergonomics [1], also known as human factors, is the scientific study of how people interact with their work environment. It aims to redesign tasks, tools, and spaces to fit the worker, rather than forcing the worker to adapt to them, thereby enhancing well-being, reducing the risk of injuries, and preventing work-related illnesses, such as WMSDs. Ergonomics is inherently a multidisciplinary field and includes knowledge from medicine, physiology, biomechanics, psychology, and sociology [6], to name a few.

The adoption of human-centered work design is therefore closely connected to sustainability and ergonomics, both of which are essential for the long-term viability of modern industrial enterprises. In this sense, protecting the health, safety, and well-being of the workforce is not a secondary requirement or a regulatory burden, but a core design objective. A production system that causes physical harm to workers because of poor ergonomic design is, by nature, unsustainable [7]. Hence, within the Industry 5.0 vision, advanced tools for the proactive assessment of ergonomic risk are necessary to support designers and engineers in creating safer and more sustainable industrial systems.

1.2 Motivation and Problem Statement: The Challenge of Ergonomic Assessment in Industry

Advanced technologies such as sensors, AI, Computer Vision (CV), and data analytics can support worker well-being by making ergonomic assessment more objective and continuous. However, their value for this thesis must be understood from the perspective of manufacturing ergonomics rather than as a purely computational problem. A major issue in the manufacturing and assembly sector is the high incidence of Work-related Musculoskeletal Disorders (WMSDs), particularly those affecting the hand and wrist as a consequence of repetitive, forceful, or awkward manual tasks [8]. According to Global Burden of Disease 2019 [2], in Italy, WMSDs are the most prevalent occupational disease, with an increase of 88% reported between 2010 and 2017, and are estimated to account for 50% of work absences and 60% of

[1] `https://iea.cc/about/what-is-ergonomics/`

[2] `https://ergosante.fr/it/chiffre-accidentologie-tms-italie/`

permanent work-related disabilities. The incidence is higher among men than among women, and certain occupational groups are particularly affected. Therefore, ergonomic assessment is not only a procedural requirement, but also a basis for safe, sustainable, and human-centered production.

Despite the relevance of ergonomics in manufacturing, industrial risk assessment is still often carried out manually through observational methods, especially in hand-intensive and repetitive work [9]. In many cases, experts observe the worker, identify critical actions or postures, and then apply established ergonomic methods to estimate risk [9, 4]. This approach remains essential in practice, but it is time-consuming and depends strongly on expert interpretation, particularly when the task involves fast, repetitive, or fine hand actions [9].

Only a limited number of digital tools have been proposed to support this process in industrial settings, and many of them primarily focus on whole-body posture and general ergonomic monitoring rather than detailed hand activity [4]. This is a major limitation for assembly work, where grasping, manipulation, tool use, and repetitive finger movements are highly relevant to biomechanical risk assessment [9].

Sensorized gloves can capture hand motion accurately and are useful for gesture analysis, but wearable solutions may be intrusive, alter natural interaction with tools, and are often more suitable for laboratory studies than for routine use in real industrial workplaces [10]. For this reason, vision-based Hand Pose Estimation (HPE) is an attractive alternative, since it can potentially observe hand motion without constraining the worker.

Advanced HPE models such as MediaPipe Hands [11] and OpenPose [12] have demonstrated the ability to detect and track hand keypoints from RGB images or videos [13]. In principle, this capability could support continuous, objective, and scalable analysis of hand-intensive industrial tasks. Nevertheless, there is a significant gap between the performance of these models in controlled or consumer-oriented scenarios and their reliability in real manufacturing environments. Existing models are generally not trained on industrial data that include safety gloves, tools, workpieces, assembly postures, variable lighting, and frequent occlusions. As a result, their outputs can become unreliable precisely in the conditions that are most relevant for ergonomic assessment. This thesis formulates the core problem as follows:
State-of-the-art vision-based HPE models achieve strong results in controlled or consumer-oriented settings, but in their current form they are not sufficiently robust for industrial ergonomic analysis,

especially when safety gloves, tools, and persistent occlusions alter the visual appearance of the hand. This technological gap limits the use of vision-based hand analysis as a practical support tool for manufacturing ergonomics and for the design of human-centered industrial systems.

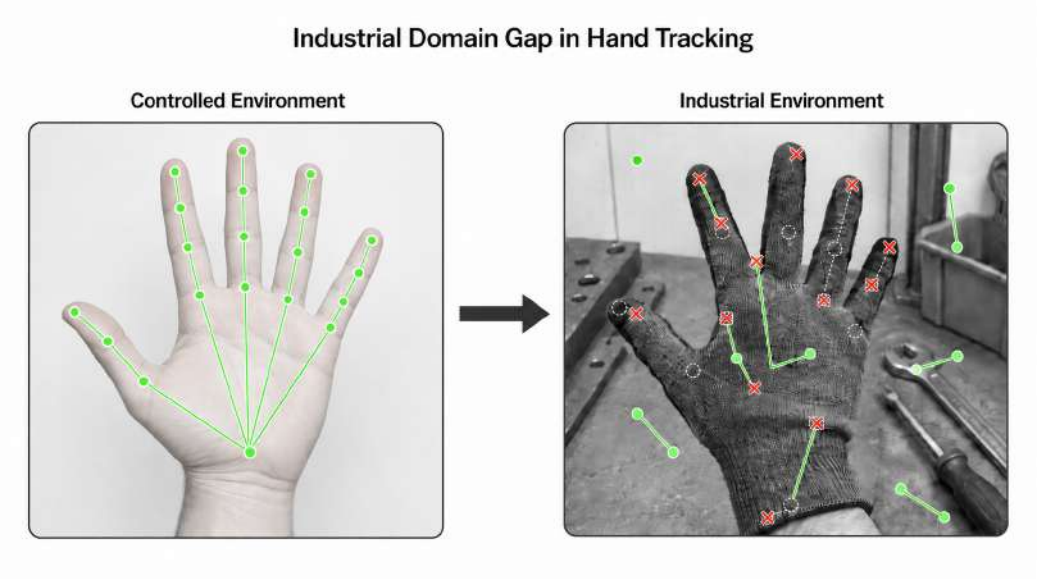


Figure 1: Industrial domain gap in hand tracking: a generic HPE model correctly estimates keypoints on a bare hand in a controlled environment (left), but fails to locate landmarks on a gloved hand in an industrial workstation (right), where red crosses indicate missed or misplaced keypoints.

1.3 Technical Challenges in Vision-Based Ergonomic Assessment under Industrial Constraints

The limitations of current HPE models in industrial contexts are not arbitrary; they stem from a fundamental mismatch between their training data and their target operational environment. To bridge the application gap identified in the previous section, it is essential to move beyond the direct use of existing models and develop a robust solution specifically engineered for industrial realities. The path toward this solution begins with understanding the core technical failure points.

The challenge to apply existing HPE models for industrial ergonomics purposes is twofold:

1. **Invalidated Visual Priors due to Gloves**: State-of-the-art HPE models are, by design, sophisticated feature detectors. Their deep neural networks have been trained on large image collections in which the hand is often visible as a bare hand, and they learn to identify keypoints by relying on visual priors such as skin texture, shadows that define knuckles, anatomical contours, and phalanx creases [14]. In industrial environments, however, manual tasks are rarely performed with bare hands. Depending on the activity and safety requirements, operators may use cut-resistant gloves, thermal gloves, mechanical protection gloves, or thick safety gloves. Each glove type can change the apparent shape, texture, colour, thickness, and visibility of the hand. A standard industrial safety glove can therefore act as a visual filter that hides the information the network was trained to recognize, leading to unstable or incorrect keypoint predictions.

2. **Data Incompleteness from Tool Occlusions**: The second critical challenge is the frequent and unavoidable partial occlusion of the hand. During industrial activities, occlusion is not a fixed visual disturbance but a task-dependent and dynamic condition. Operators continuously grasp and release parts, switch grip types, manipulate tools, and perform quick or varied movements. As a result, the hand can be repeatedly hidden by tools, workpieces, fixtures, or the other hand. When faced with incomplete visual data, HPE models may fail to predict the occluded keypoints or may infer physically implausible positions from partial evidence. For quantitative ergonomic analysis, such unreliable output can lead to misleading conclusions about the posture and movement of the worker.

Therefore, overcoming these limitations requires more than the direct application of general-purpose HPE models. It requires an industry-oriented approach aimed at creating a robust model for hand tracking under the visual constraints of manufacturing tasks. Such a model must be designed to tolerate the absence of skin-based visual features and the frequent loss of information caused by task-dependent occlusions. This thesis proposes a direct path to this objective by developing a methodology to generate high-quality, specialized training data for industrial hand analysis.

1.4 Objectives and Thesis Outline

To address the critical gap between the potential of AI in ergonomics and its practical limitations in industry, this thesis proposes a systematic, industry-oriented approach. The research is not merely about applying an existing computer vision tool, but about developing and validating a solution that can support ergonomic task analysis in realistic manufacturing conditions. The principal aim of this research is:

Creating and validating a robust Hand Pose Estimation model to support task analysis during industrial tasks, considering also the use of safety gloves and tool-occlusion issues.

This aim is guided by the following research question:

> *Can a vision-based model, trained on industrial data, provide sufficiently reliable hand-motion information to serve as input for ergonomic risk assessment in real assembly environments — even when operators wear safety gloves and experience frequent tool-driven occlusions?*

To break down the goal into smaller parts, the research will first identify and then pursue specific objectives.

1. **Measure the failure of existing models**: Establish a performance baseline by experimentally evaluating state-of-the-art models such as MediaPipe, OpenPose, and WiLoR in simulated industrial scenarios.

2. **Design a multi-modal data acquisition protocol**: Create a system that synchronizes kinematic ground-truth data from a Manus sensor glove with visual RGB-D data from a ZED stereo camera.

3. **Develop a spatio-temporal fusion algorithm**: Implement and validate a procedure, using the Kabsch algorithm, for aligning the data from the two sensor systems into a single coordinate frame.

4. **Generate a specialized industrial dataset**: Produce a high-quality, accurately annotated dataset of hand movements during an assembly task, including both bare and gloved hands.

5. **Train and evaluate a robust HPE model**: Use the newly created dataset to train a specialized model and evaluate its reliability against established baselines.

These objectives are not pursued as ends in themselves. Established ergonomic assessment methods, such as OCRA and HAL, require continuous and reliable hand-motion data as their quantitative input. In real industrial settings, however, no current off-the-shelf model can provide such data dependably, particularly when safety gloves and tool-driven occlusions are present. The work of this thesis therefore addresses the step that must precede any automated ergonomic evaluation: making hand tracking sufficiently robust to be trusted in assembly environments.

The main contribution of this thesis is an industry-oriented pipeline for manufacturing ergonomics, designed to enable more reliable hand analysis during assembly tasks and to support the design of safer and more human-centered industrial systems.

This contribution includes a dedicated data-acquisition protocol, a new industrial dataset, and a vision-based model tailored to the constraints of real assembly environments. In this way, the thesis positions HPE as a support technology for ergonomic assessment and industrial system design, rather than as an isolated computer vision problem. The remainder of this thesis is organized as follows:

- Chapter 2 provides a comprehensive review of the relevant literature, covering state-of-the-art Hand Pose Estimation techniques, existing datasets, and current practices in ergonomic risk assessment.

- Chapter 3 details the methodology of the research. It describes the experimental setup, the hardware used (Manus glove, ZED camera), the data acquisition protocol, the spatio-temporal alignment algorithm, and the architecture of the proposed deep learning model.

- Chapter 4 presents the experimental results, including the performance evaluation of baseline models and the results achieved by the newly trained robust model.

- Chapter 5 discusses the findings from the perspective of manufacturing ergonomics, industrial applicability, and the limitations of the proposed approach.

- Chapter 6 concludes the thesis with a summary of the main outcomes and suggestions for promising avenues for future work.

2 State of the Art

2.1 Introduction

Hand Pose Estimation (HPE), referring to the task of identifying hand joints or keypoints in images and videos, has become an important enabling technology for analysing manual activity. Although it is usually introduced as a computer-vision problem, its relevance in this thesis is mainly related to manufacturing ergonomics: reliable information about hand posture and movement can support the assessment of repetitive, forceful, or awkward manual tasks. In the transition from Industry 4.0 to the human-centric vision of Industry 5.0 [15], this type of information is valuable because it can help designers and engineers identify ergonomic criticalities and reduce the risk of work-related musculoskeletal disorders [16].

The field of HPE has been strongly advanced by deep learning and large multimodal datasets; recent models based on Convolutional Neural Networks (CNNs) [17] and Transformer architectures [18] have reached high levels of accuracy for both 2D and 3D hand tracking [19]. However, this accuracy is usually demonstrated on benchmark datasets and controlled or consumer-oriented scenarios. It cannot be assumed that the same level of reliability is automatically transferred to industrial assembly tasks, where hands are often gloved, partially hidden by tools, and observed under non-ideal lighting.

Despite this technical and theoretical progress, there is still a gap between the performance of HPE models under controlled laboratory conditions and their dependability in practical industrial settings. Real manufacturing environments can adversely affect HPE performance because of both their intrinsic characteristics and the non-specificity of the training datasets. The most relevant factors are:

- the use of thick safety gloves, mandatory by law (D.Lgs. 81/08 [3]), which hide the skin texture and change the hand geometry;
- the presence of occlusions caused by tools, workpieces, or machinery;
- variable or inadequate illumination [20].

This chapter reviews the State of the Art (SOTA) regarding HPE architectures and related datasets. First, it highlights the main techniques and deep

[3] https://biblus.acca.it/download/testo-unico-sicurezza-dlgs-81-2008/

learning architectures used in this domain; then, it reviews popular datasets and discusses their shortcomings for ergonomic evaluation in industrial contexts. Finally, it identifies the research gap addressed by this thesis: the need for a robust, industry-specific HPE pipeline that can support hand-focused task analysis in realistic manufacturing conditions.

A critique of representative SOTA models, including MediaPipe Hands and OpenPose, is presented in Sections 2.5.1 and 2.5.2. Most importantly, the chapter does not treat these models only as abstract computer-vision methods, but evaluates whether their assumptions are compatible with manufacturing ergonomics, where the relevant question is whether the output is stable enough to support task analysis and workstation design.

2.2 Ergonomic Assessment of Manual Work: Methods and Limitations

Work-related musculoskeletal disorders (WMSDs) affecting the hand and upper limb remain among the most prevalent occupational health problems in manufacturing and assembly [8]. They arise from prolonged exposure to repetitive movements, forceful exertions, awkward postures, and contact stress — conditions that are intrinsic to many industrial assembly tasks. Identifying and quantifying this exposure before it causes injury is therefore a primary objective of industrial ergonomics.

To support this objective, several standardised observational methods have been developed and validated for use in occupational settings. The Rapid Upper Limb Assessment (RULA) [21] provides a rapid postural scoring system for evaluating upper limb loading during work tasks. The Occupational Repetitive Action (OCRA) index [22] extends this to the quantification of repetitive upper limb exposure over a work shift, and is the method recommended by the ISO 11228-3 standard for repetitive handling tasks[4]. The Hand Activity Level (HAL), a complementary measure specifically targeting hand-intensive repetitive work, is used alongside normalised peak force to define threshold limit values for the upper limb [9].

Despite their widespread use and regulatory recognition, these methods share a fundamental limitation: they rely on manual observation by a trained expert who watches the worker — live or from video — and assigns scores based on judgement [9]. This process is time-consuming, difficult to scale

[4] `https://www.iso.org/standard/76522.html`

across multiple workstations or shifts, and subject to inter-observer variability, particularly when the task involves fast or fine finger movements that are difficult to observe directly. Recent work has demonstrated that continuous, objective motion data can reduce this variability and extend ergonomic monitoring to longer, uninterrupted task sequences [5, 23]. Vision-based Hand Pose Estimation is therefore of direct interest to manufacturing ergonomics: if it can provide reliable, continuous information about hand posture and movement in real industrial conditions, it can serve as the quantitative data source that current observational methods lack.

2.3 Overview of Hand Pose Estimation

HPE can be essentially understood as a computer vision task that aims at deducing the spatial layout of a human hand from the given sensory data. One of the main goals in most hand pose estimation systems is the accurate identification of a set of anatomical landmarks, or keypoints, that reflect the joints (e.g., wrist, knuckles) and fingertips. Usually, this skeletal representation comes with 21 keypoints that serve as a minimal representation of the hand's articulation [24, 25]. The problem domain of HPE is complex, as its object of study, the hand, is perhaps one of the most complex anatomical structures in the human body: the usual human hand model contains around 20 to 26 degrees of freedom (DoF) for the fingers alone; when wrist and palm motions are included, some models report up to 31 DoF [24], which results in a very large, high-dimensional state space of possible poses [26].

HPE techniques can be divided into various groups based on the sources of their input and forms of their output; in fact the source of the information highly determines the approach and the level of precision that can be reached. The most typical sources are:

- monocular RGB images that can be found everywhere and are quite cheap but have the problem of depth ambiguity;
- depth maps (created with technologies like Time-of-Flight or structured light) that provide clear geometric information but no texture;
- stereo or RGB-D data that merge the visual appearance with the geometric depth [27, 28].

In terms of output:

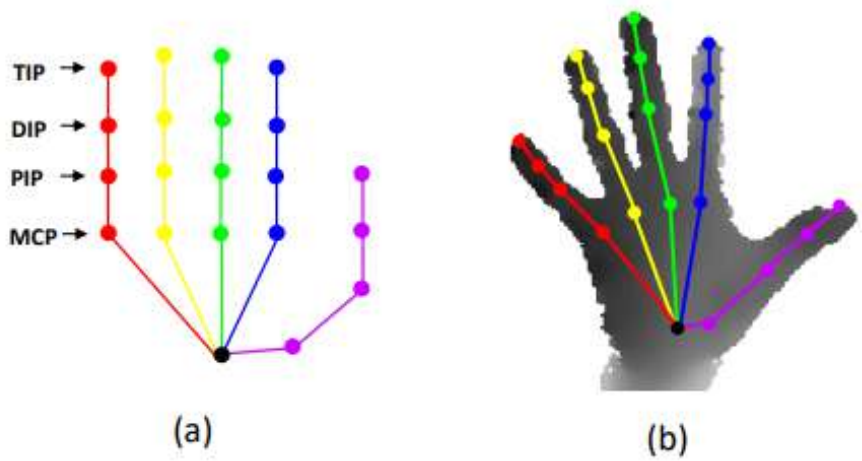


Figure 2: Hand model definition. (a) Our hand model has 21 joints and moves with 31 degrees of freedom (dof). (b) Model fitted to hand shape [24].

- systems that might infer 2D locations on the image plane;
- system that might infer 3D locations in the metric camera or world space.

Although 2D pose is enough for some screen-based interactions, accurate 3D estimation is indispensable for the applications that require real-world spatial understanding, for instance, robotic grasping or ergonomic analysis in physical environments [29].

