Accelerating industrial vision: Systematic robot-assisted dataset preparation for object detection and pose estimation

Engineering Applications of Artificial Intelligence 176 (2026) 114741

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Research paper

Accelerating industrial vision: Systematic robot-assisted dataset preparation for object detection and pose estimation☆

Hao Wang [a,*], Gonzalo Urbanos Uriel [b,c], Karim El-Nahass [d,e], Sven Ekered [a], Björn Johansson [a]

[a] Department of Mechanical Engineering, Chalmers University of Technology, Hörsalsvägen 7A, Gothenburg, 412 96, Sweden
[b] Department of Electrical Engineering, Chalmers University of Technology, Hörsalsvägen 9–11, Gothenburg, 412 96, Sweden
[c] Swisslog GmbH, Ezzestraße 4–6, Dortmund, 44379, Germany
[d] Department of Physics, Chalmers University of Technology, Kemigården 1, Gothenburg, 412 96, Sweden
[e] Department of Computer Science and Engineering, University of Gothenburg, Rännvägen 6B, Gothenburg, 412 58, Sweden

ARTICLE INFO

Keywords:
Data collection automation
Robotic data acquisition
Automatic data annotation
Robot vision
Collaborative robotic automation

ABSTRACT

The creation of large-scale, high-quality training datasets continues to present a significant challenge for the implementation of artificial intelligence in engineering and industrial robotics. This study introduces a collaborative robot-assisted pipeline that automates data acquisition and annotation, thereby accelerating dataset preparation for object detection and six-degree-of-freedom pose estimation. The proposed system integrates robotic kinematics and image processing to generate vision datasets with multimodal ground-truth labels, such as two-dimensional and three-dimensional bounding boxes, segmentation masks, six-degree-of-freedom poses, and point clouds, within a unified artificial intelligence-driven workflow. To demonstrate the pipeline's capacity to reduce manual effort and efficiently generate large-scale training datasets for industrial vision applications, an automotive wire harness connector dataset was experimentally prepared using the proposed pipeline. This method achieved annotation speeds approximately 150 times faster than traditional manual techniques and produced high-quality training data for deep learning models. Evaluation with deep learning-based object detection and pose estimation algorithms confirms the effectiveness of the proposed pipeline in preparing datasets for the development of industrial intelligent vision systems. By minimizing human intervention and ensuring systematic viewpoint coverage during dataset preparation, the proposed approach facilitates scalable adoption of artificial intelligence-powered vision systems in industrial automation. The proposed method and code are available at https://github.com/HWANG7308/AutoTrainingDataPrepare.

1. Introduction

The integration of vision systems has become increasingly prevalent in modern manufacturing to support digital transformation and automation (Yang et al., 2021; Guang et al., 2025). Typical applications include quality inspection, which verifies the presence and type of small components and detects defects (Jha and Babiceanu, 2023); assembly verification, which confirms correct placement and orientation prior to downstream operations (Pang et al., 2023); and part recognition for robotic bin-picking, which requires accurate six-degree-of-freedom (6DoF) pose estimation to obtain part position and orientation to enable reliable manipulation (Li et al., 2022). Vision-based perception offers significant advantages in these contexts by enabling non-contact, information-rich sensing (Sharma et al., 2023). Additionally, it can be reconfigured through software to accommodate changes in product variants and operating conditions. In contrast, traditional approaches such as dedicated fixtures, gauges, jigs, or contact-/probe-based inspection typically require task-specific hardware design and retooling, and they do not naturally provide the dense spatial information needed for object detection and 6DoF pose estimation (Yousif et al., 2025).

Recent advances in deep learning have further enhanced the ability of vision systems to perform recognition, inspection, and pose estimation in complex industrial environments (LeCun et al., 2015; Liu et al., 2025). Nevertheless, the limited availability of high-quality, large-scale, domain-specific annotated datasets remains a fundamental constraint on the development and deployment of learning-based perception models (Sun et al., 2017; Zhang et al., 2025). This limitation is particularly acute in domain-specific contexts where annotated training

☆ This article is part of a Special issue entitled: 'AI in Control' published in Engineering Applications of Artificial Intelligence.

* Corresponding author.
E-mail addresses: haowang@chalmers.se (H. Wang), gonurbanos@gmail.com (G. Urbanos Uriel), karimnah01@hotmail.com (K. El-Nahass), sven.ekered@chalmers.se (S. Ekered), bjorn.johansson@chalmers.se (B. Johansson).

https://doi.org/10.1016/j.engappai.2026.114741
Received 4 October 2025; Received in revised form 20 February 2026; Accepted 4 April 2026
Available online 15 April 2026

data are scarce (Li et al., 2025) and manual data collection is labor-intensive, time-consuming, and demands specialized expertise (Gygli and Ferrari, 2020; Lyu et al., 2024). As a result, the adoption of these advanced vision systems is often constrained by the difficulty of collecting datasets that capture relevant viewpoints and operational variations in practice (Roh et al., 2021; Zhang et al., 2026; Zhou et al., 2023).

To mitigate dataset scarcity, researchers have investigated both synthetic and real-world approaches for dataset generation (Roh et al., 2021). Synthetic data offers scalability and flexibility (Patki et al., 2016), but it frequently exhibits a simulation-to-reality gap and requires domain-specific modeling and real-world validation (Jakobi et al., 1995). Real-world data collection provides higher fidelity, yet it is commonly associated with complex and costly acquisition setups, including specialized gantries, turntables, multi-camera rigs, manual manipulation, and advanced control software (Seitz et al., 2006; Kasper et al., 2012; Singh et al., 2014; Hodaň et al., 2017). These factors constrain scalability and increase the overhead of dataset creation, especially when rapid iteration across different part variants is required.

Recent work has introduced robotic systems to automate aspects of training data preparation (Jensen et al., 2014; Kiyokawa et al., 2021; Koch et al., 2023; Zürn et al., 2024). While these approaches reduce hardware complexity and enable flexible data acquisition, they often provide only limited or ad hoc viewpoint coverage. Furthermore, they generally lack an end-to-end workflow for systematically generating the multimodal labels required by contemporary detection and 6DoF pose estimation methods.

To overcome these limitations, this study introduces a robot-assisted pipeline for real-world dataset preparation targeting object detection and 6DoF pose estimation. The primary contribution is the integration of systematic viewpoint planning with an automated multimodal annotation workflow within a unified pipeline. The pipeline consists of two primary modules. The first is a robotic data acquisition module that captures red-green-blue-depth (RGB–D) data from camera poses distributed on a hemisphere around the target object. Observation poses are generated in spherical coordinates to ensure uniform coverage, while allowing practitioners to adjust viewpoint density according to task requirements. The second module is an automated data annotation system that combines image processing with robot kinematics to generate multimodal labels, including two-dimensional (2D) and three-dimensional (3D) bounding boxes (BBoxes), segmentation masks, 6DoF object poses, and point clouds. Unlike many existing real-world acquisition pipelines (Seitz et al., 2006; Kasper et al., 2012; Singh et al., 2014; Jensen et al., 2014; Hodaň et al., 2017; Kiyokawa et al., 2021; Koch et al., 2023; Zürn et al., 2024), the proposed workflow reduces manual labeling effort and eliminates the need for additional hardware such as turntables or multi-camera rigs, while maintaining comprehensive annotations suitable for training contemporary perception models.

A case study on automotive wire harness connectors is conducted to demonstrate the efficiency and effectiveness of the proposed pipeline. In this domain, the absence of annotated datasets has impeded progress toward automated wire harness assembly (H. Wang et al., 2024). An RGB–D connector dataset is collected using the proposed pipeline. Subsequently, deep learning models are trained and evaluated for 2D object detection and 6DoF pose estimation, demonstrating the practical utility and adaptability of the approach.

Overall, the proposed approach reduces the primary costs of dataset creation, including repeatable multi-view acquisition and labor-intensive annotation, without the need for complex multi-camera calibration, turntable synchronization, or extensive manual labeling. Human involvement is limited to object placement and initiation of the acquisition process. The workflow also supports rapid re-collection when new variants are introduced or operating conditions change. The robotic system can be repurposed for other tasks after data acquisition, thereby improving equipment utilization. Additionally, the modular design facilitates generalization to other object categories and application scenarios.

2. Related work

Preparing real-world datasets typically involves two coupled stages: (i) data acquisition and (ii) ground-truth annotation. Numerous studies have proposed methods to simplify or automate one or both stages, with the objective of reducing manual effort and improving scalability. However, existing approaches often exhibit persistent limitations in terms of hardware complexity, viewpoint coverage, and annotation completeness. These challenges motivate the end-to-end pipeline proposed in this study.

2.1. Automated real-world data acquisition

Hardware-intensive acquisition systems with systematic viewpoint coverage. Early hardware-assisted systems, such as the Stanford Spherical Gantry,[1] employed computer-controlled gantries for sensor and lighting positioning (Seitz et al., 2006). Later systems simplified the design using turntables with either a single camera moving on a circular rail (Kasper et al., 2012) or multiple static cameras arranged on quarter-circular arcs (Singh et al., 2014; Kimble et al., 2022). These configurations facilitate automated data collection from viewpoints systematically distributed across a hemisphere centered on the object. Although these systems provide systematic viewpoint coverage, they typically require precise mechanical construction and calibration. The complexity and associated costs of these hardware systems constrain their scalability. In addition, multi-camera rigs may suffer from cross-camera interference that degrades data quality (Kimble et al., 2022).

Reduced-hardware and robot-assisted acquisition systems. To reduce hardware overhead, Hodaň et al. (2017) introduced a manually adjustable jig with triplet sensors, replacing both circular rails (Kasper et al., 2012) and multi-camera rigs (Singh et al., 2014; Kimble et al., 2022). Kiyokawa et al. (2021) further automated a related configuration by moving the camera using a robotic arm while a turntable rotates the object, thereby covering both polar and azimuth directions. Recent approaches have replaced both camera rigs and turntables with robotic systems. For example, Jensen et al. (2014), Koch et al. (2023), and Zürn et al. (2024) employed robotic arms to position cameras around objects at predefined camera poses (positions and orientations). Grenzdörffer et al. (2020) mounted 3D cameras on a UR5 robot to follow predefined trajectories, and Lee et al. (2021) used a robotic arm to acquire views from a horizontal plane above objects for detection. Although these systems increase flexibility and reduce reliance on dedicated hardware, viewpoint selection is frequently ad hoc or customized for specific tasks. Systematic viewpoint coverage is often neither ensured nor documented.

Object manipulation as an alternative to camera motion. Some methods achieve pose diversity by manipulating the object rather than moving the camera. Pattar et al. (2023) used mobile robots to reposition objects, while Chen et al. (2023) used a robotic arm to rotate objects relative to a fixed camera. This strategy reduces camera system complexity but may result in a limited range of object poses (Pattar et al., 2023) or introduce self-occlusions caused by the robot in the captured data (Chen et al., 2023).

Acquisition systems that record camera poses for downstream labeling. Robotic acquisition systems also enable recording camera poses through kinematic data, which can facilitate annotation. Elsharkawy and Kim (2022) recorded annotations from tracked human hand motion. Grenzdörffer et al. (2020), De Gregorio et al. (2020), and Ilin et al. (2021) used eye-in-hand configurations to propagate labels from manually annotated keyframes. Koch et al. (2023) and Zürn et al. (2024) exploited

[1] http://www.graphics.stanford.edu/projects/gantry/

robot kinematics to enable 6DoF object pose annotation. However, many robot-assisted acquisition methods lack a unified and systematic framework for viewpoint generation that both ensures scalable coverage and remains adaptable to practical constraints.

Summary and gap. Prior work reveals a recurring trade-off: hardware-intensive systems provide systematic viewpoint coverage but are costly and challenging to scale, whereas robot-assisted systems offer greater flexibility but frequently lack systematic viewpoint planning or comprehensive end-to-end dataset preparation workflows. This observation motivates the acquisition component of the proposed pipeline, which integrates configurable and systematic viewpoint planning with robot-assisted capture in a practical setup.

2.2. Automated real-world data annotation

Human-in-the-loop and learning-based annotation. To minimize manual effort, researchers have explored enhanced annotation interfaces (da Silva et al., 2020; Lyu et al., 2024), gamification strategies (Kavasidis et al., 2013; Kiyokawa et al., 2025), and the use of extended reality (Wirth et al., 2019). Semi-automatic annotation tools integrate algorithmic suggestions with human refinement (Benenson et al., 2019; Adhikari et al., 2021; Stumpf et al., 2021). The effectiveness of these methods depends substantially on user experience and operational efficiency (Pande et al., 2022). Learning-based automatic labeling methods have also been studied; however, their reliability typically relies on the availability of high-quality training data for label generation models (Wong et al., 2015; Alshehri et al., 2022; Geiß et al., 2023; X. Wang et al., 2024).

Automating 2D/3D labels via background removal and multimodal sensing. Object-background separation is a fundamental prerequisite for generating segmentation masks and bounding boxes. Chroma keying with uniform backgrounds, typically green (Sapp et al., 2008; Kiyokawa et al., 2021; Zanella et al., 2021; Kimble et al., 2022) or white (Singh et al., 2014), is widely adopted. However, this method is sensitive to lighting conditions and to color similarity between the object and background. Lee et al. (2021) mitigated this issue using an LCD monitor with adjustable background colors. Alternative approaches employ scene differencing. Suchi et al. (2019) introduced objects incrementally to detect foreground changes, whereas Kleeberger et al. (2019) removed objects sequentially in bin-picking scenarios. Koch et al. (2023) used background subtraction based on images captured before and after object placement. Additional sensing modalities, such as depth and point clouds, further support segmentation by enabling 3D data cropping and projection of masks onto 2D images (Chen et al., 2023). However, accurate annotation of small objects often necessitates high-resolution sensors, which can increase costs and may not always be accessible (Cop et al., 2021).

Approaches to facilitate 6DoF pose annotation. Annotating 6DoF object poses presents significantly greater challenges compared to annotating 2D or 3D bounding boxes and masks. Many studies rely on artificial fiducial markers (e.g., ArUco Garrido-Jurado et al., 2014, 2016) to support pose labeling (Hinterstoisser et al., 2013; Hodaň et al., 2017; Romero-Ramirez et al., 2018; Grenzdörffer et al., 2020; Chen et al., 2022; Viviers et al., 2024). Høffer et al. (2023) attached markers directly to objects for automatic pose labeling, while Pattar et al. (2023) introduced customized invisible markers. Kiyokawa et al. (2019b,a) used visual markers with designed placements and later exploited marker-based camera pose estimation to annotate object poses (Kiyokawa et al., 2021). Although marker-based approaches are effective, they introduce additional operational steps and can limit practical deployability.

