

Seeing ahead through a state-of-the-art review of visual perception and predictive intelligence for collaborative robots

Mahmoud Salem  · Markus Reischl · Veit Hagenmeyer · Katja Nau · Ahmed Elkaseer

Received: 22 September 2025 / Accepted: 10 September 2026
© The Author(s) 2026

Abstract

The rate at which collaborative robots (cobots) are evolving from isolated, pre-programmed manipulators into collaborative partners that operate productively alongside humans, other robots, and machines is accelerating rapidly. Advances in computer vision is enabling cobots, even in dynamic and cluttered settings, to detect, recognise and assess both environment and actions. However, achieving higher detection accuracy and real-time localization in dynamic systems remains a challenge, as most existing approaches assume uncluttered backgrounds with limited scene complexity. Closing this gap is critical for both research and industrial deployment. This paper reviews the development of visual perception in cobots, surveys emerging computer vision algorithms, and examines a holistic perception pipeline in which raw vision signals, vision language models (VLMs), and augmented/virtual reality interfaces jointly power a unified sense-think-act loop. The review finds that cobots in high-mix, reconfigurable factories face serious challenges: occlusions in fields of view, fluctuating illumination, ambiguous backgrounds, few- or zero-shot object scenarios, high-speed motion, and computational problems in fusing heterogeneous data streams in real time. Cobots also need to manipulate damaged items, requiring vision algorithms that infer geometry from sparse cues and support adaptive grasping. Finally, the review identifies three complementary research directions to unlock plug-and-play cobots: (1) information aware sensor-fusion pipelines transmitting only credible voxels to reduce bandwidth while preserving sub-millimetre accuracy; (2) language conditioned planning and control networks mapping plain-language goals to smooth, jerk-bounded trajectories; and (3) pre-trained few-shot reward heads that score task progress without bespoke metrics, closing the reinforcement learning (RL) loop. Combined, these advances promise cobots that can be commissioned with relatively few demonstrations.

Keywords Collaborative robots · Computer vision · Large-language-models · Augmented/virtual reality · Robot learning · Reinforcement learning

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.



1 Introduction

Collaborative robots, known as cobots, represent a significant and growing revolution in the interactions between humans and robots, robots with each other, and robots with machines in shared work environments. Central to this revolution is the integration of advanced computer vision algorithms and perception technologies to substantially enhance the capabilities of cobots. The pressures for increased productivity and greater efficiency in sectors such as manufacturing, logistics and healthcare are driving the development of cobots which are fundamentally different from traditional robots, being designed to operate alongside humans and other robots both safely and efficiently (Borboni et al. 2023). Cobots are proving transformative, particularly in manufacturing, requiring important changes in task scheduling, safety protocols, and system design to achieve greater production efficiency with resulting greater profitability (He et al. 2020).

To better understand the significance of cobots in industry, it is necessary to consider traditional robots. Usually, traditional robots are designed to handle hazardous materials and perform physically demanding jobs. They are usually isolated occupying a separate workspace having limited interaction with the environment (Javaid et al. 2022a, b). Cobots, however, see Fig. 1 (a), are designed to work alongside and interact with other robots, machines, and humans, engaging in high levels of interaction to perform complex tasks (Krüger et al. 2010, Zhang et al. 2018, Manoharan et al. 2020). They tend to be much less massive than traditional industrial robots, with greater and more flexible mobility which enables easier movement within their operating environments (Feng et al. 2020). A key advantage of cobots over industrial robots is their flexibility; they are capable of performing multiple tasks and can adapt to changes in the work environment and task requirements. These advantages are

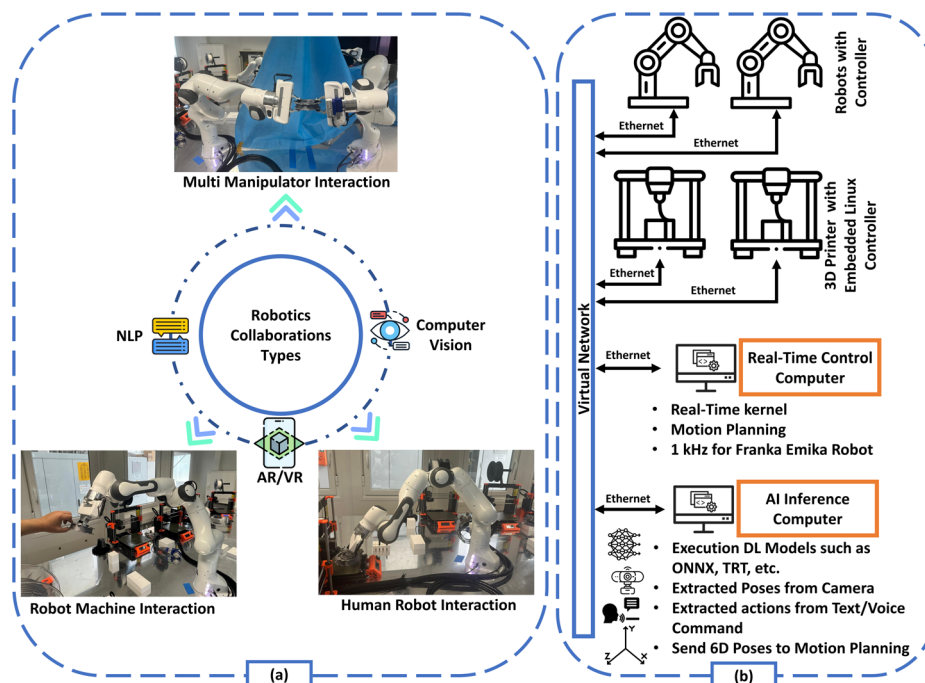


Fig. 1 Collaborative-robot types and architecture of the cobot system setup in the robotics lab at IAI-KIT. **a** Three representative collaboration modes are illustrated: Multi Manipulator Interaction where robot-robot hand-over uses a dual-arm Franka Panda setup (top), Robot Machine Interaction where robot-machine synergy for automated assembly parts with an end-effector spindle (bottom-left), and Human Robot Interaction with human–robot cooperative assembly (bottom-right). **b** A deterministic ethernet network connects two dedicated hosts: a Real-Time Control Computer for controlling the robots, and an AI Inference Computer which is GPU-equipped that runs vision, language and policy networks. It is unlike an isolated industrial robot, a cobot shares its workspace and task with humans and machines (a), and runs its vision, language, and policy models on a networked AI-inference computer working alongside a real-time controller (b)

mainly due to their advanced perception capabilities, such as vision-based object tracking (Bascetta et al. 2011), visual-based avoidance of collisions (Flacco et al. 2012), human-activity monitoring for safety in collaborative tasks (Halme et al. 2018), multimode sensor fusion of vision and haptics for human–robot collaboration (HRC) in tasks such as handing-over and assembling (Bdiwi 2014), and haptic-based perception (Tannous et al. 2020).

Cobots have been, and continue to be widely studied across several applications which shows a dynamic field shaped by evolving technologies and interdisciplinary approaches. Research shows that safe and effective human–robot collaboration depends on accurate control (Hameed et al. 2023), and recent research points to AI-driven control approaches as an increasingly critical part of that toolkit (see Fig. 1b) (Borboni et al. 2023). Recent studies have underlined an important future direction for cobots is to integrate advanced computer vision and computer graphics techniques (Liu et al. 2024a, b, c, d, e). The role of cobots in enhancing human interaction in shared industrial workspaces has been one focus, emphasising their complementarity compared to traditional robots (Vicentini 2020). Recent research focuses on how machine learning (ML) algorithms can enhance cobot performance in industrial settings by enabling safer and more adaptable operations (Shah et al. 2025). Other research emphasizes the integration of cobots into assembly and disassembly lines within the Industry 4.0 framework, revealing their potential for increased flexibility and compatibility with human operatives (Keshvarparast et al. 2024). This research also highlights the promising use of machine learning (ML) to predict human positioning and optimise task scheduling, providing both critical insights for management and identifying future research opportunities.

A major challenge in developing cobots is the creation of systems that can accurately recognize and respond to their surroundings. The production of computer vision systems that enable cobots to accurately interpret visual data is key to solving this challenge, enabling recognition of objects and navigation of complex spaces, to safely and accurately interact visually (Wang et al. 2025a, b, c, d, e) and linguistically (Xia et al. 2025) with other robots and human workers. Such innovative technological functions as advanced object detection, localization, instance segmentation, and pose estimation are essential for complex assembly tasks requiring accurate manipulation of multiple components (Chai et al. 2021). Also, enhanced computer vision improves a cobots' ability to interpret human gestures and movements by estimating human skeletal actions. This improves communication and enables synchronized collaborative actions. Crucially, perception in a collaborative setting is not confined to the physical scene. A cobot shares both its workspace and its task with human operators, machines, and/or robots. Therefore, it must also perceive those partners as agents, which means recognising and estimating the action a partner is performing. Also, this means that estimating the intention, goal, and plan behind an action allows cobots to anticipate and coordinate within the environment rather than only react. This agent-directed perception includes skeleton, pose, and gesture recognition, as well as action and motion-intention prediction. It also includes goal level and plan level inference. It is therefore discussed throughout this review as a complementary axis to the object directed perception of the surrounding scene. In addition to visual capabilities, other sensing methods such as haptic sensing, accelerometers, and force-torque sensing, can provide essential feedback on the cobot's movements and interactions. This feedback enhances cobots' ability to operate appropriately in highly dynamic environments. The integration of augmented and virtual reality (AR and VR, respectively) with computer vision further increases cobots' opportunities to perform different tasks effectively in mixed-reality environments. Also, computer vision enhances scene perception for mobile manipulator robots, which improves their ability to navigate smoothly and safely in complex environments with greater adaptability and awareness. Overall, these methods allow cobots to improve their perception, communication, and interaction abilities, making them more efficient and versatile across numerous applications. It should be noted, however, that although computer vision's capabilities are often motivated by analogy with human perception, machine vision differs fundamentally from human visual perception in its core mechanisms. Throughout this review comparisons between the two are intended to be illustrative rather than to imply functional equivalence.

The large body of existing studies in the area of computer vision for cobot perception as presented by (Liu et al. 2024a, b, c, d, e), (Keshvarparast et al. 2024) and (Shah et al. 2025) comprehensively covers limited topics within visual perception as related to cobots, presenting an overview of the role of vision-language models and

the potential of AR/VR in enhancing cobot perception but omits outlining future directions that could significantly impact both research and industry. This paper, however, attempts to bridge these gaps by addressing the following critical questions:

- What is the current state of visual perception in cobots?
- Which recently developed algorithms, models, and datasets in computer vision could positively impact visual perception in cobots?
- What are the most important current challenges in visual perception for cobots?
- What future directions and potential solutions could overcome these challenges?

This review aims to deliver a comprehensive, forward-looking synthesis of visual-perception and predictive-intelligence techniques for cobots by focusing on language and AR/VR for vision systems. Firstly, it reports the current state of the art in computer-vision algorithms, sensor-fusion frameworks, vision-language methods, and multimodal sensing for safe and efficient human–robot collaboration. Secondly, it critically evaluates emerging lines of investigation, such as event-based vision, large-language-model-driven decision support, and AR/VR-mediated supervision, assessing their potential for overcoming real-world limits in dynamic industrial environments. Thirdly, it reports research and development in predictive safety, zero- and few-shot object perception, and deformation-awareness, areas that could enhance a cobot’s adaptability and performance in complex, rapidly evolving manufacturing settings.

Realising this potential, however, depends on closing a persistent gap between what visual perception for cobots can already demonstrate in controlled settings and what dependable, such as plug-and-play collaboration on the factory floor actually requires. Across the studies surveyed here, this gap recurs through a small number of broad challenges. These include, maintaining robust perception under occlusion, clutter, variable illumination, and part-to-part variability in real industrial scenes. They also include integrating heterogeneous and multi-robot sensing in real time without prohibitive computational or communication costs, and moving from reactive safety toward genuinely predictive, human-aware safety. Further challenges concern is to enable learning that adapts to new tasks in a stable and data-efficient manner, such as learning with adequate shared datasets and benchmarks. Then, deploying these capabilities in a trustworthy and secure manner as the underlying computing and sensing technologies continue to develop. These challenges arise from the analysis and are refined in the concluding synthesis, with each point following directly from the preceding sections. They are prefigured here so that the distance between what cobot perception can already achieve and what truly plug-and-play collaboration still demands is clear from the outset.