No matter what the modality is, HPE is always under pressure to solve its significant intrinsic problems. Self-occlusion is a typical and intensive problem; as the hand moves, fingers naturally cover one another, thus making very difficult the localization of the joints that are occluded [30]. Moreover, the hands differ significantly in the shape, size, and skin color of the individuals. Furthermore, fast hand movements cause motion blur, and changes in lighting and viewpoints make the problem of detection and tracking even more difficult. In terms of methodology, to handle these problems, the community has moved from model-based (generative) to data-driven(discriminative) landmark approaches. Traditional model-based methods depended on intricate 3D kinematic hand models, which in a recurrent manner, optimized parameters that brought the projected model in line with the observed image features such

as edges or silhouettes [31]. Although being logically valid, the methods in question were frequently criticized for their performance issues and sensitivity to the start position. The current HPE is largely influenced by the data-driven deep learning methods, mainly CNNs. These models achieve direct, complex transformations from unprocessed input data to pose coordinates by means of a great volume of annotated training data, and thus they are much more robust and accurate in varying conditions than the older ones [32, 33].

2.4 Existing Datasets for Hand Pose Estimation

The effectiveness and stability of purely data-driven HPE models depend heavily on the quality, variation, and quantity of the data contained in the datasets used in training these models. The change in HPE benchmarks reflects the development of the field from tightly controlled lab conditions to "in-the-wild" scenarios. Many different public datasets have been made available to cover various facets of the hand pose challenge and differ in terms of input modality (RGB, Depth, Egocentric), annotation type (2D vs. 3D), and scenario complexity [34].

The first datasets created were mainly focused on depth data that had been obtained in controlled environments to solve the problems of segmentation and lighting that are the main issues in RGB images. The NYU Hand Pose Dataset [35] recorded by Microsoft Kinect was the first step that provided the aligned RGB and depth frames from multiple views, however, most of the time it was used for single-view depth-based evaluation. The ICVL Dataset [36] also helped in this manner by providing a large number of depth frames with accurate annotations. This research was mainly focused on the different orientations of the hand; to do this the backgrounds used were removed and the hands were isolated. To overcome the problem related to the data requirements of deeper neural network and accurate 3D estimation, the next work went on to exponentially increase the scale of the dataset. They achieved considerable progress when BigHand2.2M [24] was released; it provided over 2.2 million depth maps with accurate 6D magnetic tracking annotations that depict the wrist and the finger articulations in the expansive range.
Turning their attention to the commonly-used RGB cameras, FreiHand [37] set the standard for monocular 3D HPE. It employs a multi-camera system to produce high-quality 3D annotations for RGB images of hands taken against

different nature backgrounds, some with a few handheld objects.

As a matter of fact, real-world hand movements are rarely without the support of other limbs or the interplay with the environment, therefore, recently the main focus of the researchers has been on the intricacies and complexities of this interaction. By providing a rigorous dataset of interacting two-hand scenarios with considerable inter-hand occlusion, InterHand2.6M [38] solves the toughest challenge of close interaction between two hands. Additionally, datasets like HO-3D [39] that emphasize hand-object interaction only, deliver 3D poses not only for the hand but also for the manipulated object, thus bringing the benchmarks to the level of realistic tasks.

Even though there are enough resources to use, a thorough investigation of them reveals that they have a long way to go to get industrial applications. In fact, most of the well-known public datasets described in the literature mainly contain bare hands. Furthermore, the datasets that explicitly mention the gloved hands, a safety necessity in industrial sectors, are few and far between in the public domain. Besides that, only a few of those datasets allow for object manipulation, and in those datasets, the occlusions caused by large tools and complex machinery are rarely severe and sustained, and also the specific harsh lighting conditions that can be found on actual manufacturing assembly lines are not considered. Hence, the current models that achieve top performance only on these existing benchmarks usually fail to generalize well to industrial environments with specific domain constraints.

2.5 State-of-the-Art AI Models for Hand Tracking

Deep learning methods mainly dominate the current scenario of HPE, and they have already outperformed traditional methods by using large-scale datasets and powerful computational architectures [19]. Present-day SOTA models largely rely on end-to-end trainable deep neural networks, mainly CNNs, and more recently, Vision Transformers, to directly predict hand keypoint locations from input images. Based on how they handle the complexity of pose estimation, these methods can be considered as classes of strategies.
In fact, the fundamental difference of approach is:

- **Top-Down Methods**: these detection-based methods first isolate the hand area with a bounding box detector and then determining the pose

in that cropped region;

- **Bottom-Up Methods**: these regression-based methods locate all possible keypoints in an image and then link them to generate consistent hand skeletal structures, which are usually also capable of multiple hand simultaneous processing [40, 41].

Besides that, these leading models lie far apart in their architectural designs, their orientation towards either 2D or 3D pose estimation, their computational power requirements for real-time usage scenarios on different hardware, and their ability to resist the difficult environmental factors already mentioned. The subsequent subsections review three representative pose estimation frameworks evaluated in this thesis:

- MediaPipe 2.5.1,
- OpenPose 2.5.2,
- WiLoR 2.5.3.

In addition, Section 2.5.4 covers YOLO, which is not a pose estimation model but a real-time object detector used in this thesis as the hand-detection stage that precedes keypoint regression.

2.5.1 MediaPipe

Developed by Google, MediaPipe Hands[5] is arguably the most prominent and widely adopted framework for real-time hand tracking, particularly renowned for its efficiency on resource-constrained edge devices like mobile phones and CPUs [42]. It operates as a lightweight, top-down pipeline comprising two distinct models working in tandem. First, a palm detector model (BlazePalm), optimized for speed, operates on the full image to localize hands and generate bounding boxes. Second, a hand landmark model receives the cropped image region defined by the bounding box and regresses the precise 3D coordinates of 21 predefined skeletal keypoints within that region.

A key design strategy of MediaPipe is its reliance on the initial palm detector; the computationally heavier landmark model runs only when a palm is successfully detected. This ensures high frame rates in scenarios where

[5] `https://google.github.io/mediapipe/`

hands are absent. The model was trained on a large, diversified dataset containing roughly 30K real-world hands images, augmented with a high-quality synthetic dataset to improve generalization across varied rotation, scaling, and lighting conditions.

Despite its impressive performance in general-purpose applications like gesture control and AR effects, MediaPipe Hands exhibits significant limitations when translated to rigorous industrial environments. Its architecture is heavily dependent on recognizing distinct visual features of the bare hand, such as skin tone gradients and typical finger geometry, for both initial palm detection and subsequent landmark regression. Consequently, as indicated by previous studies and our preliminary analysis, its tracking robustness degrades sharply under severe occlusions or when the hand's visual appearance is fundamentally altered by safety gloves, leading to frequent tracking loss and jittery keypoint predictions in such scenarios.

2.5.2 OpenPose

OpenPose[6] developed at Carnegie Mellon University's Perceptual Computing Lab is a landmark contribution to the field, being among the first real-time multi-person systems to jointly detect human body, foot, face, and hand keypoints on single images [43]. While MediaPipe follows a top-down approach, OpenPose goes bottom-up. In other words, it initially finds all anatomical keypoints for all people in the image at once without a person detector. After that, it figures out the associations between these keypoints to build the skeleton of each person. Without a doubt, the main idea that enables this bottom-up method to work is the invention of Part Affinity Fields (PAFs): PAFs are 2D vector fields that indicate both the position and the direction of the connection between two body parts (e.g., the connection between a wrist and an elbow). Along with the confidence maps for keypoint locations, the network also predicts these affinity fields. OpenPose is capable of matching body parts even in complex scenes where there are multiple overlapping people by simply utilizing the directional information from the PAFs to sort the detected keypoints. As for the hand tracking, the model normally predicts 21 keypoints per hand similar to other standard models.

[6] `https://github.com/CMU-Perceptual-Computing-Lab/openpose`

On the one hand, OpenPose is praised for its extreme accuracy in multi-person situations and its capability of handling the partial occlusion that is more difficult in previous methods, while on the other hand, it has some limitations for certain industrial applications. Though the bottom-up architecture of the OpenPose is effective, it is much more computationally demanding than the top-down methods like MediaPipe; thus, the real-time performance usually requires substantial GPU power, which could be a problem for embedded industrial systems. Besides that, just like other vision-based models, which are trained mainly on datasets of bare hands, its performance will decline when it is confronted with the situation of severe, long-term occlusion of the hand by a tool and the visual ambiguity caused by thick gloves in the manufacturing environment.

2.5.3 WiLoR

WiLoR which is an acronym for "Wild Local Refinement" is a more recent, advanced concept to make 3D hand pose estimation more precise and robust especially for unconstrained, "in-the-wild" cases. WiLoR, a method by Potamias et al. [44], mitigates the shortcomings of a typical top-down pipeline that usually loses track of the exact localization if the initial bounding box is wrong or if the hand is partially occluded. The main breakthrough of WiLoR is its iterative refinement strategy.

Rather than producing a single final 3D hand pose estimate from a cropped image, WiLoR utilizes a feedback loop. The loop results in a first rough estimation of hand pose and shape (in many cases a MANO mesh is used). The very important thing here is that the method projects the initial 3D estimate back to the 2D image plane to find the differences. The network then pinpoints the very small, localized regions where the registration is bad: in other words, it is "zooming in" on trouble spots, so to refine the pose estimation in the next iteration. By doing this iteratively, the model can gradually fix the errors it made before and reach higher accuracy, even if the hand articulations are complicated or self-occluded.

WiLoR, however, is not without its weaknesses from the point of view of industrial safety even though it shows very impressive results at cutting-edge benchmarks from the public domain (e.g., FreiHand, and HO-3D). This is because the method depends on the availability of visual cues for the iterative

refinement of the hand pose. The glove-wearing hand with the glove without texture and the hand that is heavily obscured by a tool are good examples where the initial coarse estimation might be extremely erroneous, and the subsequent local refinement steps may not end at the correct pose as the visual evidence for correction is absent or misleading.

2.5.4 YOLO as Hand Detector

Unlike the pose estimation frameworks discussed above (MediaPipe, OpenPose, WiLoR), YOLO (You Only Look Once) is a real-time object detector, not a keypoint regression model. In the pipeline proposed in this thesis, a fine-tuned YOLOv8 model serves as the hand-detection stage: it localises the hand in the frame and produces a bounding box, which is then passed to a separate model for joint-coordinate regression. YOLO is reviewed here because its detection reliability under industrial conditions directly determines whether the downstream pose estimation step can run at all. YOLO revolutionized object detection by framing it as a single regression problem, straight from image pixels to bounding box coordinates and class probabilities, rather than a complex multi-stage pipeline [45].

The architecture of YOLO splits the input image into a grid and predicts bounding boxes and probabilities for each grid cell simultaneously. This unified architecture enables extremely fast inference speeds, making it highly attractive for real-time industrial monitoring applications where latency is a concern. Various iterations of YOLO (v3, v5, v8, etc.) have consistently improved accuracy while maintaining this speed advantage.

However, the efficacy of YOLO in our specific industrial context is heavily dependent on the training data. Standard off-the-shelf YOLO models are typically pre-trained on large-scale general datasets like COCO, where the "hand" or "person" class predominantly features visible skin textures and typical anatomical shapes. Consequently, when applied to industrial scenarios, the model faces a significant domain shift. As demonstrated by our preliminary experiments (detailed in Section 2.6), the visual abstraction caused by safety gloves—which mask skin color and alter the hand's silhouette—often leads the model to completely miss the presence of the hand. This results in a high rate of False Negatives, rendering the subsequent pose estimation step impossible because the system fails to "see" the hand in the first place.

2.6 Identifying the Research Gap

The review presented in Sections 2.3 to Section 2.5 shows that current Hand Pose Estimation (HPE) methods have reached impressive performance on standard benchmarks and controlled scenarios. Datasets such as BigHand2.2M, FreiHAND, InterHand2.6M and HO-3D provide millions of annotated frames and have enabled highly accurate 2D/3D pose estimation for bare hands interacting with simple objects or in egocentric views. However, these datasets rarely include safety gloves, heavy tools or cluttered industrial backgrounds, which are typical of real manufacturing environments. As a consequence, state-of-the-art models like MediaPipe Hands, WiLoR, and YOLO-based detectors are inherently biased towards the visual statistics of bare hands and "clean" scenes.

At the same time, the ergonomic assessment tasks considered in Industry 4.0 and 5.0 applications require exactly the opposite setting: operators wear thick gloves, frequently occlude their hands with tools and components, and work under non-ideal illumination conditions. This mismatch between training data and operational context suggests a serious domain gap, but the literature provides only limited quantitative evidence of how severe this problem becomes in real assembly tasks. The core aim of this section is therefore to make this gap explicit, both conceptually and empirically, and to motivate the need for the new data-collection and modelling strategy developed in this thesis.

2.6.1 Formulation of the research gap

Bringing together the findings of the literature review, the research gap that motivates this thesis can now be stated more precisely. An empirical confirmation of these gaps is provided in Chapter 4, where the benchmark results of off-the-shelf models under industrial conditions — including pairwise agreement matrices, recall and FNR metrics, and qualitative failure modes — are reported and discussed.

- **Dataset gap.** Existing public datasets offer large quantities of labelled hand poses, but they almost exclusively feature bare hands, mild occlusions and non-industrial backgrounds. They do not represent

workers wearing safety gloves, handling tools and interacting with complex machinery under realistic factory lighting. No dataset was found that jointly provides RGB/RGB-D images and high-precision kinematic ground truth for gloved hands in industrial assembly.

- **Model gap.** Current state-of-the-art methods—MediaPipe, OpenPose, WiLoR, YOLO and related architectures—are trained and optimised on these biased datasets. Their internal visual priors rely heavily on skin appearance and clear contour information [42, 43, 44]. As the benchmark later reported in Chapter 4 confirms, once these priors are invalidated by gloves and occlusions, the models exhibit high false-negative rates and inconsistent behaviour across methods. None of the surveyed approaches exploits multi-modal sensing to compensate for missing visual information.

- **Evaluation gap.** In the ergonomics literature, there is still no agreed-upon protocol for evaluating HPE systems in realistic industrial scenarios, where the relevant metric is not only pixel-level accuracy on a benchmark image but the stability and continuity of tracking over long task sequences. The simple hand/no-hand benchmark introduced here already suggests that off-the-shelf models are not robust enough even at this coarse level, let alone for precise joint-angle estimation.

These gaps point towards the need for a dedicated, industry-specific pipeline that tackles the problem from the data level upwards. The approach developed in the following chapters addresses this need by:

1. designing a multi-modal acquisition protocol that combines a Manus sensor glove with a ZED stereo camera to generate reliable ground truth for gloved hands under occlusion;

2. constructing a new industrial dataset based on this protocol; and

3. training and validating a robust deep learning model tailored to these conditions, with the explicit goal of enabling continuous ergonomic assessment in real assembly tasks.

In other words, the contribution of this thesis is not only to show that existing models fail in industrial contexts, but to propose a concrete, reproducible route towards overcoming this limitation and closing the gap between academic HPE research and human-centred manufacturing practice.

3 Materials and Methods

Chapter Overview

This chapter explains how the research was carried out, starting from the industrial assembly task and moving step by step toward the creation of the dataset and the development of the proposed AI pipeline. The methodological path is organized to keep the industrial use case, the physical experimental setup, the software architecture, and the user testing protocol clearly separated. First, the Bonfiglioli assembly task is introduced as a representative manufacturing use case; then, the workstation, cameras, hardware, and protective equipment are described as part of the experimental setup. The chapter then presents the data acquisition and processing architecture, the testing procedure with users, the preprocessing and ground-truth generation steps, the dataset creation strategy, and the proposed learning pipeline. In this way, the chapter connects the practical reality of manual industrial work with the technical requirements needed to train and evaluate a robust hand-tracking model for ergonomic analysis.

3.1 Industrial Use Case

The industrial use case selected for this thesis is a manual assembly task based on a Bonfiglioli gearbox reducer. This task was chosen because it reflects several conditions that are common in real industrial work, including repetitive hand movements, tool use, bimanual coordination, and partial occlusions caused by components, tools, and the operator's own hands. At the same time, it remains controlled and repeatable, making it suitable for comparing different recordings and evaluating the robustness of hand-tracking methods for ergonomic analysis.

3.1.1 Bonfiglioli Reducer Assembly Task

The industrial case study is based on the manual assembly of a small Bonfiglioli gearbox reducer, using a didactic kit derived from a real industrial product (Figure 3). The kit includes the main housing, shafts and gears, bearings, spacers, a cover, screws, and small auxiliary components arranged on a dedicated base. This configuration makes it possible to repeat the same task across different sessions and operators without damaging the parts, while still

preserving the mechanical logic of a real assembly activity.

At the beginning of each trial, all components are restored to their initial positions and the reducer is returned to a fully disassembled state. The operator is then asked to complete one full assembly cycle by following a predefined sequence of actions. Before recording starts, this sequence is illustrated through slides and short demonstrative videos. In practical terms, the procedure includes orienting the housing, inserting shafts and gears, positioning bearings and spacers, closing the housing with the cover, and tightening the screws with a hand tool such as a screwdriver or a hex key. The cycle can be repeated several times within the same session, which makes it possible to collect multiple executions of the same task under comparable conditions.

From both an ergonomic and a computer-vision perspective, this assembly task is particularly suitable for the aims of the thesis. It combines fine manipulations with larger upper-limb motions, requires continuous bimanual coordination, and naturally generates a wide range of grasp types, finger configurations, and wrist postures. At the same time, the repeated use of tools and the interaction with mechanical parts create frequent self-occlusions and object-driven occlusions, which are exactly the visual difficulties that challenge existing hand-tracking systems in industrial settings.

Throughout the trials, operators wear certified industrial safety gloves. This choice is important for two reasons: first, it reflects real manufacturing practice, where hand protection is part of the normal working condition rather than an optional add-on. Second, it deliberately introduces the domain shift that motivates this thesis: once the hand is covered, skin color, texture, and several anatomical cues are no longer directly visible, which makes the task a meaningful stress test for robust, industry-oriented AI models.

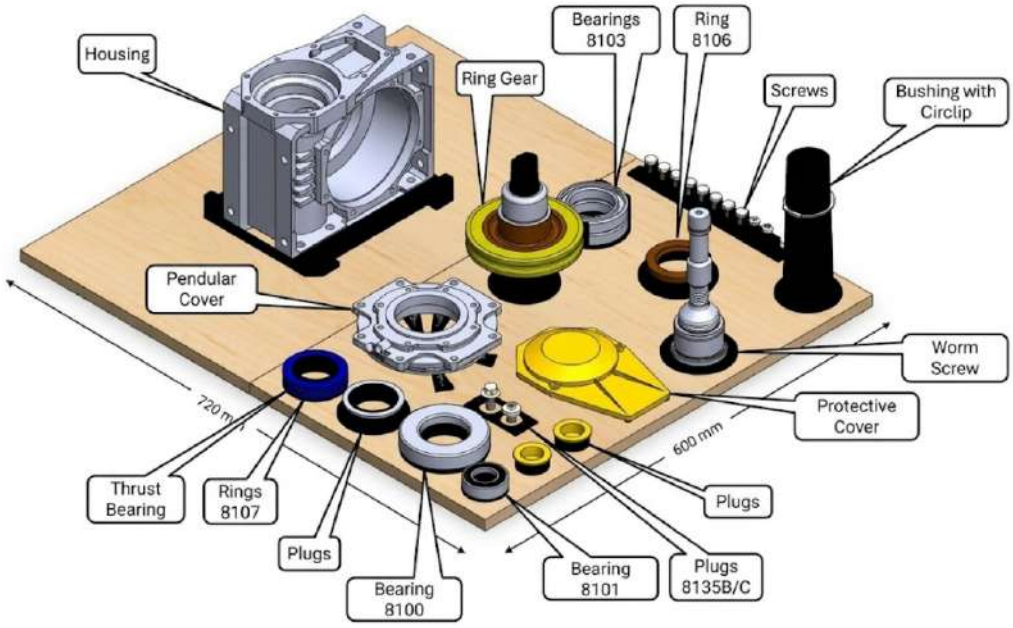


Figure 3: Bonfiglioli gearbox assembly kit.