Another line of work propagates labels across frames using camera tracking (Elsharkawy and Kim, 2022; Caporali et al., 2023). These methods often require manual labeling of initial frames and accurate camera pose estimation, obtained via markers (Shuichi and Manabu, 2019), robot kinematics (De Gregorio et al., 2020), or Simultaneous Localization and Mapping (SLAM) (Elsharkawy and Kim, 2022). More recently, kinematics-based approaches derive pose labels directly from known transformations among the robot, camera, and object. Koch et al. (2023) and Zürn et al. (2024) annotated 6DoF poses using robot kinematics under predefined observation poses. However, these approaches do not provide a systematic viewpoint generation framework that guarantees comprehensive coverage or maintains a practical, configurable workflow.

Summary and gap. Current annotation methods generally require trade-offs among flexibility, scalability, and completeness of annotation. Marker-based approaches facilitate pose labeling but introduce operational overhead. In contrast, kinematics-based approaches reduce reliance on markers but are often not integrated with systematic viewpoint planning or comprehensive multimodal label generation. This motivates the design of the annotation component in the proposed pipeline, which integrates lightweight image processing with robot kinematics to generate multimodal labels while minimizing human intervention.

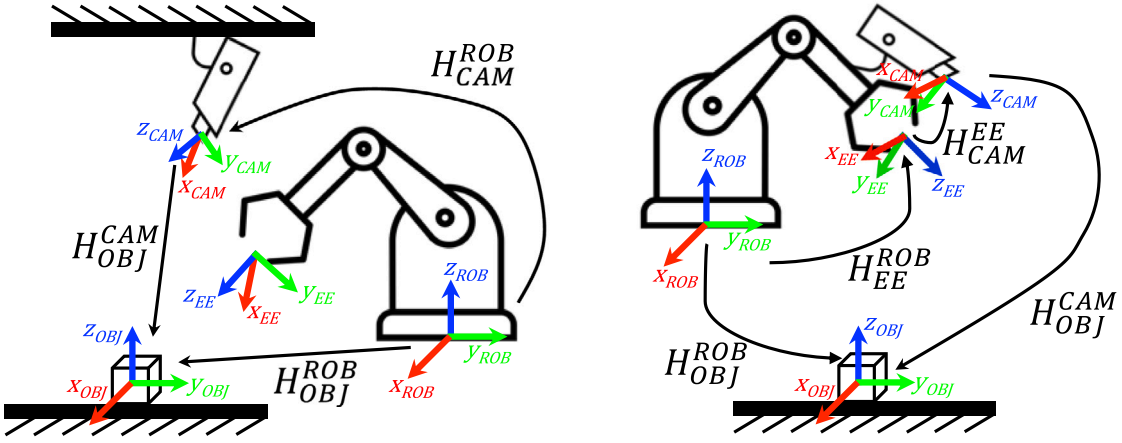
2.3. Positioning of the present work

Previous research has achieved systematic viewpoint acquisition using hardware-intensive rigs, improved flexibility through robot-assisted capture without guaranteeing comprehensive viewpoint coverage, or automated annotation processes that continue to rely on markers, manual key frames, or incomplete label sets. The present work addresses this gap by introducing an end-to-end robot-assisted dataset preparation pipeline that integrates configurable, systematic viewpoint planning with automatic multimodal annotation, including 2D and 3D BBoxes, masks, and 6DoF poses. This approach utilizes robot kinematics and spherical coordinates to systematically define observation viewpoints and applies image processing techniques to generate multimodal annotations, thereby removing the need for turntables, multi-camera rigs, or external fiducial markers.

3. Research methodology

The Design Science Research Methodology (DSRM) (Hevner and Chatterjee, 2010) is utilized to address the challenge of efficiently and scalably preparing datasets for industrial robotic vision systems. In design science research (DSR), researchers, as designers, contribute scientific knowledge by creating innovative artifacts that are both useful and fundamental to understanding and addressing human problems (Hevner and Chatterjee, 2010). Artifacts include constructs, models, methods, and instantiations that facilitate the transition from the current state to a desired state (Gregor and Hevner, 2013; Hevner et al., 2004). Following the DSRM framework (Peffers et al., 2007), this study is structured to focus on the design and development of a robot-assisted pipeline for generating real-world datasets, followed by demonstration and evaluation of the pipeline's performance. This design-oriented approach ensures that the developed artifact achieves both practical relevance and undergoes rigorous evaluation. The key components of the methodology are summarized below.

Problem identification and motivation. Prior research identifies the lack of high-quality, domain-specific datasets as a significant barrier to deploying deep learning-based object detection and pose estimation in industrial automation (Zhou et al., 2023). Manual dataset preparation is time-consuming and labor-intensive (Lyu et al., 2024), particularly for small, texture-less objects such as automotive wire harness connectors (Wang and Johansson, 2023). Current automation efforts in dataset preparation frequently require trade-offs among flexibility (Singh et al., 2014), scalability (Koch et al., 2023), and annotation comprehensiveness (Lee et al., 2021). These limitations motivate research into enhanced dataset generation pipelines to advance the development of vision systems, particularly for industrial applications.



(a) Eye-to-hand (b) Eye-in-hand

Fig. 1. Two typical configurations for the camera's location in a vision-guided robotic system.

Objective definition. Based on the identified problem and motivation, the objective is to design a dataset preparation pipeline that reduces human effort, streamlines hardware requirements, and enables systematic data acquisition and annotation for object detection and 6DoF pose estimation.

Design and development. The proposed solution comprises a robotic data acquisition module and an automatic data annotation module. The robotic data acquisition module employs a collaborative robotic arm equipped with an eye-in-hand camera to capture multi-view images from systematically generated observation poses. The automatic data annotation module labels acquired data using image processing and robotic kinematics. The pipeline produces multimodal ground-truth labels, including 2D and 3D BBoxes, segmentation masks, 6DoF poses, and point clouds. Further details are provided in Sections 4 and 5.

Demonstration and evaluation. To demonstrate the effectiveness of the pipeline, a dataset containing twenty types of wire harness connectors will be generated. Deep learning models for object detection and 6DoF pose estimation will be trained and evaluated using standard metrics to assess the performance of the proposed pipeline. Section 6 presents an experiment involving the preparation of an automotive wire harness connector dataset to demonstrate and evaluate the performance of the proposed pipeline. The advantages and limitations of the pipeline are discussed in Section 7.

4. Robot system and workflow

4.1. Robot vision system

Robotic manipulation typically requires the object pose relative to the robot base frame (H_{OBJ}^{ROB}).[2] This pose may be specified a priori or estimated from sensor data. In vision-guided settings, 6DoF object pose estimation aims to recover the object pose in the camera frame (H_{OBJ}^{CAM}) (Marullo et al., 2023). Fig. 1 illustrates two standard camera placement configurations for these systems.

4.1.1. Eye-to-hand configuration

Fig. 1(a) illustrates an eye-to-hand vision configuration, where the camera is fixed within the workspace and functions as a global sensor independent of the robot. In this configuration, the object pose relative to the robot base (H_{OBJ}^{ROB}) is computed using Eq. (1). The camera-to-robot-base transformation (H_{CAM}^{ROB}) is constant and obtained through hand–eye calibration, while the object pose relative to the camera (H_{OBJ}^{CAM}) is estimated via 6DoF pose estimation.

$$H_{OBJ}^{ROB} = H_{CAM}^{ROB} \cdot H_{OBJ}^{CAM} \tag{1}$$

This relationship motivates the use of kinematics-assisted dataset generation. Specifically, if the object is rigidly attached to the robot end effector such that H_{OBJ}^{EE} is constant, then for a given robot pose (H_{EE}^{ROB}) the corresponding object pose in the camera frame can be computed using Eq. (2).

$$H_{OBJ}^{CAM} = {H_{CAM}^{ROB}}^{-1} \cdot H_{EE}^{ROB} \cdot H_{OBJ}^{EE} \tag{2}$$

A practical advantage of eye-to-hand setups is stable imaging, since the camera is decoupled from robot motion. However, maintaining a uniform background across varying object elevations is more difficult, which can complicate foreground–background separation. Incorporating depth or point-cloud cues can alleviate this, but reliable segmentation of small objects may require higher-precision sensing.

4.1.2. Eye-in-hand configuration

Fig. 1(b) shows an eye-in-hand configuration, in which the camera is mounted on the robot wrist and rigidly attached to the end effector with a constant transform (H_{CAM}^{EE}, estimated via hand–eye calibration). This setup facilitates the use of a uniform background and reduces the likelihood of robot self-occlusion during image acquisition. Therefore, the proposed pipeline adopts an eye-in-hand configuration.

In this setting, the object pose in the robot base frame (H_{OBJ}^{ROB}) is computed using Eq. (3), where H_{EE}^{ROB} is provided by the robot controller and H_{OBJ}^{CAM} is obtained from 6DoF pose estimation.

$$H_{OBJ}^{ROB} = H_{EE}^{ROB} \cdot H_{CAM}^{EE} \cdot H_{OBJ}^{CAM} \tag{3}$$

This formulation underlies prior kinematics-based 6DoF object pose labeling approaches (Koch et al., 2023; Zürn et al., 2024), which capture images at a predefined set of observation poses and compute the corresponding object poses using Eq. (4). However, these methods typically rely on heuristic or manually specified pose sets, which may limit viewpoint coverage and scalability (Koch et al., 2023; Zürn et al., 2024).

$$H_{OBJ}^{CAM} = {H_{CAM}^{EE}}^{-1} \cdot {H_{EE}^{ROB}}^{-1} \cdot H_{OBJ}^{ROB} \tag{4}$$

In contrast, the proposed pipeline first specifies the desired acquisition viewpoints (i.e., target object poses relative to the camera) and then computes the corresponding robot end-effector poses and joint configurations required to realize them. The systematic viewpoint generation procedure is described in Section 5.1. This design enables configurable and scalable sampling while ensuring that the object is consistently captured in the camera's field of view with the intended pose.

[2] Unless otherwise specified, transformations between coordinate systems are represented by 4 × 4 homogeneous transformation matrices H.

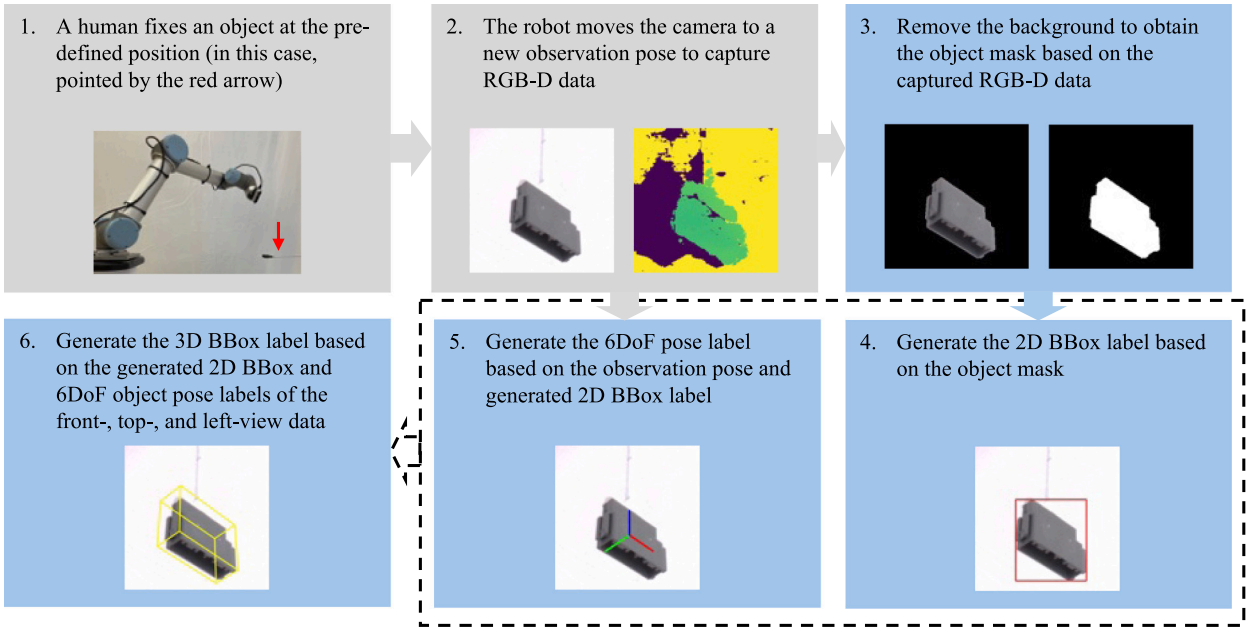


Fig. 2. The proposed robot-assisted data collection pipeline (step 1 to 2 for robotic data acquisition and step 3 to 6 for automatic data annotation).

4.2. Conceptual design of the robot system

Traditional gantry-based systems acquire multi-view visual data by sampling viewpoints distributed over a hemisphere centered on an object placed at the hemisphere's center (Seitz et al., 2006; Kasper et al., 2012; Singh et al., 2014; Kimble et al., 2022). This configuration is conceptually analogous to using a robotic arm to maneuver an eye-in-hand camera through a series of viewpoints distributed on a hemisphere centered on the target object. Accordingly, configuring the robotic acquisition system involves first selecting the hemisphere radius (r), which defines the observation distance, followed by positioning the hemisphere at a feasible location within the robot workspace. The minimum observation distance is constrained by the camera's operating range, whereas the maximum distance is limited by the robot's reach.

The hemisphere placement further determines the spatial relationship between the object and the robot. Notably, the object placement for data acquisition is not predetermined; instead, it is influenced by the choice of robot and camera, as well as the practitioner-specified observation distance.

To support automatic annotation via chroma keying, the robot cell is enclosed with a uniform background, as described in Section 5. Fig. 10 presents an example of an acquisition system configured according to this conceptual design. Section 6 details the implementation settings used in the industrial case study.

4.3. Workflow for dataset preparation

Fig. 2 presents the proposed pipeline for systematic preparation of real-world datasets with a robotic system. The pipeline consists of two primary components: robotic data acquisition and automatic data annotation.

The robotic data acquisition component involves two sequential steps. Initially, an operator positions the target object at a predefined location and ensures it remains stationary relative to the robot base. After data acquisition is initiated, a set of acquisition viewpoints (observation poses) is systematically generated. The robot subsequently maneuvers the eye-in-hand camera through these observation poses, capturing a pair of aligned RGB–D images at each pose to produce a multi-view collection of raw data. To prevent self-occlusion, robot motion is constrained so that no manipulator link enters the viewpoint-sampling hemisphere, which is the region from which images are acquired. This constraint is implemented by specifying permissible joint-angle ranges and verifying joint-limit constraints before executing each motion to the next observation pose.

The automatic data annotation component consists of five steps. First, the system segments the object region and generates an object mask using image processing and depth thresholding. A uniform background enables background removal through chroma keying. Second, the system derives 2D BBox labels from the segmented object images. Third, it computes 6DoF object pose labels by combining the a priori object poses used for viewpoint generation with the resulting 2D BBox annotations. Fourth, it generates 3D BBox labels by leveraging the 2D BBox annotations from three orthographic views of the object. Finally, it reconstructs an object point cloud by fusing the depth images across viewpoints. The outputs include ground-truth labels for 2D and 3D BBoxes, segmentation masks, 6DoF poses, and point clouds, each associated with the corresponding RGB–D images.