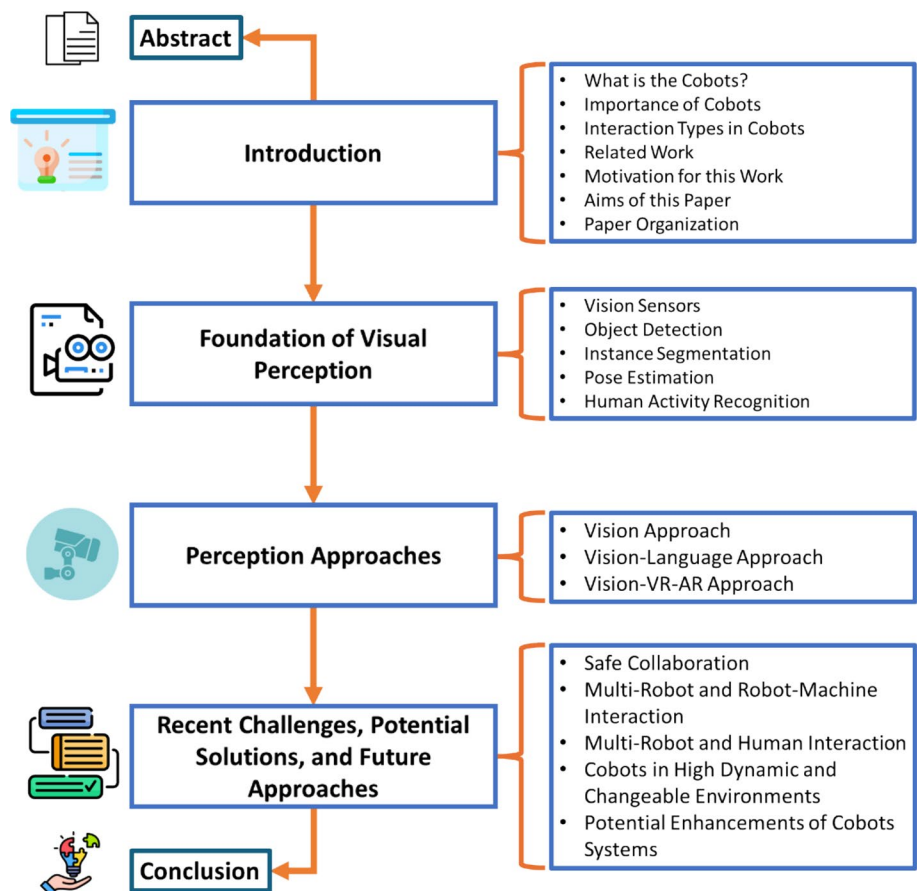
The approach used in this study to identify papers for review was based on a defined set of criteria to ensure relevance and depth. In particular, relevant works were identified through a structured search of IEEE Xplore, ScienceDirect, SpringerLink and Google Scholar, supplemented by arXiv to capture very recent vision-language and learning-based work. The search combined terms describing the collaborative setting ("cobot", "collaborative robot", "human–robot collaboration", "human–robot interaction") with terms describing the perception technology ("computer vision", "visual perception", "pose estimation", "sensor fusion", "vision-language model", "augmented reality", "virtual reality"). This initial set was extended through backward and forward citation tracking, following the reference lists and citing works of the most relevant papers. Priority was given to recent publications, reflecting the rapid pace of the field, although older works were retained where they establish the concepts and benchmarks on which current methods build. The retained studies were then grouped by perception approach such as vision, vision-language, and vision with AR/VR, and each group was examined not only for the capabilities reported but also for the operating conditions under which those capabilities were demonstrated and those under which they were left untested. Limitations recurring across the groups were consolidated into the challenges set out in Sect. 4, and the future directions proposed there were derived from those challenges rather than collected from the stated aims of the source papers, so that each direction responds to a gap evidenced by the reviewed literature.

The review is structured as shown in Fig. 2. Following the introduction, the Foundations of Visual Perception section discusses vision sensing methods and scene understanding algorithms. The Perception Approaches section reviews visual perception in collaborative robots and focuses on three main approaches: vision, vision-language and vision-AR/VR. Next, the paper explores the future direction, limitations, challenges, and potential solutions. The final section provides, as a conclusion, the key findings of the review.

2 Foundations of visual perception

Before visual perception can support collaborative behaviour, a cobot must first sense its surroundings and convert raw signals into structured scene information. This section begins the domain-general foundations on which that capability builds, and it is organised in two sub-sections. Firstly, Sub-Sect. 2.1 surveys the vision sensors that supply the input data, across 2-D, 3-D, and point-cloud modalities. Also, this subsection outlines the trade-offs that make each suited to particular tasks. Secondly, Sub-Sect. 2.2 reviews the core computer vision algorithm families that turn this sensor data into actionable scene understanding, such as object detection, instance segmentation, pose estimation, and human activity recognition. These sensors and algorithms are not specific to collaborative robots but they constitute the building blocks of modern visual perception more broadly. After that, Sect. 3 builds directly upon them to show how these components are composed and integrated into the deployed perception systems that operate within real collaborative robot settings.

Fig. 2 Review organization: It proceeds from scene understanding (Sect. 2), through the state of the art in vision, vision-language, and VR/AR (Sect. 3), to challenges and future directions (Sect. 4), before concluding



2.1 Vision sensors

Cobots use various vision sensors to improve their interactions with the environment, to enhance both safety and efficiency (Javaid et al. 2022a, b). These sensors provide either 2-dimensional (2-D) or 3-dimensional (3-D) images to represent the surroundings and can provide point clouds as well.

2-D vision systems capture visual data in two dimensions, height and width, similar to a regular camera. Due to their simplicity, these systems offer quick processing and are suitable for tasks that do not require depth perception, including reading barcodes, detecting the presence of objects, or simple manufacturing tasks. The inability to measure distance or depth of objects directly can be overcome by inferred depth from visual cues such as colour contrast in the red–green–blue (RGB) colour space (D’Emilia and Di Gasbarro 2017), or employing multiple 2-D cameras to reconstruct a 3-D vision of the scene (Bellandi et al. 2012).

Different types of 2-D vision sensors have different purposes. RGB cameras, the most basic type, capture two-dimensional images (Maxim et al. 2023) and are commonly used for basic object recognition (Cazzato et al. 2020) and visual inspection (Minz and Saini 2021). Thermal cameras detect infrared energy, making them useful for detecting and/or tracking in low-visibility conditions (Wilson et al. 2023), they have been widely applied for inspection in agricultural processes (Yun et al. 2024) and for privacy enhancement in HRC in cobots applications (Lintvedt 2023). Hyperspectral and multispectral cameras, which capture a range of wavelengths beyond the visible spectrum (Bhargava et al. 2024) are used in pharmaceutical robotics for quality inspection (Su et al. 2022). Dynamic vision sensors (DVS), also known as event cameras (Gallego et al. 2022), are designed to detect changes in light intensity at each pixel independently, providing high temporal resolution for dynamic environments and have been used successfully in cobot applications such as HRC (Duarte et al. 2024) and multi object tracing in highly dynamic robotics environments (Apps, Wang et al. 2025a, b, c, d, e).

3-D vision systems capture depth information in addition to length and width, offering a full spatial representation of the environment, which can be used in robot manipulators (Cong, Chen et al. 2023a, b) and in autonomous driving (Contreras, Jain et al. 2024). This capability enables cobots to perform tasks that require precise manipulation, navigation around obstacles, and safe collaboration with humans or other machines (Boschetti et al. 2023) including such high-precision tasks as robotic assembly (Yan et al. 2021). Stereo vision systems for depth detection use two 2-D cameras placed at different viewpoints to capture images simultaneously; depth is calculated by comparing these images which have high impact in robot trajectory planning (Zaki et al. 2023) and cobots system (Mohamed et al. 2017). Stereo vision can also be integrated with event-camera systems to assess scene geometry by accumulating sensor observations over many frames or event streams rather than single frames and fusing them together using the camera’s estimated motion/pose (Zhou et al. 2018). Time-of-flight (ToF) cameras determine the time taken for light to travel to the object and back (He and Chen 2019), enabling the construction of a depth map (Horaud et al. 2016). Structured light cameras project a pattern onto a scene and measure deformation to construct 3-D shapes (Zanuttigh et al. 2016). Finally, light detection and ranging (LiDAR) emits laser pulses and measures the reflection time to create accurate, detailed point clouds, making it ideal for applications requiring high-precision at long-distances (Zhong et al. 2021) (Fig. 3).

2.2 Cobot perception methods

2.2.1 Overview

As previously stated, different computer vision techniques can be utilised to accurately identify, locate, and differentiate among multiple objects in a scene for industrial applications. When integrated into cobots systems, these techniques enhance the cobots’ ability to perform tasks safely and efficiently alongside other robots, machines, and human workers (Zheng et al. 2025). These vision techniques include object detection based on 2D/3D vision to localize objects in an image or scene (see Fig. 4c) (Wang et al. 2025a, b, c, d, e), which can also be performed using event-based vision, a method that focuses on pixel-level intensity changes (Cho et al. 2025).

Fig. 3 Perception pipelines organized by output geometry. **a** 2-D Vision Frames: a monocular RGB path and thermal IR path **b** 3-D Vision Frames: a RGB Stereo Camera path and RGB Depth Camera path. **c** Points Clouds: LiDAR path. Where perception inputs are grouped by the geometry they produce, from 2-D frames (RGB and thermal), to 3-D frames (stereo and depth), to LiDAR point clouds

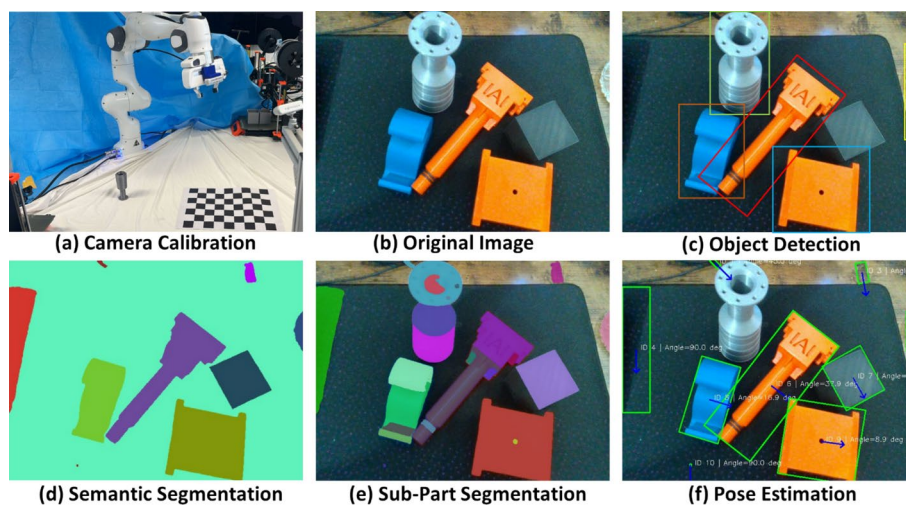
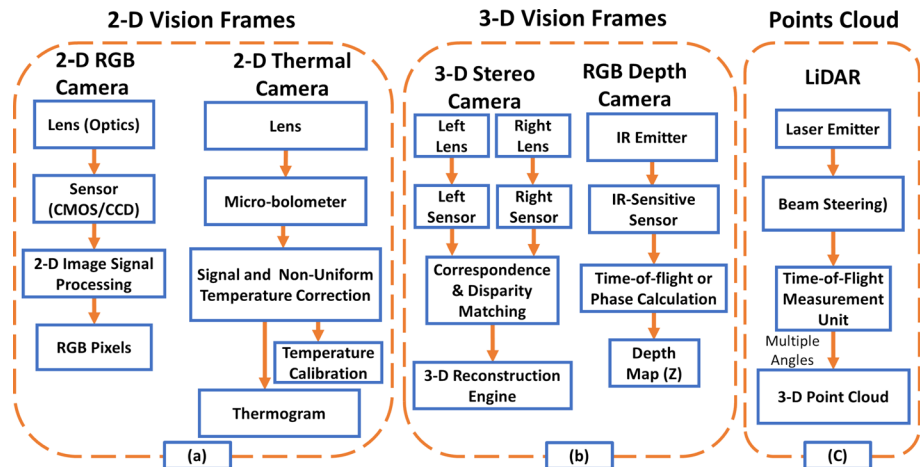
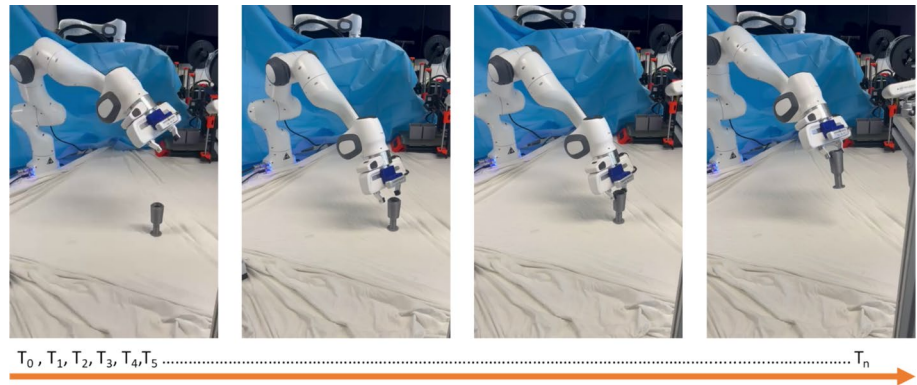