3.1.2 Task Characteristics Relevant to Hand Analysis and Ergonomics

Beyond its mechanical simplicity, the Bonfiglioli task is useful because it concentrates several conditions that matter for ergonomic analysis at the hand and wrist level. During the assembly cycle, the operator alternates between precise actions on small parts and more forceful actions involving tool use and object stabilization. As a result, the task does not generate a single stereotyped posture; instead, it produces a sequence of changing grips, wrist orientations, and coordination patterns between the two hands.

This variability is important from an ergonomic standpoint: in assembly work, risk does not emerge only from one isolated posture, but from the repetition of hand-intensive actions over time, often combined with awkward wrist positions, sustained grasping, and short but frequent exertions. The selected task reproduces these conditions in a controlled way. It therefore offers a good compromise between experimental repeatability and ecological validity: the action sequence is stable enough to compare recordings across trials, but rich enough to expose the visual and biomechanical challenges that

make industrial hand monitoring difficult.

A second aspect is the visibility of the hands. The operator repeatedly grasps tools, holds components close to the workbench, and temporarily covers one hand with the other while aligning parts. These moments are not occasional exceptions; they are part of the normal execution of the task. For this reason, the assembly cycle is well suited to studying not only whether a hand is present in the scene, but also how reliably a vision-based system can continue to track it when gloves, clutter, and occlusions are all present at the same time.

3.2 Experimental Setup and Hardware Configuration

The experimental setup was designed to reproduce, as closely as possible, a realistic hand-intensive assembly scenario while maintaining sufficient control for systematic data collection. Unlike the industrial use case described in Section 3.1, which defines the product, task, components, and assembly logic, this section focuses on the physical recording conditions: workstation layout, camera placement, sensing hardware, and protective equipment. The Bonfiglioli reducer kit described in Section 3.1.1 was placed on a standard workstation, and the sensing equipment was arranged so that both the operator and the manipulated objects could be observed from stable and complementary viewpoints.

Figure 4 provides an overview of the laboratory workstation used for the recordings. The layout was chosen to keep the task natural for the operator while ensuring that the acquisition system could capture the main visual elements needed for subsequent analysis: both hands, the workbench, the tools and the reducer components.

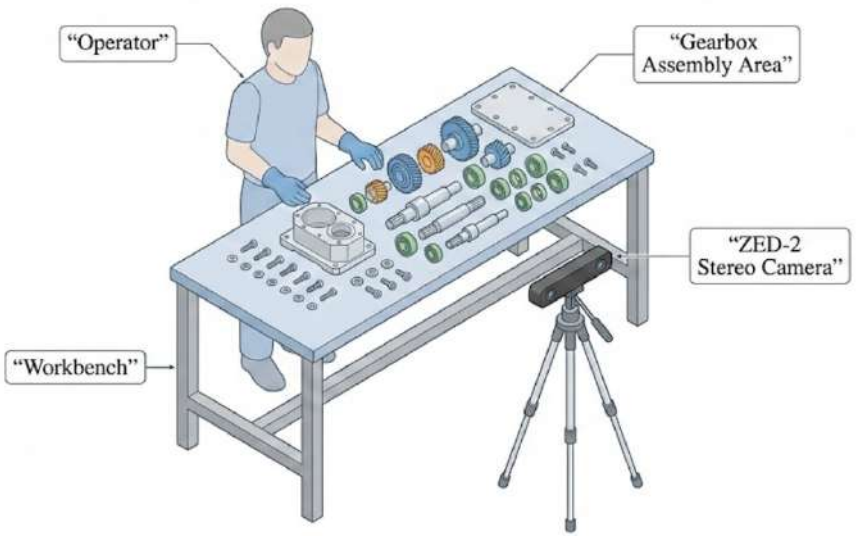


Figure 4: Laboratory workstation used for the Bonfiglioli gearbox assembly recordings, including the workbench, assembly kit, tools, and camera acquisition area.

3.2.1 Workspace Layout

The experiment was carried out on a dedicated laboratory workstation configured to resemble a small manual assembly station. The workbench height was comparable to that of typical industrial assembly lines, allowing participants to perform the task in a comfortable standing posture. The Bonfiglioli kit was placed in the center of the work surface, directly in front of the operator, so that both hands could move freely inside the main working area without requiring exaggerated reach or body displacement.

Tools and auxiliary components were arranged in fixed positions around the assembly area. At the same time, non-essential objects were removed so that the scene remained realistic without becoming unnecessarily noisy. The intention was not to create an artificially clean vision benchmark, but to preserve the visual structure of an assembly workstation while controlling avoidable sources of variability.

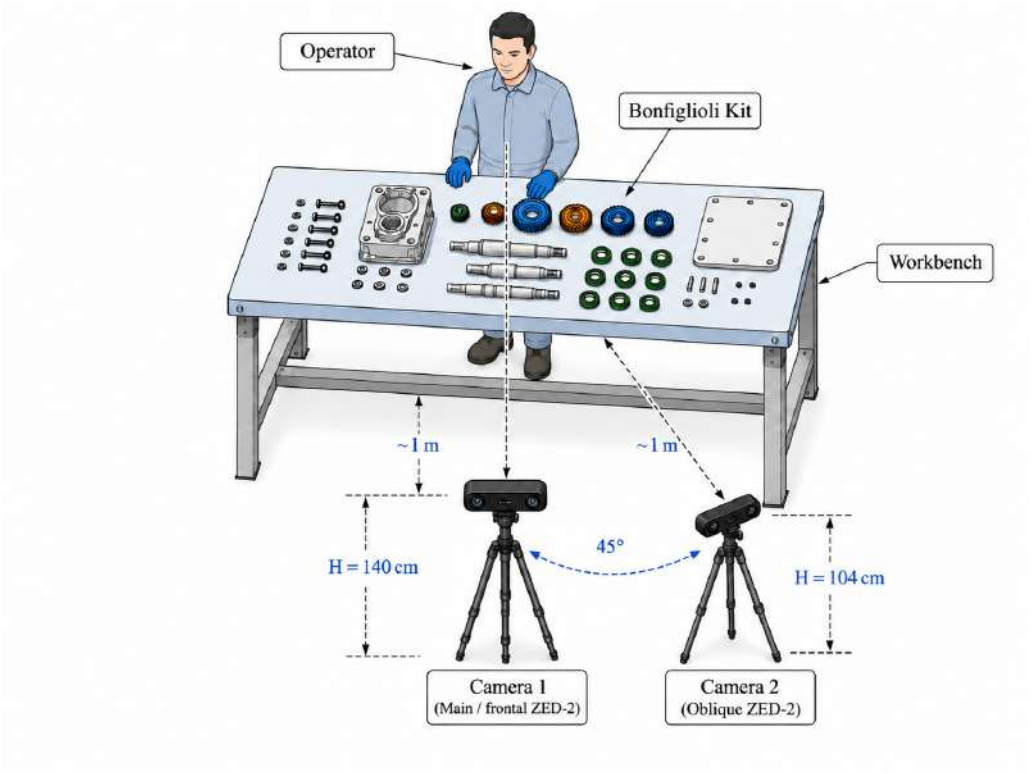


Figure 5: Schematic of the assembly workstation, showing the operator, the Bonfiglioli kit arranged on the workbench, and the placement of the two ZED-2 cameras with their approximate distance from the workstation ($\sim$1 m), relative angle (45°), and mounting heights ($H = 140$ cm for the main frontal camera and $H = 104$ cm for the oblique camera).

3.2.2 Camera Positioning

Two ZED-2 stereo cameras were mounted on tripods approximately one meter from the workstation. The main camera was positioned frontally, with its optical axis directed toward the center of the assembly area, so that the upper body of the operator, both hands, the gearbox, and the tools could be captured in a single field of view (Figure 5). A second camera was placed at approximately 45° with respect to the workstation, also at a distance of about one meter, to provide an additional view of the participant and reduce the loss of visual information caused by occlusions. The same camera placement was maintained across sessions so that changes in the data would primarily

reflect differences in task execution rather than changes in viewpoint.

3.2.3 Laboratory Recording Conditions

The recording environment was kept as stable as possible throughout the campaign. All sessions were conducted indoors, under the same artificial lighting conditions, and with the same workstation arrangement. This controlled but still realistic setup was important for two reasons: it made the dataset suitable for quantitative comparison across trials, and it ensured that the visual challenges observed in the recordings were genuinely tied to gloves, manipulation, and occlusion, rather than to uncontrolled changes in the environment.

3.2.4 Hardware and Protective Equipment

The acquisition system combines two vision-based sensors, namely the ZED-2 stereo cameras, with instrumented gloves, namely the Quantum Manus gloves, worn underneath standard industrial safety gloves. This configuration was chosen to preserve realistic working conditions while still providing access to reliable hand kinematics for ground-truth generation. In other words, the setup makes it possible to collect data under PPE constraints without giving up the ability to supervise the learning pipeline with dense hand information.

ZED-2 stereo cameras. The vision sensors used in the setup are two ZED-2 stereo cameras (Stereolabs[7]). Each device employs two synchronized 4-megapixel CMOS sensors with a fixed baseline of 120 mm and provides both RGB imagery and depth information. According to the manufacturer specifications, it supports resolutions up to $2 \times (2208 \times 1242)$ at 15 fps, as well as 1920×1080, 1280×720, and 672×376 at progressively higher frame rates. The combination of a frontal view and an oblique view makes it possible to observe the operator and the workspace from complementary perspectives.

Beyond image capture, the ZED-2 cameras also provide body tracking through the ZED SDK. In this work, the main camera is therefore used not only as an RGB-D sensor, but also as the reference source for a coarse 3D skeleton of the operator. The corresponding camera reference frame is later used as the global frame for aligning the hand keypoints obtained from

[7] `https://www.stereolabs.com/en-it/store/products/zed-2i`

the Manus gloves and the MediaPipe Hand Landmarker, according to the experimental pipeline followed. The placement of the two cameras relative to the workstation is illustrated in Figure 5.

In practice, the ZED-2 cameras are connected via USB 3.x to the acquisition PC, where the ZED SDK records color frames, depth maps, and body-tracking keypoints. The selected recording configuration balances spatial detail, temporal resolution, and storage requirements, which is important given the length and repetition of the assembly trials.

Quantum Manus gloves. Fine-grained hand and finger motion is captured using Quantum Manus gloves[8] (Manus, Quantum Metagloves line). These are professional motion-capture gloves equipped with local tracking sensors placed on the fingertips and on other relevant hand locations. The system reconstructs a biomechanical hand skeleton scaled to the user hand, allowing the acquisition of detailed hand kinematics even when parts of the hand are visually difficult to observe from the camera viewpoint.

From a practical perspective, the Quantum gloves operate at a sampling rate of 120 Hz, with low latency and sufficient battery autonomy to support recording sessions of realistic duration. Communication with the acquisition PC is handled through Manus Core, which streams the glove data for later synchronization with the ZED recordings.

In the proposed setup, the Manus gloves provide the dense 3D hand signal used for supervision. The software outputs a 21-joint skeleton for each hand, which closely follows the landmark convention commonly adopted in hand-pose estimation literature, such as the MediaPipe Hand Landmarker. Before each recording, the gloves are calibrated following the manufacturer procedure so that the digital hand model matches the user neutral posture and hand size. When the external safety gloves change across sessions, calibration is repeated to reduce systematic offsets introduced by differences in stiffness or fit.

[8] `https://www.manus-meta.com/`

Figure 6: Instrumented hand representation used in the acquisition pipeline, showing the glove configuration and the 21 hand landmarks used for ground-truth generation and model training.

Safety gloves and personal protective equipment. To reproduce realistic industrial conditions, the Manus gloves are not used as the outer layer. Each operator wears a pair of standard industrial safety gloves over the instrumented gloves. These outer gloves are representative of mechanical assembly PPE and fully cover the hand and fingers. Throughout the recording campaign, different glove models with different colors and surface textures were used to reflect the variability typically found on real shop floors.

This choice has both a practical and a methodological consequence: from the practical side, it keeps the experiment aligned with the human-centered perspective of Industry 5.0, where worker protection is non-negotiable. From the methodological side, it removes exactly the visual cues on which many off-the-shelf hand-tracking models implicitly rely, such as skin texture, knuckle

contours, and local shading patterns. Combined with the frequent occlusions produced by parts and tools, the PPE layer creates the difficult conditions under which the proposed approach is expected to demonstrate its value.

In addition to hand protection, participants operated under normal laboratory safety rules. When required, this included safety shoes and protective eyewear. These items do not directly affect the hand-tracking pipeline, but they complete the description of the experimental conditions.

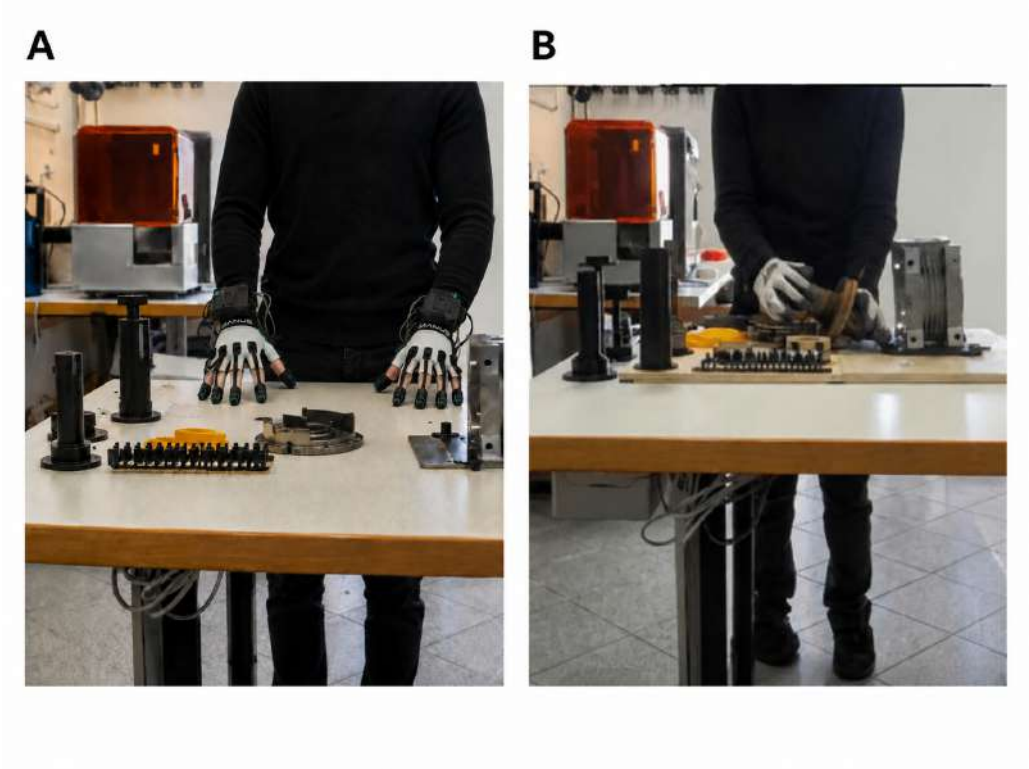


Figure 7: PPE configuration used in the experimental setup: a) operator wearing only the Quantum Manus instrumented gloves at the assembly workstation; b) the same operator wearing standard industrial safety gloves over the Manus gloves while performing the assembly task.

3.3 Data Acquisition and Processing Architecture

The sensing hardware described above is embedded in a data acquisition and processing architecture that serves two related purposes. First, it records and

synchronizes the raw streams generated by the two ZED-2 cameras and the Quantum Manus gloves. Second, it runs a set of vision-based baseline modules that are later used for comparison with the proposed approach. Grouping these elements in a dedicated section makes the workflow easier to follow, because it separates the physical setup and hardware configuration from the software and data-processing logic built on top of them.

All acquisition and processing routines were implemented in Python, using the ZED SDK for RGB-D capture and body tracking from the camera system, the Manus Core software for glove streaming, and standard scientific libraries for data handling. Figure 8 summarizes the multimodal acquisition logic adopted during the recordings.

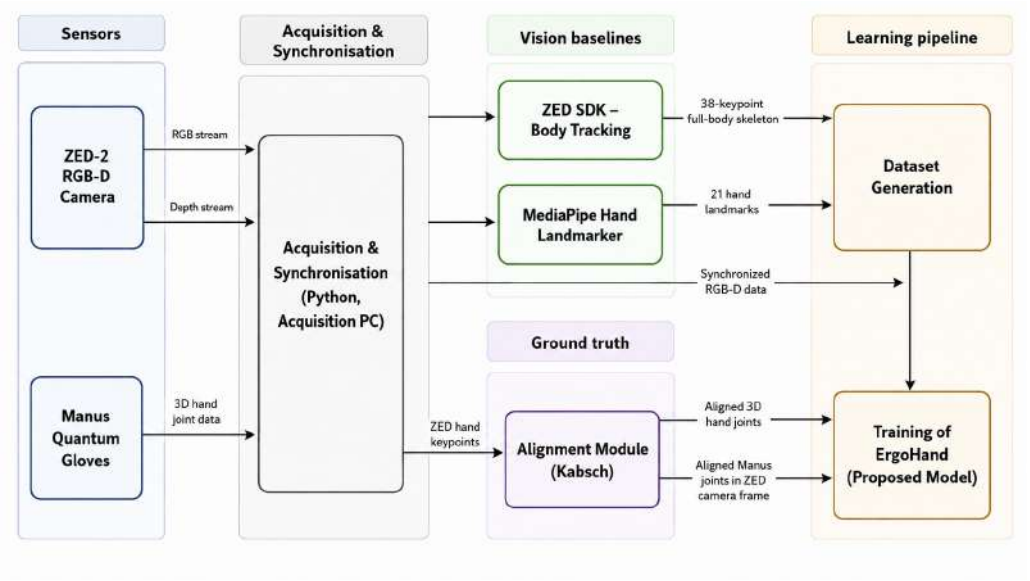


Figure 8: Overview of the multimodal acquisition pipeline used to collect synchronised RGB-D frames from the two ZED-2 cameras together with glove-based hand kinematics during the industrial assembly task. The pipeline ultimately feeds the training of the proposed **ErgoHand** model (right-hand side).

As shown on the right-hand side of Figure 8, the whole acquisition and supervision pathway ultimately feeds the training of the proposed deep-learning model, referred to throughout this thesis as **ErgoHand**. ErgoHand is the main methodological contribution of this work: a modular, computer

vision-based pipeline that learns to detect the hand, distinguish bare from gloved hands, regress the hand keypoints, and classify task-oriented grasps directly from the RGB stream of the ZED camera, under realistic industrial conditions. The ErgoHand detailed structure, components, and training strategy are presented later in this chapter in Section 3.7 and evaluated in Chapter 4. The collection of labeled samples built specifically to train the **ErgoHand** model, presented in Section 3.6, is correspondingly named the *ErgoHand dataset*.