This workflow is not fully automated. Human intervention is necessary for object placement and for initiating the acquisition and annotation software program. Additionally, before data acquisition, the camera intrinsic calibration matrix must be obtained, and observation poses must be generated to guide the robot in maneuvering the eye-in-hand camera. Several established procedures can be used to estimate the camera intrinsic calibration matrix (Lundberg et al., 2014); detailed discussion is omitted for brevity. The observation poses specify the end-effector positions and orientations required to capture the planned views and are subsequently converted into robot joint configurations. Section 5 provides further details on systematic observation-pose generation using spherical coordinates and on ground-truth label computation via image processing and robotic kinematics.

5. Implementation details

5.1. Systematic generation of observation poses

Prior robot-assisted pipelines for collecting multi-view datasets with kinematics-derived 6DoF pose labels typically execute a user-specified set of observation poses and then compute the corresponding object

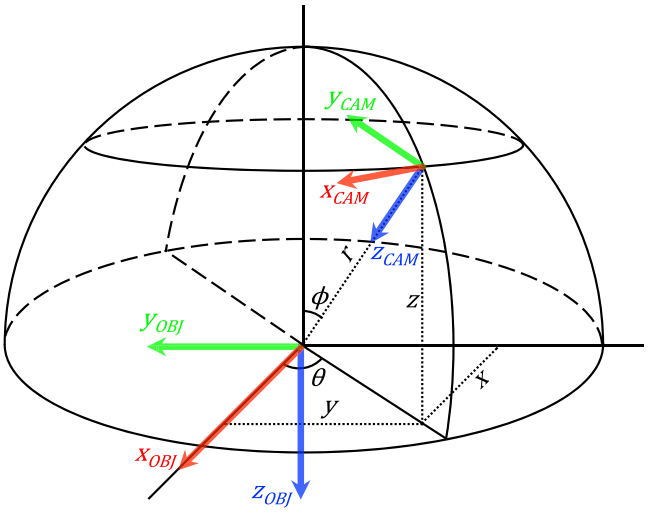


Fig. 3. Observation pose calculation using a spherical coordinate system.

poses using Eq. (4) (Koch et al., 2023; Zürn et al., 2024). Because the observation poses are usually defined heuristically or manually, viewpoint coverage and scalability across different objects and application requirements can be limited.

In contrast, this study adopts a viewpoint-first strategy. Specifically, desired object poses in the camera frame (H_{OBJ}^{CAM}) are specified first, and the required end-effector observation poses (H_{EE}^{ROB}) are then computed analytically using Eq. (5), where H_{OBJ}^{ROB} is fixed by the robot-cell configuration, H_{CAM}^{EE} is obtained via hand–eye calibration, and H_{OBJ}^{CAM} encodes the desired acquisition viewpoint. This formulation enables systematic acquisition and provides direct control of the sampled viewpoints. When a specific object pose is required, the corresponding robot pose can be computed and executed without manual trial-and-error.

$$H_{EE}^{ROB} = H_{OBJ}^{ROB} \cdot {H_{OBJ}^{CAM}}^{-1} \cdot {H_{CAM}^{EE}}^{-1} \tag{5}$$

To generate viewpoints uniformly distributed on a hemisphere centered on the object, the camera pose relative to the object frame (H_{CAM}^{OBJ}) is parameterized in spherical coordinates (Fig. 3). The radius r specifies the Euclidean observation distance. The orientation component is defined by an azimuth angle θ and a polar angle ϕ using an Euler-angle construction: a rotation of θ about the z-axis followed by a rotation of $-\phi$ about the rotated x-axis, with $\theta \in [0, 2\pi]$ and $\phi \in [0, \pi/2]$ to cover the upper hemisphere. The translation vector (x, y, z) is then computed using Eq. (6).

$$\begin{cases} x = r \cdot \sin(\phi) \cdot \sin(\theta) \\ y = -r \cdot \sin(\phi) \cdot \cos(\theta) \\ z = -r \cdot \cos(\phi) \end{cases} \tag{6}$$

Discretizing θ and ϕ with step sizes s_θ and s_ϕ yields viewpoints located at the vertices of the upper half of a UV sphere, comprising π/s_ϕ rings and $2\pi/s_\theta$ segments. This construction is analogous to turntable-based acquisition (Singh et al., 2014; Kimble et al., 2022), which can be interpreted as sampling π/s_ϕ camera elevations while rotating the object by increments of $2\pi/s_\theta$. Unlike static rigs, the proposed method allows practitioners to adjust both viewpoint density (via s_θ, s_ϕ) and observation distance (via r) to meet task requirements and hardware constraints. Fig. 4 illustrates example viewpoint sets generated with different step-size choices.

5.2. 2D bounding box generation

In 2D object detection, ground-truth labels are represented by 2D BBoxes, defined as axis-aligned rectangles that enclose target objects. Fig. 5 illustrates the procedure for generating these 2D BBoxes.

Data collection employs an RGB–D camera, which enables reliable segmentation of the object of interest using chroma keying against a uniform background. The inclusion of depth information further refines background removal and increases robustness. After this process, the object is isolated on a transparent background, and a corresponding 2D object mask is generated. This mask may also serve as ground truth for segmentation tasks.

The 2D BBox is then computed from the object mask. The leftmost, rightmost, topmost, and bottommost boundary pixels of the mask, illustrated as four red points in Fig. 5, are identified. The coordinates of these points define the extent of the axis-aligned 2D BBox in image space.

5.3. 6DoF object pose annotation

In 6DoF object pose estimation, the ground-truth label is defined as the rigid transformation from the camera frame to the object frame (H_{OBJ}^{CAM}). As detailed in Section 5.1, for a specified H_{OBJ}^{CAM}, the corresponding robot end-effector observation pose (H_{EE}^{ROB}) is calculated using Eq. (5). Accordingly, the nominal 6DoF pose associated with each captured image is predetermined and utilized to guide the robot in positioning the eye-in-hand camera.

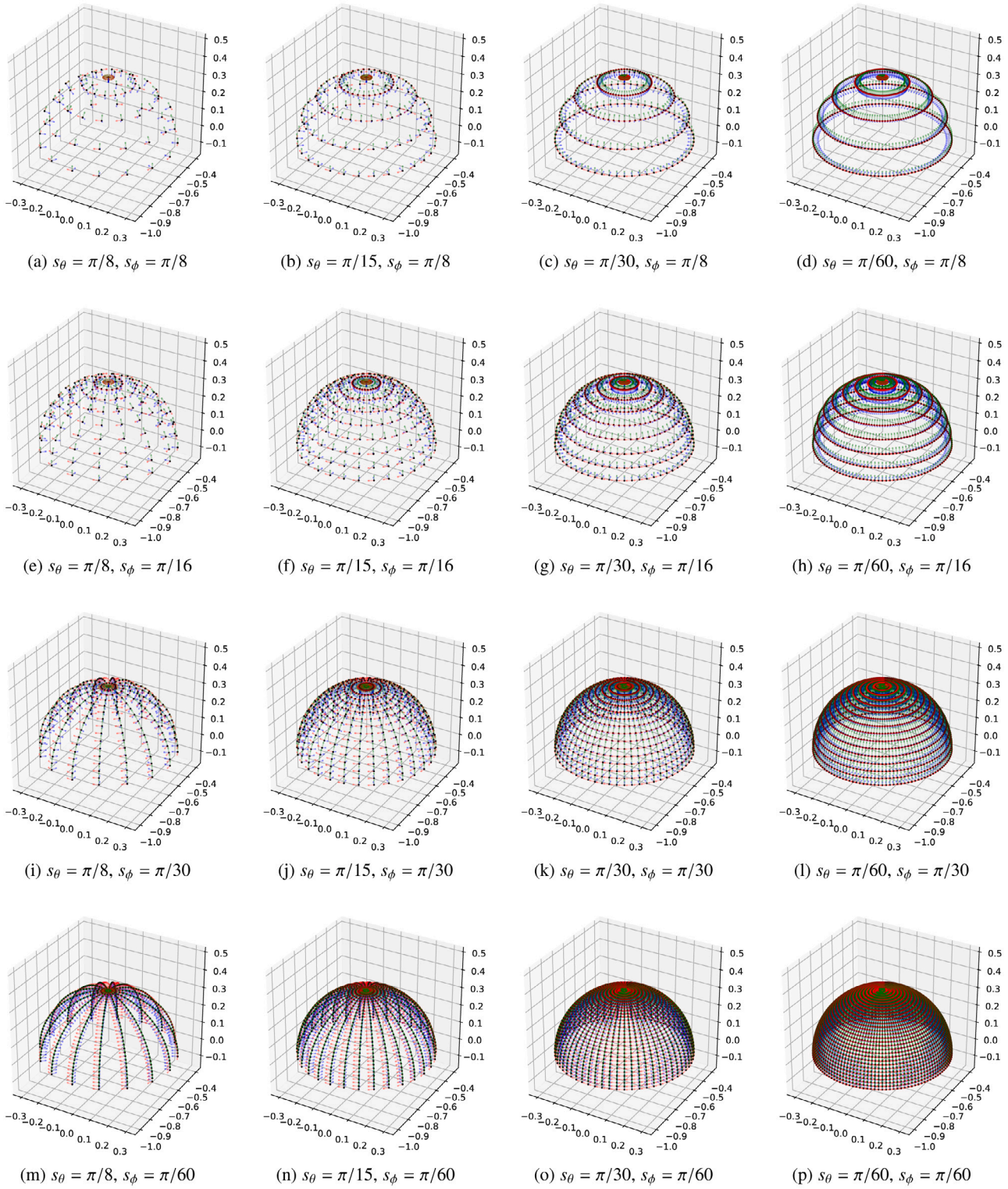
The objective of annotation is to represent the object pose as accurately as possible relative to the object centroid, defined as the object's geometric center. In practice, maintaining the centroid at an exact, predefined location during object placement presents challenges for human operators. To enhance repeatability, the center point of the object's bottom surface (yellow points in Fig. 6) is aligned with a predefined location, and observation poses are computed relative to this bottom-center reference.

Subsequently, the 6DoF pose label is refined to align with the object's geometric center. This refinement utilizes the 2D BBox annotations obtained from three orthographic views (front, top, and left), as illustrated in Fig. 6. These orthographic views are defined relative to the object frame (OBJ). Specifically, the front view is generated by observing the object along its positive z-axis, the top view along its positive y-axis, and the left view along its positive x-axis. In each view, the object centroid is projected onto the corresponding image plane and constrained to coincide with the center of the 2D BBox (purple points in Fig. 6). The three projected centers collectively determine the 3D location of the object centroid, enabling estimation of the transformation from the bottom-center reference to the geometric center (H_{OBJ-C}^{OBJ}). The offset between the yellow and purple points in the front view captures translation along the x and y axes. The offset observed in the top view provides the additional constraint required to recover translation along the z-axis. The resulting refined pose (H_{OBJ-C}^{CAM}) serves as the final 6DoF pose annotation and is computed using Eq. (7).

$$H_{OBJ-C}^{CAM} = H_{OBJ}^{CAM} \cdot H_{OBJ-C}^{OBJ} \tag{7}$$

5.4. 3D bounding box generation

Fig. 7 presents the procedure for generating 3D BBox labels, which are represented as amodal cuboids enclosing the target objects. A cuboid aligned with the object frame is reconstructed from the 2D BBoxes obtained in three orthographic views, in accordance with three-view-based 3D reconstruction principles (Hu et al., 2023; Phuong et al., 2024). The front, top, and left orthographic views produce the corresponding 2D BBoxes (Fig. 6(a), (b), and (c)). Collectively, these 2D BBoxes provide sufficient constraints to recover the extents of the 3D BBox along the object-frame axes. Specifically, the front-view BBox determines the extents along the x and y axes, while the top-view BBox provides the extent along the z-axis. The resulting cuboid is positioned and oriented using the corresponding 6DoF pose annotation to obtain the final 3D BBox in the camera frame.



(a) $s_\theta = \pi/8$, $s_\phi = \pi/8$ (b) $s_\theta = \pi/15$, $s_\phi = \pi/8$ (c) $s_\theta = \pi/30$, $s_\phi = \pi/8$ (d) $s_\theta = \pi/60$, $s_\phi = \pi/8$

(e) $s_\theta = \pi/8$, $s_\phi = \pi/16$ (f) $s_\theta = \pi/15$, $s_\phi = \pi/16$ (g) $s_\theta = \pi/30$, $s_\phi = \pi/16$ (h) $s_\theta = \pi/60$, $s_\phi = \pi/16$

(i) $s_\theta = \pi/8$, $s_\phi = \pi/30$ (j) $s_\theta = \pi/15$, $s_\phi = \pi/30$ (k) $s_\theta = \pi/30$, $s_\phi = \pi/30$ (l) $s_\theta = \pi/60$, $s_\phi = \pi/30$

(m) $s_\theta = \pi/8$, $s_\phi = \pi/60$ (n) $s_\theta = \pi/15$, $s_\phi = \pi/60$ (o) $s_\theta = \pi/30$, $s_\phi = \pi/60$ (p) $s_\theta = \pi/60$, $s_\phi = \pi/60$

Fig. 4. Observation poses generated given different step sizes for camera movement along azimuth and polar directions, respectively.

6. Experiments and system evaluation

The effectiveness of the proposed robot-assisted dataset preparation pipeline was demonstrated by generating a dataset of automotive wire harness connectors. This object category was chosen for its industrial significance and the requirement for precise perception in automating wire harness handling, a process that remains predominantly manual and presents ongoing production challenges (Salunkhe et al., 2023). The resulting dataset was subsequently employed to train and evaluate deep learning models for 2D object detection and 6DoF pose estimation.