Fig. 4 Vision perception pipeline used by a collaborative robot in robotics lab at IAI-KIT; **a** Camera and workspace calibration with a chess board target, **b** Raw RGB image of a cluttered parts tray, **c** Object-detection stage producing class-coloured bounding boxes, **d** Semantic-segmentation map assigning a label to every pixel, **e** Instance-segmentation masks isolating each individual part, **f** 6-DoF pose-estimation and grasp-proposal output, green oriented boxes and arrows denote feasible pick poses, ready for downstream motion planning. As shown, a vision pipeline turns raw images into usable scene knowledge step by step, detecting objects, segmenting them, and estimating their pose, until the result is ready for motion planning

Semantic segmentation is another key technique, as it assigns a class label to every pixel and generates masks that delineate scene elements (Kerssies et al. 2025). Also, real-time variants of semantic segmentation are particularly important for enabling fast processing in critical applications (Yang et al. 2025a, b, c, d, e). Moreover, to provide more detailed delineation, instance segmentation identifies individual object instances in a scene (see Fig. 4e) (Shi, Zhang et al. 2025), while part segmentation, also termed as object-part parsing, detects subcomponents of objects (Li et al. 2024). Furthermore, panoptic segmentation combines both semantic and instance segmentation, offering a unified representation of the scene.

Beyond recognizing and segmenting objects, other methods aim to estimate an object's 3D position and orientation, a method termed as object pose estimation (see Fig. 4f) (Shi, Zhang et al. 2025). Finally, to incorporate the temporal dimension, temporal understanding addresses time-varying data in vision systems, which is particularly important in cobot tasks (Sampieri et al. 2022). It enables applications such as pose forecasting and robot action segmentation (see Fig. 5) (Kimhi et al. 2025).

Fig. 5 Temporal data sequence for action understanding in robotics lab at IAI-KIT. Snapshots T_0 to T_n capture a cobot as it approaches, grasps, and lifts a fastener. Such time-ordered visual data enable temporal-action segmentation, motion prediction, and adaptive control. The ordering visual snapshots over time is what lets a cobot segment actions, forecast poses, and adapt its control, the temporal foundation of perceiving an agent



2.2.2 Object detection

The purpose of object detection is to identify both the class and location of objects within a scene. Object detection can be applied using traditional computer vision which relies on feature engineering and signal processing algorithms (Pagire et al. 2025). However, advances in deep learning (DL) have led to substantial progress in modern computer vision (Edozie et al. 2025) not only in object detection but also in other methods (Chai et al. 2021). These have significantly improved cobot task performance such as object grasping (Chen and Lu 2021). The effect of DL can be seen in the Regional Convolutional Neural Network (R-CNN) widely regarded as marking a shift from traditional methods to CNN-based feature extraction, significantly improving accuracy and generalization (Ren et al. 2016).

R-CNN generated a round 2,000 regions of interest (ROIs) through Selective Search, processes each with a pre-trained CNN, classified by a class-specific Support Vector Machine (SVM), and refined boxes with regression (Cervantes et al. 2020). The R-CNN approach was accurate but slow (Girshick et al. 2014). Fast R-CNN addressed this limitation via a structure that pooled regions directly from the convolutional feature map, eliminating redundant processing. This improved speed and accuracy but still relied on Selective Search, reducing real-time performance (Girshick 2015). Faster R-CNN was developed based on the integration of candidate generation, feature extraction, classification, and regression within a single CNN. This architecture enabled near-real-time speed, but the large number of architectural parameters tends to reduce detection speed (Ren et al. 2016). The Region-based Fully Convolutional Network (R-FCN) reduced the number of parameters, but it still struggled with small or densely packed objects (Rettenberger et al. 2023). Mask R-CNN extended Faster R-CNN with pixel level segmentation, which provides precise detection but at a high computational cost (K and G 2024).

In parallel, fully convolutional encoder-decoder architectures such as U-Net have been widely used for pixel level segmentation in robotics tasks, such as object grasping tasks (Rasheed et al. 2024) and cobot detection (Chmielowiec 2025). U-Net is based on a contracting path to capture context, which is termed the encoder, and an expanding path with skip connections to recover spatial detail, which is termed the decoder (Ronneberger et al. 2015). This architecture enables precise delineation of object boundaries even with limited training data (Salem et al. 2026). In detection pipelines, U-Net style architectures are often coupled with backbone features to refine masks or to provide high-resolution semantic cues that benefit small object detection and scene understanding, offering a practical balance between accuracy and efficiency in complex industrial settings. The U-Net architecture can be proposed in combination with rule-based metrics, termed as Attention U-Net, to be used in many application such as welding line detection (Lukács et al. 2025).

For real-time applications, single-stage methods such as Single Shot MultiBox Detector (SSD) and You Only Look Once (YOLO) prioritize speed. YOLO divides images into grids, predicting bounding boxes and classes simultaneously but struggles with small objects or cluttered fields of view. YOLO has achieving high precision by introducing anchor boxes and clustering for better small-object detection. Successive YOLO versions (YOLOv2

to YOLOv11) continuously improved feature extraction, data augmentation, and architecture for enhanced speed and accuracy (Sapkota et al. 2025). YOLOv11 was a single-stage, anchor-free detector that predicted classes and bounding boxes over multi-scale feature maps in one forward pass (Terven et al. 2023). The network followed the standard YOLO design with an improved backbone and neck for stronger multi-scale features, fused by a Feature Pyramid Network (PAN) and Path Aggregation Network (FPN) style neck (Khanam and Hussain 2024). A decoupled detection head made separate classification and box-regression predictions at three scales (P3-P5) to cover small-to-large objects. During training, YOLOv11 used task-aligned label assignment and a composite loss that combined focal/varifocal classification terms with IoU-based box loss and Distribution Focal Loss (DFL) for localization. At inference, per-scale predictions were decoded and filtered to produce the final detections. The same backbone and neck paired with task-specific heads for instance segmentation (via a ProtoNet mask branch), keypoints/pose, and oriented bounding boxes (OBB), which enabled a unified, high-throughput perception stack with a good speed-accuracy trade-off (Hidayatullah et al. 2025). YOLO can detect small objects in complex environments, such as underwater (Chen and Er 2024). YOLO is not restricted to work with 2-D and 3-D images but it can also work with point clouds of LiDAR and Radar (Berens et al. 2024).

SSD directly predicted class confidence and bounding box offsets using anchor boxes. It improved upon YOLOv1 by addressing small-object detection through multi-scale feature extraction and reducing parameters via fully convolutional networks (Liu et al. 2016). Feature Fusion SSD advanced SSD's capturing of context information by fusing fine details with contextual information across layers, enhancing small-object detection (Li et al. 2017). The difficulty SSD experienced with imbalanced data during training was addressed by RetinaNet, by introducing a focal loss function that emphasized harder-to-classify samples, thus enhancing detection performance (Lin et al. 2018).

To resolve the challenges posed by data imbalance and detection of objects at different scales, Libra R-CNN introduced several innovations. First, it used IoU-balanced sampling to ensure that both easy and hard training samples were represented equally, which led to an enhanced balanced learning. Second, the Balanced Feature Pyramid (BFP) was introduced to manage multi-scale features, which allowed the model to capture more details effectively across various object sizes. Lastly, balanced L1 loss was applied, which is defined as an L1/Huber-style regression loss that down-weights very small and very large errors while emphasizing medium errors. Thus, based on balancing inliers and outliers, the model could adjust training contributions from different samples to keep the learning process stable and consistent. Together, these techniques aided Libra R-CNN to reach a significant accuracy improvement in the detection of varying scales' objects (Pang, Chen et al. 2019a, b). In parallel, TridentNet introduced a multi-branch network architecture that was developed to generate distinct feature maps for different scales. Each branch captured features specific to small, medium, or large objects by adjusting the receptive field size. This architecture enabled TridentNet to detect objects of different scales more effectively in a single pass (Li, Chen et al. 2019a, b).

CNN-based object detection methods typically rely on sliding windows to detect local features within an image. However, this approach limits the model's ability to capture global context across the entire image, because it focuses on localized regions. Transformer-based methods, such as DETection TRansformer (DETR) and Vision Transformer (ViT), overcame this limitation by capturing both local and global relationships, which offered a more holistic view of the scene.

DETR adapted the Transformer architecture from natural language processing to object detection. It directly modeled relationships between objects across the whole image, which allowed it to predict object categories and locations without the need to perform additional post-processing tasks (Carion et al. 2020). Although DETR achieved accurate results, it had high computational costs and struggled with small-object detection. To address these challenges and improve efficiency, Deformable DETR introduced sparse sampling, which focused computational resources on relevant regions (Zhu et al. 2020).

Vision Transformer (ViT) applies the Transformer approach to image classification tasks. It divides an image into patches and processes each patch through a Transformer encoder. This patch-based approach allows ViT to capture global image context effectively (Dosovitskiy 2020). Data-efficient Image Transformers (DeiT) further

improved the model's performance on small datasets by using a distillation-based training approach, which makes it more effective in cases of limited data (Touvron et al. 2021).

Overall, these advancements in object detection techniques, from R-CNN to YOLO, SSD, and Transformer-based models, have enabled increasingly accurate, adaptable and efficient object detection, providing robust frameworks for a wide range of cobotic applications.

All the algorithms discussed above were initially designed for 2-D images, limiting detection, classification and localization to a plane, which is inadequate for cobots. Effective cobot interaction requires 3-D information. Extracting 3-D objects from 2-D images was achieved through monocular, feature-matching, and multi-camera approaches, while 3-D image detection relied on stereo, LiDAR, and fusion techniques (Niedermaier et al. 2025).

In what follows, the authors review six complementary 3-D perception approaches for cobots which are monocular 2-D-to-3-D inference, feature-matching with CAD, multi-camera/BEV methods, stereo-based depth, point-/voxel-based LiDAR and RGB-D, and multi-modal fusion.

Firstly, Monocular approaches employed established 2-D frameworks with geometric priorities to estimate 3-D information from a single 2-D image. Mono3-D network assumed objects lay on the ground plane, generating 3-D candidate boxes through energy minimization, filtered by prior known geometric features and refined with a convolutional network (Chen et al. 2016). To handle 3-D parameters such as position, size, and rotation within a single regression loss, MonoDIS separated the regression loss by parameter (loss disentanglement), improving training stability and reducing parameter interference. MonoDIS also introduced an IoU-based loss (signed IoU) for 2-D detection to better align optimization with the IoU evaluation metric, enhancing alignment between predicted and ground-truth boxes, enhancing generalization across object scales and shapes (Simonelli et al. 2019). MonoGRNet (Qin et al. 2019) enhanced monocular 3-D detection by dividing the task into four steps, which are 2-D detection, local feature extraction, instance depth estimation, and 3-D bounding box prediction. This modular approach utilized semantic features from 2-D images to improve inference speed by avoiding complex proposal generation. It also featured an Instance Depth Estimation module that efficiently estimated an object's centre depth, which helped reduce computational demand in real-time applications. MonoRCNN further refined distance prediction by decomposing object distance into physical and projection heights. This method improved the geometric interpretation of the object-to-camera relationship, which resulted in robust and accurate distance estimations. This made it more suitable for many real-world applications (Shi et al. 2021).