It is important to clarify the role of the Quantum Manus gloves within this architecture, because they occupy a very specific position in the workflow. The instrumented gloves are *not* part of the ErgoHand system that is meant to be deployed on the shop floor. They are used exclusively during the data-acquisition and training phase, where they provide dense and reliable 3D hand kinematics for three purposes: calibration of the hand model, reconstruction of the hand configuration even when it is visually occluded, and generation of the ground-truth supervision required to train the model. Once ErgoHand has been trained on this supervision, it operates on camera images alone, and the Manus gloves are no longer required for the subsequent use of the system. In other words, the wearable sensor acts as a temporary "teacher" during dataset creation, while the final deployed solution remains purely vision-based and non-intrusive for the operator.

3.3.1 Acquisition Software and Synchronisation

The acquisition logic differs slightly between the two experimental conditions. In the PPE condition, the ZED video streams and the Manus glove stream are recorded in parallel and associated with session- and trial-level identifiers, so that they can later be matched during dataset generation. In the bare-hand condition, the ZED video streams are synchronized with the outputs produced by the MediaPipe Hand Landmarker. In both cases, the ZED branch provides RGB frames, depth maps, and body-tracking outputs, while the main camera reference frame is used as the geometric reference for subsequent alignment. The Manus branch provides high-frequency hand kinematics with timestamps. Because the camera system and the Manus gloves do not share the same internal clock, a dedicated temporal alignment stage is required before the visual and glove streams can be used together for supervision.

In addition to raw capture, the processing pipeline includes the extraction of baseline predictions from vision-only modules. These outputs are useful

for two reasons. On the one hand, they provide qualitative examples of how generic models behave in the industrial scene. On the other hand, they form the basis for the out-of-the-box benchmark discussed in Chapter 4.

3.3.2 ZED Body Tracking Module

The first module used in the pipeline is the body-tracking component integrated in the ZED SDK. For each frame recorded by the main ZED-2 camera, it estimates a 38-keypoint skeleton in the camera coordinate frame. The full skeleton includes the main body joints and a sparse set of hand-related points, such as the wrists and a few finger extremities.

Although these hand points are much coarser than the 21-joint representation provided by the Manus gloves, they are still useful in the present work. First, they offer a rough description of hand position and orientation in 3D space. Second, they provide a set of reference landmarks that can be matched with anatomically corresponding Manus joints during the spatial alignment step. In this sense, the ZED body tracker is not used as the final supervisory source, but as an intermediate geometric anchor that helps bridge the main camera stream and the glove stream.

The most relevant ZED keypoints for this purpose are the left wrist and selected left-hand extremities (IDs 16, 30, 32, 34, and 36), together with their right-hand counterparts (IDs 17, 31, 33, 35, and 37). These points are later used in the correspondence table introduced in Section 3.5.3.

3.3.3 MediaPipe Hand Landmarker

The second baseline module is MediaPipe Hands, used through the MediaPipe Hand Landmarker solution as a representative vision-only hand-tracking model. In this thesis, MediaPipe is applied frame by frame to the RGB stream from the main ZED-2 camera. When a palm is detected, the model returns 2D landmarks in image space, normalized 3D landmark coordinates, and an associated confidence score.

MediaPipe is included because it is a widely adopted and computationally efficient model, but also because it highlights the limits of generic hand-tracking solutions in industrial scenes. Its internal assumptions were largely learned from the use of bare hands in everyday environments. Once safety gloves, tool-driven occlusions, and cluttered assembly backgrounds are introduced, these assumptions become much less reliable. For this reason,

MediaPipe serves here as an informative baseline only for bare-hand experiments.

3.3.4 Baseline Vision Models for Benchmarking

In addition to the acquisition modules, the processing architecture includes the offline execution of representative off-the-shelf vision models used for benchmarking. These include MediaPipe, OpenPose, WiLoR, YOLO-based hand detection, and a simple OpenCV-based detector. Their role is not to generate the final ground truth, but to provide a common reference for evaluating how general-purpose hand-detection and hand-pose pipelines behave when they are applied directly to the industrial assembly recordings.

This distinction is important methodologically. The ZED and Manus streams define the multimodal acquisition and supervision pathway, whereas the baseline models define the comparison pathway. Keeping both pathways inside the same data architecture makes it possible to compare all outputs on the same task, under the same glove, tool, workpiece, and occlusion conditions. The resulting benchmark is reported in Chapter 4 and is used to support the need for an industry-specific model rather than an off-the-shelf computer-vision solution.

3.4 Testing with Users

After defining the industrial task, the physical setup, and the data acquisition architecture, the next methodological step is the organization of the recording campaign with human operators. This section therefore focuses on who participated in the recordings, how the sessions were structured, and how the task was executed to produce a coherent multimodal dataset. The emphasis is on the testing procedure with users, rather than on the internal software architecture already described in the previous section.

3.4.1 Ethical Considerations and Informed Consent

All participants took part in the recording campaign on a voluntary basis and provided informed consent before each session. They were briefed on the purpose of the study, the data collected, and their right to withdraw at any time without consequence. No personal or sensitive data were collected beyond what was strictly necessary for the hand-tracking pipeline: the recordings

capture hand and upper-body motion during a standardized assembly task, and no biometric identifiers are stored in association with the visual data. The study was conducted in accordance with the ethical guidelines of the host institution.

3.4.2 Participants

The recording campaign produced a total of 18 trials covering both bare-hand and PPE gloved conditions, involving volunteer participants and different pairs of safety gloves across sessions. This design made it possible to introduce visual variability at the glove level while keeping the task itself unchanged.

The purpose of the user study was not to optimize speed or compare participants in terms of performance. Instead, the goal was to collect repeated and traceable examples of the same assembly cycle under controlled conditions, while preserving enough variability in execution style, posture transitions, and object interaction to make the recordings representative of real manual work.

3.4.3 Recording Procedure and Session Design

Each recording session followed the same general structure. Before the session started, the workstation was restored to a fixed reference layout, with the gearbox kit positioned in the center of the work area and the tools arranged in predefined locations. The two ZED-2 cameras were then checked to ensure stable framing of the upper body, both hands, and the assembly area from the frontal and oblique viewpoints.

The task sequence was reviewed with the participant using a fixed set of instructional slides and short demonstrative videos. When needed, a short familiarization run was allowed so that the participant could perform the procedure comfortably during the recorded trials. No artificial speed constraints were imposed, because the goal was to capture natural execution rather than hurried performance.

For sessions involving instrumented recording in the presence of worn PPE, the Quantum Manus gloves were fitted and calibrated before the task began. The external safety gloves were then worn over the Manus gloves to recreate realistic industrial conditions. When the outer glove model or size changed, calibration was repeated so that the glove-based kinematic reconstruction remained as consistent as possible across sessions.

3.4.4 Task Execution and Data Collection

A single trial corresponds to one complete reducer assembly cycle starting from the fully disassembled state. After each trial, the kit is reset so that the next execution begins from the same initial condition. During task execution, the ZED-2 cameras record color and depth data from complementary viewpoints, while the ZED body-tracking module outputs a coarse operator skeleton. In parallel, the Manus system streams high-frequency hand kinematics.

Figure 9: Example of an operator performing the Bonfiglioli reducer assembly while wearing the complete PPE configuration, with the camera system capturing the hands, tools, and assembly area from complementary viewpoints.

The acquired streams are logged together with session and trial identifiers, making it possible to reconstruct the temporal structure of each recording during the offline processing stage. At the end of each session, a basic integrity check is carried out to verify that the video streams and the glove stream were correctly stored before the files are transferred into the dataset structure used for subsequent processing.

3.5 Data Preprocessing and Ground-Truth Generation

This section describes how the raw multimodal recordings are converted into a supervised dataset suitable for training and evaluating hand-recognition models in industrial-like conditions. The guiding idea is to exploit the complementarity between the two sensing modalities: the two ZED-2 cameras supply realistic RGB-D observations from complementary viewpoints, while the Manus gloves provide dense 3D hand kinematics that remain informative even when the hand is visually degraded by gloves or occlusion.

Once the visual streams and the glove stream are temporally synchronized and spatially aligned, the Manus joints can be projected onto the main ZED image plane and used to define reliable labels. In this way, the dataset-generation procedure becomes a central methodological contribution of the thesis rather than a simple preprocessing step.

3.5.1 Raw Data Streams and File Structure

Each recording trial produces two main types of data streams:

- **ZED-2 RGB-D streams.** The two cameras provide color images and associated depth maps from complementary viewpoints. In addition, the ZED body-tracking module outputs a 3D skeletal estimate in the main camera reference frame, including the wrist-level hand points later used for alignment.

- **Manus glove stream.** The Manus system outputs a 21-joint hand skeleton for each hand together with timestamps, providing dense 3D keypoints for the wrist and finger joints.

The visual and glove streams are stored with explicit trial identifiers so that they can be synchronized and processed offline. This organization is important not only for implementation convenience but also for traceability, because it makes it possible to trace each exported training sample back to its source session and trial.

3.5.2 Need for ZED–Manus Alignment

Although the camera system and the Manus gloves capture information about the same task, their outputs are not directly comparable. Two issues must be addressed before the glove data can be used to label the camera data.

First, the two sensing systems operate in different coordinate frames. The main ZED-2 camera expresses 3D points in the camera reference frame, whereas the Manus gloves express hand joints in the Manus tracking frame. Without a spatial transformation between these frames, the glove joints cannot be meaningfully overlaid onto the RGB images.

Second, the camera system and the Manus gloves operate at different temporal resolutions and do not share a hardware trigger. The ZED streams are frame-based, whereas the Manus gloves provide higher-frequency kinematic samples. As a consequence, a temporal matching step is required before each camera frame can be associated with a coherent set of glove joints.

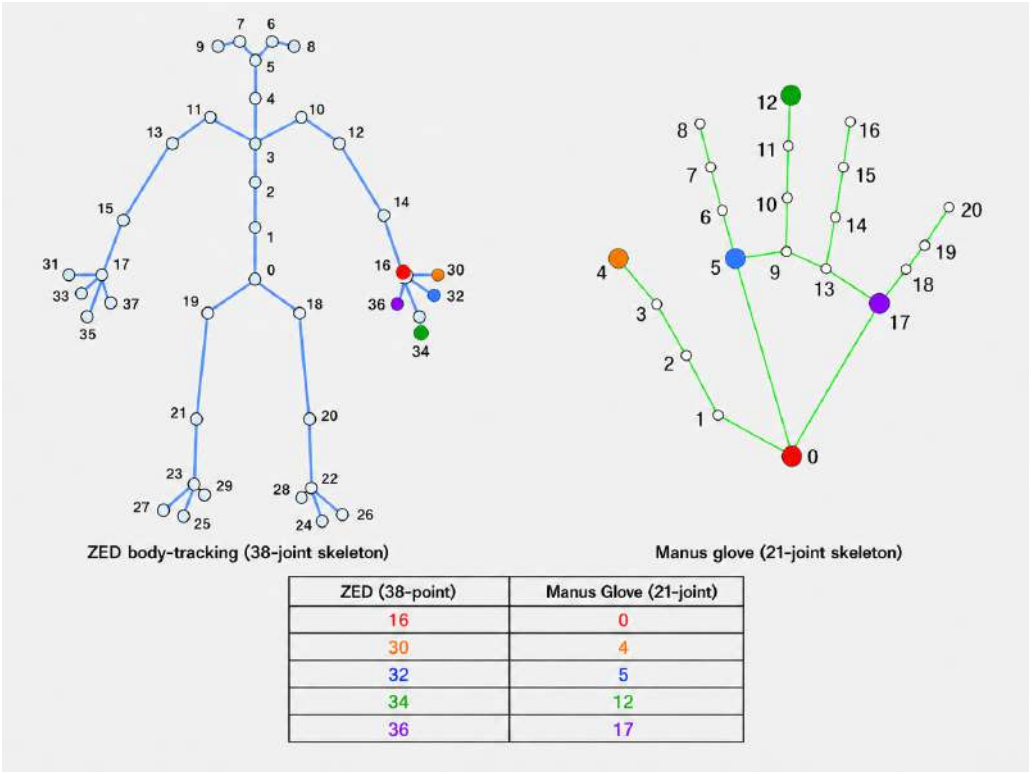
3.5.3 Spatial and Temporal Alignment Algorithm

The alignment procedure is organized in two consecutive stages: temporal synchronization and spatial alignment. Temporal synchronization associates Manus samples with the main ZED frame times. Spatial alignment estimates a rigid transformation that maps Manus 3D points into the main ZED camera frame.

Keypoint correspondences used for spatial alignment. To estimate the rigid transformation, a sparse but stable set of corresponding landmarks is used. From the ZED body skeleton, the wrist and selected hand extremities are retained. From Manus, the anatomically corresponding joints are selected from the 21-joint hand representation. The adopted correspondences are reported in Table 1 and illustrated visually in Figure 10.

Table 1: Mapping between ZED body-tracking keypoints and Manus joints used for point-set alignment.

Hand	**ZED keypoint (ID)**	**Manus joint (21-joint skeleton)**
Left	wrist (16)	wrist
Left	thumb tip (30)	thumb tip
Left	index base (32)	index MCP / base
Left	middle tip (34)	middle tip
Left	pinky base (36)	pinky MCP / base
Right	wrist (17)	wrist
Right	thumb tip (31)	thumb tip
Right	index base (33)	index MCP / base
Right	middle tip (35)	middle tip
Right	pinky base (37)	pinky MCP / base



ZED (38-point)	Manus Glove (21-joint)
16	0
30	4
32	5
34	12
36	17

Figure 10: Visual mapping between the ZED body-tracking skeleton (left, 38 joints) and the Manus glove skeleton (right, 21 joints) used for point-set alignment. The five colour-coded points indicate the corresponding landmarks selected for the rigid alignment: wrist, thumb tip, index base (MCP), middle tip, and pinky base (MCP). The example refers to the left hand (ZED IDs 16, 30, 32, 34, 36 mapped to Manus joints 0, 4, 5, 12, 17); the right hand follows the same correspondence using ZED IDs 17, 31, 33, 35, and 37.

Temporal synchronisation. The alignment algorithm implemented is described below: let $\{t_k^Z\}_{k=1}^N$ denote the timestamps of the ZED frames and $\{t_j^M\}_{j=1}^M$ the timestamps of the Manus samples. Since the two systems operate independently, their signals may contain an unknown temporal offset. To compensate for this, a global time shift Δt is estimated by maximizing the similarity between motion signals derived from the wrist trajectories of the two systems. In particular, the magnitude of the wrist velocity is used because

it is invariant to rigid translations and rotations:

$$\Delta t = \arg\max_{\tau} \ \mathrm{corr}\big(\|\dot{w}^{Z}(t)\|, \ \|\dot{w}^{M}(t+\tau)\|\big), \tag{1}$$

where w^Z denotes the wrist trajectory provided by the ZED body-tracking module and w^M the corresponding wrist trajectory measured by the Manus gloves.

After compensating for the estimated offset, the Manus signal is resampled at the main ZED frame timestamps using linear interpolation. For longer recordings, the temporal mapping can be extended to an affine model $t^Z \approx a\, t^M + b$ to absorb small clock drift effects when necessary.

Spatial alignment (Kabsch-based rigid transform). Once temporal correspondence has been established, spatial alignment is performed by estimating a rigid transformation between corresponding landmarks. Let $\{(p_i, q_i)\}_{i=1}^{n}$ be the set of paired 3D points, where p_i is a Manus joint expressed in the Manus frame and q_i is the anatomically corresponding ZED keypoint expressed in the camera frame. The goal is to estimate a rotation $R \in SO(3)$ and a translation $t \in \mathbb{R}^3$ such that

$$(R^\star, t^\star) = \arg\min_{R \in SO(3),\, t} \sum_{i=1}^{n} \|R p_i + t - q_i\|^2. \tag{2}$$

This optimization problem is solved using the Kabsch algorithm: after centering both point sets, the optimal rotation is obtained from the singular value decomposition of the cross-covariance matrix, and the translation is recovered from the centroids. To improve robustness, only frames with sufficiently reliable ZED keypoints (Table 1) are retained for the final fit, and large residual outliers can be discarded before estimating the final transformation.

Because the main camera remains fixed during a session and the Manus reference frame is stable within that session, the resulting rigid transform is assumed constant for the duration of the trial set.

Alignment quality control. After estimating $(R^\star, t^\star)$, the quality of the alignment is evaluated in two complementary ways. First, the residual fitting error is computed on the 3D landmarks used for alignment. Second, the aligned Manus joints are projected onto the RGB image plane and a subset of frames is visually inspected to verify that the projected landmarks remain

coherent with the hand motion across the workspace. The qualitative examples derived from this check are discussed in Chapter 4.

3.5.4 Ground-Truth Generation and Label Definition

After spatial alignment, each Manus joint position x^M is mapped into the ZED camera frame according to

$$x^Z = R^\star x^M + t^\star. \tag{3}$$

Given the intrinsic calibration matrix K of the main ZED camera, each aligned 3D point $x^Z = (X, Y, Z)$ is projected onto the image plane as

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \sim K \begin{bmatrix} X/Z \\ Y/Z \\ 1 \end{bmatrix}. \tag{4}$$

This procedure produces a dense set of 2D landmarks for each main ZED frame. These projected landmarks are used in two complementary ways: first, as a visual reference for qualitative inspection of the alignment; second, as a geometric basis for defining frame-level labels and extracting spatially coherent regions of interest from the RGB data.

Binary label definition (Hand / No Hand). The dataset is labeled independently for the left and right hand at the frame level. For a given hand:

- **Hand** indicates that the wrist landmark and a minimum subset of associated joints, obtained by superposition of ZED and Manus keypoints, are projected within the image boundaries and correspond to valid depth values. This includes both bare-hand and gloved-hand recordings.
- **No Hand** indicates that the projected landmarks lie outside the image boundaries or correspond to invalid depth values, indicating that the hand is not observable from the camera viewpoint in that frame.

This definition is consistent with the evaluation protocol introduced in Chapter 2, where the relevant task is framed as reliable recognition of the presence or absence of the hand in images centered on the wrist.

3.6 Dataset Creation

The raw data gathered according to the acquisition protocol described in Section 3.4.3 were converted into a supervised dataset designed to train and evaluate the **ErgoHand** model introduced in Section 3.3. Since ErgoHand is organized as a modular pipeline, the dataset was also structured into task-specific subsets. Each subset supports one step of the model, from detecting the hand in the frame to estimating keypoints and classifying task-oriented hand gestures.

In particular, the dataset was designed to support five complementary tasks:

1. **Hand Detection**: detecting hands in complex industrial scenes, including cases with occlusions and protective gloves.

2. **Hand vs Glove Classification**: distinguishing between bare hands and hands covered by PPE gloves, providing contextual information for the following keypoint-regression stage.

3. **3D Keypoints Regression**: estimating the 3D coordinates of the 21 anatomical hand keypoints for each visible hand.

4. **Left vs Right Hand Classification**: classifying whether the reconstructed hand structure corresponds to the operator's left or right hand.

5. **THGs Classification**: classifying the task-oriented hand gesture associated with the estimated 3D hand structure.

Therefore, the **ErgoHand** dataset was organized to address these technical needs in a traceable way, as summarized in Figure 11.