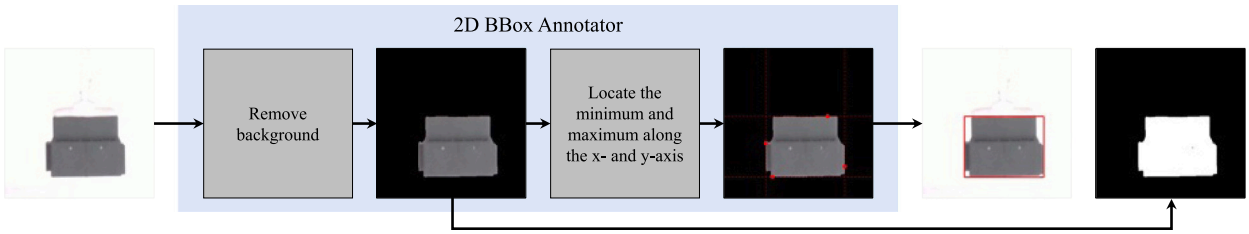
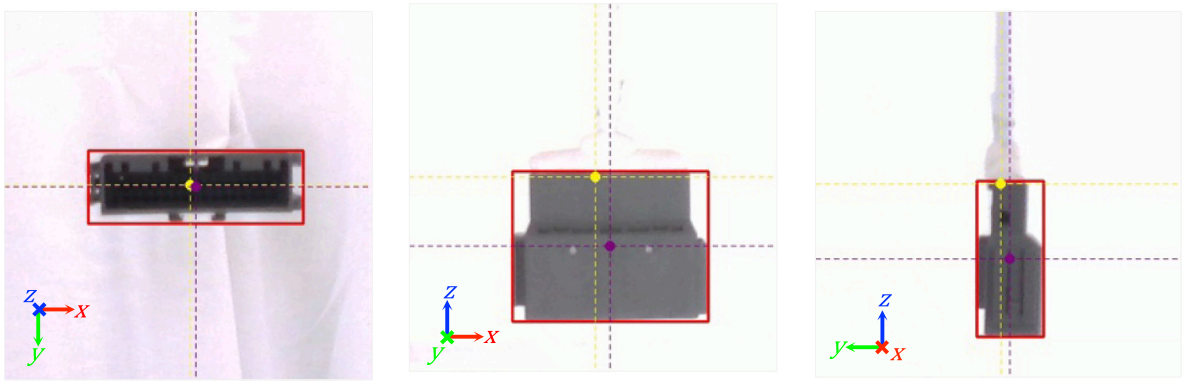


Fig. 5. The generation process of 2D bounding boxes.



(a) Front view (b) Top view (c) Left view

Fig. 6. 2D bounding boxes for three orthographic views (front, top, and left) of an object, used to refine the 6DoF pose annotation and to derive the 3D bounding box. The coordinate system shown in the bottom-left of each subfigure denotes the corresponding object frame.

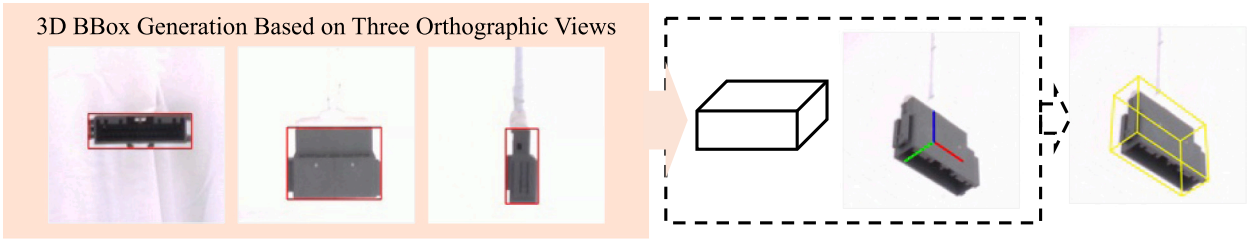


Fig. 7. The generation process of 3D bounding boxes.

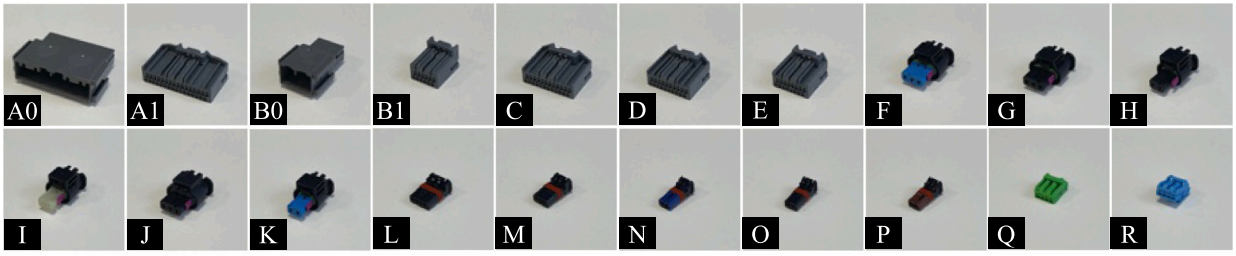


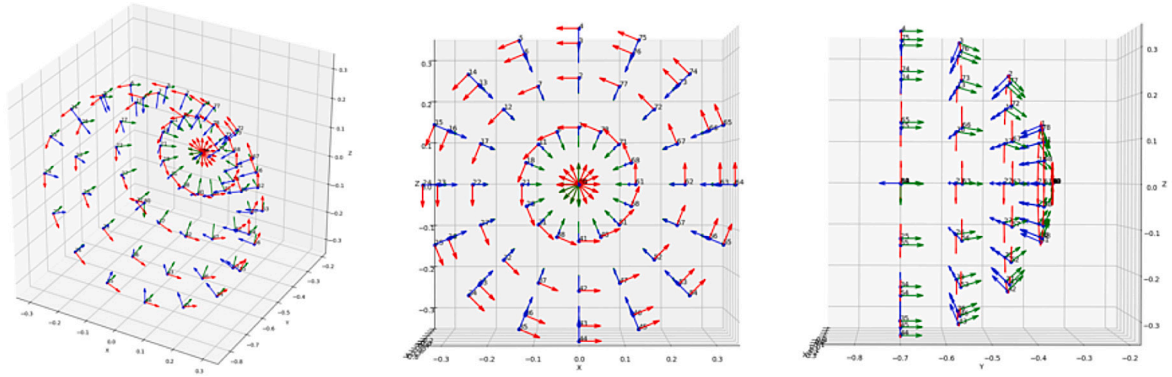
Fig. 8. The twenty types of connectors obtained for generating a real-world dataset using the proposed pipeline. The class of each sample is simplified and labeled at the bottom-left corner of each image.

Experimental results confirm that the pipeline produces high-quality training data suitable for learning-based robotic vision in real-world manufacturing environments.

6.1. Objects of interest

Fig. 8 presents the 20 connector types used in this study to demonstrate real-world dataset generation using the proposed robot-assisted pipeline.

Wire harness connectors represent primary targets for object detection and pose estimation within robotic wire harness assembly (Wang et al., 2023). Previous research on visual recognition of wire harness connectors has identified significant detection challenges requiring more comprehensive datasets and multi-view training data (Wang and Johansson, 2023). The discriminative features of these connectors are generally visible only from particular viewpoints and under optimal lighting conditions. Furthermore, these features are subtle and frequently occluded, which complicates recognition and necessitates fine-grained visual analysis (Wei et al., 2022). Additionally, the used connectors are predominantly texture-less and small, with dimensions ranging from 45.9 mm × 33.8 mm × 14.5 mm (Class A0) to 14.6 mm ×



(a) Angular view (b) Front view (c) Left view

Fig. 9. The generated observation poses (with a step size of $\pi/8$ for movement along both azimuth and polar direction) distributed uniformly on a hemisphere centered on the object.

11.7 mm × 8.7 mm (Classes Q and R). These characteristics make the dataset especially appropriate for research on detection and 6DoF pose estimation of small, texture-less industrial components.

6.2. Observation pose generation

The observation distance r was set to 300 mm to generate the wire harness connector dataset using the proposed pipeline. This value was selected based on the effective operating range of the chosen camera and the reachable workspace of the robotic system. Angular step sizes were set to $s_\theta = s_\phi = \pi/8$, yielding 80 camera positions uniformly distributed over the upper half of a UV sphere centered on the object, corresponding to 8 rings and 16 segments, as illustrated in Fig. 9.

6.3. Robot cell configuration

Fig. 10 shows the robotic data acquisition setup used to collect the connector dataset, serving as a representative implementation of the conceptual design described in Section 4.2. An Intel RealSense D435 RGB–D camera was used to acquire aligned RGB–D data. The camera was mounted to the robot flange via a 3D-printed adapter, forming an eye-in-hand configuration. A UR5 robot (Universal Robots) then positioned the camera at pre-generated viewpoints for multi-view acquisition. The robot's pose repeatability (<0.1 mm) supports consistent viewpoint execution and reliable kinematic pose propagation, enabling systematic capture of visual data at known camera poses.

The UR5 and D435 were chosen to demonstrate and validate the proposed pipeline with widely available hardware that allows for straightforward integration. The UR5 offers a mature software ecosystem and high repeatability, while the D435 provides synchronized RGB–D measurements that facilitate end-to-end validation of multimodal annotation in a laboratory setting. Notably, the proposed pipeline is hardware-agnostic and can be integrated with alternative manipulators and sensors, including higher-resolution RGB cameras, macro optics, or industrial depth or structured-light systems, when finer spatial resolution or improved depth quality is required for specific applications. Practitioners should select hardware according to the target object size, accuracy requirements, and deployment constraints, and subsequently apply the proposed pipeline to configure the acquisition system and generate the corresponding dataset.

During acquisition, the object is placed along the negative y axis of the robot base frame at a distance of 700 mm, with its front view oriented toward the robot base. Here, the side of the connector that interfaces with its mating counterpart (the insertion side) is defined as the front side. Images of the connector's back side are not collected for two reasons. First, in practical wire harness settings, the back side is often occluded by wires entering the connector. Second, in production environments, human operators can adjust connector orientation prior to assembly to improve visibility. Manual adjustment is generally more practical than relying exclusively on robotic autonomy to recognize heavily occluded connectors. Accordingly, H_{OBJ}^{ROB} is configured with a translation $(0, -700, 0)$ mm and an Euler-angle rotation $(\pi/2, 0, \pi)$ (XYZ, rad).

A rack equipped with a supporting rod is used to secure the object and maintain a stable pose relative to the robot base. Alternatively, objects may be placed on a flat table to improve stability, provided that the robotic arm is mounted laterally or that a manipulator with extended reach is available.

To facilitate background removal, the robot workspace is enclosed with white fabric and the supporting rack is painted white, providing a consistent background at all elevations. A white background was selected to avoid color conflicts with green and blue connector samples (e.g., Classes F, K, Q, and R in Fig. 8), thereby enabling efficient chroma-keying during data annotation. Although depth-based background removal can be effective in eye-to-hand configurations (Chen et al., 2023) or in more complex scenes, the small size of the connectors places higher demands on depth accuracy to reliably separate the object from the background (Cop et al., 2021). Nevertheless, depth-based background removal remains a viable alternative when higher-precision sensors are available or when larger objects are considered (Chen et al., 2023).

6.4. Camera calibration

Accurate intrinsic and extrinsic camera parameters must be obtained through camera calibration prior to data collection. In this study, the intrinsic parameters were obtained from the factory calibration provided by the sensor. The extrinsic parameters, which define the camera pose in terms of rotation and translation relative to a reference coordinate system, were estimated using hand–eye calibration.

In the eye-in-hand configuration, calibration determines the rigid transformation between the robot end-effector frame and the camera frame. Hand–eye calibration was conducted using a 9 × 6 OpenCV chessboard to estimate the end-effector-to-camera transformation (H_{CAM}^{EE}).

6.5. Automotive wire harness connector dataset

Visual data for each connector are collected using the systematic procedure described in previous sections. At each observation pose, the camera captures the RGB–D data, which are subsequently processed by the automatic annotation module to generate multimodal ground-truth labels. Specifically, the raw RGB image is initially converted to grayscale, and pixels with grayscale intensities in $[170, 255]$ are

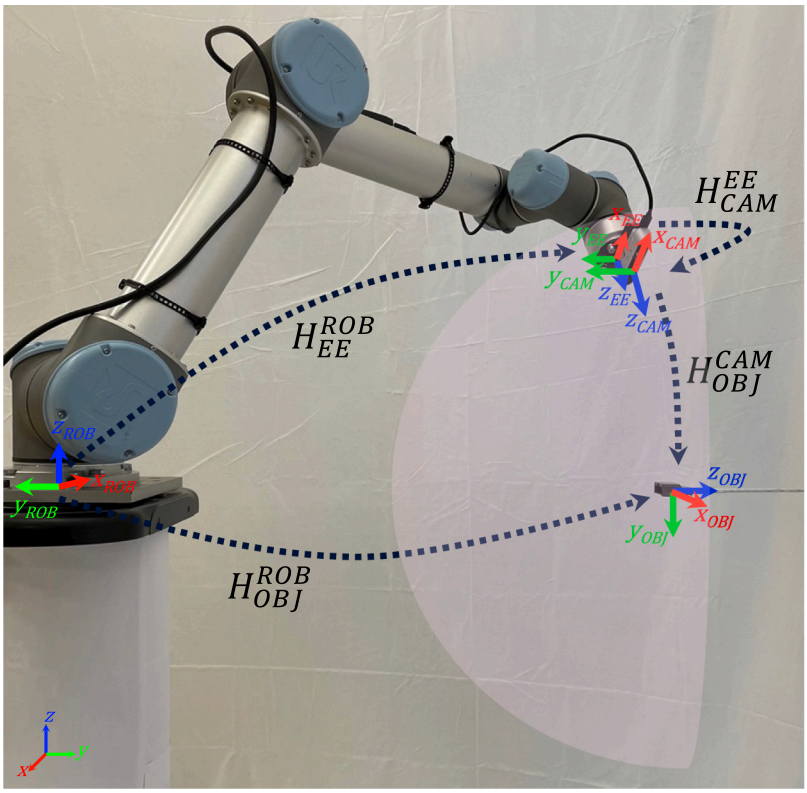


Fig. 10. Robot-assisted data acquisition setup used for collecting visual data of connectors. The translucent, light-violet hemisphere represents the viewpoint sampling surface where camera viewpoints are distributed. The coordinate system in the bottom-left corner indicates the world frame (W), which serves as the fixed reference. Additional coordinate frames are depicted at their respective physical locations: the robot base frame (ROB) at the robot base, the end-effector (tool) frame (EE) at the tool center point (TCP), the camera frame (CAM) rigidly attached to the end-effector (eye-in-hand), and the object frame (OBJ) attached to the target connector. Dotted arrows illustrate the kinematic transformations between coordinate frames.

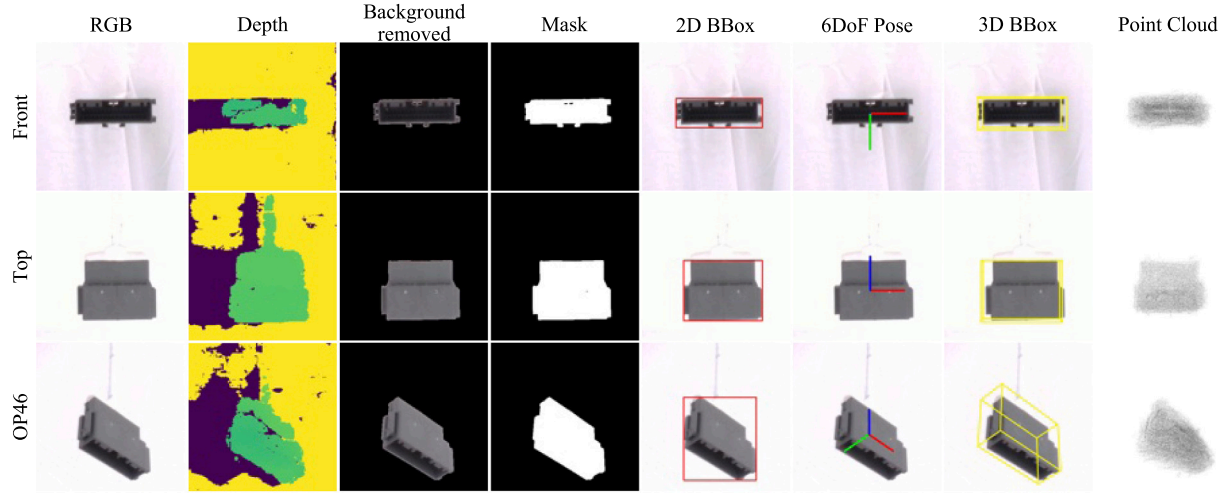


Fig. 11. Examples of collected RGB data, depth data, and corresponding annotation results. The point clouds in the last column are oriented according to their respective 6DoF pose for visualization purposes. The last row displays the image captured at observation pose 46 (OP46).

designated as the white background for chroma keying. To enhance robustness, depth-assisted background removal is performed following chroma keying, using a depth threshold of 350 mm. This threshold is determined by the observation distance and anticipated depth measurement noise.