Secondly, in robotics, feature matching-based approaches combined with computer-aided design (CAD) data have proven highly effective for object detection and pose estimation. These methods used 3-D CAD data to identify and localize objects in real environments, integrating DL with 3-D geometry. Local surface embeddings aligned 2-D image points with 3-D CAD points, and models such as Mask R-CNN when paired with Random SAmple Consensus (RANSAC) achieved accurate 6D pose predictions (Pitteri et al. 2020). In complex environments with occlusion, an unsupervised method was used to learn dense 2-D/3-D correspondences, which is per-pixel/per-point matches across the image and object surface, not just sparse keypoints, without prior knowledge of visual ambiguities, enabling the effective sampling, scoring, and refining of pose hypotheses. Using a contrastive loss function which pulls embeddings of matched pairs together and pushes mismatched pairs apart with a margin in latent space, these correspondences were represented in latent spaces through a dual model structure, accurately encoding complex visual information even in the presence of ambiguities (Haugaard and Buch 2022). Additionally, and with minimal training, these algorithms generalized well on new data, allowing object detection and pose estimation without prior 3-D model knowledge or object-specific training. A single image input could be used to predict relative poses for future images via a model trained to generate discriminative embeddings with U-Net architecture and attention mechanisms, facilitating generalization to new objects (Nguyen et al. 2024). This approach could also detect unseen objects or partially occluded items by learning local representations from CAD models (Nguyen et al. 2022) or depth models (Örnek et al. 2025) for template matching, effectively achieving pose estimation for new objects with minimal training. A so-called Siamese neural network has been used to detect objects by comparing two image inputs, one from the observed scene and one from a reference, a capability

that is particularly useful in robotic applications (Cabrera et al. 2024), but is also widely used in high speed object tracking tasks of unmanned aerial vehicles (Fu et al. 2023).

Thirdly, multi-camera-based approaches enabled enhanced 3-D feature prediction from 2-D images by estimating object bounding boxes from the multiple camera inputs. These methods utilized Bird's Eye View (BEV) representations to provide a global view of the scene that highlights object locations and scales, mitigates occlusion, and supports temporal and spatial fusion for multi-frame integration. The BEVDet model (Huang, Huang et al. 2021) improves 3-D detection by projecting 2-D image features into a BEV format, enhancing spatial comprehension, while data augmentation and an upgraded Non-Maximum Suppression strategy (Hosang et al. 2017) further enhances robustness, reduces overfitting, and refines bounding box accuracy. Additionally, BEV methods use learnable queries to map 2-D features into BEV sequences, enabling 3-D predictions to be made using a decoder. For greater adaptability BEVFormer introduced deformable sparse attention, enabling BEV queries to focus on relevant image regions, enhancing efficiency and reducing reliance on precise calibration (Li et al. 2022). BEVFormer v2 optimised temporal fusion by aligning historical and current BEV features, supporting better model performance in dynamic environments (Yang, Chen et al. 2023a, b), while BEVerse integrated 3-D detection, semantic mapping, and motion prediction into a unified framework through spatio-temporal BEV representations and multi-task learning (Zhang et al. 2022). Multi-2-D-frame methods such as BEVDepth optimised depth prediction by integrating depth supervision and spatial alignment, improving accuracy by using direct depth data (Li et al. 2023). BEVStereo enhanced depth perception in dynamic scenarios by using temporal stereo, and processing sequences of multi-view 2-D images (Li et al. 2023).

Fourthly, stereo-based approaches utilized stereo camera frames to enhance depth perception and reduce the effect of background noise through additional geometric constraints. As with 2-D detection, Stereo R-CNN extends the Faster R-CNN model for 3-D tasks, providing faster processing with added accuracy by introducing a keypoint prediction branch and photometric alignment, which enhance 3-D positioning precision (Li, Chen et al. 2019a, b). The Deep Stereo Geometry Network (DSGN) provides an accurate assessment of 3-D scenes from stereo images, bypassing costly LiDAR sensors, by constructing a 3-D volume from stereo pairs, estimating depth and detecting objects in a single-stage pipeline. An improved version, DSGN++, refines this with depth-wise plane sweeping for enhanced 2-D-stereo volume connections and Dual-view Stereo Volume to combine front- and top-view features for diversely shaped objects (Chen et al. 2022).

Inspired by the YOLO model, YOLOStereo3-D avoided traditional disparity estimation, expanding 2-D detection with stereo features for efficient 3-D detection. YOLOStereo3-D used pixel-level correlations to balance stereo and monocular features. It incorporated a densely connected ghost module and a hierarchical fusion structure. This structure enabled accurate 3-D detection without LiDAR (Liu, Wang et al. 2021). To reduce depth instabilities, the Continuous Disparity Network replaced conventional regression loss with a Wasserstein distance-based loss function, which allowed the model to cope with diverse disparity values. This approach stabilized gradients and improved depth estimation accuracy, which provided a robust solution for 3-D feature extraction (Garg et al. 2020).

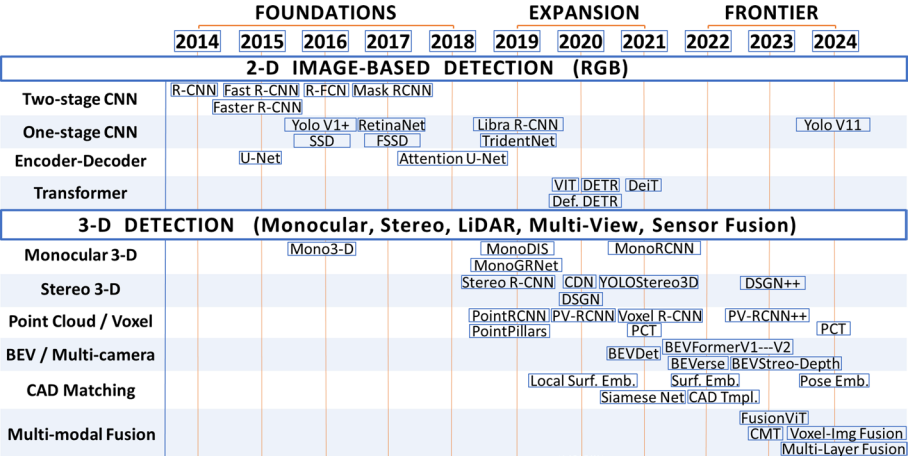
Fifthly, the introduction of point-cloud and voxel-based methods marked a significant advancement in the detection precision of 3-D objects. This was achieved using LiDAR and RGB-D sensors, which have gained popularity for their ability to provide direct and precise 3-D spatial information. In contrast to traditional 2-D data, which require complex computations to infer depth, 3-D sensor signals are rich in geometric information. This richness has enabled the development of sophisticated detection frameworks. Two primary approaches have emerged for 3-D data representation, which are voxel-based and point-based. Voxel-based methods efficiently extract features by dividing data into grids, but at the expense of accuracy. In contrast, point-based methods retain precise point positions and give greater accuracy but require more computation due to the unordered storage of point-level features. Building on the R-CNN architecture, Voxel R-CNN demonstrated that precise point positioning was not essential for high-performance detection. Instead, coarse voxel granularity could achieve comparable results with correspondingly lower computational costs (Deng, Shi et al. 2021). While voxel-based methods were efficient, point-based techniques such as PointRCNN (Shi et al. 2019) directly utilized raw point-cloud data,

employing a two-stage framework that segmented point clouds into foreground and background using both local and global features to refine proposals (Shi et al. 2019). To merge the strengths of both voxel- and point-based methods, PV-RCNN and PV-RCNN++ integrated features from both, improving 3-D detection by encoding scene information through voxel grids and refining key points with point-based features (Shi et al. 2020; 2023). An important addition to these approaches was the Point Pillars method, which sought to balance speed and accuracy by organizing point cloud data into vertical columns, or "pillars" and using a PointNet-based network for effective feature detection (Lang et al. 2019). This encoding, processed with 2-D convolutional architecture, created a highly efficient pipeline with both speed and accuracy, surpassing other methods then available (Qi et al. 2017). The accuracy of point-cloud processing has been improved by using Vision Transformer (ViT) architectures, commonly known as Point cloud transformer (PCT) (Guo et al. 2021). These models learn high-quality feature representations (Wang et al. 2024a, b, c, d, e).

Finally, fusion-based approaches which integrate various vision sensor types, combine multi-modal data such as 2-D RGB images and point clouds, to enhance 3-D object detection, improving representation learning and detection accuracy. The FusionViT model exemplified this approach by adopting a hierarchical, Transformer-only framework that embedded both images and point clouds for more robust learning (Xiang and Zhang 2023). Similarly, the Cross Modal Transformer processed images and point cloud tokens (each of which contain fine-grained information) into precise 3-D bounding boxes without explicit view transformations, aligning multi-modal data by encoding 3-D points into unified features (Yan, Liu et al. 2023a, b). This approach includes a Voxel Enhancement Module and an Image Enhancement Module, which update each other to bridge the modality gap, enhancing LiDAR and camera-based spatial awareness and adaptively selecting modality-specific features for a unified 3-D representation (Song and Wang 2024). Moreover, sensor fusion proved effective for the detection of small objects without limiting detection of larger items. A multi-layer fusion network applied a dense fusion approach, merging features mapped across layers to retain more useful information. An adaptive fusion module was further incorporated to minimize background interference, enhancing small object detection by adaptively weighting and fusing feature layers (Guo and Hu 2024).

Figure 6 closes this section with a chronological overview of the object-detection methods surveyed above; the corresponding full catalogue, method, input, key contribution and reference, is provided in Table A.1 in Appendix. The methods are placed on a shared 2014–2024 timeline and grouped into swim-lanes by family, separating the two-dimensional detectors, such as two-stage, one-stage, encoder-decoder and transformer, from the six three-dimensional families such as monocular, stereo, point-/voxel-based, bird's-eye-view multi-camera, CAD-matching and multi-modal fusion. Read from left to right, the diagram makes the maturation of the field explicit: research begins almost exclusively with two-stage 2-D detectors and progressively expands until, in the most recent years, nearly every family is active, evidencing a clear movement from RGB-only 2-D detection towards 3-D, multi-view and sensor-fusion perception better suited to the spatial demands of collaborative

Fig. 6 Chronological overview of the object-detection methods reviewed in Sect. 2.2.2, arranged on a 2014–2024 timeline and grouped into swim-lanes by family (two-dimensional detectors above; the six three-dimensional families below). Horizontal position denotes the original publication year, solid markers indicate landmark methods, and the period brackets summarise the dominant research theme of each interval. Full method details and references are listed in Table A.1



robotics. Landmark methods are emphasized within each lane to indicate the principal milestones, and the period annotations capture the dominant research theme of each interval. Readers are referred to Table 1 for the detailed attributes of each method represented in the figure.

2.2.3 Instance segmentation

Instance segmentation is focused on identifying all instances of a specific category within an image, marking the exact pixels associated with each instance. Unlike traditional bounding box detection, which outlined objects within boxes, instance segmentation required a detailed separation of each individual object instance rather than simply recognizing object classes at the pixel level.

Alongside the existing bounding box recognition branch, the addition of Mask R-CNN (see Sect. 2.2.2) serves as an effective method for instance segmentation of a branch to predict object masks. This dual-purpose model has been efficient at detecting objects while generating high-quality segmentation masks for each instance. Training Mask R-CNN is straightforward, adding only minimal computational overhead when running at 5 fps (frames per second), and can be generalized to include other tasks, such as human pose estimation, within the same framework (He et al. 2017). Another model, MaskLab (Chen et al. 2018), based on the Faster RCNN object detector (see Sect. 2.2.2), addressed the instance segmentation problem by refining object detection through semantic and directional features, producing three outputs: box detection, semantic segmentation, and direction prediction. This approach accurately localized object instances using predicted boxes and combined semantic segmentation with directional cues to separate instances within the same semantic class (Chen et al. 2018). Also, Mask RCNN outperforms U-Net in segmenting overlapping object, such as cells in biomedical engineering, and achieves comparable performance if objects do not intersect (Rettenberger et al. 2023).

Further advancements included the Hybrid Task Cascade, inspired by Mask R-CNN (Kijdech and Vongbun-yong 2024) and Cascade R-CNN (Cai and Vasconcelos 2018), which enhanced instance segmentation by utilizing the reciprocal relationship between detection and segmentation (Chen, Pang et al. 2019). Another innovation, Point-based Rendering (PointRend), draws on computer graphics algorithms which approach segmentation as a rendering problem, addressing over- and under-sampling challenges in pixel labelling. PointRend performs point-based predictions at adaptively selected locations using an iterative subdivision algorithm, which makes it suitable for both instance and semantic segmentation tasks (Kirillov et al. 2020).

In contrast to Mask R-CNN, which detects bounding boxes first, TensorMask employs a dense sliding-window method to treat each spatial location's output as a 4D tensor. This architecture allows TensorMask to use instance segmentation as a means of tensor prediction by utilizing novel operators specifically designed for this approach (Chen et al. 2019). To achieve real-time performance, the YOLACT and YOLACT++ models introduced a method that decomposed instance segmentation into two parallel tasks, which are generating prototype masks and predicting per-instance mask coefficients (Bolya, Zhou et al. 2019a, b). This process provided high-quality masks with temporal stability at a minimal computational cost (Bolya, Zhou et al. 2019a, b).