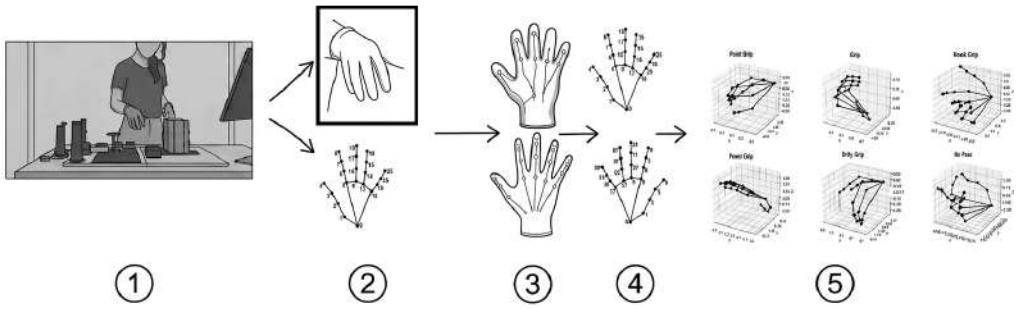


Figure 11: Schematic representation of the **ErgoHand** dataset structure: initial frames from the Bonfiglioli gearbox assembly task are processed into task-specific subsets for hand detection, hand/glove classification, keypoint regression, left/right hand classification, and THGs classification.

3.6.1 Quantitative Characteristics of the Dataset

As described in the acquisition protocol, the recording campaign produced two complementary sets of trials:

1. **Bare-hand** trials: a group of volunteer participants performed the assembly procedure without PPE gloves. These recordings were retained to represent the visually cleaner condition and to provide a reference point for comparison.

2. **Gloved-hand** trials: a further set of trials was recorded under PPE conditions, involving different volunteer participants and different pairs of safety gloves across the sessions. This subset was designed to introduce variability in glove appearance while keeping the task itself unchanged.

Table 2 reports the number of processed frames, the spatial resolution, and the nominal frame rate for the two sets.

Overall, the corpus contains **18 assembly trials** in total, covering both **bare-hand** and **gloved-hand** recording conditions. All trials were acquired at $\mathbf{1920 \times 1080}$ resolution with a nominal frame rate of **25 FPS**. Across trials, the processed frame counts show a clear difference between the two subsets:

Video	**Bare Hands Frames N**	**Gloved Hands Frames N**	**Resolution**	**FPS**
1	6372	3180	1920 X 1080	25
2	8199	2621	1920 X 1080	25
3	5603	4220	1920 X 1080	25
4	8709	1628	1920 X 1080	25
5	5782	2301	1920 X 1080	25
6	4800	2490	1920 X 1080	25
7	6809	3109	1920 X 1080	25
8	6728	2658	1920 X 1080	25
9	8854	2944	1920 X 1080	25
10	6245	2731	1920 X 1080	25
11	5140	2784	1920 X 1080	25
12	5515	2433	1920 X 1080	25
13	5336	2655	1920 X 1080	25
14	5865	1582	1920 X 1080	25
15	7746	2574	1920 X 1080	25
16	7851	1567	1920 X 1080	25
17	5229	2308	1920 X 1080	25
18	7144	1778	1920 X 1080	25

Table 2: Information about the number of frames and technical details of recording for the **Bare Hands** and **Gloved Hands** data pool.

- **Bare Hands:** from 4,800 to 8,854 frames, with a mean of 6,551.50 and a standard deviation of 1,276.13.
- **Gloved Hands:** from 1,567 to 4,220 frames, with a mean of 2,531.28 and a standard deviation of 653.06.

This imbalance is expected. The gloved-hand recordings were tailored to the ground-truth generation workflow and therefore tend to be shorter and more controlled, whereas the bare-hand trials were collected with a broader participant pool and longer execution variability.

3.6.2 Qualitative Characteristics of the Dataset

Basic counts alone do not capture why this dataset is relevant for the thesis. What makes the corpus useful is the fact that it systematically reproduces the visual conditions that generic hand-tracking pipelines find most difficult. In particular, the recordings repeatedly include:

- **safety gloves**, which suppress skin texture and modify the apparent contour of the hand;
- **tool-driven and task-driven occlusions**, generated when operators grasp tools, stabilize components, or partially cover one hand with the other;
- **industrial background clutter**, including parts, fixtures, tools, and repeated geometric structures around the workbench;
- **dynamic and complex experimental conditions**, including variations in tool use, local hand visibility, and illumination that make the recordings closer to a real industrial context.

These properties make the corpus more informative than a clean hand dataset collected in isolation. The difficulty is not concentrated in one exceptional frame, but distributed across the normal execution of the task. This is precisely the condition that matters for ergonomic monitoring, where reliability must be maintained across an entire manipulation sequence rather than on a few carefully selected images.

3.6.3 Dataset Organisation and Traceability

To keep the pipeline reproducible, the dataset-generation stage is organized around a clear separation between sample export, data loading, image-processing utilities, and model training. In practice, the implementation is packaged in a dedicated training module whose components can be re-used whenever a preprocessing assumption or a labeling rule is updated.

A representative organization is shown below:

```
Training/
  ai_model/                  # model definition/import
                               (e.g., ResNet-50)
  data_loading_functions/    # data loaders and sampling
                               utilities
  dataset/                   # dataset creation (hand / no_hand)
  img_proc_functions/        # image processing functions
  main.py                    # training entry point
  requirements.txt           # pinned dependencies for
                               reproducibility
```

This structure keeps the origin of each exported sample explicit and makes it easier to reproduce the training setup under fixed software dependencies. It also supports one of the practical goals of the thesis: making the proposed workflow transparent enough to be re-run, checked, and extended rather than treated as a one-off experiment.

3.6.4 Split Strategy for Training, Validation, and Test Sets

When the final supervised dataset is created, special care is taken to avoid overly optimistic estimates caused by temporal correlation. Consecutive frames from the same trial are often almost identical, which means that a random frame-level split would leak near-duplicate content across subsets.

For this reason, the split is performed at the trial level. All frames belonging to the same assembly cycle are assigned to a single subset only. When recordings from multiple days are available, the preferred strategy is to keep the split session-aware as well, so that the test subset includes sessions and execution conditions not observed during training. This better reflects the intended deployment scenario, where the model must generalize across operators, glove appearances, and small differences in task execution.

In practice, the dataset is divided into three disjoint subsets:

- **Training set**, used to optimize the model parameters;
- **Validation set**, used for model selection and hyperparameter tuning;
- **Test set**, reserved for the final evaluation only.

The split is kept approximately stratified with respect to the binary hand/no-hand label so that each subset contains a representative mixture of positive and negative samples, together with a reasonable diversity of glove appearances and occlusion patterns. To ensure reproducibility, the list of trial identifiers assigned to each subset is stored in a dedicated split file (for example, `splits.json`) and used consistently across all training runs.

For both bare-hand and gloved-hand conditions, an equal number of additional frames are created in which the hand is masked, simulating its absence (Section 3.7).

3.6.5 Hand Detection Dataset

The data collection and preprocessing pipeline described in Sections 3.4.3 and 3.5.4 produces the operator frames and the corresponding hand keypoints in pixel coordinates.

From these data, a bounding box was generated around each visible hand according to the following strategy:

1. **Frame and keypoint overlay**: the corresponding hand keypoints were projected onto the RGB frame in pixel coordinates.
2. **Bounding-box creation**: using the keypoint coordinates, a bounding box (**BBox**) was generated around the hand by adding a fixed margin to the minimum and maximum x and y coordinates. This procedure was performed for both the left and right hand when visible in the frame.
3. **Data export**: the **BBox** coordinates and the corresponding cropped image region were saved in dedicated directories for subsequent training.

This dataset focused on monitoring the presence or absence of a hand within the complex experimental environment; however, it did not yet address aspects

such as glove usage, keypoint positions, hand laterality (right or left), or the type of THGs performed.

	Crop/BBox n°	Crop/BBox sx n°	Crop/BBox dx n°
Hands	18105	7104	11001

Table 3: Technical details of the **Hand Detection Dataset**.

Table 3 reports the main technical details of the **Hand Detection Dataset**.

3.6.6 Hand vs Glove Classification Dataset

To allow the **ErgoHand** model to distinguish between a bare hand and a hand wearing a glove, each cropped image was assigned a corresponding class label. The images cropped as described in Section 3.6.5 were saved together with the label inherited from the source recording (**hand** or **glove**).

To increase the variability of the gloved-hand data, a glove-augmentation strategy was used to generate synthetic glove appearances from real hand images while preserving the original hand geometry and pose:

1. Starting from the detected Quantum Manus Metagloves keypoints extracted using the alignment algorithm described in Section 3.5.3, a convex hull mask was generated around the keypoints and refined with Gaussian blur to produce softened edges. This mask isolates the hand region only, enabling localized augmentation while preserving the original background unchanged.

2. To simulate different glove appearances, a color-based augmentation strategy was applied to the original frames with probability $n = 0.3$. The method uses K-means clustering on the hand-region pixels to identify the dominant color distributions, which are then remapped to predefined target colors (red, green, blue, and white) while preserving the original brightness characteristics. This process maintains realistic illumination and shading effects in the augmented gloves. Finally, the recolored hand region was blended with the original frame through alpha compositing, generating natural-looking synthetic glove variations.

3. Additionally, small random perturbations were added to the landmark coordinates during CSV export, introducing slight spatial noise that improves model robustness against minor tracking inaccuracies.

Technical details about the **Hand vs Glove Classification Dataset** are reported in Table 4.

Modality	Crop/BBox n°	Crop/BBox sx n°	Crop/BBox dx n°
Bare Hands	8995	3584	5411
Gloved Hands	9110	3520	5590

Table 4: Technical details of the **Hand vs Glove Classification Dataset**.

3.6.7 3D Keypoints Regression Dataset

One of the key functions of the **ErgoHand** model is to estimate the 3D coordinates of the hand keypoints, which are required for the subsequent THGs classification task.

Depending on the analyzed modality, the 3D keypoints were obtained from the MediaPipe-based baseline for **Bare Hand** recordings (Section 3.3.3) and from the ZED–Manus alignment procedure for **Gloved Hand** recordings (Section 3.5.3).

To express the keypoints in the coordinate system of the cropped image, the original keypoint coordinates were translated by subtracting the x and y offsets of the corresponding **BBox** (Section 3.6.5).

3.6.8 Left vs Right Hand Classification Dataset

The **ErgoHand** model must distinguish between the operator's right and left hand, including cases with occlusions and worn gloves. The laterality label (left vs right) was saved during the acquisition protocol described in Section 3.4.3. For the **Bare Hand** modality, this information is provided by the MediaPipe Hand Landmarker, while for the **Gloved Hand** modality it is obtained from the Quantum Manus Metagloves sensors (see Table 4 for more details).

3.6.9 THGs Classification Dataset

The Taxonomy of Hand Grasps (THGs) classifies manual actions into biomechanically distinct grasp categories that carry different ergonomic risk profiles.

In the context of this thesis, five classes are used:

- **No Pose (np)**: the hand is not performing a specific grasp — typically a transitional or resting state with low biomechanical load.
- **Power Grip (pg)**: the fingers wrap around an object and the palm exerts force — common during tool use and associated with high wrist and forearm extensor loading.
- **Pinch (pn)**: the object is held between the thumb and one or more fingertips — involves elevated contact stress on the fingertip pads and is associated with increased risk of finger and wrist disorders under repetitive conditions.
- **Hook Grip (hk)**: the fingers curl to form a hook without thumb involvement — sustained finger flexor loading, relevant for carrying and pulling actions.
- **Dirty Grip (dg)**: a grasp that does not fit the standard categories, often transitional or ergonomically ambiguous.

Classifying these grasp types automatically and continuously is directly relevant to manufacturing ergonomics: methods such as OCRA and HAL require knowledge of hand activity type and intensity over time, and the THGs taxonomy provides exactly the categorical description of manual actions that feeds into such assessments.

The dataset used for THGs classification comes from previous work by Borghi et al. [46]. It includes normalized 3D coordinates of 21 hand keypoints obtained with MediaPipe Hands from videos of manual operations, partly collected from YouTube and partly recorded in real industrial contexts.

The 21-keypoint structure used in that previous work is consistent with the representation adopted in the present thesis. As reported in Table 5, the 3D coordinates are normalized according to the same strategy described in Section 3.6.7 and are separated for left and right hands.

The THGs labels were assigned frame by frame by a professional ergonomist, following the iterative active-learning protocol described by Borghi et al. [46].

Hand	np	dg	hk	pg	pn
sx	62999	16	2016	1610	19897
dx	22796	254	743	911	6048

Table 5: Technical details of the **THGs Classification Dataset** (np = no pose; dg = dirty grip; hk = hook grip; pg = power grip; pn = pinch).

3.7 Proposed ErgoHand Pipeline and Training Strategy

The analysis presented in Chapter 2 showed that generic, vision-based hand pipelines degrade sharply once PPE gloves and tool-driven occlusions are introduced. In practical terms, this often leads to missed detections, unstable keypoints, or unreliable gesture labels, which is problematic when the downstream objective is ergonomic analysis. For this reason, the thesis introduces **ErgoHand**, a modular deep learning pipeline trained on the industrial dataset generated through the methodology described in the previous sections.

The model is designed as a sequence of specialized modules rather than as a single monolithic architecture. This choice reflects the complexity of the task: hand localization, glove recognition, keypoint regression, hand laterality, and THGs classification require different types of information and different learning strategies. By decomposing the problem into connected stages, the output of each upstream module can simplify the input of the following one, reducing background noise and improving the reliability of the full pipeline.

Manus-based supervision is used only during offline dataset generation and training. At deployment time, the pipeline remains vision-based: it receives RGB images from the acquisition system, detects the hands, estimates their structure, and then uses the recovered hand representation for the subsequent classification tasks.

The training pipeline was organized as follows (see Figures 12 and 13):

1. **Frame capture**: an RGB frame acquired by the ZED camera system is provided to the **ErgoHand** model for evaluation.

2. **Hand Detection**: a YOLOv8s model, fine-tuned on the **Hand Detection Dataset** (Section 3.6.5), detects hands in the analyzed frame regardless of whether PPE gloves are worn and creates a bounding box around each detected hand. The resulting crop becomes the input of the following modules, allowing them to focus on the hand region rather than on noisy background information.

3. **Hand vs Glove Classification**: a ResNet-50 model, fine-tuned on the **Hand vs Glove Classification Dataset** (Section 3.6.6), classifies whether the detected hand is bare or covered by PPE gloves. This information is useful in itself for safety-related interpretation and also provides contextual cues for the subsequent 3D keypoint-regression task.

4. **3D Keypoints Regression**: the same ResNet-50 backbone used for hand/glove classification is extended with a regression head to estimate the 3D coordinates of the 21 keypoints of each hand. The input dataset for this model is the **3D Keypoints Regression Dataset** described in Section 3.6.7.

5. **Left vs Right Hand Classification**: the 3D graph structures obtained from the estimated hand keypoints are used to classify whether the detected hand is the operator's right or left hand. A dedicated Graph Neural Network (GNN) is fine-tuned on the dataset described in Section 3.6.8; its output is then used to route the sample to the appropriate gesture-classification model.

6. **THGs Classification**: two separate GNN models, one for the `left` hand and one for the `right` hand, are trained on the dataset described in Section 3.6.9 to classify the corresponding THGs.

To avoid over-optimistic results caused by the strong similarity between consecutive frames, the dataset split follows the trial-level strategy introduced in Section 3.6.4. The final partition follows an 80%/10%/10% scheme for `train`, `validation`, and `test`, respectively. When class imbalance is observed, for example across left/right hands, hand/glove labels, or THGs classes, dedicated augmentation strategies are applied. This choice makes the evaluation more realistic, since the model must generalize across different executions, glove appearances, and occlusion patterns rather than memorize nearly identical frames from the same recording.

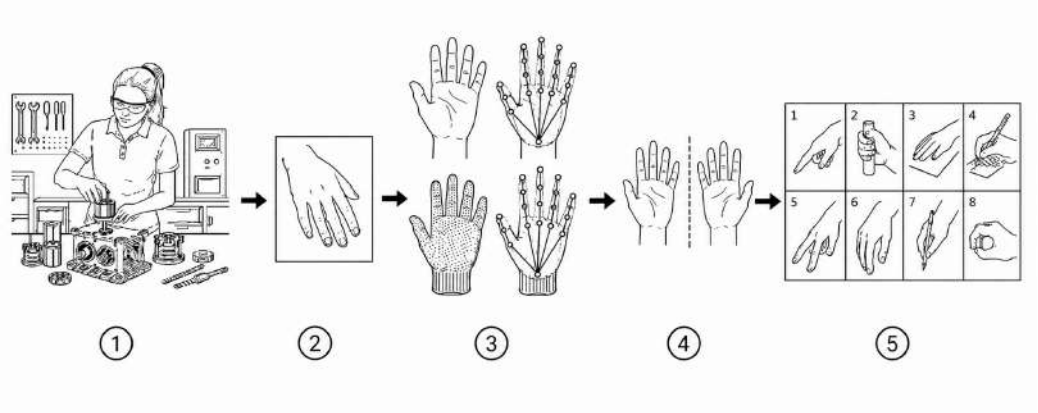


Figure 12: Schematic representation of the **ErgoHand** model functions: **1**: Frame capture; **2**: Hand Detection; **3**: Hand vs Glove Classification and Keypoints Regression; **4**: Left vs Right Hand Classification and **5**: THGs Classification.

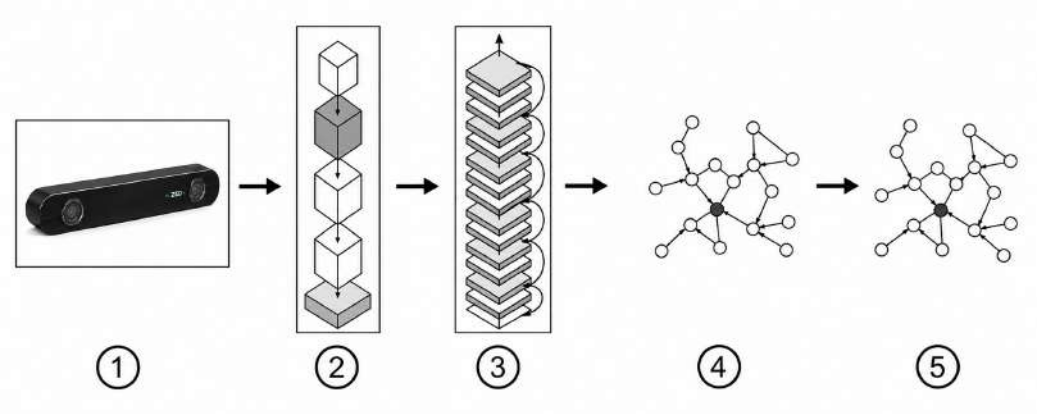


Figure 13: Schematic representation of **ErgoHand** architecture: **1**: Image acquisition system; **2**: YOLOv8s; **3**: ResNet-50; **4**: GNN model (sx vs dx) and **5**: GNN model (THGs classification).

The training details of each model will be reported in the following sections.

3.7.1 Hand Detection Pipeline

To complement vision-based analysis of hand activity in the industrial assembly scenario, a dedicated YOLO-based hand detection pipeline was im-

plemented, with the purpose of training a detector able to localize hands in complex RGB images under both bare-hand and safety-glove conditions. The task of this model can be summarized as: *Is there a hand in this image? If so, where?* Rather than treating bare hands and gloved hands as two separate semantic categories, both data sources were mapped to a single class named `hand`. At this stage, the relevant requirement is to localize the operator's hand reliably, independently of whether the hand is visually covered by PPE.

For the training dataset, each frame was paired with its associated BBox file. Samples were excluded when the BBox was missing, when the image or annotation file could not be loaded, or when the annotation file did not contain the four coordinates required to define the BBox.