The resulting dataset comprises 1600 RGB–D image pairs. Each image pair is annotated with an object mask, 2D and 3D BBoxes, a 6DoF object pose, and a corresponding point cloud. The raw images and generated annotations were visually inspected to verify the accuracy of the annotations. The dataset specifically targets small objects with diameters less than 60 mm, thereby increasing its relevance for small-object detection (Cheng et al., 2023) and fine-grained image analysis (Wei et al., 2022). Fig. 11 shows representative samples of the acquired data alongside visualizations of the corresponding ground-truth labels.

A manually labeled dataset was also constructed to serve as an independent baseline for comparison. The first three authors, each possessing over 100 h of experience in 2D BBox and 6DoF pose annotation, manually annotated the robot-acquired raw data. The resulting manual

Table 1
Per-class Average Precision (AP) on the Manually Annotated (MA) and Automatically Annotated (AA) datasets.

Model	Dataset	AP									
		A0	A1	B0	B1	C	D	E	F	G	H
Faster R-CNN (Ren et al., 2015)	MA	0.798	0.980	0.932	0.928	0.975	0.979	0.985	0.880	0.774	0.981
	AA	0.812	0.978	0.956	0.916	0.979	0.983	0.984	0.871	0.786	0.984
YOLOv5 (Jocher, 2020)	MA	0.968	0.991	0.970	0.989	0.988	0.792	0.854	0.972	0.901	0.970
	AA	0.971	1.000	0.973	1.000	0.998	0.785	0.884	0.969	0.866	0.970
Model	Dataset	AP									
		I	J	K	L	M	N	O	P	Q	R
Faster R-CNN (Ren et al., 2015)	MA	0.916	0.786	0.916	0.871	0.782	0.917	0.632	0.824	0.892	0.903
	AA	0.935	0.729	0.928	0.821	0.782	0.902	0.665	0.871	0.917	0.894
YOLOv5 (Jocher, 2020)	MA	0.968	0.956	0.969	0.724	0.981	0.965	0.845	0.935	0.964	0.986
	AA	0.973	0.955	0.970	0.969	0.983	0.972	0.867	0.972	0.973	0.971

Table 2
The Mean Average Precision (mAP) on the Manually Annotated (MA) and Automatically Annotated (AA) datasets.

Model	Dataset	mAP_{50}	$mAP_{50:95}$
Faster R-CNN (Ren et al., 2015)	MA	0.937	0.883
	AA	0.938	0.885
YOLOv5 (Jocher, 2020)	MA	0.981	0.947
	AA	0.988	0.971

2D BBox and 6DoF pose labels form the baseline dataset. For clarity, the dataset with automatically generated annotations is referred to as AA, while the manually annotated dataset is referred to as MA, hereinafter, unless otherwise specified.

6.6. Performance of various datasets

Experiments were conducted to assess the effectiveness of the proposed robot-assisted pipeline in generating real-world datasets for industrial 2D object detection and 6DoF pose estimation. All experiments were performed on a workstation with an Intel Core i9-13900KF CPU (3.0 GHz), 64 GB RAM, and an NVIDIA GeForce RTX 4090 GPU. Both AA and MA datasets were divided into training, validation, and test sets using an 80:10:10 split. Stratified sampling ensured balanced class distributions across the subsets, and validation and test images were randomly selected within each class.

To complete the end-to-end workflow from dataset preparation to model training and evaluation, widely used baseline methods were adopted. Faster R-CNN (Ren et al., 2015) and YOLOv5 (Jocher, 2020) were selected as representative two-stage and one-stage detectors, respectively, to illustrate the trade-off between accuracy and inference speed. DenseFusion (Wang et al., 2019) and YOLOv5-6D (Viviers et al., 2024) served as representative 6DoF pose estimators, exemplifying RGB–D fusion-based and direct regression-based paradigms, respectively. Notably, the proposed dataset preparation pipeline remains model-agnostic, allowing practitioners to pair it with alternative detection and pose estimation methods according to specific application requirements.

6.6.1. Evaluation on 2D object detection

The performance of a two-stage detector was compared to that of a one-stage detector. Both models were trained, validated, and tested using training, validation, and test sets, respectively.

The two-stage detector utilized Faster R-CNN (Ren et al., 2015) with a ResNet (He et al., 2016) backbone and a Feature Pyramid Network (FPN) (Lin et al., 2017). Baseline settings and hyperparameters were adopted from the publicly available Detectron2 (Wu et al., 2019) implementation. Training was conducted using stochastic gradient descent (SGD) with a batch size of 8 and a learning rate of 0.00025. Model weights were initialized from the Detectron2 checkpoint *Faster R-CNN R101-FPN 3x* (Wu et al., 2019). During inference, detections with confidence scores below a threshold of 0.7 were treated as background and discarded.

The one-stage detector was implemented using YOLOv5 (Jocher, 2020) with the official open-source configuration. Training utilized SGD with an initial learning rate of 0.01, weight decay of 0.0005, momentum of 0.937, and a batch size of 16. Model weights were initialized from the pretrained *yolov5x* checkpoint (Jocher, 2020).

Detection performance was evaluated using per-class average precision (AP) and mean average precision (mAP) across all classes on the test sets. Tables 1 and 2 present the AP and mAP results, demonstrating that YOLOv5 (Jocher, 2020) outperforms Faster R-CNN (Ren et al., 2015) on the connector dataset under the specified settings. The results further indicate that training on the automatically annotated dataset (AA) and the manually annotated dataset (MA) yields comparable performance, with only minor differences in AP and mAP. Notably, Classes A1 and B1 achieve 100% AP despite their small size and largely texture-less appearance. This outcome is likely due to their relatively low intra-class variation and visually distinctive geometry under controlled acquisition conditions, which reduces confusion with other connector types and enables reliable detection.

6.6.2. Evaluation on 6DoF object pose estimation

The performance of 6DoF object pose estimation methods was evaluated on both the automatically annotated (AA) and manually annotated (MA) datasets using DenseFusion (Wang et al., 2019) and YOLOv5-6D (Viviers et al., 2024). DenseFusion (Wang et al., 2019) performs 6DoF pose estimation by extracting and fusing features from RGB and depth measurements. In contrast, YOLOv5-6D (Viviers et al., 2024) is a single-shot framework that performs joint 2D detection and 6DoF pose estimation from RGB input. This approach, originally demonstrated on both RGB and X-ray imagery, eliminates the need for depth data or precomputed 3D object models.

Both models were trained on the combined training and validation sets of the connector dataset and evaluated on the test set. Training protocols followed the original implementations (Wang et al., 2019; Viviers et al., 2024). Performance was assessed using standard pose metrics: (i) the Average Distance of Model Points (ADD) success rate with a threshold of 10% of the object diameter (ADD-0.1d) (Hinterstoisser et al., 2013); (ii) the *cm-degree* success rate with thresholds of 5 centimeters (cm) translation error and 5 degrees (deg) rotation error (5cm-5deg) (Shotton et al., 2013); and (iii) the 2D projection success rate with a 5-pixel threshold (Prj-5) (Brachmann et al., 2016).

Tables 3 and 4 report the results on the MA and AA datasets, respectively. Similar performance trends across both datasets indicate that the automatically generated annotations provide sufficient reliability for training and evaluating 6DoF pose estimators. Among the evaluated methods, YOLOv5-6D (Viviers et al., 2024) achieves higher success rates under the stricter Prj-5 and 5cm-5deg criteria, indicating

Table 3
Evaluation results on 6DoF object pose estimation on the Manually Annotated dataset (MA).

Object	DenseFusion (Wang et al., 2019)			YOLOv5-6D (Viviers et al., 2024)		
	ADD-0.1d	5cm-5deg	Prj-5	ADD-0.1d	5cm-5deg	Prj-5
A0	0.498	0.125	0.002	0.875	0.775	0.739
A1	0.739	0.200	0.079	0.738	0.475	0.800
B0	0.625	0.088	0.088	0.500	0.725	0.739
B1	0.673	0.088	0.138	0.546	0.650	0.700
C	0.709	0.135	0.070	0.500	0.475	0.800
D	0.731	0.075	0.125	0.500	0.550	0.850
E	0.715	0.100	0.150	0.550	0.750	0.863
F	0.756	0.125	0.175	0.550	0.800	0.640
G	0.698	0.088	0.021	0.725	0.725	0.738
H	0.751	0.100	0.075	0.688	0.550	0.739
I	0.565	0.075	0.100	0.375	0.640	0.725
J	0.809	0.100	0.125	0.325	0.500	0.625
K	0.743	0.125	0.079	0.875	0.750	0.640
L	0.687	0.125	0.253	0.375	0.475	0.750
M	0.721	0.100	0.375	0.215	0.400	0.700
N	0.550	0.088	0.200	0.375	0.500	0.756
O	0.539	0.125	0.185	0.333	0.275	0.750
P	0.683	0.075	0.225	0.550	0.500	0.788
Q	0.631	0.100	0.200	0.250	0.325	0.700
R	0.718	0.125	0.375	0.275	0.214	0.800
All	0.677	0.108	0.152	0.506	0.553	0.742

Table 4
Evaluation results on 6DoF object pose estimation on the Automatically Annotated dataset (AA).

Object	DenseFusion (Wang et al., 2019)			YOLOv5-6D (Viviers et al., 2024)		
	ADD-0.1d	5cm-5deg	Prj-5	ADD-0.1d	5cm-5deg	Prj-5
A0	0.538	0.100	0.000	0.875	0.763	0.800
A1	0.750	0.275	0.088	0.750	0.463	0.900
B0	0.575	0.063	0.075	0.546	0.600	0.739
B1	0.650	0.100	0.125	0.500	0.663	0.763
C	0.713	0.138	0.075	0.500	0.488	0.863
D	0.725	0.150	0.138	0.688	0.588	0.863
E	0.738	0.150	0.150	0.688	0.800	0.850
F	0.775	0.088	0.188	0.625	0.750	0.738
G	0.700	0.113	0.013	0.313	0.688	0.713
H	0.750	0.100	0.100	0.750	0.725	0.813
I	0.563	0.075	0.088	0.563	0.550	0.700
J	0.813	0.125	0.100	0.333	0.480	0.640
K	0.750	0.100	0.088	0.938	0.725	0.750
L	0.688	0.138	0.288	0.313	0.410	0.756
M	0.725	0.113	0.325	0.200	0.429	0.625
N	0.588	0.075	0.225	0.438	0.513	0.700
O	0.563	0.100	0.175	0.250	0.338	0.700
P	0.675	0.050	0.250	0.563	0.513	0.750
Q	0.613	0.163	0.213	0.250	0.288	0.788
R	0.700	0.100	0.325	0.250	0.234	0.818
All	0.680	0.116	0.151	0.517	0.550	0.764

strong 2D alignment and threshold-based pose accuracy. DenseFusion (Wang et al., 2019) attains higher accuracy under ADD-0.1d, indicating superior average 3D alignment. This outcome aligns with DenseFusion's RGB–D fusion and point-cloud alignment strategy, which reduces mean 3D model-point error even if small residual rotation or translation errors persist. However, the comparatively lower Prj-5 and 5cm-5deg scores indicate challenges in consistently achieving stricter pose thresholds and maintaining precise 2D reprojection alignment.

Multiple factors may contribute to this discrepancy. DenseFusion (Wang et al., 2019) relies on iterative refinement driven by depth-based alignment, which effectively minimizes ADD but does not always converge to poses that meet strict translation and rotation constraints. Additionally, performance is sensitive to depth noise at close range. At a 300 mm acquisition distance, small connectors occupy few pixels, and fine geometric details are often poorly resolved in low-resolution depth measurements. Although the connectors are intentionally asymmetric to support poka-yoke assembly, many discriminative features are subtle and small-scale, presenting challenges for depth-driven alignment. DenseFusion (Wang et al., 2019), originally developed for larger and more textured objects, is therefore susceptible to small pose drift, which can result in failures under stricter metrics such as 5cm-5deg and Prj-5. In contrast, YOLOv5-6D (Viviers et al., 2024) directly regresses rotation and translation from RGB input. This approach can yield more consistent performance under threshold-based metrics and improved 2D alignment, even if overall 3D alignment measured by ADD is lower. These findings reinforce that accurate perception of wire harness connectors requires high-resolution sensing and fine-grained visual cues (Wei et al., 2022; Cheng et al., 2023). This highlights the importance of pose estimation methods that are robust to small object scale, low texture, and subtle geometric features in precision-critical industrial applications.

Table 5
Influence of viewpoint density for 2D object detection.

N	Faster R-CNN (Ren et al., 2015)		YOLOv5 (Jocher, 2020)	
	mAP_{50}	$mAP_{50:95}$	mAP_{50}	$mAP_{50:95}$
10	0.450	0.397	0.792	0.764
20	0.541	0.471	0.884	0.825
40	0.732	0.675	0.917	0.902
80	0.938	0.883	0.989	0.969

Table 6
Influence of viewpoint density for 6DoF object pose estimation.

N	DenseFusion (Wang et al., 2019)			YOLOv5-6D (Viviers et al., 2024)		
	ADD-0.1d	5cm-5deg	Prj-5	ADD-0.1d	5cm-5deg	Prj-5
10	0.175	0.000	0.000	0.125	0.000	0.000
20	0.313	0.025	0.013	0.280	0.063	0.275
40	0.498	0.088	0.088	0364	0.231	0.325
80	0.673	0.100	0.150	0.506	0.550	0.756

6.7. Ablation study

This subsection reports ablation studies on the collected connector dataset to evaluate the impact of (i) viewpoint sampling density and (ii) background removal strategy, specifically comparing RGB-only chroma keying and depth-assisted segmentation, on subsequent 2D detection and 6DoF pose estimation performance.