The Segmenting Objects by Locations (SOLO) method transformed instance segmentation into a classification task by assigning "instance categories" to each pixel based on location and size. This helped to simplify the segmentation process while maintaining accuracy. SOLO's framework mapped input images to instance masks without post-processing or bounding-box detection (Wang et al. 2020). BlendMask combined top-down and bottom-up methods, integrating instance-level information with fine-grained semantic details for enhanced mask prediction. High-level instance data from a top-down perspective provides context about object boundaries and general structure, but the bottom-up layer contributes pixel-level details for fine-grained segmentation (Chen et al. 2020).

All the algorithms discussed above were designed for 2-D images and generated 2-D bounding boxes, limiting detection to classification and localization on a plane, which proved inadequate for cobots. Effective robot interaction requires 3-D information. In what follows, we review six approaches: firstly, monocular 2-D-to-3-D inference; secondly, CAD-based matching; thirdly, multi-camera/BEV methods; fourthly, depth-camera (RGB-D)

approaches; fifthly, 3-D point/voxel-based segmentation and detection; and finally, multi-modal fusion of RGB and LiDAR.

Firstly, monocular methods infer 3-D structure and human/object pose from a single RGB image. Switching from 2-D to 3-D vision: the Shape-Aware Monocular 3-D Object Detection method combines an instance segmentation head with 3-D bounding box projections, concentrating on visible shapes in single RGB images, effectively reducing interference from irrelevant regions and improving detection of occluded objects (Chen et al. 2023a, b). This detection method also introduces the Average Depth Similarity (ADS) metric to compare predicted and ground-truth depths which shown improvements overlooked by more traditional IoU-based evaluations. By uniting instance segmentation with ADS, this method achieves greater accuracy with robust 3-D detection in challenging test cases. In order to reduce or even remove camera dependency, the MonoCInIS approach separates intrinsic parameters from the core learning process (Heylen et al. 2021). It begins with instance segmentation, predicting 3-D reference points at the pixel level to enhance detection precision across various viewpoints. The model employs camera-independent geometric reasoning to estimate 3-D poses and so does not rely on intrinsic camera parameters. In post-processing, camera parameter 2-D reference points are mapped into 3-D space for accurate localization. MonoCInIS uses diverse datasets to outperform earlier methods based on the KITTI Vision benchmark (Geiger et al. 2013), setting a new standard of camera independence. A cross-dimensional 3-D instance segmentation (CD3IS) approach has achieved 3-D representation using only monocular RGB images, bypassing the need for specialized 3-D devices and extensive labelling (Tang, Chen et al. 2024a, b). It starts by a Double-Snake Multitask Network estimated depth while performing instance segmentation, using depth cues to refine object delineation. Then, an Adaptive Point Cloud Filtering Algorithm generated and refined 3-D point clouds from single images, guided by depth and colour data to better distinguish objects. Validated on datasets such as Cityscapes and SOP, CD3IS demonstrated a strong performance and practical feasibility in real workshop environments.

Secondly, CAD-based matching has been used to align real-world imagery with accurate 3-D object representations utilizing CAD models for novel object segmentation and shape reconstruction to obtain robust 3-D insights for minimal manual effort. One such algorithm is the CAD-based Novel Object Segmentation (CNOS) which employs a three-stage pipeline to segment previously unseen objects in RGB images using only their CAD models and without retraining (Nguyen et al. 2023). In the Proposal Generation stage, it uses Fast Segment Anything to produce multiple segmentation proposals, each represented by a binary mask. During the Descriptor Creation stage, CNOS used the vision foundation model DINOv2 to generate feature vectors for both the segmentation proposals and the CAD models, enabling accurate matching. In the Matching and Segmentation stage, the model compared proposal descriptors with CAD model descriptors, assigning object identities to the most likely matches. It was claimed that this approach surpassed existing methods by almost 20% in Average Precision across the seven core BOP challenge datasets (Hodaň et al. 2018), establishing a new benchmark for CAD-based novel object segmentation. Mask2CAD has been used to reconstruct 3-D object structures from 2-D images using large-scale 3-D model libraries (Kuo et al. 2020). After detecting and segmenting objects in real RGB images Mask2CAD projects both 2-D segments and 3-D CAD models into a shared embedding space to identify the best-matching CAD template for each object. After retrieving the most suitable model, Mask2CAD refines the object pose, optimizing scale, rotation and translation, to achieve optimum alignment. This CAD-based representation offers a lightweight method for modelling object shapes, facilitating tasks such as content creation and interactive applications. Evaluations on Pix3-D confirmed Mask2CAD's enhanced 3-D shape prediction, and additional experiments on the more varied ScanNet dataset confirmed its robustness in complex, real-world scenarios (Sun et al. 2018).

Thirdly, using multiple-camera inputs to enhance 3-D feature predictions from 2-D images by estimating object bounding boxes. One such is the BEVSegFormer used with autonomous vehicles to produce BEV semantic segmentation using arbitrary camera setups. The process begins by encoding multi-camera image features through a shared backbone, then refining them with a deformable transformer-based encoder. A BEV transformer decoder module generates the semantic segmentation outputs, aided by a multi-camera deformable attention unit

Table 1 Computer-vision, vision-language and VR/AR methods reviewed for collaborative-robot perception, with representative references

Category	Methods	Representative References
Vision Methods	Mask RCNN	(He et al. 2017)
	Faster-Mask RCNN	(Ren et al. 2016)
Object detection	U-Net	(Salem et al. 2026)
	Single Shot Detector	(Li et al. 2017)
	You Only Look Once (YOLO)	(Khanam and Hussain 2024)
	Transformer	(Carion et al. 2020)
	Deformable DETR	(Zhu et al. 2020)
	Stereo-based 3-D	(Chen et al. 2022)
	Point and voxel	(Shi et al. 2023)
	Multi-modal fusion	(Guo and Hu 2024)
	Region-based and cascaded	(Rettenberger et al. 2023)
	Rendering-/point-based and dense sliding-window	(Chen et al. 2019, Kirillov et al. 2020)
Object Segmentation	Real-time or prototype-based and grid-based	(Chen et al. 2020)
		(Wang et al. 2020)
	3-D instance segmentation	(Heylen et al. 2021)
	CAD-based novel object segmentation	(Kuo et al. 2020)
	Multi-camera	(Peng, Chen et al. 2023a, b)
	RGB-D semantic segmentation	(Zhang et al. 2023)
	Point cloud	(He et al. 2021)
	Fusion	(Bai et al. 2022)
	Instance-level	(Labbé et al. 2020)
	Category-level	(Irshad et al. 2022)
Human Pose and Action Recognition	Unseen/novel objects	(Örnek et al. 2025)
	Pose estimation	(Sampieri 2022))
	Activity recognition	(Liu et al. 2024a, b)
	Gesture-driven interaction	(Subašić et al. 2025)
	Gaze estimation and tracking	(Belcamino et al. 2024)
	Motion intention	(Wong et al. 2024)
	LLM + vision	(Choi et al. 2024, Ju et al. 2024a, b, c, d, e)
	Task reasoning and high-level control	(Luo et al. 2024a, b, c)
	LLM access to robot logs	(Sobrin et al. 2024)
	Complex task planning via LLMs	(Wächter et al. 2018)
Richer vision-linguistic grounding	Language to robot actions	(Gkoumelos et al. 2024)
	Transformer/diffusion and multimodal fusion policy	(Ju, Wang et al. 2024a, b, c, d, e)
	Embodied QA and observation-to-action mapping with large VLMs	(Zitkovich et al. 2023, Luo et al. 2024a, b, c)
	VLA1 (long-term navigation with retrieved visuals and chat describer)	(Luo et al. 2024a, b, c)
Vision-Language Method	Large/domain-specific LLMs for robots	

Table 1 (continued)

Category	Methods	Representative References
Code synthesis	Automated code generation Offline programming Chat-to-action frameworks	(Luo et al. 2024a, b, c) (Luo et al. 2024a, b, c) (Shirai et al. 2024)
LLMs for multi-robot collaboration	Robot-to-robot instruction and collaboration Prompting for multi-robot task planning Application to AGVs and smart-manufacturing fleets	(Liu et al. 2025) (Chen et al. 2024) (Wang et al. 2024a, b, c, d, e)
Vision-VR/AR Method		
Safe teleoperation and mixed-reality control	MR-HMD with RL for state capture, communication and control	(Li et al. 2022a, b)
Brain-computer interface	BCI-mediated cobot control	(Hu 2024)
Programming by demonstration with AR overlays	PbD with AR overlays	(Chu and Weng 2024)
AR-based interface for multiple robots	Digital twin Physical synchronisation through AR Immersive DT layout, programming and safety tooling	(Langås et al. 2025) (Zhang et al. 2024a, b) (Dianafar et al. 2021)

that maps between BEV and image views. Queries were organized in a BEV grid and sampled for final predictions. Evaluated using the nuScenes data set (Caesar et al. 2020) and an author-collected dataset, BEVSegFormer showed what the authors described as promising results, with an ablation study confirming the effectiveness of each component (Peng, Chen et al. 2023a, b). The ODIN (Omni-Dimensional INstance segmentation) approach challenges the idea that 2-D and 3-D perception required distinct models by introducing a Transformer architecture that jointly segmented 2-D RGB images and 3-D point clouds (Jain et al. 2024). The model alternated between fusing within-view 2-D and cross-view 3-D features, encoding positional information through tokens for both pixel and 3-D coordinates. The authors claim ODIN achieved state of the art results using the ScanNet200, Matterport3-D and AI2THOR data sets, and performed at least as well as similar models using ScanNet (Rozenberszki et al. 2022), S3-DIS (Armeni et al. 2016), and COCO (Lin et al. 2014). ODIN excelled at handling sensed 3-D point clouds over mesh-sampling and, when deployed as a 3-D perception engine in an embodied agent, set new performance benchmarks on the TEACH dataset (Padmakumar et al. 2022) when performing an action-from-dialogue task. The Instant NeRF approach introduced one of the earliest learning-based pipelines for 3-D instance segmentation in Neural Radiance Fields (NeRF) (Liu et al. 2023). A NeRF model pre-trained on multi-view RGB images, learned an instance field that generated discrete 3-D instance masks through a proposal-based network. These masks were projected to align with 2-D segmentations presenting multiple views, which served as supervision. In addition to producing consistent 2-D segmentation from novel viewpoints, Instant NeRF could interrogate instance information at any 3-D point, thus significantly enhancing NeRF's object manipulation capabilities and demonstrating superior performance on synthetic and real-world indoor NeRF datasets, outperforming earlier NeRF segmentation methods and top 2-D segmentation models on novel-view segmentation tasks.

Fourthly, the depth camera-based approach, increased the efficiency of RGB-D semantic segmentation by means of an adaptive context aware network using spatial information, a lightweight encoder-decoder network utilized cross-modal correlations which minimized computational costs by reducing the computational parameters (Zhang et al. 2023). To mitigate noise from depth data, the network employed channel and spatial fusion attention modules to capture multi-level RGB-D features. High-level representations were enriched using a globally guided local affinity context module, and a lightweight residual unit in the decoder fused short- and long-distance information. Validated on NYUv2 (Silberman et al. 2012), SUN RGB-D (Song et al. 2015), and Cityscapes (Cordts et al. 2016), this method demonstrated superior accuracy, inference speed, and parameter efficiency compared to other approaches.

Recently the DiPFormer approach has been used to enhance RGB-D scene understanding by addressing attention misalignment problems in depth maps (Chen et al. 2024). DiPFormer improved road detection and semantic segmentation by focusing on more precise spatial attention. Depth Spatial-Aware Optimization (Depth SAO) was introduced to capture real-world spatial relationships, and Depth Linear Cross-Attention (Depth LCA) to distinguish spatial differences at the pixel level. A multi-scale MLP Decoder then fused features while maintaining real-time performance. This technique achieved state of the art scores on the benchmarks KITTI (Geiger et al. 2013), KITTI-360 (Liao et al. 2023), and Cityscapes (Cordts et al. 2016) datasets.