Valid BBoxes, originally expressed as $(x_{\min}, y_{\min}, x_{\max}, y_{\max})$, were converted into the normalized YOLO format:

$$x_c = \frac{x_{\min} + x_{\max}}{2W}, \qquad y_c = \frac{y_{\min} + y_{\max}}{2H}, \tag{5}$$

$$w_b = \frac{x_{\max} - x_{\min}}{W}, \qquad h_b = \frac{y_{\max} - y_{\min}}{H}, \tag{6}$$

where W and H denote the image width and height.

The dataset was split into 80% *training*, 10% *validation*, and 10% *test* subsets. The YOLO configuration was stored in a `data.yaml` file using a single class name, `hand`.

The training stage was implemented using the Ultralytics YOLO framework and a pretrained YOLOv8s model, allowing transfer learning from a general object-detection representation to the specific industrial hand-detection domain of **ErgoHand**. The detector was fine-tuned for 20 epochs using an input image size of 640 pixels, a batch size of 16, and GPU acceleration on an **NVIDIA GeForce RTX 3070 Laptop GPU**. The corresponding evaluation metrics are reported in Chapter 4.

The resulting detector can be used as a vision-only front-end module for localizing hands in industrial frames before subsequent hand-pose estimation, ergonomic assessment, or THGs classification.

3.7.2 Hand vs Glove Classification Pipeline

The cropped hand region produced by the YOLO-based detector (Section 3.7.1) was used as the input of the second stage of the pipeline. This stage was designed as a multi-task deep learning module aimed at distinguishing between

bare hands and gloved hands while also learning a structured representation of the hand through anatomical keypoints. This design was adopted because the two tasks are closely related: the same visual evidence that helps the model recognize a safety glove, such as texture, contour, and finger appearance, also provides useful information about the underlying hand structure.

The system is built around a ResNet-50 visual backbone initialized from a pretrained checkpoint. Each input sample was in the form of a cropped RGB hand image stored as a NumPy array and paired with two types of supervision: a binary hand/glove label and a set of 3D keypoint annotations, organized as CSV files. In fact, the dataset described in Section 3.6.6 was organized into two classes: `glove`, corresponding to a cropped hand wearing PPE, and `hand`, corresponding to a cropped bare hand. To reduce the effect of class imbalance, dynamic oversampling was applied within the data loader so that both classes were represented with comparable frequency during training.

During preprocessing, each crop was resized to 224×224 pixels and scaled to the $[0, 1]$ range before being passed to the network. A lightweight image-level augmentation strategy is applied during training to increase robustness to small variations in appearance and scale. Since the same module also learns keypoint positions, these transformations were kept moderate in order to preserve the geometric consistency of the cropped hand representation.

The ResNet-50 backbone acts as a shared feature extractor, where the resulting feature representations (hand vs glove label and keypoints) were then passed to two task-specific heads:

1. **a classification head**, which performs binary classification between *bare hand* and *gloved hand*;

2. **a keypoint-regression head**, which predicts a 63-dimensional vector corresponding to the 3D coordinates of 21 anatomical hand landmarks.

This multi-task formulation encourages the backbone to learn features that are not limited to texture-based glove recognition, but also encode the spatial organization of the hand. In other words, the model is encouraged to understand not only what the surface of the hand looks like, but also how the hand is geometrically structured.

The classification branch provides explicit information about whether the detected hand is visually covered by PPE. This is relevant both for practical safety-related interpretation and for the following stages of the pipeline, because gloved hands may hide several visual cues that are normally used by

generic hand-pose models. The keypoint-regression branch is described in more detail in the next section.

3.7.3 3D Keypoints Regression Pipeline

In addition to the hand/glove classification task, the same ResNet-50-based module performs a regression task to estimate the 3D configuration of the hand. The target representation consists of 21 anatomical landmarks, each described by three coordinates. Therefore, the regression head predicts a 63-dimensional vector:

$$\mathbf{k} = [x_1, y_1, z_1, \ldots, x_{21}, y_{21}, z_{21}] \in \mathbb{R}^{63}. \tag{7}$$

This representation provides a compact description of the hand structure and is later used by the left/right hand classifier and by the THG classification stage.

The keypoint annotations were provided through structured CSV files containing normalized image coordinates and depth values for each of the 21 hand joints. During data loading, the normalized image-plane coordinates are first mapped back to the original frame resolution and then shifted according to the upper-left corner of the corresponding hand bounding box. This makes the landmark targets consistent with the cropped hand region used as network input. For the i-th landmark, the implemented transformation can be written as:

$$\hat{x}_i = \frac{W x_i^n - x_{\min}}{S}, \qquad \hat{y}_i = \frac{H y_i^n - y_{\min}}{S}, \tag{8}$$

where (x_i^n, y_i^n) are the normalized image coordinates of the i-th landmark, W and H are the width and height of the original RGB frame, $(x_{\min}, y_{\min})$ is the upper-left corner of the hand bounding box, and S is the input-size normalization factor used by the network.

The depth coordinate is scaled independently as:

$$\hat{z}_i = \alpha z_i^n, \tag{9}$$

where z_i^n is the normalized depth value and α is the depth scaling factor. In this way, the model learns the hand geometry in a local representation associated with the detected hand crop, rather than learning directly from the full image frame.

Training of this model was performed with a joint optimization objective that combines the classification and regression tasks. The classification branch

is optimized using cross-entropy loss with label smoothing, while the keypoint-regression branch is optimized using Smooth L1 loss. The total loss is defined as:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{kp}, \tag{10}$$

where $\mathcal{L}_{cls}$ is the hand/glove classification loss, $\mathcal{L}_{kp}$ is the keypoint-regression loss, and $\lambda = 0.5$ in the implemented training configuration. This weighting keeps the classification objective central while still encouraging the network to learn a geometrically consistent representation of the hand.

The model was trained for 30 epochs using the AdamW optimizer with a learning rate set to 10^{-4}, a weight decay of 10^{-4}, and a batch size of 8. At each epoch, the classification branch is evaluated using accuracy, whereas the regression branch is evaluated using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). As a result, this module only requires the RGB hand crop produced by the YOLO detector, preserving the objective of the pipeline: extracting a PPE-aware hand structure from visual data without requiring wearable sensors during deployment.

3.7.4 Left vs Right Hand Classification Pipeline

The present pipeline adopts a graph-based approach to hand representation, using as input the keypoints obtained from the model described in Section 3.7.3, where the classification task is performed directly on the structured keypoint coordinates.

This allows the evaluation to depend on the spatial hand structure rather than on the visual appearance of the hand. Specifically, the model distinguishes between two classes, left and right hand, encoded as *sx* and *dx*, using a GNN that operates on the spatial configuration of the hand joints.

Each hand sample, obtained from the dataset created in Section 3.6.8, is represented as a graph with 21 nodes corresponding to anatomical keypoints and edges defined according to the natural hand skeleton topology. These connections encode both finger chains and cross-links between fingers, allowing the model to capture relational dependencies between joints rather than treating them independently.

The input data were loaded from CSV files containing 3D coordinates for each keypoint. The coordinates were normalized by using the wrist joint as the reference origin, and the entire hand was scaled based on the distance to a central joint. This makes the representation invariant to translation and scale. In addition, a directional feature is appended to each node to encode

the relative horizontal orientation of the hand, which helps break symmetry and supports left/right classification.

To improve generalization, optional data augmentation is applied in the form of small random rotations, Gaussian noise, and scaling perturbations directly on the 3D space. This preserves the structural nature of the data while introducing pose variability for under-represented classes.

The core model, **HandGNN**, based on the work of Borghi et al. [46], consists of multiple *graph convolutional layers* (GCNConv) with batch normalization, leaky ReLU activations, and dropout. Residual connections were used between layers to stabilize training and improve feature propagation. After node-level feature extraction, global pooling operations (both max and mean) aggregate information across the entire graph, producing a compact representation of the hand structure.

The final classification was performed through a fully connected layer followed by a log-softmax activation: training was optimized using negative log-likelihood loss, with class weights computed dynamically to handle potential class imbalance. Performance was evaluated using accuracy and F1-score, providing a balanced view of classification quality.

Overall, this pipeline leverages the inherent structure of hand keypoints by modeling them as a graph, enabling the network to learn relational and geometric patterns that are difficult to capture with standard vector-based approaches.

3.7.5 THGs Classification Pipeline

This pipeline extends the previous graph-based approach to a more complex multi-class hand gesture classification task, where each sample was assigned to one of eight semantic classes. The system leverages the same Graph Neural Network architecture (HandGNN) presented in Section 3.7.4, but adapts the data processing and training strategy to handle richer annotations and increased class variability.

The input data consists of coupled hand 3D spatial keypoints and label annotations stored in CSV files (Section 3.6.9). For each sequence, hand coordinates (21 keypoints in 3D) were merged with frame-level labels, ensuring alignment between geometric information and class annotations. Invalid or ambiguous labels (e.g., "not found") were filtered out to maintain data quality: a label encoder was then used to convert textual class labels into numerical indices, enabling multi-class classification.

Each hand is processed independently by separating samples into left (sx) and right (dx) subsets, so as to generate one model specific to each hand. For every frame, the 3D keypoints were extracted and normalized by translating the wrist joint to the origin: as in the previous pipeline, a directional feature was appended to each node to encode horizontal orientation and resolve symmetry ambiguities. The resulting node features therefore consist of four dimensions (`x`, `y`, `z`, direction) for each of the 21 graph nodes.

The dataset was then split into `train`, `validation`, and `test` sets using a stratified 80%/10%/10% partition.

The model architecture is the same as the one used previously in Section 3.7.4: a stack of graph convolutional layers with residual connections extracts node-level features, which are then aggregated via global pooling to produce a graph-level representation and the last classification layer.

Training was performed using negative log-likelihood loss, with class weights computed dynamically based on the training distribution to further mitigate imbalance.

To address class imbalance, the following data augmentation strategy was applied: random scaling with factors ranging from 0.8 to 1.2, 3D rotations around the wrist joint (keypoint 0) using randomly selected angles between -30° and +30°, positional translations within ±0.1 units, and Gaussian noise with a standard deviation of 0.01 to increase data variability in underrepresented classes.

In order to make the model robust to occlusions, the same curriculum learning strategy applied in Borghi et al. [46] was used. This strategy featured progressive difficulty through anatomy-aware masking, where hand keypoints were masked based on the following probability distribution: wrist=20%, fingertips=5%, intermediate joints=30-55%, with masking intensity gradually increasing over 70% of the training period.

This strategy simulates partially occluded hands and encourages the system to remain robust when classifying THGs under incomplete observations.

The model was optimized using AdamW, and performance was monitored at each epoch using both accuracy and macro F1-score, the latter being particularly important in multi-class settings with potential class imbalance.

Model selection was based on the best validation F1-score, ensuring that the chosen model generalizes well across all classes rather than favoring dominant ones, and separate models were generated for the **left** and **right** hands.

Overall, this pipeline demonstrates how structured hand keypoint data can

be effectively leveraged for multi-class gesture recognition using graph-based learning, with careful attention to normalization, symmetry handling, and class balancing playing a critical role in performance.

4 Results

Chapter Overview

This chapter presents the empirical results obtained from the methodology described in Chapter 3. The focus is on two aspects: first, how off-the-shelf hand-tracking models behave in industrial conditions, and second, how the proposed ErgoHand pipeline is configured on the generated supervised dataset.

For clarity, the chapter is organized into three parts:

- **Benchmark results of off-the-shelf models**, reporting their reliability under PPE gloves, occlusions, and industrial clutter.
- **Configuration of the ErgoHand model**, summarizing the training setup adopted for the proposed approach.
- **Summary of key results**, highlighting the main findings discussed in Chapter 5.

4.1 Benchmark Results of Off-the-Shelf Models in Industrial Conditions

This section reports the outcome of applying representative off-the-shelf hand-tracking and hand-pose pipelines to the recorded assembly trials. The goal is not to tune these models to the new domain, but to quantify their **out-of-the-box** behavior under PPE gloves, manipulation-driven occlusions, and industrial clutter.

4.1.1 Evaluation Criterion: Hand-Presence Reliability

To make heterogeneous models comparable, the evaluation follows the binary formulation introduced in Chapter 2. For each frame and for each hand, every method is reduced to a single decision: *hand* or *no hand* in the expected wrist-centred region.

The benchmark dataset was acquired from the operator's viewpoint during the industrial task described in Section 3.1. The recordings were obtained using a Tobii eye-tracking device[9], which in this context was used as a head-mounted RGB camera providing an egocentric first-person view of the

[9] `https://www.tobii.com/`

assembly workspace — not for gaze analysis. This viewpoint was selected because it naturally captures the operator's hands, tools, and components from a perspective that is highly relevant to continuous ergonomic monitoring. The main technical characteristics of the resulting recordings are reported in Table 6.

Table 6: Technical details of the operator-viewpoint recordings used for benchmarking the off-the-shelf hand-tracking models.

Video	**Frames**	**Duration (s)**	**Size (GB)**	**Resolution**	**FPS**
1	1538	1706	1.25	1920×1080	25
2	1980	2242	1.68	1920×1080	25
3	482	967	0.59	1920×1080	25
4	1647	2182	1.30	1920×1080	25
5	1370	1436	0.88	1920×1080	25
6	1061	1183	0.72	1920×1080	25
7	1560	1663	0.99	1920×1080	25
8	1531	1631	0.99	1920×1080	25
9	1911	1996	1.19	1920×1080	25
10	1345	1409	0.86	1920×1080	25
11	1229	1539	0.94	1920×1080	25
12	1334	1404	0.86	1920×1080	25
13	1213	1526	0.93	1920×1080	25
14	1408	1466	0.90	1920×1080	25
15	1762	2085	1.24	1920×1080	25
16	1830	2747	1.63	1920×1080	25
17	1176	1227	0.75	1920×1080	25
18	1616	1853	1.10	1920×1080	25

This choice may appear simpler than full pose estimation, but it is directly relevant to ergonomic monitoring. If a pipeline cannot reliably establish that a hand is present, then any downstream estimate of landmarks, joint angles, or temporal exposure becomes fragmented and difficult to trust. Ground truth is derived from the projected Manus-based labels described in Section 3.5.4, which remain available even when the hand is visually degraded by gloves or partly hidden by the task.

Table 7: Benchmark performance of off-the-shelf models on the industrial assembly recordings under the hand/no-hand formulation (test split). The exported CSVs encode miss detections, enabling Recall and FNR computation. N denotes the number of evaluated hand instances (left+right).

Model	**Recall**	**FNR**	N
MediaPipe	0.0506	0.9494	52,024
OpenPose	0.0636	0.9364	5,012
WiLoR	0.3881	0.6119	52,024

4.1.2 Quantitative Comparison Using Exported Metrics

Table 7 provides the clearest quantitative summary of the benchmark. The dominant failure mode is a high **false negative rate**: the model outputs *no hand* even when the hand is present. In industrial scenes, this happens mainly because gloves suppress the visual cues learned on general-purpose datasets and because sustained occlusions interrupt the hand contour.

In the currently available export format, the most reliable quantities are those related to miss detections. For this reason, **recall** and **false negative rate (FNR)** are reported. From an ergonomic perspective, these are the most critical metrics, because missed detections create discontinuities in the signal and make long task sequences difficult to analyse.

These results are displayed in Figure 14, which reports, for each off-the-shelf model, the recall (dark bars) next to the false negative rate (light bars). Because the benchmark is formulated as a binary hand/no-hand decision over the hand instances that are actually present, recall and FNR are complementary by construction, and their sum equals one. The chart therefore makes the trade-off immediately readable: the height of each light bar corresponds exactly to the fraction of hands that the model fails to detect, while the dark bar next to it represents the fraction it manages to recover. The visual pattern is unambiguous. For MediaPipe and OpenPose, the FNR bars almost reach the top of the chart while the recall bars are barely visible, meaning that these pipelines miss the large majority of hands under the present industrial conditions, with recall values below 0.07. WiLoR is the only model whose recall bar rises to a non-negligible level, which suggests that its refinement strategy offers some resilience to difficult scenes; nevertheless, its FNR bar still clearly exceeds its recall bar, confirming that even the

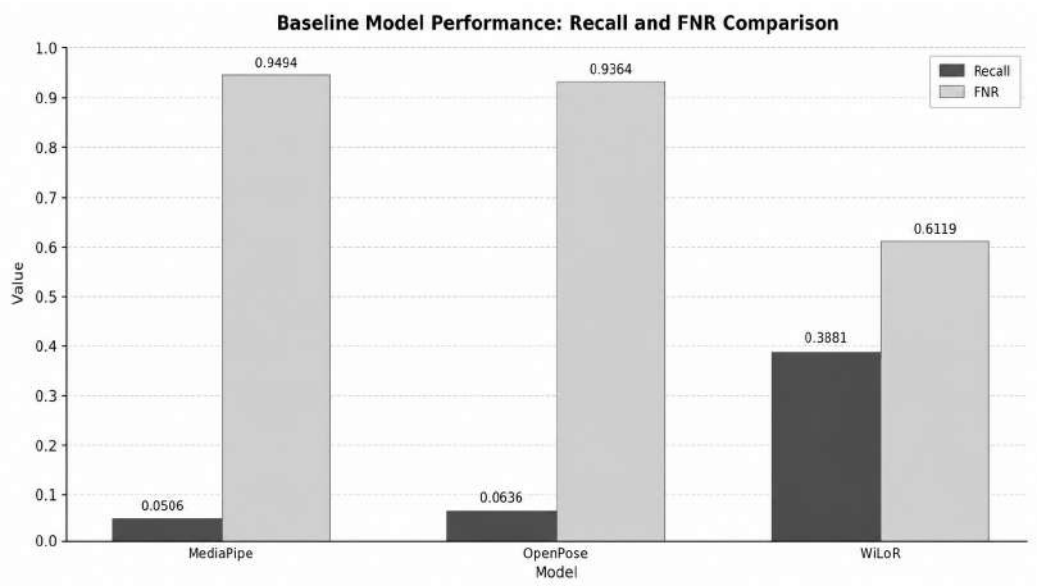


Figure 14: Recall and FNR comparison of off-the-shelf models under industrial conditions. MediaPipe and OpenPose miss more than 93% of hands; WiLoR improves recall to 0.39 but still fails on the majority of hand instances.

best-performing off-the-shelf option fails to recover the majority of the hands that are actually present.

To appreciate what these numbers mean in practice for ergonomic monitoring: a recall of 0.39 implies that in a typical assembly sequence of 10 minutes, more than 6 minutes of hand data would be lost due to missed detections. Methods such as OCRA and HAL require continuous observation of hand activity over complete work cycles to produce reliable risk estimates; a signal with gaps of this magnitude would make any such assessment impossible without manual intervention to fill the missing intervals.

4.1.3 Detailed Pairwise Agreement Analysis

To complement the summary metrics above, the agreement between models was analysed pairwise, separately for the right (DX) and left (SX) hand. For every pair of models and every frame, the two binary hand/no-hand decisions were compared. The most informative quantity is the *joint hand-detection rate*: the percentage of frames in which both models of a pair detect the hand at the same time. This indicator is directly relevant to ergonomic monitoring,

because a dependable continuous signal would require different methods to agree on the presence of the hand rather than each detecting it in different, non-overlapping frames. Alongside the pose-oriented pipelines (MediaPipe and WiLoR), the comparison also includes the detector-oriented baselines (YOLO and an OpenCV-based detector), so that failures can be compared across families of methods. The results are summarised in Figure 15.