6.7.1. Influence of viewpoint density

The effect of viewpoint density was assessed by systematically varying the number of viewpoints (N) employed during image acquisition. Specifically, N was increased from 10 to 80 in two-fold increments ($N \in \{10, 20, 40, 80\}$). For each configuration, a corresponding dataset was generated and partitioned into training, validation, and test subsets using an 80:10:10 ratio. Models were trained according to the protocols outlined in Sections 6.6.1 and 6.6.2. To establish an independent benchmark and reduce potential bias from automatically generated labels, evaluation was conducted on the manually annotated test set (MA_{test}). Tables 5 and 6 present the results for 2D object detection and 6DoF pose estimation, respectively. Overall, performance consistently improves as N increases, suggesting that denser viewpoint coverage results in more robust detection and pose estimation.

6.7.2. Influence of background removal strategy

In addition to analyzing viewpoint density, the impact of the background removal strategy implemented in the data annotation module was evaluated. The proposed pipeline utilizes RGB-based chroma keying and, in its default configuration, further refines segmentation through depth thresholding. To quantify the contribution of depth-assisted background removal, an additional automatically annotated dataset was generated by running the annotation module on raw

Table 7
Influence of background removal strategy for 2D object detection.

Background removal		Faster R-CNN (Ren et al., 2015)		YOLOv5 (Jocher, 2020)	
Chroma key (RGB)	Depth threshold	mAP_{50}	$mAP_{50:95}$	mAP_{50}	$mAP_{50:95}$
✓		0.937	0.871	0.988	0.962
✓	✓	0.938	0.883	0.989	0.969

Table 8
Influence of background removal strategy for 6DoF object pose estimation.

Background removal		DenseFusion (Wang et al., 2019)			YOLOv5-6D (Viviers et al., 2024)		
Chroma key (RGB)	Depth threshold	ADD-0.1d	5cm-5deg	Prj-5	ADD-0.1d	5cm-5deg	Prj-5
✓		0.668	0.100	0.116	0.500	0.513	0.750
✓	✓	0.673	0.100	0.150	0.506	0.550	0.756

connector images without the depth-thresholding step, that is, using chroma keying only.

For a controlled comparison, this dataset was divided into training, validation, and test sets using the same 80:10:10 split protocol. Models were trained according to the procedures described in Sections 6.6.1 and 6.6.2, and evaluated on the manually annotated test set (MA_{test}) to mitigate potential bias from automatically generated labels. Tables 7 and 8 present the results for 2D object detection and 6DoF pose estimation, respectively. Overall, incorporating depth-threshold assistance results in modest yet consistent performance gains for both tasks, suggesting that depth-assisted background removal enhances segmentation quality and thereby refines the training labels. Although chroma keying alone is effective under the controlled acquisition conditions of the connector dataset, the addition of a depth constraint provides measurable benefits, particularly under more stringent pose evaluation criteria.

6.8. Time costs analysis

The potential time savings of the proposed system were assessed by quantifying the time costs of robot-assisted dataset generation, including both robotic data acquisition and automatic annotation. Table 9 summarizes the measured time required by each component of the pipeline. During acquisition, the robot tool speed was set to 1 meter per second for translation and 180 degree per second for rotation. Because the same observation poses and motion sequence were used for all connector samples, the acquisition and annotation time per sample was consistent. On average, generating a single sample, which includes the raw RGB–D data and corresponding annotations for 2D BBoxes, 3D BBoxes, and 6DoF object poses, required 2.644 s. The values reported in Table 9 may vary with robot speed settings, motion-planning policies, and the viewpoint configurations selected to meet application-specific requirements.

To facilitate direct comparison, the time required for manual data acquisition and annotation was measured using procedures consistent with those applied in the connector dataset. The manually annotated dataset was generated through crowd-sourced labeling by three annotators. The reported manual annotation time represents the average across these individuals, each of whom annotated 500 images. For manual operation, the setup time was approximately 60 s, capturing a single sample required about 2 s, and manual annotation required approximately 40 s per sample. In contrast, robot-assisted data collection required approximately 600 s for system setup and 2.390 s per sample for acquisition. Annotation time was reduced to 0.254 s per sample for generating the required bounding boxes and object pose labels. Overall, the proposed pipeline achieves a comparable acquisition speed and demonstrates an approximately 150-fold improvement in annotation efficiency compared to the manual process.

Fig. 12 presents a comparison of the total time costs for robot-assisted and manual workflows. The results indicate that the robot-assisted approach becomes advantageous in terms of time and human effort when the number of samples exceeds 14. These results suggest that the proposed pipeline can substantially improve the efficiency of generating real-world datasets for object detection and pose estimation, particularly in industrial settings that require large-scale data collection and streamlined labeling workflows.

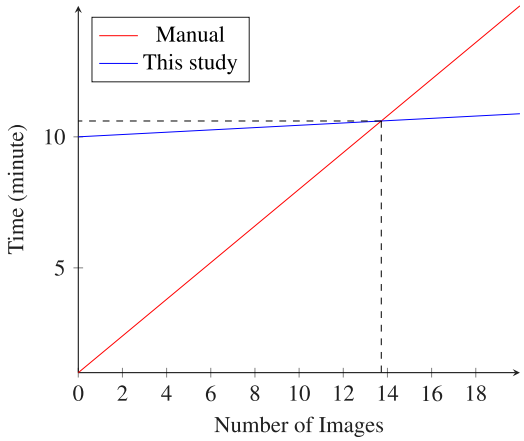


Fig. 12. Time costs of generating a real-world dataset with ground-truth labels of 2D BBoxes, 3D BBoxes, and 6DoF object poses manually and using the proposed pipeline with robotic data acquisition and automatic data annotation.

7. Discussion

7.1. Advantages of the proposed pipeline

Utilizing a robotic arm for data acquisition facilitates flexible and scalable generation of observation poses. This approach supports systematic multi-view capture while adhering to practical workspace and setup constraints. In comparison to traditional gantry-based systems (Seitz et al., 2006; Kasper et al., 2012), the proposed approach reduces hardware complexity and can be implemented on cost-effective robotic platforms, including collaborative robots, according to specific application requirements (Gusan and Ţîţu, 2021; Sahan et al., 2023; Shah et al., 2025). Furthermore, robot-assisted acquisition removes the need for auxiliary mechanisms such as turntables and enables automated annotation, particularly through kinematics-based 6DoF pose labeling. This process eliminates dependence on external camera-tracking aids such as marker boards (Grenzdörffer et al., 2020).

Manufacturers can directly utilize the resulting annotated datasets to develop, deploy, and maintain perception modules for various production tasks. For instance, the generated 2D labels facilitate the training of object detectors for connector recognition and inspection. In contrast, the 6DoF pose annotations enable pose-aware applications, including assembly verification such as presence and orientation checks,

Table 9
Time cost (Second) of preparing a connector dataset using the proposed pipeline.

Object	Data acquisition				Data annotation				Total
	Robot	Camera	Storage	Sum	2D BBox	3D BBox	6DoF Pose	Sum	
A0	192.603	0.871	3.212	196.686	9.572	1.003	9.637	20.212	216.898
A1	189.352	0.878	3.562	193.792	9.895	1.175	9.941	21.011	214.803
B0	188.346	0.895	3.743	192.984	9.928	1.287	9.947	21.162	214.146
B1	189.549	0.869	3.598	194.016	10.053	1.211	9.867	21.131	215.147
C	183.409	0.863	3.436	187.708	9.892	1.133	9.826	20.851	208.559
D	187.164	0.903	3.527	191.594	9.955	1.162	1.162	12.279	203.873
E	184.590	0.869	3.617	189.076	9.987	1.215	9.848	21.050	210.126
F	184.091	0.909	3.264	188.264	9.816	1.023	9.637	20.476	208.740
G	183.426	0.904	3.439	187.769	9.918	1.130	9.892	20.940	208.709
H	183.558	0.871	3.076	187.505	9.714	0.893	9.550	20.157	207.662
I	184.597	0.879	3.351	188.827	9.850	1.083	9.693	20.626	209.453
J	185.652	0.869	3.615	190.136	10.015	1.223	9.871	21.109	211.245
K	190.724	0.869	2.869	194.462	9.654	0.820	9.492	19.966	214.428
L	187.109	0.886	3.499	191.494	9.988	1.165	9.701	20.854	212.348
M	185.918	0.847	3.460	190.225	9.977	1.161	9.789	20.927	211.152
N	184.024	0.873	3.414	188.311	9.920	1.111	9.804	20.835	209.146
O	184.414	0.893	3.541	188.848	10.046	1.191	9.791	21.028	209.876
P	188.912	0.871	3.249	193.032	9.862	1.018	9.718	20.598	213.630
Q	189.645	0.862	3.236	193.743	9.841	1.010	9.727	20.578	214.321
R	192.053	0.879	3.277	196.209	9.838	1.001	9.675	20.514	216.723
All	3739.136	17.560	67.985	3824.681	197.721	22.015	186.568	406.304	4230.985
Avg.-Obj.	186.957	0.878	3.399	191.234	9.886	1.101	9.328	20.315	211.549
Avg.-Img.	2.337	0.011	0.042	2.390	0.124	0.014	0.117	0.254	2.644

as well as robotic handling tasks like bin picking and pose-guided grasping. When new part variants are introduced or operating conditions change, manufacturers can efficiently re-collect multi-view data and regenerate annotations using the same pipeline. This capability enables rapid model updates with minimal additional manual labeling. In summary, the proposed pipeline offers an end-to-end and scalable workflow for generating multimodal ground-truth labels comparable to those in datasets such as Grenzdörffer et al. (2020). This approach does not require pre-existing object models, thereby enhancing practical applicability in industrial environments.

7.1.1. Efficiency compared to manual dataset preparation

The proposed pipeline offers a scalable and efficient approach to dataset preparation, particularly benefiting industrial robotic applications that demand precise 6DoF object pose annotations. The findings demonstrate that generating datasets for object detection and pose estimation using the proposed pipeline significantly reduces manual intervention and greatly enhances annotation efficiency, achieving annotation speeds approximately 150 times faster than a fully manual workflow. Fig. 12 further highlights this advantage by comparing the time required for the proposed pipeline with that of manual data preparation. The results suggest that the proposed pipeline is especially effective for generating large-scale annotated datasets, in which manual labeling constitutes a significant portion of the overall cost.

While integrating robotic hardware requires initial setup effort and associated costs, the subsequent savings in human labor and time render the approach both advantageous and scalable for real-world dataset generation in industrial environments. Furthermore, the robotic platform may be repurposed for additional automation tasks following data collection, thereby enhancing overall hardware utilization.

7.1.2. Benefits over gantry-based approaches

In comparison to traditional gantry-based approaches, the proposed pipeline offers increased flexibility, simplified configuration, enhanced cost efficiency, and more effective utilization of hardware resources. For instance, configuring a robotic arm equipped with an eye-in-hand camera is typically less complex and more cost-effective than constructing large-scale gantry systems such as the Stanford Spherical Gantry (Seitz et al., 2006). The proposed pipeline further demonstrates superior scalability and adaptability compared to systems utilizing pre-fabricated multi-camera racks with turntables (Singh et al., 2014; Kimble et al., 2022). These multi-camera systems are inherently limited by fixed viewpoint layouts, making expansion or reconfiguration challenging. For example, the multi-camera rack with a turntable in Singh et al. (2014) was restricted to the set of viewpoints illustrated in Fig. 4(d).

Previous research has also investigated single-camera configurations combined with turntables for data collection (Kasper et al., 2012; Hodaň et al., 2017). Although these systems are capable of acquiring multi-view imagery with a single sensor, they frequently necessitate specialized hardware designs (Kasper et al., 2012) or manual camera repositioning (Hodaň et al., 2017). In contrast, the proposed pipeline facilitates dynamic, programmable, and systematic generation of viewpoints through the use of a robotic arm equipped with an eye-in-hand camera. This design reduces hardware overhead and enables comprehensive viewpoint coverage without the need for manual intervention or supplementary acquisition mechanisms.

7.1.3. Improvements over existing robot-assisted approaches

Previous research has investigated robot-assisted methods to enhance the efficiency and quality of real-world dataset generation for industrial vision applications. However, these approaches frequently encounter challenges related to high hardware complexity, limited viewpoint coverage, and insufficient automation of multimodal annotation. The proposed pipeline addresses these challenges by offering a streamlined, scalable, and adaptable workflow for dataset preparation. Specifically, the pipeline reduces hardware requirements, facilitates systematic viewpoint generation, and supports comprehensive multimodal ground-truth annotation. These features enhance its applicability in practical industrial environments.

Hardware complexity. Many current systems depend on auxiliary devices, including turntables (Kiyokawa et al., 2021) and marker boards (Grenzdörffer et al., 2020). For instance, Kiyokawa et al. (2021) integrated a collaborative robot with a turntable and augmented reality markers. In this configuration, the object was positioned at the center of the turntable and rotated around the azimuth axis. Simultaneously, the robot moved an eye-in-hand camera along the polar direction to capture multi-view data. Pose labels were determined using predefined transformations among the camera, markers, and object. Similarly, Grenzdörffer et al. (2020) employed a collaborative robot in

conjunction with a marker board containing ArUco markers (Garrido-Jurado et al., 2014, 2016; Romero-Ramirez et al., 2018) to estimate camera poses. The annotation workflow included initial pose estimation with PoseCNN (Xiang et al., 2018), manual refinement on a reference frame, and subsequent label propagation based on the camera pose estimated with the marker board. In contrast, the proposed pipeline utilizes only a single collaborative robot and an eye-in-hand camera, thereby removing the requirement for turntables and marker boards. This reduction in hardware and software complexity enhances scalability and supports deployment in industrial environments.

Viewpoint coverage. Many existing robot-assisted data acquisition systems rely on predetermined robot trajectories (Jensen et al., 2014; Grenzdörffer et al., 2020) or require manual object manipulation (Koch et al., 2023). Jensen et al. (2014) used an industrial robot to traverse predefined paths generated by heuristics for multi-view stereopsis evaluation. However, the viewpoints were limited mainly to frontal aspects and were not determined through a systematic planning approach. Koch et al. (2023) incorporated viewpoints from both frontal and rear aspects, but required a heuristic to define observation poses and demanded manual rotation of the object to capture rear views. In contrast, the proposed pipeline systematically generates observation poses using a spherical coordinate parameterization. This approach enables comprehensive coverage of both frontal and rear perspectives without manual intervention. This design supports flexible and configurable viewpoint planning, thereby enhancing adaptability to a wide range of objects and data acquisition requirements.

Annotation automation. In the context of automated annotation, particularly for 6DoF pose labels, previous methods frequently rely on external markers (Grenzdörffer et al., 2020; Kiyokawa et al., 2021) or pre-acquired object models (Grenzdörffer et al., 2020). Grenzdörffer et al. (2020) derived 3D BBoxes from known object-model dimensions and generated 2D BBoxes and segmentation masks by rendering synthetic depth images of these models. Kiyokawa et al. (2021) utilized augmented reality markers to infer 6DoF object poses based on robotic kinematics. Lee et al. (2021) employed an LCD screen to vary background colors for chroma keying; however, their approach was limited to 2D BBox generation. The proposed pipeline extends previous work by enabling fully automated annotation through image processing and robotic kinematics, eliminating the need for object models, auxiliary markers, or manual refinements. Furthermore, the pipeline generates multimodal ground-truth labels, including 2D and 3D BBoxes, segmentation masks, and 6DoF poses, within a unified workflow.