Fifthly, the use of 3-D semantic segmentations with deep neural networks (DNNs) have been remarkably successful, showing significant increases in performance. However, the small number of categories used in current 3-D segmentation is insufficient to represent the range found in real environments and to better reflect reality. The 3-D semantic segmentation on ScanNet was expanded to 200 categories (Rozenberszki et al. 2022), this substantially larger vocabulary introduced class imbalance, posing new challenges. To cope with sparse training data, the study aligned under-represented classes and used language-driven pre-training methods (Point cloud and voxel-based methods) to better cope with the complexities and broader image semantics of real images encouraging robust 3-D features. The DyCo3-D approach (He et al. 2021) has been used to investigate point cloud instance segmentation by replacing resource-heavy bottom-up strategies with a dynamic, proposal-free method. Grouping points with similar semantic categories and centroid votes simplified the segmentation pipeline assisting the learning of instance-specific convolution kernels. A lightweight Transformer module enlarged the receptive field, capturing long-range dependencies in sparse 3-D data. This streamlined approach boosted performance on such

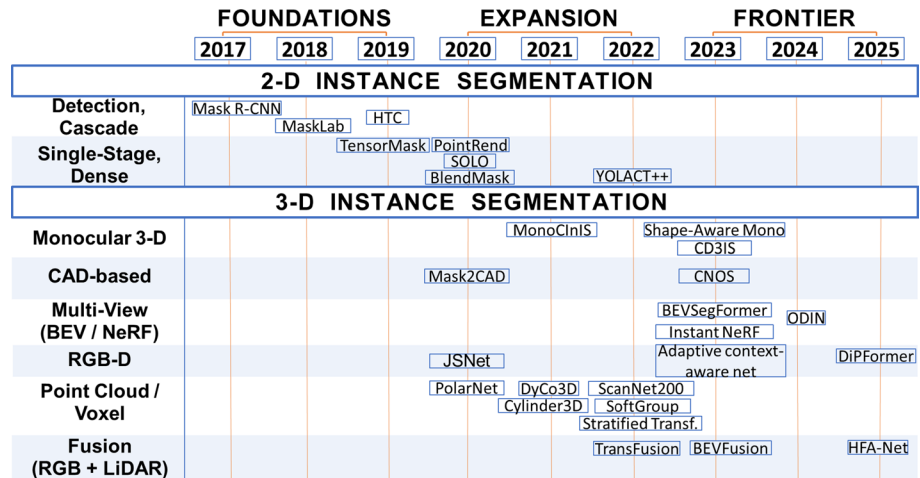
datasets as ScanNetV2 (Dai et al. 2017) and S3-DIS (Armeni et al. 2016), showing improved inference speed compared to previous methods. Further, the JSNet approach (Zhao and Tao 2020) jointly performed instance and semantic segmentation of raw 3-D point clouds within a unified network by fusing multi-level backbone features, mapping semantic information into an instance embedding space, and feeding instance context back into the semantic domain. Mean-shift clustering then grouped points belonging to the same object, eliminating the need for object proposals.

PolarNet approached the need for near-real-time, fine-grained semantic segmentation of single-scan LiDAR by introducing a polar BEV representation (Zhang, Zhou et al. 2020). PolarNet balanced point distributions along the radial axis, assisting the segmentation network to cope with the long-tailed distribution of LiDAR points more effectively. The SoftGroup approach attempted to resolve 3-D instance segmentation issues stemming from hard semantic predictions by applying a bottom-up soft grouping step and top-down refinement (Vu et al. 2022). Each point was allowed to hold multiple semantic class possibilities, mitigating error propagation in final instance assignments. The use of transfer opened a new direction in this area and the Stratified Transformer approach (Lai et al. 2022) improved 3-D point cloud segmentation with a key sampling strategy that gathered local information densely and distant sparsely, thus expanding the field of interest at reduced computational cost. First-layer point embedding was used to produce a more stable convergence, and contextual relative position encoding to cope with irregular point distributions. The Cylinder3-D approach avoided 2-D projection losses by processing LiDAR points in 3-D space, partitioning them into cylindrical grids (Zhou, Zhu et al. 2020). A dimension-decomposition context module was used to capture high-level geometric information progressively, better preserving 3-D topology and structure.

Finally, multi-data fusion of RGB and LiDAR. The TransFusion approach addressed the challenge of LiDAR-RGB camera fusion in adverse image conditions, such as poor lighting or sensor misalignment, by employing a transformer decoder with a soft association between LiDAR points and image features (Bai et al. 2022). The first decoder layer predicts bounding boxes from the LiDAR data, while the second layer adaptively fuses these boxes with relevant image features using spatial and contextual cues. This design ensures resilience to calibration errors, provides near optimal performance on large-scale datasets, and ranked first on the nuScenes (Caesar et al. 2020) 3-D tracking leader board. The BEVFusion approach (Liu et al. 2023) introduced a generic multi-task fusion framework that unifies LiDAR and camera data within a shared BEV representation, thus preserving both geometric and semantic information. By optimizing the BEV pooling process, BEVFusion dramatically reduces latency and achieved then state of the art results on nuScenes (Caesar et al. 2020), improving 3-D object detection and BEV segmentation while lowering computational costs. The method's task-resilient architecture supports a range of 3-D perception tasks without significant modifications. HFA-Net segments large-scale point clouds by effectively fusing geometric and colour inputs. It integrates a Hybrid Feature Extraction Component (HFEC) with a Hybrid Bilateral Enhancement Component (HBAC) to jointly capture and enhance geometric, colour, and semantic cues, together with Ternary-Distance Attention Pooling (TDAP), which uses trilateral distances to focus on both salient local details and broader neighbourhoods. Stacked within dense residual blocks, these components widen the receptive field and yield stronger feature aggregation (Wen, Zhang et al. 2025).

To summarize the progression of instance segmentation research, Fig. 7 arranges the reviewed methods on a 2017–2025 timeline. These methods are organized into swim-lanes that distinguish the two-dimensional approaches, such as detection-based, cascade-based, single-stage and dense, from the three-dimensional approaches such as monocular, CAD-based, multi-view including BEV and NeRF, RGB-D, point-/voxel-based and RGB–LiDAR fusion. The complete listing of methods and references appears in Table A.2 in Appendix. The temporal layout reveals the same broadening trend observed for detection: 2-D mask prediction dominates the earliest period, point-cloud and voxel methods proliferate around 2020–2022, and the most recent work is increasingly oriented towards foundation-model, multi-view and multi-modal segmentation. Representative methods are highlighted in each family to mark the key advances, and the era annotations indicate the prevailing emphasis of each phase.

Fig. 7 Timeline of the instance-segmentation methods reviewed in Sect. 2.2.3 (2017–2025) organized into swim-lanes separating the two-dimensional approaches from the six three-dimensional families. Horizontal position denotes the original publication year, solid markers indicate representative methods, and the period brackets summarise the prevailing emphasis of each phase. Full method details and references are listed in Table B.1



2.2.4 Pose estimation

A fundamental task in computer vision is estimation of an object's pose which requires prediction of the orientation and position of objects in 3-D space (Du et al. 2021) in order to perform inverse kinematic calculations for the robot to reach or grasp the object (Sun et al. 2024). Pose estimation at the instance level focuses on known, specific object instances, often utilizing detailed 3-D models of those objects to determine their poses with high precision. Category-level object pose estimation extends beyond specific instances, generalizing to broader object categories, such as chairs, cars, or bottles, where the system aims to infer an established pose for any item within a category, even if the precise 3-D model is not available. Finally, unseen object pose estimation extends the process to the handling of objects that the model never encountered during training, testing the system's ability to transfer knowledge about geometric and semantic features from known objects to new ones.

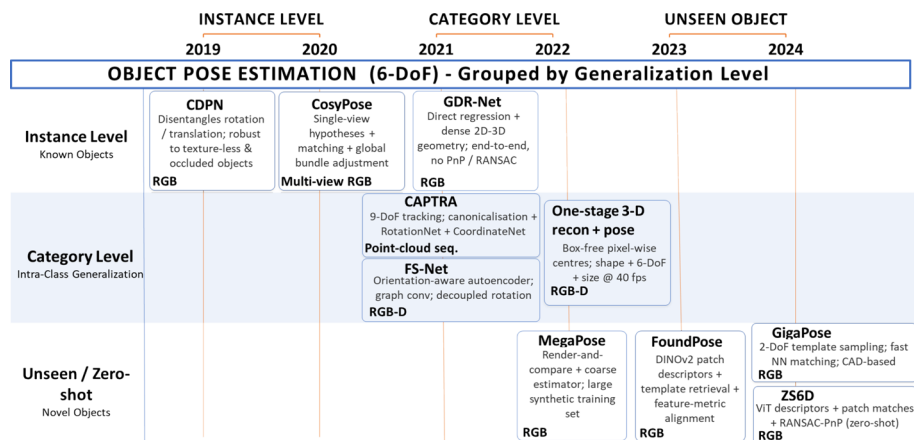
First, we consider instant level pose estimation which focuses on known object instances, often utilizing detailed 3-D models of those objects to determine their poses. There are many approach in this area such as GDR-Net which combines direct regression and geometry-based methods for 6D pose estimation from a single RGB image (Wang et al. 2021). It utilizes dense 2-D-3-D correspondence-style geometric representations in an end-to-end trainable network, overcoming the limitations of two-stage pipelines that rely on Perspective-n-Point (PnP) or RANSAC. Experiments using Linemod (LM) (Hinterstoisser Lepetit et al. 2013), Linemod-Occluded (LM-O) (Brachmann Krull et al. 2014), and YCB-Video (YCB-V) (Xiang Schmidt et al. 2017) demonstrated significant performance improvements over then-existing methods. Also, the Coordinates-Based Disentangled Pose Network (CDPN) approach which proposed a disentangled framework for 6-DoF object pose estimation, treating rotation and translation separately to enhance accuracy and robustness (Li et al. 2019). By regressing these components independently from a single RGB image, CDPN could effectively cope with texture-less and occluded objects, achieving then state of the art results on LM (Hinterstoisser et al. 2013) and occlusion (Brachmann et al. 2014) benchmarks. CosyPose extended pose estimation to multi-view settings where the viewpoints of the camera were not known (Labbé et al. 2020). CosyPose first generated single-view 6D pose hypotheses, then matched and refined these hypotheses across multiple input images. This process culminated in a global bundle adjustment that jointly optimised camera and object poses. CosyPose achieved superior performance on YCB-V (Xiang et al. 2017) and T-LESS (Hodan et al. 2017) datasets, significantly surpassing the then state of the art methods.

Second is Category-Level Object Pose Estimation which progresses from specific instances to category-level object pose estimation generalized to broader object categories such as chairs, cars or bottles where the system aims to deduce a standard pose for any item within a category, even if the precise 3-D model is not available. The authors developed an approach which attempted category-level pose tracking of rigid and articulated objects from point cloud sequences (Weng, Wang et al. 2021), presenting a pose as a 9DoF state (6D object pose plus 3-D size)

using a novel end-to-end pipeline. The pipeline comprised (1) a pose canonicalization module that normalized the input point cloud's pose, (2) RotationNet for regressing small interframe rotational changes, and (3) CoordinateNet for predicting normalized coordinates and segmentation. This design allowed analytical computation of size and translation, avoided non-differentiable RANSAC, and enabling small pose changes to combine dense coordinate prediction with direct rotation regression. The FS-Net approach emphasized efficient feature extraction using RGB-D data to address category-level 6D pose and size estimation (Chen et al. 2021). This approach used an orientation-aware auto-encoder with 3-D graph convolutions for learning shift- and scale-invariant latent features. A novel decoupled rotation mechanism used two decoders to extract complementary rotational information, with translation and size predicted via residuals relative to the mean object size and position. An online box-cage 3-D deformation strategy further enhanced generalization. Validations on two benchmarks showed FS-Net's performance was then state of the art in both category- and instance-level 6D object pose estimation. A subsequent study developed a one-stage 3-D reconstruction and pose estimation method to develop, from single-view RGB-D inputs and without available CAD models, targeted simultaneous multi-object 3-D reconstruction and 6D pose/size estimation (Irshad et al. 2022). Instead of multi-stage pipelines, it employed a bounding-box-free strategy that identified each object instance through pixel-wise "spatial centres," predicting complete shapes, 6D poses, and sizes in real time (40 fps).