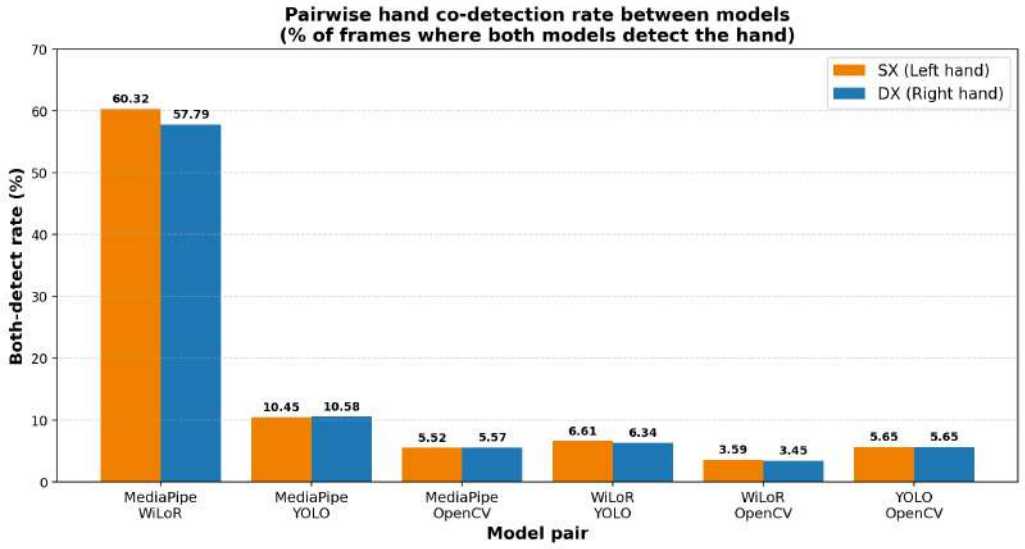


Figure 15: Pairwise hand co-detection rate between models: percentage of frames in which both models of a pair detect the hand, reported for the left (SX) and right (DX) hand. Only the MediaPipe–WiLoR pair reaches a substantial level of joint detection; every pair involving YOLO or the OpenCV detector remains below 11%.

The pattern in Figure 15 is unambiguous. Only the two pose-oriented models agree to a meaningful extent: MediaPipe and WiLoR jointly detect the hand in about 57.79% of the frames for the right hand and 60.32% for the left hand. Every other pair sits far lower. All combinations that involve the detector-oriented baselines remain below 11%: MediaPipe–YOLO reaches only about 10.6% (DX) and 10.5% (SX), while pairs such as WiLoR–OpenCV fall to roughly 3.5%. This confirms that YOLO and the OpenCV detector almost never recognise the hand at the same time as the other methods, which is consistent with their strong bias toward the negative (no-hand) class under PPE gloves and occlusion.

Two practical conclusions follow. First, the detector-oriented baselines do not merely lose accuracy; they collapse toward missed detections so systematically that they rarely overlap with any other method, and combining them with the pose pipelines would therefore add little reliable information. Second, even the best-agreeing pair, MediaPipe–WiLoR, only reaches about 58–60% joint detection, which means that in roughly four out of ten frames at least one of the two strongest models still loses the hand. This level of agreement, although clearly higher than for the detectors, remains far below what would be required to treat off-the-shelf outputs as a dependable continuous signal for ergonomic analysis. The behaviour is consistent across both hands, indicating that the limitation is systematic rather than specific to one side.

4.1.4 Qualitative Failure Modes

A qualitative inspection of the model outputs confirms that the benchmark failures are not random. They cluster around predictable events that are intrinsic to the task. The most recurrent patterns are:

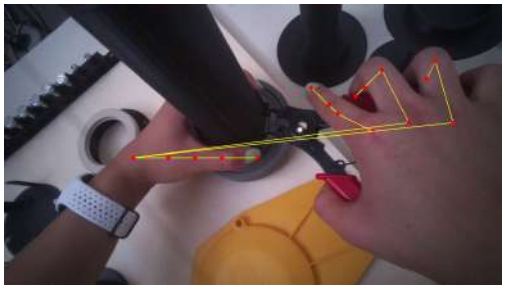
- **Glove-induced misses.** Thick or visually uniform safety gloves cause missed detections even when the hand is fully inside the camera field of view.
- **Tool-driven occlusions.** Sustained contact with tools or workpieces often interrupts tracking for tens or hundreds of consecutive frames.
- **Keypoints snapped to nearby objects.** When visibility becomes partial, some methods place landmarks on tool handles, component edges, or other rigid structures, producing outputs that may look plausible at first sight but are anatomically incorrect.

Figure 16 illustrates a representative failure case of this type. Although the hand is clearly present in the RGB frame, OpenPose detects only a small subset of keypoints, leading to an incomplete and unreliable output for any downstream ergonomic interpretation.

4.2 Validation of the ZED–Manus Alignment Pipeline

A central methodological contribution of this thesis is the multimodal alignment procedure that maps Manus glove kinematics into the ZED camera

(a) Detected hand keypoints returned by OpenPose in a challenging industrial frame.



(b) Hand skeleton reconstructed by OpenPose for the same frame.

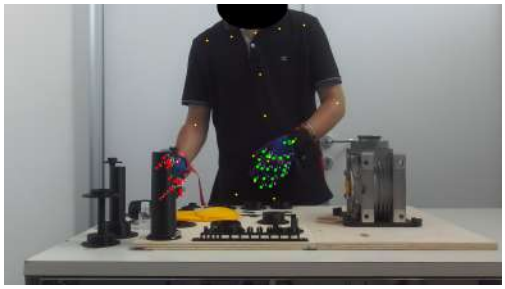
Figure 16: Example of OpenPose failure under industrial conditions. The figure shows inaccurate and incomplete hand-pose estimation caused by occlusions, close hand-object interaction, and the visual complexity of the assembly scene.

coordinate frame (Section 3.5.3). Without a reliable alignment, the projected hand landmarks cannot be used as ground-truth labels, and the entire supervised dataset loses its validity. This section presents qualitative evidence that the alignment pipeline produces coherent and anatomically consistent results across the recorded assembly trials.

Figure 17 shows two representative frames in which the aligned Manus keypoints have been projected onto the main ZED RGB image. In both cases, the 21 projected landmarks follow the visible hand geometry closely, including finger tips and MCP joints, despite the presence of safety gloves and partial occlusions caused by tools and assembly components. The spatial consistency of the projected landmarks across frames and operators confirms that the Kabsch-based rigid transformation generalises well within a recording session and that the temporal synchronisation step successfully compensates for the clock offset between the two acquisition systems.

This result is significant from an ergonomic standpoint: it demonstrates that accurate, dense hand-motion data can be obtained even when the operator wears safety gloves, by exploiting the complementarity between the wearable Manus system and the vision-based ZED camera. The projected labels form the supervisory signal on which the ErgoHand model is trained, and their quality directly determines the reliability of the downstream pipeline.

(a) Frame 1001: projected Manus landmarks overlaid on the ZED RGB image during gloved-hand assembly.



(b) Frame 1012: projected landmarks under a different hand pose and tool-occlusion condition.

Figure 17: Qualitative validation of the ZED–Manus alignment pipeline. The projected 21-joint hand skeleton remains anatomically consistent with the visible hand geometry in both frames, confirming the reliability of the alignment procedure as a ground-truth generation method.

4.3 Training Configuration of the ErgoHand Model

This section reports the training configuration adopted for the ErgoHand model. In line with the dataset-creation strategy described in Section 3.6, the samples used to train this module were balanced across the four target classes: **hand**, **no_hand**, **glove_hand**, and **no_hand_glove**. Therefore, no additional class-balancing procedure was required before training.

To improve model robustness, the following data augmentation strategy was applied to the training set with a probability of 80%:

1. A random rotation was applied by sampling the rotation angle uniformly between $-12°$ and $+12°$, reducing sensitivity to small orientation variations.

2. A random zoom was applied with a scale factor between 0.9 and 1.1. When zooming in, the image was centrally cropped to preserve the original size; when zooming out, empty regions were filled using reflection padding.

Before augmentation, images were resized to a fixed resolution of 224×224 pixels and normalized to the [0,1] range. The augmentations were applied after converting tensors from CHW to HWC format for OpenCV operations, and the augmented samples were subsequently converted back to CHW format.

The ResNet-50 backbone was fine-tuned for 30 epochs using the AdamW optimizer with a learning rate of 10^{-4} and a weight decay of 10^{-4} to improve generalization. A batch size of 2 was used because of memory constraints and the computational cost of processing the input images.

To stabilize optimization, automatic mixed precision (AMP) was enabled using *torch.amp.autocast* together with a gradient scaler, reducing the risk of numerical underflow while accelerating GPU training. The loss function was cross-entropy loss with label smoothing set to 0.1, which reduces prediction overconfidence and can improve robustness in multi-class classification tasks. During training, performance was evaluated at the end of each epoch on the validation set by monitoring both loss and accuracy. The weights associated with the lowest validation loss were saved as the best checkpoint, and early stopping with a patience of 5 epochs was used to limit overfitting.

The design, implementation, and training setup of the ErgoHand pipeline represent the methodological contribution of this thesis. The final quantitative evaluation of the trained model — including test-set accuracy, F1-score, and comparison against the off-the-shelf baselines reported in Section 4.1 — is the subject of ongoing work conducted in collaboration with the research group and is discussed as a planned next step in Chapter 5. The reproducible training configuration documented here is intended to ensure that this evaluation step can be executed and verified independently.

4.4 Summary of Key Results

The main outcomes of this chapter can be summarized as follows.

1. A realistic corpus of **18 assembly trials** was collected, covering both bare-hand and gloved-hand conditions, and acquired under the visual constraints that are most relevant to the thesis: PPE gloves, clutter, and frequent occlusions.

2. The ZED–Manus alignment pipeline produces a **usable supervisory signal** that remains coherent even in visually difficult frames, making it possible to export a structured dataset for hand-presence recognition.

3. The final supervised dataset is **balanced and traceable**. At the level of each hand side, it contains 8,874 positive and 8,874 negative samples; taken together, the right-hand and left-hand branches amount to 35,496 hand-centred samples.

4. Off-the-shelf hand-tracking pipelines show **systematic reliability limits** under industrial conditions. MediaPipe and OpenPose achieve recall below 0.07, while WiLoR improves recall to 0.3881 but still misses more than half of the hands that are actually present.

5. The pairwise agreement analysis shows that these failures are not isolated. The dominant pattern is repeated collapse toward the *no hand* label, together with substantial disagreement between models even when they belong to the same broad family of hand-analysis methods.

5 Discussion

This chapter interprets the results presented in Chapter 4 in relation to the research gap defined in Chapters 1 and 2. The central question is not whether a vision model can occasionally recognise a hand in a difficult frame, but whether it can provide a signal that remains stable enough to support ergonomic reasoning over time. This distinction is important in manufacturing contexts, where ergonomic risk emerges from repeated manual actions, constrained postures, and accumulated exposure across an assembly cycle rather than from isolated images. The discussion therefore treats the proposed pipeline as a support technology for manufacturing ergonomics and human-centered industrial system design, not as an isolated computer-vision exercise.

5.1 Key Takeaways from Chapter 4

The empirical results lead to four main takeaways.

1. **The industrial domain gap is substantial.** The benchmark confirms that generic hand-tracking pipelines degrade sharply when PPE gloves, tool interaction, and cluttered workbench scenes are introduced. This is a reliability problem because missed detections interrupt the temporal continuity required for ergonomic analysis.

2. **The multi-modal supervision strategy is feasible.** The ZED–Manus alignment pipeline makes it possible to generate viewpoint-consistent supervision in frames where vision-only methods are unreliable. This is a relevant outcome because it turns a difficult annotation problem into a reproducible data-generation workflow.

3. **Hand presence is a necessary first milestone.** The hand/no-hand formulation is simpler than full pose estimation, but it addresses a prior requirement: a system that cannot robustly establish whether the hand is visible cannot be expected to support trustworthy joint estimation or ergonomic scoring.

4. **The results have ergonomic and design implications.** Even before complete pose estimation is achieved, the work identifies when standard pipelines fail, how supervision can be strengthened, and how future

monitoring systems should account for real industrial constraints rather than laboratory assumptions. This information is relevant for workstation design, camera placement, and the selection of complementary sensing strategies.

5.2 Interpreting the Baseline Failure Modes

The benchmark results show that the observed failures are not random. They cluster around conditions that are intrinsic to the assembly task: safety gloves, object manipulation, partial occlusion, and visual clutter. This means that the poor baseline performance should not be interpreted as a few isolated errors, but as evidence of a mismatch between public-domain model assumptions and industrial working conditions.

Gloves create a visual domain shift. Most off-the-shelf pipelines are trained primarily on scenes where hands are visible through skin texture, anatomical contours, and local shading. Industrial gloves weaken or remove these cues. As a result, the detector stage becomes unstable, and once detection fails, downstream hand-pose estimates also become unreliable.

Occlusion is a temporal problem. In assembly work, occlusions are not only brief single-frame events. Operators grasp tools, stabilise parts, rotate components, and perform bimanual actions. Tracking interruptions can therefore persist across many consecutive frames. From an ergonomic perspective, this is critical because fragmented tracking makes it difficult to reason about repetition, posture continuity, and exposure duration.

Clutter can produce misleading outputs. Tools, metallic edges, and component contours may visually compete with gloved fingers when the hand is only partly visible. This helps explain why some models not only miss the hand, but may also produce anatomically implausible outputs near nearby objects. Such errors are problematic because they may appear visually plausible while remaining biomechanically incorrect.

The agreement analysis supports this interpretation. The pairwise co-occurrence matrices show that the baselines do not fail in exactly the same way. This indicates that the issue is not limited to one weak model. At

the same time, the limited agreement between methods suggests that simply combining off-the-shelf outputs would not be sufficient without domain-specific adaptation or supervision.

WiLoR improves recall but does not solve the problem. Among the tested baselines, WiLoR achieves the highest recall, indicating that stronger model design can improve robustness. However, its recall remains below the level required for dependable continuous monitoring. The practical conclusion is therefore balanced: architecture matters, but industrially relevant data and supervision remain necessary.

5.3 Ground-Truth Quality and Supervisory Reliability

A central contribution of this work is the generation of supervision by aligning Manus glove kinematics with ZED RGB(-D) recordings. This strategy is valuable because glove kinematics remain available even when the visual appearance of the hand is degraded by PPE or occlusion. By projecting the aligned data into the camera viewpoint, the labels remain connected to the visual scene used by the learning pipeline.

It is important to make explicit the role that the Manus gloves play in this strategy, to avoid any ambiguity. The instrumented gloves are a *development and training* instrument: they are used offline, during dataset creation, only to generate and validate the ground-truth supervision on which the model is trained. They are *not* part of the analysis that the system performs once it is in use. After training on the resulting labels, the hand-monitoring pipeline operates from the camera stream alone, so no wearable sensor is worn by the operator and no glove data is required during the actual ergonomic analysis of the task. In other words, the Manus gloves serve to *build* the model, not to *run* it.

For the present hand-presence formulation, the supervisory signal does not need millimetre-level pose accuracy to be useful. What matters is that the projected information is coherent enough to identify whether the target hand is present in the relevant image region. This makes the approach appropriate for a first learning target while avoiding a stronger claim that the current pipeline already solves dense 2D or 3D hand-pose estimation.

Some sources of error must still be acknowledged. Residual temporal offsets, imperfect spatial alignment, noise in the ZED body keypoints, and ambiguous frames near image borders can all affect label quality. These issues

justify a conservative export strategy: ambiguous samples should be filtered out rather than retained with uncertain labels. Although this reduces dataset size, it improves the credibility of the resulting supervised corpus.

The strength of the pipeline is therefore not only the labels it produces, but also its reproducibility. The traceable structure described in Chapter 3 means that the dataset can be regenerated, checked, refined, or extended when new filtering rules or additional recordings become available.

5.4 Implications for Ergonomic Monitoring and Industrial System Design

Beyond the specific numbers, it is worth asking what it would take for a system like this to be adopted on a real shop floor, and what concrete value it would bring once installed. The following points summarise the practical implications from that point of view.

What the system would actually do on the line. In a realistic deployment, one or more fixed cameras observe a manual workstation while the operator works normally, wearing the usual safety gloves and handling the usual tools. The system tracks the hands automatically and converts the recordings into the continuous hand-activity information that established ergonomic methods such as OCRA, HAL, or RULA require. Today this information is gathered by an expert who watches the task and fills in checklists by hand, which is slow and can only be done occasionally; a camera-based tool could provide the same information continuously and at a very low cost per workstation.

Continuity is what really matters. For ergonomic scoring, briefly losing the hand several times is more damaging than a small error in a single frame, because the score is computed over whole work cycles rather than on isolated images. A usable system therefore has to keep following the hand even during tool use and gloved manipulation. The benchmark in this thesis shows that off-the-shelf models do not yet reach this stability under real industrial conditions, which is precisely why a domain-specific model such as ErgoHand is needed before such a tool can be trusted on the line.

A simple "is the hand visible?" function is already useful. Even before full keypoint estimation is solved, a reliable hand-presence module has immediate value. It can tell an ergonomist which parts of a task are clearly visible from the installed cameras and which are not, so that only the trustworthy segments are scored automatically while the rest are flagged for a second camera or a manual check. In an adoption scenario this honesty about coverage is important: the tool reports how much of the task it can actually see, instead of producing confident but unreliable numbers.

Installation is part of the solution, not just the model. The results also show that camera placement, field of view, expected occlusion zones, glove colour, and workstation clutter strongly affect how well the hands can be monitored. For a company this means that adopting such a system is not only a matter of choosing a good model, but also of positioning the cameras and organising the workstation so that the hands stay observable. This is encouraging for adoption, because these are practical, low-cost decisions that a plant can control at installation time.

Fit with the Industry 5.0 vision. Finally, these implications align with the Industry 5.0 perspective of Chapter 1. The approach is *human-centric*, because it adapts to workers wearing real PPE instead of assuming ideal bare-hand conditions and asks nothing extra of the operator at run time. It supports *sustainability*, because continuous monitoring helps catch ergonomic problems early and reduce the long-term cost of work-related musculoskeletal disorders. And it improves *resilience*, because a traceable, automatic pipeline lets a plant react to risky or poorly observable tasks instead of relying on occasional manual inspection.

5.5 Limitations and Threats to Validity

Several limitations should be considered when interpreting the results.

- **Current empirical scope.** The thesis establishes the industrial reliability gap and the feasibility of the supervisory pipeline, but it does not yet provide a complete end-to-end demonstration of robust keypoint-level industrial hand-pose estimation.

- **Task specificity.** The evaluation is based on one assembly case study, the Bonfiglioli reducer task. Although this task includes bimanual actions, tool use, and occlusions, it cannot represent the full diversity of industrial manual work.
- **Controlled recording environment.** The laboratory setup improves traceability and comparison, but it may under-represent variability found in real factories, such as illumination changes, reflective surfaces, larger operator movement, or different workstation layouts.
- **Participant and glove distribution.** The bare-hand and gloved-hand subsets were designed for different purposes. The gloved subset is useful for studying appearance shift, but it does not fully capture all inter-operator strategies and anthropometric differences.
- **Single-view observability.** With a fixed camera viewpoint, some occlusions are physically unavoidable. In these cases, the limitation is not only model performance but also the absence of visual information.
- **Dependence on alignment quality.** The label-generation pipeline depends on temporal synchronisation and spatial alignment. If either stage drifts, the projected supervision may degrade, making quality control an essential part of the workflow.