Pose planning via a reversed kinematic strategy. An additional distinguishing aspect of the proposed pipeline is its reversed approach to kinematics-based 6DoF pose labeling. In contrast to previous methods that record robot poses, which are initially generated based on heuristics, during image capture and subsequently compute object poses (Koch et al., 2023; Zürn et al., 2024), the proposed pipeline begins with predefined object poses and calculates the corresponding robot poses to guide camera movement. This approach offers explicit control over viewpoint generation and selection, facilitates scalable and systematic viewpoint coverage, and supports data acquisition across a broader range of poses, including both frontal and rear perspectives, without the need for manual object manipulation.

7.2. Limitations and future work

While the proposed robot-assisted dataset preparation pipeline facilitates efficient and scalable training data generation, several limitations persist. These limitations suggest future improvements in three primary areas: automation and scalability, annotation quality and robustness, and adaptability to more complex acquisition and deployment scenarios.

7.2.1. Enhancing automation and scalability

The current pipeline requires manual object placement and swapping, which can become an operational bottleneck when processing a large number of object types or variants. Such manual intervention increases process variability and constrains throughput, especially in high-volume industrial contexts. Incorporating automated object handling, such as a dual-arm robotic cell (Chen et al., 2023), would reduce human involvement and facilitate continuous, high-throughput dataset generation.

While the pipeline supports scalable viewpoint generation and data acquisition for various industrial applications, it has not yet been integrated with real-time production systems. Integrating the pipeline into manufacturing environments would enable on-demand dataset generation and incremental model updates, thereby supporting continuous learning and adaptation to changing operational conditions.

7.2.2. Improving annotation robustness

The accuracy of kinematics-based pose annotation depends on the quality of robot calibration and motion repeatability. Mechanical drift, joint backlash, and calibration misalignment can introduce systematic errors, particularly in applications that require high-precision 6DoF pose labels. Applying post-processing and refinement strategies, such as fiducial-marker-based correction or optimization-based alignment, can improve annotation reliability.

Additionally, the chroma-keying segmentation method in the proposed pipeline assumes a static, uniform background, which reduces effectiveness when object colors are similar to the background. Utilizing more adaptive background designs, such as programmable LCD displays (Lee et al., 2021), can improve segmentation robustness across a broader range of object appearances and illumination conditions.

Moreover, practical deployment often requires converting annotations into formats compatible with widely used deep learning models. Developing a standardized export module that supports multiple dataset formats would facilitate integration with existing frameworks and improve reproducibility.

7.2.3. Extending to complex and dynamic scenarios

The current pipeline is optimized for rigid, opaque objects in isolated and visually simple scenes. It does not yet accommodate deformable objects or those with challenging optical properties, such as transparent or reflective surfaces. Additionally, it is not designed for cluttered environments that involve occlusions and inter-object interactions. Expanding the pipeline to multi-object scenarios, including occlusion-aware viewpoint planning and annotation, as well as strategies for complex materials, would significantly enhance its applicability to real-world industrial contexts.

Furthermore, lighting conditions were not strictly controlled during data acquisition in this study, with the aim of promoting robustness to illumination variation in the trained models (Chen et al., 2023). In practical applications, however, practitioners may deliberately adjust illumination to replicate specific deployment conditions, thereby facilitating the collection of task-specific datasets for particular production environments.

8. Conclusion

This paper introduces a robot-assisted pipeline designed to address the complex task of preparing real-world datasets for the development of deep learning-based robotic vision systems, with a particular focus on industrial applications of object detection and pose estimation. The integration of robotic data acquisition and automatic data annotation enables rapid generation of high-quality training data for object detection and pose estimation, significantly reducing human labor requirements.

A collaborative robot equipped with an eye-in-hand RGB–D camera, combined with a systematic observation-pose generation strategy based on spherical coordinates, facilitates flexible and scalable

multi-view data acquisition. The automatic annotation module utilizes image processing and robotic kinematics to generate comprehensive multimodal ground-truth labels, thereby enhancing the utility of the resulting dataset.

A case study involving automotive wire harness connectors demonstrates the practical relevance of the proposed pipeline in industrial settings where precise visual perception is essential. The findings show reduced hardware complexity and labor requirements, with average acquisition and annotation times of 2.390 s and 0.254 s per sample, respectively. Overall, annotation is approximately 150 times faster than a fully manual workflow.

The modular design and minimal calibration requirements enhance adaptability and cost efficiency, making the pipeline especially suitable for industries that already utilize robotic systems. This work advances the efficiency of large-scale real-world dataset preparation for robotic vision and underscores the potential for broader adoption to support industrial applications that depend on reliable object detection and precise 6DoF pose estimation.

CRediT authorship contribution statement

Hao Wang: Writing – original draft, Writing – review & editing, Visualization, Software, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Gonzalo Urbanos Uriel:** Writing – review & editing, Software, Investigation, Formal analysis, Data curation. **Karim El-Nahass:** Writing – review & editing, Software, Investigation, Formal analysis, Data curation. **Sven Ekered:** Writing – review & editing, Validation, Resources. **Björn Johansson:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Funding acquisition.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used Grammarly and OpenAI's GPT-5 in order to enhance the readability and language of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the EWASS project, funded by the Swedish innovation agency, Vinnova, and the strategic innovation program, Produktion 2030, with grant number 2022-01279. This work was also supported by the MAXBATT project, funded by Region Västra Götaland, with grant number MRU2024-00381. The work was carried out within Chalmers Production Area of Advance. Wiretronic AB and Volvo Car Corporation provided the automotive wire harness connectors used in this work. The support is gratefully acknowledged.

Data availability

Data will be made available on request.

References

Adhikari, B., Rahtu, E., Huttunen, H., 2021. Sample selection for efficient image annotation. In: 2021 9th European Workshop on Visual Information Processing (EUVIP). pp. 1–6. http://dx.doi.org/10.1109/EUVIP50544.2021.9484022.

Alshehri, A., Taileb, M., Alotaibi, R., 2022. DeepAIA: An automatic image annotation model based on generative adversarial networks and transfer learning. IEEE Access 10, 38437–38445. http://dx.doi.org/10.1109/ACCESS.2022.3165077.

Benenson, R., Popov, S., Ferrari, V., 2019. Large-scale interactive object segmentation with human annotators. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 11692–11701. http://dx.doi.org/10.1109/CVPR.2019.01197.

Brachmann, E., Michel, F., Krull, A., Yang, M.Y., Gumhold, S., Rother, C., 2016. Uncertainty-driven 6D pose estimation of objects and scenes from a single RGB image. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 3364–3372. http://dx.doi.org/10.1109/CVPR.2016.366.

Caporali, A., Pantano, M., Janisch, L., Regulin, D., Palli, G., Lee, D., 2023. A weakly supervised semi-automatic image labeling approach for deformable linear objects. IEEE Robot. Autom. Lett. 8 (2), 1013–1020. http://dx.doi.org/10.1109/LRA.2023.3234799.

Chen, H., Wan, W., Matsushita, M., Kotaka, T., Harada, K., 2023. Automatically prepare training data for YOLO using robotic in-hand observation and synthesis. IEEE Trans. Autom. Sci. Eng. 1–17. http://dx.doi.org/10.1109/TASE.2023.3304420.

Chen, L., Yang, H., Wu, C., Wu, S., 2022. MP6D: An RGB-D dataset for metal parts' 6D pose estimation. IEEE Robot. Autom. Lett. 7 (3), 5912–5919. http://dx.doi.org/10.1109/LRA.2022.3154807.

Cheng, G., Yuan, X., Yao, X., Yan, K., Zeng, Q., Xie, X., Han, J., 2023. Towards large-scale small object detection: Survey and benchmarks. IEEE Trans. Pattern Anal. Mach. Intell. 45 (11), 13467–13488. http://dx.doi.org/10.1109/TPAMI.2023.3290594.

Cop, K.P., Peters, A., Žagar, B.L., Hettegger, D., Knoll, A.C., 2021. New metrics for industrial depth sensors evaluation for precise robotic applications. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems. IROS, pp. 5350–5356. http://dx.doi.org/10.1109/IROS51168.2021.9636322.

De Gregorio, D., Tonioni, A., Palli, G., Di Stefano, L., 2020. Semiautomatic labeling for deep learning in robotics. IEEE Trans. Autom. Sci. Eng. 17 (2), 611–620. http://dx.doi.org/10.1109/TASE.2019.2938316.

Elsharkawy, A., Kim, M.S., 2022. Human–robot labeling framework to construct multitype real-world datasets. IEEE Access 10, 131166–131180. http://dx.doi.org/10.1109/ACCESS.2022.3229864.

Garrido-Jurado, S., Muñoz-Salinas, R., Madrid-Cuevas, F., Marín-Jiménez, M., 2014. Automatic generation and detection of highly reliable fiducial markers under occlusion. Pattern Recognit. 47 (6), 2280–2292. http://dx.doi.org/10.1016/j.patcog.2014.01.005.

Garrido-Jurado, S., Muñoz-Salinas, R., Madrid-Cuevas, F., Medina-Carnicer, R., 2016. Generation of fiducial marker dictionaries using mixed integer linear programming. Pattern Recognit. 51, 481–491. http://dx.doi.org/10.1016/j.patcog.2015.09.023.

Geiß, M., Wagner, R., Baresch, M., Steiner, J., Zwick, M., 2023. Automatic bounding box annotation with small training datasets for industrial manufacturing. Micromachines 14 (2), http://dx.doi.org/10.3390/mi14020442.

Gregor, S., Hevner, A.R., 2013. Positioning and presenting design science research for maximum impact. MIS Q. 37 (2), 337–355. http://dx.doi.org/10.25300/MISQ/2013/37.2.01.

Grenzdörffer, T., Günther, M., Hertzberg, J., 2020. YCB-M: A multi-camera RGB-D dataset for object recognition and 6DoF pose estimation. In: 2020 IEEE International Conference on Robotics and Automation. ICRA, pp. 3650–3656. http://dx.doi.org/10.1109/ICRA40945.2020.9197426.

Guang, R., Li, X., Lei, Y., Yang, B., Li, N., 2025. Dynamic vision-based machine vibration sensing and fault diagnosis with signal alignment and feature clustering. Eng. Appl. Artif. Intell. 162, 112445. http://dx.doi.org/10.1016/j.engappai.2025.112445.

Gusan, V., Ţîţu, A.M., 2021. Management of cost reduction and process improvement. Implementation of industrial robots versus colaborative robots. Rev. Manag. Econ. Eng. 20 (3), 195–209. http://dx.doi.org/10.71235/rmee.107.

Gygli, M., Ferrari, V., 2020. Efficient object annotation via speaking and pointing. Int. J. Comput. Vis. 128 (5), 1061–1075. http://dx.doi.org/10.1007/s11263-019-01255-4.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 770–778. http://dx.doi.org/10.1109/CVPR.2016.90.

Hevner, A., Chatterjee, S., 2010. Design science research in information systems. In: Design Research in Information Systems: Theory and Practice, vol. 22, Springer US, pp. 9–22. http://dx.doi.org/10.1007/978-1-4419-5653-8_2.

Hevner, A.R., March, S.T., Park, J., Ram, S., 2004. Design science in information systems research. MIS Q. 28 (1), 75–105. http://dx.doi.org/10.2307/25148625.

Hinterstoisser, S., Lepetit, V., Ilic, S., Holzer, S., Bradski, G., Konolige, K., Navab, N., 2013. Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes. In: Computer Vision – ACCV 2012. pp. 548–562. http://dx.doi.org/10.1007/978-3-642-37331-2_42.

Hodaň, T., Haluza, P., Obdržálek, Š., Matas, J., Lourakis, M., Zabulis, X., 2017. T-LESS: An RGB-D dataset for 6D pose estimation of texture-less objects. In: 2017 IEEE Winter Conference on Applications of Computer Vision. WACV, pp. 880–888. http://dx.doi.org/10.1109/WACV.2017.103.

Høffer, M.F., Koldkjær, K.-E.S., Andersen, D.S., Herlev, V.D., Abrahamsen, S.D., Nyboe, F.F., Malle, N.H., Ebeid, E., 2023. Robotics framework for object tracking using FPGA with novel automatic image labelling. In: IEEE EUROCON 2023 - 20th International Conference on Smart Technologies. pp. 782–787. http://dx.doi.org/10.1109/EUROCON56442.2023.10199079.

Hu, W., Zheng, J., Zhang, Z., Yuan, X., Yin, J., Zhou, Z., 2023. PlankAssembly: Robust 3D reconstruction from three orthographic views with learnt shape programs. In: 2023 IEEE/CVF International Conference on Computer Vision. ICCV, pp. 18449–18459. http://dx.doi.org/10.1109/ICCV51070.2023.01695.

Ilin, V., Kalinov, I., Karpyshev, P., Tsetserukou, D., 2021. DeepScanner: a robotic system for automated 2D object dataset collection with annotations. In: 2021 26th IEEE International Conference on Emerging Technologies and Factory Automation. ETFA, pp. 01–08. http://dx.doi.org/10.1109/ETFA45728.2021.9613396.

Jakobi, N., Husbands, P., Harvey, I., 1995. Noise and the reality gap: The use of simulation in evolutionary robotics. In: Advances in Artificial Life. pp. 704–720. http://dx.doi.org/10.1007/3-540-59496-5_337.

Jensen, R., Dahl, A., Vogiatzis, G., Tola, E., Aanæs, H., 2014. Large scale multi-view stereopsis evaluation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. pp. 406–413. http://dx.doi.org/10.1109/CVPR.2014.59.

Jha, S.B., Babiceanu, R.F., 2023. Deep CNN-based visual defect detection: Survey of current literature. Comput. Ind. 148, 103911. http://dx.doi.org/10.1016/j.compind.2023.103911.

Jocher, G., 2020. YOLOv5. https://github.com/ultralytics/yolov5.

Kasper, A., Xue, Z., Dillmann, R., 2012. The KIT object models database: An object model database for object recognition, localization and manipulation in service robotics. Int. J. Robot. Res. 31 (8), 927–934. http://dx.doi.org/10.1177/0278364912445831.

Kavasidis, I., Spampinato, C., Giordano, D., 2013. Generation of ground truth for object detection while playing an online game: Productive gaming or recreational working? In: 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 694–699. http://dx.doi.org/10.1109/CVPRW.2013.105.

Kimble, K., Albrecht, J., Zimmerman, M., Falco, J., 2022. Performance measures to benchmark the grasping, manipulation, and assembly of deformable objects typical to manufacturing applications. Front. Robot. AI 9, http://dx.doi.org/10.3389/frobt.2022.999348.

Kiyokawa, T., Katayama, H., Tatsuta, Y., Takamatsu, J., Ogasawara, T., 2021. Robotic waste sorter with agile manipulation and quickly trainable detector. IEEE Access 9, 124616–124631. http://dx.doi.org/10.1109/ACCESS.2021.3110795.