Finally, unknown instant level object pose estimation aims to cope with objects that the model didn't encounter during training, thus testing the system's ability to transfer knowledge about geometric and semantic features from known objects to new ones according its CAD shape (Labbé et al. 2022). There have been a number of studies in this area such as MegaPose which addressed 6D pose estimation for unseen objects by combining a render-and-compare strategy for pose refinement with a coarse pose estimator trained to classify whether the pose error could be corrected by the refiner. Magapose introduced a large, wide ranging synthetic dataset of thousands of objects, enabling generalization to novel objects in real images without retraining. Evaluations on different datasets such as YCB-Video (Yan et al. 2020) showed MegaPose achieved then state of the art speed compared to methods requiring access to target objects during training. The GigaPose approach provided a fast and robust CAD-based solution for novel object pose estimation in RGB images. It sampled rendered templates in a two-degree-of-freedom space, significantly accelerating nearest-neighbor feature matching (Nguyen et al. 2024). Extensive benchmarking in the 2024 open BOP Challenge (Model-Based and Model-Free 6D Object Pose Estimation) demonstrated its state of the art accuracy and simple integration with existing pose refiners. GigaPose also showed potential for pose estimation using 3-D models generated by single-image reconstruction, easing the reliance on CAD models. Further, the ZS6D approach eliminated task-specific fine-tuning for zero-shot novel object pose estimation by employing visual descriptors from pre-trained Vision Transformers (ViT) (Ausserlechner et al. 2024). ZS6D using local patch matches established 2-D-3-D correspondences between rendered templates and query images, it then used RANSAC-based PnP to compute 6D poses. Tested on LMO (Brachmann et al. 2014), YCB-Video (YCB-V) (Xiang et al. 2017), and TLESS (Hodan et al. 2017) datasets, ZS6D improved average recall over two state of the art methods without expensive data rendering or specialized network training. The FoundPose approach estimated the 6D pose of novel objects from a single RGB image without object- or task-specific pre-training (Örnek et al. 2023). FoundPose matched patch descriptors from the self-supervised DINOv2 model between the image and pre-rendered object templates to establish 2-D-3-D correspondences, then retrieved the most similar template to achieve computational efficiency. The additional feature metric alignment accounted for coarse patch sampling discrepancies, and experiments on different BOP core datasets confirmed superior pose estimation performance (Nguyen et al. 2025). Combining FoundPose with render-and-compare refinement led to a new, current state of the art in purely RGB-based pose estimation. Figure 8 consolidates the object-pose-estimation methods discussed in this section along a 2019–2024 timeline, with each method shown as a card stating its key idea and input modality; the corresponding source table, including full references, is given in Table C.1 (Appendix C). In contrast to the family-based organization used elsewhere, the lanes here follow the level of generalization, such as instance-level (known objects), category-level (intra-class) and unseen/zero-shot (novel objects), ordered to expose the field's trajectory. Because instance-level methods cluster in the

Fig. 8 Object-pose-estimation methods reviewed in Sect. 2.2.4, shown on a 2019–2024 timeline with lanes ordered by level of generalisation (instance-level, category-level, and unseen/zero-shot). Each card states the method's key idea and input modality, solid cards mark representative methods, and the diagonal clustering of cards reflects the field's progression towards generalisation to novel objects. Full method details and references are listed in Table C.1



earlier years and unseen-object methods in the later years, the populated cards form a diagonal progression that visualizes the dominant trend: a steady move from pose estimation for known instances towards generalization to unseen objects, increasingly enabled by vision foundation models. A representative method is highlighted within each level, and the period annotations summarize the emergent emphasis of each interval. The figure offers a high-level synthesis only, and the reader is referred to Table 3 for the detailed characterization of each method.

2.2.5 Human activity recognition

Whereas the methods reviewed so far perceive the objects in a cobot's workspace, the methods in this subsection perceive the agents that share it. For a robot that must carry out a task jointly with a human, perceiving the partner involves more than locating a body in space: it extends from recognising the action currently being performed, to anticipating the action that will follow, and ultimately to inferring the partner's underlying intention, goal, and plan, so that the robot can act proactively and align its own plan with that of the human. The remainder of this subsection follows this progression — from skeleton, pose, and gesture recognition, through action recognition, to real-time motion-intention and plan-level prediction.

Detection and recognition of the human skeleton (Yu, Zhang et al. 2024a, b), pose (Liu and Wang 2021), and hand gestures (Liu and Wang 2018) are crucial for cobots as accurate determination of human motions and actions are essential for efficient and safe interactions between humans and the cobots (Hetto et al. 2023).

Activity recognition for workers could be achieved using different vision methods (Shafizadegan et al. 2024), but the using multicamera system enables more accuracy where it covers the working space from different view-points. Such approaches are not restricted to industrial workplaces but could be extended to, for example, construction sites. The multicamera system could utilize a particle filter algorithm to fuse 2-D joint locations from multiple camera viewpoints to estimate a 3-D human pose. A Long-Short Term Memory (LSTM) network is then used to recognize ten activities common to human–robot collaborative (HRC) tasks (Jang et al. 2023), which enhanced HRC in assembly tasks by interpreting implicit communication cues from the operator using computer vision and DL. This approach utilized a combination of a residual convolutional neural network (ResNet-34) and LSTM to perform action recognition based on the operator's tasks (Moutinho et al. 2023).

Human–Robot Interaction (HRI) has recently introduced a new term “human–robot teaming” (HrT) which explores how to improve team coordination and communication in HRI. This idea could be implemented by using ML-based visual action recognition to monitor human actions in industrial settings specifically designed for human action recognition in HRC assembly tasks (Mehak et al. 2024). It has been suggested that the time of training for human activity recognition could be reduced by providing an intuitive method to encode domain knowledge and automatically construct a probabilistic estimation model from this knowledge. The system used had two main components, a predicate estimator and an activity estimator, which utilizes the Region Connection

Calculus (RCC) and the Planning Domain Definition Language (PDDL) to facilitate intuitive encoding and automatic construction of the activity recognition model which enhances the HRI approach (Lee et al. 2019).

In addition, the integration of action recognition with multi-scale control increases the scalability and enables a high level of adaptability to widely different scenarios. By defining multiple basic gestures for robot motion control, constructing a control instruction library containing 70 different commands covering a wide range of HRC needs, utilizing static gesture, dynamic action process, and YOLOv5 object recognition and positioning technology (Hidayatullah et al. 2025), a foundation has been provided for significantly improving HRC systems in flexible production lines (Wang et al. 2025a, b, c, d, e). Also the understanding of hand gestures is not limited to a single input but can become more efficient with multimodal sensors such as RGB and infrared images (Wang et al. 2024a, b, c, d, e), which are particularly useful in high-illumination environments.

In terms of safety, DL and CNN enable real-time detection of workers' motions during collaborative tasks in assembly systems (Chiurco et al. 2022), which is essential for human safety in environments shared with robots, providing a distance-based algorithm to confirm workers are at a safe distance from the robot's motions. Advanced developments in this area provide efficient path planning for the robot to avoid collisions while reaching the target positions as promptly as possible. The method developed incorporates a transfer learning-based human pose recognition algorithm with low computational requirements to further optimise the assembly process (Liu and Wang 2021).

To maintain high level safety in HRI/HRC, the system should not be limited to understanding human motion taking place but should be extended to estimate, in real-time, the next action. This method could be termed motion prediction or human action/motion intention. Estimating what a partner will do next, and ultimately the intention behind it, is the agent-level perception on which fluent collaboration depends; the intent-recognition methods of Sect. 3 and the joint task-planning discussed in Sect. 4 build directly on this capability. This approach could be realised by utilizing of combination of the Kinectv2 sensor for capturing human joint positions and applying a Gaussian Mixture Model (Reynolds 2009) and Expectation–Maximization algorithm (Jin et al. 2022) to cluster estimated motion patterns. By labelling different clusters, the approach enables the estimation of human motion trajectories, which allows robots to adjust their movements accordingly and avoid collisions, even in emergency situations (Kang et al. 2018). The utilization of a Recurrent Neural Network (RNN) to analyse visual observations of human actions in real time, enables a cobot to plan and execute actions pro-actively, to improve efficiency and avoid collisions. The RNN captures the evolutionary motion patterns of the human body parts and their coordination, and incorporating probabilistic inference, using Monte-Carlo dropout to minimize uncertainty-related robot mis-triggers, further improves reliability (Zhang et al. 2020a, b, c, d).

Use of a simple NN structure could help in area of HRI/HRC, a multi-stage layer incremental perceptron (MSLP) network designed to improve human motion prediction for safe and efficient human–robot interaction. MSLP provides an efficient and smooth approach by incorporating two key components: a multi-channel feature extraction and enhancement block which can integrate multi-dimensional action features to enhance action node representation, and a temporal feature extraction module which can captured contextual relationships between actions over time (Zou 2024). Also, HRI/HRC can be approached via a memory graph neural network to capture both short- and long-term spatio-temporal characteristics, which enables a behavioural conditional probabilistic diffusion model to capture the variability of human motion distributions and predicts missing connecting motions the better to cope with the unpredictable nature of human motion, (Gui, Li et al. 2024a, b, c, d).

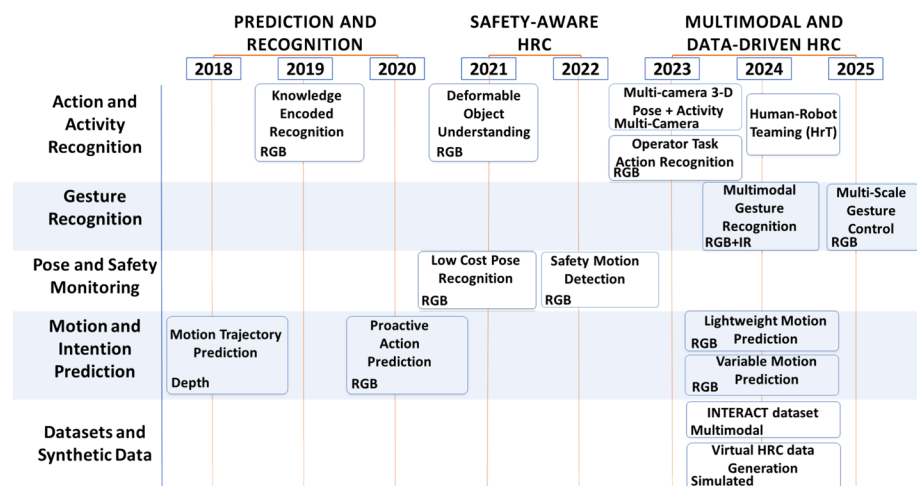
To overcome the current dataset limitations and lack of generalizable algorithms for predicting human motion, a comprehensive new dataset, INTERACT (Yasar et al. 2024), has been recently developed for human-cobot interaction. This multimodal dataset includes 3-D Skeleton, RGB-Depth, eye gaze tracking (MoGaze), and robot joint data. This approach addresses the low sampling efficiency and high cost associated with traditional real-world data collection by simulating the human musculoskeletal dynamics and calculating the metabolic energy consumption to assess the realism of the generated virtual data. A digital model, consisting of a digital human upper limb, a digital robotic model, and an admittance controller, was developed to produce virtual HRC data (Yao et al. 2024).

Complex scenarios in cobots, such as the in-air rope-knotting problem, require more advanced visual understanding, such as Convolutional Auto-Encoders (Pei, Chen et al. 2024a, b) and LSTM (Tang, Dong et al. 2024) networks can be utilized to better interpret image frames as a time series. Convolutional Auto-Encoders extract image features and perform accurate identification of, e.g., the rope and the human, and their states, enabling more precise and adaptive task execution (Suzuki et al. 2021). Finally, Fig. 9 summarises the human-activity-recognition approaches relevant to human–robot collaboration on a 2018–2025 timeline; the underlying methods, models, sensing modalities and references are listed in full in Table A.4 in Appendix. As this body of work comprises diverse, application-driven contributions rather than a single evolving architecture, the methods are organised by HRC perception function, action and activity recognition, gesture recognition, pose and safety monitoring, motion and intention prediction, and datasets and synthetic-data generation, with a chip on each card denoting the sensing modality. The figure supports two observations: first, the individual functions emerge at different points in time, with motion prediction and action recognition appearing earliest, safety-oriented monitoring in the intermediate period, and multimodal gesture recognition together with dedicated HRC datasets emerging most recently; and second, the sensing chips reveal a strong reliance on RGB input, with depth, infrared, multimodal and simulated data appearing only in isolated cases. A representative method is marked within each function, and the period annotations describe the prevailing focus of each interval.

3 Perception approaches

While Sect. 2 examined visual perception at the level of individual sensors and algorithms, this section shifts vision algorithms and camera setups from isolated components to integrated robotic systems. This section reviews the state of the art in how these vision algorithms and camera setups are combined and deployed within real collaborative robot applications such as human–robot interaction (HRI), robot-robot coordination and robot-machine collaboration. The review is organised around three increasingly capable approaches. The vision approach (Sub-Sect. 3.1) relies on perception alone to drive recognition, tracking, and learning from demonstration. The vision-language approach (Sub-Sect. 3.2) couples perception with natural language understanding, so that cobots can be instructed and can interact through reasoning directly. The vision-AR/VR approach (Sub-Sect. 3.3) adds an interface layer that supports intuitive supervision, such as digital-twin mapping and real-time feedback. These approaches trace the progression from perception as an isolated capability towards the tightly integrated, multimodal systems that characterise current cobot deployments.