These limitations define the boundary of what can be claimed with confidence and what should remain part of future research.

5.6 Future Work

The present work opens several concrete directions for continuation.

1. **Complete the empirical evaluation of the proposed pipeline.** The full test-set evaluation of the ErgoHand model — including accuracy, F1-score, and direct comparison against the off-the-shelf baselines — is currently being carried out in collaboration with the research group, using the reproducible training setup and traceable dataset splits defined in Chapters 3 and 4.
2. **Extend the supervision target beyond presence recognition.** Once the alignment and filtering stages are further consolidated, the

same multi-modal pipeline can support wrist-centred localisation, 2D landmark estimation, and eventually 3D hand-pose reconstruction under PPE conditions.

3. **Incorporate temporal modelling.** Future models should exploit frame-to-frame continuity rather than treating each frame independently. Temporal smoothing, recurrent modules, or transformer-based sequence models could help reduce flicker during prolonged occlusions.

4. **Increase external validity.** Additional recordings across more operators, glove types, workstations, lighting conditions, and manual tasks would strengthen the conclusions and separate appearance-related effects from task-specific effects.

5. **Move closer to ergonomic decision support.** A longer-term direction is to connect visual monitoring to ergonomic indicators such as repetition patterns, visibility coverage, and joint-angle-based risk indices once pose estimation becomes sufficiently robust. This would make the pipeline more directly useful for designers who need to evaluate and redesign industrial workstations.

5.7 Concluding Remarks

In summary, the results discussed in this chapter confirm that reliable hand monitoring in industrial environments cannot be achieved by transferring generic vision models to the shop-floor context without adaptation. At the same time, they show that the thesis provides a solid foundation: the industrial domain gap is made explicit, the multi-modal supervision pipeline is feasible, and a presence-first learning strategy is a defensible intermediate step toward robust hand analysis.

The main message is therefore that progress in industrial hand monitoring requires data, models, and evaluation criteria grounded in real working conditions. The contribution of this thesis lies in establishing that foundation for future ergonomic assessment tools and for the design of industrial systems where worker protection is treated as a design requirement rather than an afterthought.

6 Conclusions

This thesis started from a practical problem in industrial ergonomics and human-centered manufacturing: hand-tracking methods that perform well in controlled or public benchmark settings often become unreliable when transferred to real assembly work. Safety gloves, tools, cluttered workstations, and sustained occlusions change the visual appearance of the hand and interrupt the continuity required for ergonomic analysis. In this context, the main difficulty is not only detecting a hand in isolated frames, but maintaining a dependable signal across a complete task sequence, so that manual work can be analysed in a way that is useful for workstation and process design.

To address this gap, the thesis developed and evaluated a multi-modal strategy based on ZED-2 RGB(-D) recordings and Quantum Manus glove kinematics. Through temporal synchronisation and spatial alignment, the two data streams were brought into a common reference frame and used to generate visual supervision under conditions where vision-only methods are unreliable. This made it possible to move from raw industrial recordings to a structured, traceable dataset for hand-presence analysis, while keeping the final objective focused on vision-based monitoring that could support ergonomic assessment without requiring intrusive wearable sensors during deployment.

The results support three main conclusions. First, the industrial domain gap is substantial: under the adopted hand/no-hand benchmark, MediaPipe and OpenPose achieved recall values below 0.07, while WiLoR improved recall to 0.3881 but still missed a large portion of the hands that were actually present. Second, the proposed data-generation workflow is feasible and produces a supervised corpus of 35,496 hand-centred samples when both hand sides are considered together. Third, hand-presence recognition is a defensible first learning target, because visual observability must be established before moving toward reliable keypoint-level pose estimation, joint-angle analysis, or ergonomic risk indicators.

6.1 Answer to the Research Question

The research question posed in Chapter 1 asked whether a vision-based model, trained on industrial data, can provide sufficiently reliable hand-motion information to serve as input for ergonomic risk assessment in real assembly environments, even when operators wear safety gloves and experience frequent

tool-driven occlusions.

Based on the work presented in this thesis, the answer is *partially affirmative at the foundational level*. The evidence shows that off-the-shelf models cannot currently provide this reliability: recall values below 0.07 for MediaPipe and OpenPose, and below 0.40 for WiLoR, confirm that existing pipelines cannot sustain the continuous signal required by methods such as OCRA or HAL in industrial conditions. At the same time, the thesis demonstrates that the prerequisite conditions for building such a model are achievable: the ZED–Manus alignment pipeline produces coherent supervision even under PPE and occlusion, and the resulting dataset provides a structured foundation for training a domain-specific classifier. Whether the trained ErgoHand model fully closes this gap will be confirmed by the evaluation currently in progress. What the thesis establishes with confidence is that the path toward reliable industrial hand monitoring is technically feasible and that the data infrastructure to pursue it is in place.

6.2 Main Contributions

The main contributions of the thesis can be summarised as follows:

1. **An empirical characterisation of the industrial reliability gap.** A quantitative benchmark of off-the-shelf models under the hand/no-hand formulation showed that generic pipelines cannot be assumed reliable in assembly scenarios involving PPE gloves, object interaction, and persistent occlusions. MediaPipe and OpenPose reached recall values of only 0.05 and 0.06, missing more than 93% of the hands that were actually present, while the best-performing baseline, WiLoR, reached a recall of 0.39 (FNR = 0.61) and therefore still missed roughly six out of every ten hands. Put in practical terms, a recall of 0.39 means that more than half of a hand-intensive task would go unobserved, which is incompatible with continuous methods such as OCRA or HAL. These concrete results directly support the thesis argument that industrial ergonomic monitoring requires domain-specific validation rather than the simple transfer of generic models.

2. **A Kabsch-based spatio-temporal alignment pipeline.** A dedicated algorithm was developed and validated to synchronise Manus glove kinematics with ZED-2 RGB(-D) recordings in time and to map

the glove joints into the camera coordinate frame using a Kabsch-based rigid transformation. This alignment procedure is the core technical contribution that makes it possible to generate reliable visual supervision for gloved hands under occlusion, without requiring manual annotation.

3. **A reproducible dataset-generation workflow.** The thesis establishes a traceable path from recorded trials to supervised hand-centred samples, with explicit dataset structure and split management.

4. **A staged, presence-first approach to industrial hand analysis.** Instead of attempting full hand-pose estimation under all industrial conditions at once, the thesis argues for solving the problem in stages, starting from the most basic question: *is the hand visible in the frame at all?* Reliably answering this hand-presence question is treated as a necessary first step, because there is little value in estimating finger joints or computing ergonomic indices on frames where the hand cannot even be located. This presence-first step therefore provides the foundation on which the later stages — keypoint estimation, joint-angle analysis, and ergonomic risk scoring — can be built.

These contributions address the first four research objectives defined in Chapter 1. The fifth objective, training and validating a robust HPE model, is addressed at a foundational level rather than as a complete final solution. The thesis defines the data logic, supervision mechanism, and training configuration needed for that objective, while the full test-set evaluation and extension toward keypoint-level pose estimation remain future work. This framing is intentionally conservative: the thesis does not present the AI model as an end in itself, but as a step toward reliable digital support for manufacturing ergonomics.

6.3 Final Perspective

From the perspective of manufacturing ergonomics, the value of this work lies in shifting the focus from isolated benchmark performance to reliable observation under real working conditions. Even before full pose estimation is achieved, the intended hand-presence module can help identify phases of poor visibility, quantify how much of a task is visually trackable, and guide decisions about camera placement or complementary sensing once its test-set performance is fully evaluated. These outputs are relevant for mechanical and

industrial engineers because they inform how a workstation, sensing layout, or manual task could be redesigned to make ergonomic assessment more feasible and less dependent on subjective observation alone.

The thesis therefore provides a concrete foundation for future ergonomic assessment tools: industrial recordings, aligned supervision, a traceable dataset, and a first domain-oriented learning target. The next step is to complete the empirical evaluation of the trained classifier and extend the same framework toward localisation, landmark estimation, and eventually robust hand-pose analysis under PPE and occlusion. In the longer term, this direction can support the three pillars of Industry 5.0: *human-centricity*, by adapting monitoring tools to workers and their protective equipment; *sustainability*, by helping preserve worker health and reduce the cost of work-related musculoskeletal disorders; and *resilience*, by enabling industrial systems to detect and react to risky or poorly observable manual-task conditions.

References

[1] İ. Yilmaz, D. Aygün, and Z. Tanrikulu. "Social Media's Perspective on Industry 4.0: A Twitter Analysis". In: *Social Networking* 6.4 (2017), pp. 251–261. DOI: 10.4236/sn.2017.64017. URL: https://www.scirp.org/journal/paperinformation.aspx?paperid=78450.

[2] M. Breque, L. De Nul, and A. Petridis. *Industry 5 — Towards a Sustainable, Human-Centric and Resilient European Industry.* Tech. rep. Directorate-General for Research and Innovation, The Publications Office of the European Union, 2021.

[3] Maija Breque, Lars De Nul, and Athanasios Petridis. *Industry 5.0: Towards a Sustainable, Human-Centric and Resilient European Industry.* Brussels: European Commission, Directorate-General for Research and Innovation, 2021.

[4] Vittorio Villani et al. "Ergonomic Human-Robot Collaboration in Industry: A Review". In: *Frontiers in Robotics and AI* 9 (2023), p. 813907. DOI: 10.3389/frobt.2022.813907.

[5] Eugenio Monari et al. "Physical Ergonomics Monitoring in Human–Robot Collaboration: A Standard-Based Approach for Hand-Guiding Applications". In: *Machines* 12.4 (2024), p. 231. DOI: 10.3390/machines12040231.

[6] John R Wilson. "Fundamentals of ergonomics in theory and practice". In: *Applied ergonomics* 31.6 (2000), pp. 557–567.

[7] Praveen Kumar Reddy Maddikunta et al. "Industry 5.0: A survey on enabling technologies and potential applications". In: *Journal of Industrial Information Integration* 26 (2022), p. 100257. ISSN: 2452-414X. DOI: https://doi.org/10.1016/j.jii.2021.100257. URL: https://www.sciencedirect.com/science/article/pii/S2452414X21000558.

[8] "Work-related musculoskeletal disorders: the epidemiologic evidence and the debate". In: *Journal of Electromyography and Kinesiology* 14.1 (2004). State of the art research perspectives on muscoskeletal disorder causation and control, pp. 13–23. ISSN: 1050-6411. DOI: https://doi.org/10.1016/j.jelekin.2003.09.015.

[9] Teresia Nyman et al. “Reliability and Validity of Six Selected Observational Methods for Risk Assessment of Hand Intensive and Repetitive Work”. In: *International Journal of Environmental Research and Public Health* 20.8 (2023), p. 5505. DOI: 10.3390/ijerph20085505.

[10] Weiya Chen et al. “A Survey on Hand Pose Estimation with Wearable Sensors and Computer-Vision-Based Methods”. In: *Sensors* 20.4 (2020), p. 1074. DOI: 10.3390/s20041074.

[11] Fan Zhang et al. “Mediapipe hands: On-device real-time hand tracking”. In: *arXiv preprint arXiv:2006.10214* (2020).

[12] Ginés Hidalgo Martınez. “Openpose: Whole-body pose estimation”. In: *Ph. D. dissertation* (2019).

[13] Shashwat Vyas et al. “Human Activity Recognition in MMH: Pilot Study on Lifting and Lowering Classification”. In: *IFAC-PapersOnLine* 59.10 (2025). 11th IFAC Conference on Manufacturing Modelling, Management and Control MIM 2025, pp. 763–768. ISSN: 2405-8963. DOI: https://doi.org/10.1016/j.ifacol.2025.09.130. URL: https://www.sciencedirect.com/science/article/pii/S2405896325008912.

[14] Christian Zimmermann and Thomas Brox. “Learning to Estimate 3D Hand Pose from Single RGB Images”. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 4913–4921. DOI: 10.1109/ICCV.2017.525.

[15] “Industry 5.0: Prospect and retrospect”. In: *Journal of Manufacturing Systems* 65 (2022), pp. 279–295. ISSN: 0278-6125. DOI: https://doi.org/10.1016/j.jmsy.2022.09.017.

[16] “Exploring the potential of Operator 4.0 interface and monitoring”. In: *Computers & Industrial Engineering* 139 (2020), p. 105600. ISSN: 0360-8352. DOI: https://doi.org/10.1016/j.cie.2018.12.047.

[17] Zewen Li et al. “A survey of convolutional neural networks: analysis, applications, and prospects”. In: *IEEE transactions on neural networks and learning systems* 33.12 (2021), pp. 6999–7019.

[18] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).

[19] Xingyu Chen et al. “Camera-Space Hand Mesh Recovery via Semantic Aggregation and Adaptive 2D-1D Registration”. In: *CoRR* abs/2103.02845 (2021). arXiv: 2103.02845. URL: https://arxiv.org/abs/2103.02845.

[20] “Vision-based hand pose estimation: A review”. In: *Computer Vision and Image Understanding* 108.1 (2007). Special Issue on Vision for Human-Computer Interaction, pp. 52–73. ISSN: 1077-3142. DOI: https://doi.org/10.1016/j.cviu.2006.10.012.

[21] Lynn McAtamney and E. Nigel Corlett. “RULA: a survey method for the investigation of work-related upper limb disorders”. In: *Applied Ergonomics* 24.2 (1993), pp. 91–99. DOI: 10.1016/0003-6870(93)90080-S.

[22] Enrico Occhipinti. “OCRA: a concise index for the assessment of exposure to repetitive movements of the upper limbs”. In: *Ergonomics* 41.9 (1998), pp. 1290–1311. DOI: 10.1080/001401398186315.

[23] “Validation of markerless vision-based motion capture for ergonomics risk assessment”. In: *International Journal of Industrial Ergonomics* 107 (2025), p. 103734. ISSN: 0169-8141. DOI: https://doi.org/10.1016/j.ergon.2025.103734.

[24] Shanxin Yuan et al. “BigHand2.2M Benchmark: Hand Pose Dataset and State of the Art Analysis”. In: *CoRR* abs/1704.02612 (2017). arXiv: 1704.02612. URL: http://arxiv.org/abs/1704.02612.

[25] Umar Iqbal et al. “Hand Pose Estimation via Latent 2.5D Heatmap Regression”. In: *CoRR* abs/1804.09534 (2018). arXiv: 1804.09534. URL: http://arxiv.org/abs/1804.09534.

[26] Adriana-Violeta Savescu et al. *A 25 degrees of freedom hand geometrical model for better hand attitude simulation*. Tech. rep. SAE Technical Paper, 2004.

[27] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit. “Hands Deep in Deep Learning for Hand Pose Estimation”. In: *CoRR* abs/1502.06807 (2015). arXiv: 1502.06807. URL: http://arxiv.org/abs/1502.06807.

[28] Franziska Mueller et al. “Real-time Hand Tracking under Occlusion from an Egocentric RGB-D Sensor”. In: *CoRR* abs/1704.02201 (2017). arXiv: 1704.02201. URL: http://arxiv.org/abs/1704.02201.

[29] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. "V2V-PoseNet: Voxel-to-Voxel Prediction Network for Accurate 3D Hand and Human Pose Estimation from a Single Depth Map". In: *CoRR* abs/1711.07399 (2017). arXiv: 1711.07399. URL: http://arxiv.org/abs/1711.07399.

[30] Adrian Spurr et al. "Cross-modal Deep Variational Hand Pose Estimation". In: *CoRR* abs/1803.11404 (2018). arXiv: 1803.11404. URL: http://arxiv.org/abs/1803.11404.

[31] Martin de La Gorce, David J. Fleet, and Nikos Paragios. "Model-Based 3D Hand Pose Estimation from Monocular Video". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33.9 (2011), pp. 1793–1805. DOI: 10.1109/TPAMI.2011.33.

[32] Liuhao Ge et al. "Hand PointNet: 3D Hand Pose Estimation Using Point Sets". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 8417–8426. DOI: 10.1109/CVPR.2018.00878.

[33] Xiaoming Deng et al. "Hand3D: Hand Pose Estimation using 3D Neural Network". In: *CoRR* abs/1704.02224 (2017). arXiv: 1704.02224. URL: http://arxiv.org/abs/1704.02224.

[34] Weiya Chen et al. "A Survey on Hand Pose Estimation with Wearable Sensors and Computer-Vision-Based Methods". In: *Sensors* 20.4 (2020). DOI: 10.3390/s20041074. URL: https://www.mdpi.com/1424-8220/20/4/1074.

[35] Jonathan Tompson et al. "Real-Time Continuous Pose Recovery of Human Hands Using Convolutional Networks". In: 33.5 (2014). ISSN: 0730-0301. DOI: 10.1145/2629500. URL: https://doi.org/10.1145/2629500.

[36] Danhang Tang et al. "Latent Regression Forest: Structured Estimation of 3D Articulated Hand Posture". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 3786–3793. DOI: 10.1109/CVPR.2014.490.

[37] Christian Zimmermann et al. "FreiHAND: A Dataset for Markerless Capture of Hand Pose and Shape from Single RGB Images". In: *CoRR* abs/1909.04349 (2019). arXiv: 1909.04349. URL: http://arxiv.org/abs/1909.04349.

[38] Gyeongsik Moon et al. “InterHand2.6M: A Dataset and Baseline for 3D Interacting Hand Pose Estimation from a Single RGB Image”. In: *CoRR* abs/2008.09309 (2020). arXiv: 2008.09309. URL: https://arxiv.org/abs/2008.09309.

[39] Shreyas Hampali et al. “HO-3D: A Multi-User, Multi-Object Dataset for Joint 3D Hand-Object Pose Estimation”. In: *CoRR* abs/1907.01481 (2019). arXiv: 1907.01481. URL: http://arxiv.org/abs/1907.01481.

[40] Andreas Doering, Umar Iqbal, and Juergen Gall. “Joint Flow: Temporal Flow Fields for Multi Person Tracking”. In: *CoRR* abs/1805.04596 (2018). arXiv: 1805.04596. URL: http://arxiv.org/abs/1805.04596.

[41] Yujun Cai et al. “Exploiting Spatial-Temporal Relationships for 3D Pose Estimation via Graph Convolutional Networks”. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 2272–2281. DOI: 10.1109/ICCV.2019.00236.

[42] Fan Zhang et al. “MediaPipe Hands: On-device Real-time Hand Tracking”. In: *CoRR* abs/2006.10214 (2020). arXiv: 2006.10214. URL: https://arxiv.org/abs/2006.10214.

[43] Zhe Cao et al. “OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.1 (2021), pp. 172–186. DOI: 10.1109/TPAMI.2019.2929257.

[44] Rolandos Alexandros Potamias et al. *WiLoR: End-to-end 3D Hand Localization and Reconstruction in-the-wild*. 2025. arXiv: 2409.12259 [cs.CV]. URL: https://arxiv.org/abs/2409.12259.

[45] Joseph Redmon et al. “You Only Look Once: Unified, Real-Time Object Detection”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 779–788. DOI: 10.1109/CVPR.2016.91.

[46] Simone Borghi et al. “Automating Ergonomics: Scalable AI for Technical Hand Grip Classification”. In: *International Workshop on Annotation of Real-World Data for Artificial Intelligence Systems*. Springer. 2025, pp. 3–15.