Kiyokawa, T., Shirakura, N., Katayama, H., Tomochika, K., Takamatsu, J., 2025. Efficiently collecting training dataset for 2D object detection by online visual feedback. J. Robot. Mechatronics 37 (2), 270–283. http://dx.doi.org/10.20965/jrm.2025.p0270.

Kiyokawa, T., Tomochika, K., Takamatsu, J., and, T.O., 2019a. Efficient collection and automatic annotation of real-world object images by taking advantage of post-diminished multiple visual markers. Adv. Robot. 33 (24), 1264–1280. http://dx.doi.org/10.1080/01691864.2019.1697750.

Kiyokawa, T., Tomochika, K., Takamatsu, J., Ogasawara, T., 2019b. Fully automated annotation with noise-masked visual markers for deep-learning-based object detection. IEEE Robot. Autom. Lett. 4 (2), 1972–1977. http://dx.doi.org/10.1109/LRA.2019.2899153.

Kleeberger, K., Landgraf, C., Huber, M.F., 2019. Large-scale 6D object pose estimation dataset for industrial bin-picking. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems. IROS, pp. 2573–2578. http://dx.doi.org/10.1109/IROS40897.2019.8967594.

Koch, P., Schlüter, M., Thill, S., Krüger, J., 2023. Towards robot-assisted data generation with minimal user interaction for autonomously training 6D pose estimation in operational environments. Procedia CIRP 120, 249–254. http://dx.doi.org/10.1016/j.procir.2023.08.045.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444. http://dx.doi.org/10.1038/nature14539.

Lee, W.-C., Zhang, J.-Y., Wei, C.-C., 2021. Using an LCD monitor and a robotic arm to quickly establish image datasets for object detection. IEEE Access 9, 131006–131019. http://dx.doi.org/10.1109/ACCESS.2021.3111314.

Li, X., Cao, R., Feng, Y., Chen, K., Yang, B., Fu, C.-W., Li, Y., Dou, Q., Liu, Y.-H., Heng, P.-A., 2022. A sim-to-real object recognition and localization framework for industrial robotic bin picking. IEEE Robot. Autom. Lett. 7 (2), 3961–3968. http://dx.doi.org/10.1109/LRA.2022.3149026.

Li, X., Lin, Y., Yang, S., Zhang, W., 2025. Intelligent domain-generalized second-life EV battery state-of-health estimation. J. Energy Storage 140, 118989. http://dx.doi.org/10.1016/j.est.2025.118989.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 936–944. http://dx.doi.org/10.1109/CVPR.2017.106.

Liu, C., Li, X., Chen, X., Khan, S., 2025. Neuromorphic computing-enabled generalized machine fault diagnosis with dynamic vision. Adv. Eng. Inform. 65, 103300. http://dx.doi.org/10.1016/j.aei.2025.103300.

Lundberg, I., Björkman, M., Ögren, P., 2014. Intrinsic camera and hand-eye calibration for a robot vision system using a point marker. In: 2014 IEEE-RAS International Conference on Humanoid Robots. pp. 59–66. http://dx.doi.org/10.1109/HUMANOIDS.2014.7041338.

Lyu, H., Bai, Y., Liang, X., Das, U., Shi, C., Gong, L., Li, Y., Sun, M., Ge, M., Ma, X., 2024. FARPLS: A feature-augmented robot trajectory preference labeling system to assist human labelers' preference elicitation. In: Proceedings of the 29th International Conference on Intelligent User Interfaces. pp. 344–369. http://dx.doi.org/10.1145/3640543.3645145.

Marullo, G., Tanzi, L., Piazzolla, P., Vezzetti, E., 2023. 6D object position estimation from 2D images: a literature review. Multimedia Tools Appl. 82 (16), 24605–24643. http://dx.doi.org/10.1007/s11042-022-14213-z.

Pande, B., Padamwar, K., Bhattacharya, S., Roshan, S., Bhamare, M., 2022. A review of image annotation tools for object detection. In: 2022 International Conference on Applied Artificial Intelligence and Computing. ICAAIC, pp. 976–982. http://dx.doi.org/10.1109/ICAAIC53929.2022.9792665.

Pang, J., Zheng, P., Li, S., Liu, S., 2023. A verification-oriented and part-focused assembly monitoring system based on multi-layered digital twin. J. Manuf. Syst. 68, 477–492. http://dx.doi.org/10.1016/j.jmsy.2023.05.008.

Patki, N., Wedge, R., Veeramachaneni, K., 2016. The synthetic data vault. In: 2016 IEEE International Conference on Data Science and Advanced Analytics. DSAA, pp. 399–410. http://dx.doi.org/10.1109/DSAA.2016.49.

Pattar, S.P., Killus, T., Hirakawa, T., Yamashita, T., Sawanobori, T., Fujiyoshi, H., 2023. Automatic data collection for object detection and grasp-position estimation with mobile robots and invisible markers. Adv. Robot. 37 (4), 241–256. http://dx.doi.org/10.1080/01691864.2022.2136504.

Peffers, K., Tuunanen, T., Rothenberger, M.A., Chatterjee, S., 2007. A design science research methodology for information systems research. J. Manage. Inf. Syst. 24 (3), 45–77. http://dx.doi.org/10.2753/MIS0742-1222240302.

Phuong, T.N., Sakaino, H., Duy, V.N., 2024. IsoGAN: A gan based method for isometric view images generation from three orthographic views contour drawings. In: 2024 IEEE International Conference on Image Processing Challenges and Workshops. ICIPCW, pp. 4116–4120. http://dx.doi.org/10.1109/ICIPCW64161.2024.10769167.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, vol. 28, pp. 1–9.

Roh, Y., Heo, G., Whang, S.E., 2021. A survey on data collection for machine learning: A big data - AI integration perspective. IEEE Trans. Knowl. Data Eng. 33 (4), 1328–1347. http://dx.doi.org/10.1109/TKDE.2019.2946162.

Romero-Ramirez, F.J., Muñoz-Salinas, R., Medina-Carnicer, R., 2018. Speeded up detection of squared fiducial markers. Image Vis. Comput. 76, 38–47. http://dx.doi.org/10.1016/j.imavis.2018.05.004.

Sahan, A.M., Kathiravan, S., Lokesh, M., Raffik, R., 2023. Role of cobots over industrial robots in industry 5.0: A review. In: 2023 2nd International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation. ICAECA, pp. 1–5. http://dx.doi.org/10.1109/ICAECA56562.2023.10201199.

Salunkhe, O., Quadrini, W., Wang, H., Stahre, J., Romero, D., Fumagalli, L., Lämkull, D., 2023. Review of current status and future directions for collaborative and semi-automated automotive wire harnesses assembly. Procedia CIRP 120, 696–701. http://dx.doi.org/10.1016/j.procir.2023.09.061.

Sapp, B., Saxena, A., Ng, A.Y., 2008. A fast data collection and augmentation procedure for object recognition. AAAI, In: Proceedings of the 23rd National Conference on Artificial Intelligence, vol. 3, pp. 1402–1408.

Seitz, S., Curless, B., Diebel, J., Scharstein, D., Szeliski, R., 2006. A comparison and evaluation of multi-view stereo reconstruction algorithms. CVPR'06, In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 519–528. http://dx.doi.org/10.1109/CVPR.2006.19.

Shah, R., Doss, A.S.A., Lakshmaiya, N., 2025. Advancements in AI-enhanced collaborative robotics: towards safer, smarter, and human-centric industrial automation. Results Eng. 27, 105704. http://dx.doi.org/10.1016/j.rineng.2025.105704.

Sharma, H., Kumar, H., Gupta, A., Shah, M.A., 2023. Computer vision in manufacturing: a bibliometric analysis and future research propositions. Int. J. Adv. Manuf. Technol. 127 (11), 5691–5710. http://dx.doi.org/10.1007/s00170-023-11907-y.

Shotton, J., Glocker, B., Zach, C., Izadi, S., Criminisi, A., Fitzgibbon, A., 2013. Scene coordinate regression forests for camera relocalization in RGB-D images. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition. pp. 2930–2937. http://dx.doi.org/10.1109/CVPR.2013.377.

Shuichi, A., Manabu, H., 2019. Semi-automatic training data generation for semantic segmentation using 6dof pose estimation. In: Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. pp. 607–613. http://dx.doi.org/10.5220/0007568706070613.

da Silva, J.L., Tabata, A.N., Broto, L.C., Cocron, M.P., Zimmer, A., Brandmeier, T., 2020. Open source multipurpose multimedia annotation tool. In: Image Analysis and Recognition. pp. 356–367. http://dx.doi.org/10.1007/978-3-030-50347-5_31.

Singh, A., Sha, J., Narayan, K.S., Achim, T., Abbeel, P., 2014. BigBIRD: A large-scale 3D database of object instances. In: 2014 IEEE International Conference on Robotics and Automation. ICRA, pp. 509–516. http://dx.doi.org/10.1109/ICRA.2014.6906903.

Stumpf, D., Krauß, S., Reis, G., Wasenmüller, O., Stricker, D., 2021. SALT: A semi-automatic labeling tool for RGB-D video sequences. In: Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2021) - Volume 4: VISAPP. pp. 595–603. http://dx.doi.org/10.5220/0010303005950603.

Suchi, M., Patten, T., Fischinger, D., Vincze, M., 2019. EasyLabel: A semi-automatic pixel-wise object annotation tool for creating robotic RGB-D datasets. In: 2019 International Conference on Robotics and Automation. ICRA, pp. 6678–6684. http://dx.doi.org/10.1109/ICRA.2019.8793917.

Sun, C., Shrivastava, A., Singh, S., Gupta, A., 2017. Revisiting unreasonable effectiveness of data in deep learning era. In: 2017 IEEE International Conference on Computer Vision. ICCV, pp. 843–852. http://dx.doi.org/10.1109/ICCV.2017.97.

Viviers, C.G.A., Filatova, L., Termeer, M., de With, P.H.N., van der Sommen, F., 2024. Advancing 6-DoF instrument pose estimation in variable X-Ray imaging geometries. IEEE Trans. Image Process. 33, 2462–2476. http://dx.doi.org/10.1109/TIP.2024.3378469.

Wang, H., Johansson, B., 2023. Deep learning-based connector detection for robotized assembly of automotive wire harnesses. In: 2023 IEEE 19th International Conference on Automation Science and Engineering. CASE, pp. 1–8. http://dx.doi.org/10.1109/CASE56687.2023.10260619.

Wang, H., Salunkhe, O., Quadrini, W., Lämkull, D., Ore, F., Despeisse, M., Fumagalli, L., Stahre, J., Johansson, B., 2024. A systematic literature review of computer vision applications in robotized wire harness assembly. Adv. Eng. Inform. 62, 102596. http://dx.doi.org/10.1016/j.aei.2024.102596.

Wang, H., Salunkhe, O., Quadrini, W., Lämkull, D., Ore, F., Johansson, B., Stahre, J., 2023. Overview of computer vision techniques in robotized wire harness assembly: Current state and future opportunities. Procedia CIRP 120, 1071–1076. http://dx.doi.org/10.1016/j.procir.2023.09.127.

Wang, X., Wei, G., Chen, S., Liu, J., 2024. An efficient weakly semi-supervised method for object automated annotation. Multimedia Tools Appl. 83 (3), 9417–9440. http://dx.doi.org/10.1007/s11042-023-15305-0.

Wang, C., Xu, D., Zhu, Y., Martín-Martín, R., Lu, C., Fei-Fei, L., Savarese, S., 2019. DenseFusion: 6D object pose estimation by iterative dense fusion. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 3338–3347. http://dx.doi.org/10.1109/CVPR.2019.00346.

Wei, X.-S., Song, Y.-Z., Aodha, O.M., Wu, J., Peng, Y., Tang, J., Yang, J., Belongie, S., 2022. Fine-grained image analysis with deep learning: A survey. IEEE Trans. Pattern Anal. Mach. Intell. 44 (12), 8927–8948. http://dx.doi.org/10.1109/TPAMI.2021.3126648.

Wirth, F., Quehl, J., Ota, J., Stiller, C., 2019. PointAtMe: Efficient 3D point cloud labeling in virtual reality. In: 2019 IEEE Intelligent Vehicles Symposium. IV, pp. 1693–1698. http://dx.doi.org/10.1109/IVS.2019.8814115.

Wong, Y.-S., Chu, H.-K., Mitra, N.J., 2015. SmartAnnotator An interactive tool for annotating indoor RGBD images. Comput. Graph. Forum 34 (2), 447–457. http://dx.doi.org/10.1111/cgf.12574.

Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., Girshick, R., 2019. Detectron2. https://github.com/facebookresearch/detectron2.

Xiang, Y., Schmidt, T., Narayanan, V., Fox, D., 2018. PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes. In: Proceedings of Robotics: Science and Systems. pp. 1–10. http://dx.doi.org/10.15607/RSS.2018.XIV.019.

Yang, J., Wang, C., Jiang, B., Song, H., Meng, Q., 2021. Visual perception enabled industry intelligence: State of the art, challenges and prospects. IEEE Trans. Ind. Inform. 17 (3), 2204–2219. http://dx.doi.org/10.1109/TII.2020.2998818.

Yousif, I., Burns, L., El Kalach, F., Harik, R., 2025. Leveraging computer vision towards high-efficiency autonomous industrial facilities. J. Intell. Manuf. 36 (5), 2983–3008. http://dx.doi.org/10.1007/s10845-024-02396-1.

Zanella, R., Caporali, A., Tadaka, K., De Gregorio, D., Palli, G., 2021. Auto-generated wires dataset for semantic segmentation with domain-independence. In: 2021 International Conference on Computer, Control and Robotics. ICCCR, pp. 292–298. http://dx.doi.org/10.1109/ICCCR49711.2021.9349395.

Zhang, W., Hao, H., Zhang, Y., Yang, H., Li, X., 2026. State of charge prediction for lithium-ion batteries in electric aircraft based on self-supervised informer. Appl. Soft Comput. 186, 114283. http://dx.doi.org/10.1016/j.asoc.2025.114283.

Zhang, W., Jiang, N., Yang, S., Li, X., 2025. Federated transfer learning for remaining useful life prediction in prognostics with data privacy. Meas. Sci. Technol. 36 (7), 076107. http://dx.doi.org/10.1088/1361-6501/ade552.

Zhou, L., Zhang, L., Konz, N., 2023. Computer vision techniques in manufacturing. IEEE Trans. Syst. Man Cybern.: Syst. 53 (1), 105–117. http://dx.doi.org/10.1109/TSMC.2022.3166397.

Zürn, M., Dzubba, M., Reiff, C., Ajdinović, S., Lechler, A., Verl, A., 2024. Cobot for automated vision data: Streamlining production with automated annotations for machine learning. In: 2024 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications. ACDSA, pp. 1–6. http://dx.doi.org/10.1109/ACDSA59508.2024.10468034.