Fig. 9 Human-activity-recognition approaches for human–robot collaboration reviewed in Sect. 2.2.5, arranged on a 2018–2025 timeline and grouped by HRC perception function. Each card names the approach with a chip denoting its sensing modality, and solid cards mark representative methods. The figure highlights both the differing emergence times of each function and the field's predominant reliance on RGB sensing. Full method details and references are listed in Table D.1



3.1 Vision approach

Cobot learning spans imitation learning (IL), shared autonomy/teleoperation, and vision-centric perception pipelines for action recognition and object/state tracking. It is complemented by data-efficient methods such as few-shot, one-shot, or synthetic data and by multimodal fusion under safety-aware intent and motion prediction. The following text instantiates these concepts in human–robot interaction (HRI), single- and dual-arm scenarios, robot-robot and robot-machine coordination, and online adaptation.

One of the best-case scenarios for HRC requires robots to learn directly from human actions, known as Learning from Demonstration (LfD) a subset of Imitation Learning (IL) (Zare et al. 2024). Such tasks, especially those that require high levels of safety and coordination, may be carried out using a single robot or multiple robots using motion planning (Srinivasan et al. 2021) or path planning (Chen et al. 2023a, b), or a combination of robots and machines, each process being integrated with human interactions (Langås et al. 2025). Here IL represents a learning algorithms that uses demonstrations, while LfD represents how the demonstrations are collected or encoded on a real robot (Gavenski et al. 2024). These latter include new tele-operation interfaces to provide better trajectory (Liu, Zheng et al. 2024) and uncertainty-awareness for learned and post-hoc scalar actions (Xu, Nguyen et al. 2025). LfD aims to enhance the cobot’s ability to define problems and perform tasks via human interaction. In this scenario the structure of the vision system could consist of two components: action recognition that captures subtle human hand movements during demonstrations and object trajectory tracking that monitors the motion of manipulated objects using a 2-D CNN based setting of two 2-D cameras (Hwang et al. 2020) or 3-D CNN based on a single 3D camera (Hwang et al. 2022) (see Fig. 10). Dual robot arms can function similarly to human hands for specific tasks, utilizing vision-based observation can assist in estimating hand motions and replicating similar movements, such as carrying boxes (Qu et al. 2019).

Dual-arm tasks can be learned by cobots through human interaction using Deep Imitation Learning (DIL). One example of a complex task is banana peeling, where two robots successfully collaborate despite the deformable nature of the object. Computer vision plays a crucial role in this process through the application of a gaze-based

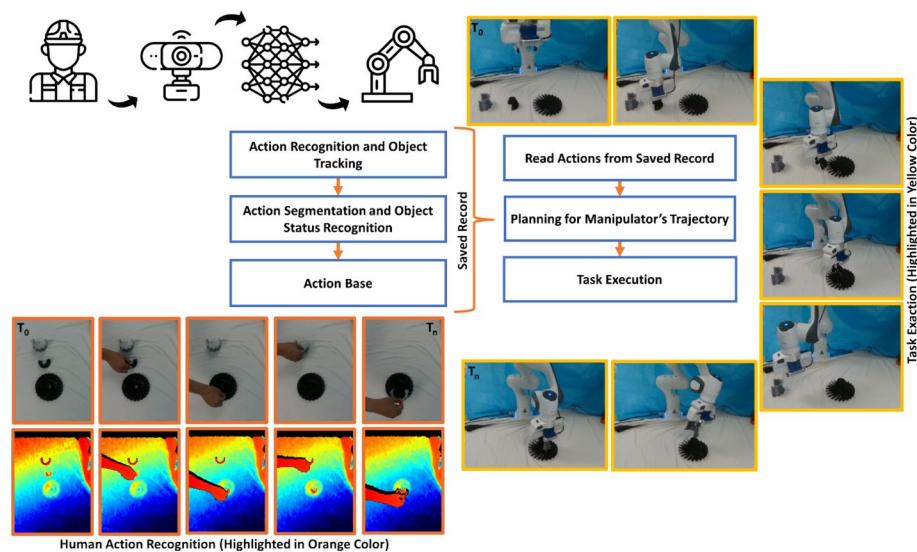


Fig. 10 Vision-guided Learning-from-Demonstration (LfD) system at IAI-KIT. Left (orange panels, “teacher”): an overhead camera records a human demonstration from T_0 to T_n ; the perception stack performs Action Recognition and Object Tracking, then Action Segmentation and Object Status Recognition, and stores results in the Action Base. Right (yellow panels, “robot”): the cobot executes Read Actions from Saved Record, then Planning for Manipulator’s Trajectory and Task Execution, with snapshots at T_0 and T_n . The two-stage pipeline converts recorded demonstrations into executable robot skills. Taken together, the key takeaway is that in Learning from Demonstration, pairing human-action recognition with object-trajectory tracking turns recorded demonstrations into skills the robot can execute

algorithm. Gaze-based algorithm enables the cobots to identify the most important/likely areas containing the required object by capturing the human gaze patterns which significantly reduces training time by narrowing the area of focus and eliminating unwanted objects or background, making the scene more comprehensible to the robot. While manual segmentation may still be required, it greatly minimizes the effort involved (Kim et al. 2024a, b).

A vision system can also facilitate dual-arm robot interaction without the need for human involvement. Collaboration in this case occurs between two robots performing specialized tasks, such as assembly, which requires subtasks such as part positioning, mating, and manipulation. The vision system enables object identification and localization, while also assisting in fault diagnosis and handling, thereby enhancing the efficiency and accuracy of the process (Fang et al. 2016). Additionally, depth vision systems can assist with complex tasks beyond assembly, such as Deformable Linear Objects (DLOs), in-air rope knotting (Suzuki et al. 2021), or cloth manipulating (Longhini et al. 2024) where the object is deformable, and sequences may vary. These scenarios require more advanced vision understanding, with Convolutional Auto-Encoders (Pei, Chen et al. 2024a, b) and LSTM networks (Tang, Dong et al. 2024) utilized to better interpret image frames as a time series. Convolutional Auto-Encoders extract image features and perform accurate identification of objects, such as the rope, and its state, i.e., whether they are slack, tangled, or knotted. This capability enables more precise and adaptive task execution. (Suzuki et al. 2021).

Dual-arm robots can utilize enhanced vision systems for better scene understanding and faster interaction. For instance, they can replace traditional 3-D scanners by using one arm to manipulate the part while the camera captures frames, enabling the reconstruction of its 3-D structure. This can also be via multi view image for the operation (Saikia et al. 2025), or cobots can move around an object to acquire key features to better recognize it (Kobayashi et al. 2023).

IL can also be achieved by capturing the human skeleton using a depth camera and mapping it to real-time where the dual-arm robot's joints move in correspondence with human motions (Zhang et al. 2024a, b). The Kalman filter facilitates these real-time updates, ensuring smooth and accurate motion replication (Liu et al. 2019). The Kalman filter enables real-time pose tracking for both object (Li, Xu et al. 2024a, b, c, d) and humans (Martini et al. 2024) where a plug is inserted into a paired socket (Chang et al. 2021). An alternative approach would be to guide the robot by hand using vision-in-the-loop IL, which combines a frozen YOLO detector with a Neural-ODE (NODE) motion module. The NODE compensates for the 6-DoF pose components that YOLO's axis-aligned bounding boxes cannot capture and is integrated as a secondary objective beneath a large-projection task-priority controller, thereby preserving stability. Live visual features are then fed into the pre-learned NODE, allowing the robot to adapt to novel object positions without further human intervention (Auddy et al. 2025).

In situations with limited data, synthetic data plays an important role in minimizing the time and cost associated with real-world training, which can use in robotics perception (Singh, Liu et al. 2024a, b, c, d, e) and cobot learning as well (Kim, Choi et al. 2024) so that a dual-arm robot can achieve precise, coordinated manipulation from a single kinesthetic demonstration. A three-stage visual-serving pipeline first aligns both end effectors with their respective targets; once aligned, replaying the recorded trajectory completes the task. This one-shot approach requires no stored object models, extra data, or post-demonstration training, yet still keeps the two arms perfectly synchronized (Wang and Johns 2025). The same approach can be used in HRI where the data is limited in cases such as assembly tasks (Wang, Liu et al. 2024a, b, c, d, e). Moreover, in HRI, understanding facial expression and analysing human speed during operations based on vision algorithms provide valuable feedback to the robots (Yahyaoui et al. 2025). This allows the robots to adapt their motions to the human's actions, rather than the other way around, thereby increasing productivity (Sadrifaridpour and Wang 2018). It is more efficient to develop a mode which can work with a facial emotion detection interface in robotics environment (Yahyaoui et al. 2025). Few-shot learning addresses this issue by introducing a temporal CrossTransformer to handle the fusion of data from human body skeletons, RGB images, and depth maps. Temporal CrossTransformer (TRX) is a few-shot video model that forms query-conditioned class prototypes by applying cross-attention between a query sequence and support sequences over ordered frame tuples; by matching relevant sub-sequences in time (rather

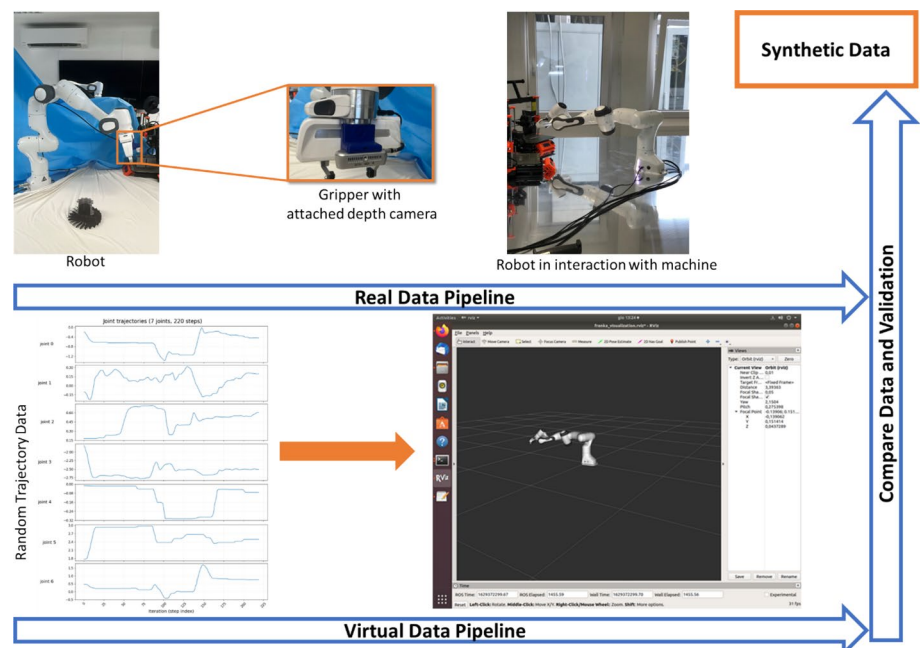
than single frames), it handles temporal misalignment and variable action speed, then classifies by comparing the query to these prototypes (Perrett et al. 2021). This idea originates from TRX for few-shot action recognition and can be used after fusing RGB, skeleton, and depth features in HRI. This approach effectively resolves problems related to similar industrial actions and body occlusions (Wang, Liu et al. 2024a, b, c, d, e). In the case of limited data, virtual data can be generated such as human motion and compared with actual motion. This approach helps optimise time by determining whether the predicted and observed motions align, thus improving cobot interaction with environments such as a 3-D printer (see Fig. 11) (Yao et al. 2024).

In cobots, enabling adaptive and synchronized human–robot collaboration is crucial. To enhance this, CNNs combined with inertial measurement units (Mohsen et al. 2021) and skeleton tracking data can be utilized for human action recognition (Gao et al. 2023). Additionally, a LSTM network can predict the remaining time for each action in a task, allowing the robot to perform real-time planning and synchronization of future actions (Male and Martinez-Hernandez 2023). During this interaction, real-time and accurate detection of various objects enables dual-arm mobile robots to grasp them. After localization, the Denavit-Hartenberg method (Zenhari et al. 2024) is applied to perform the required kinematic motions for gripping the relevant part (Chi et al. 2024).

Multi-camera or multi-view 2-D imaging can provide a more comprehensive understanding of the human body by estimating its full shape and motion using DL models, which also enable action recognition while incorporating noise filtering. The multi-camera approach enhances segmentation accuracy, reducing the risk of collisions. It can outperform 3-D cameras in scenarios involving occlusion, a common challenge in highly dynamic robotic environments (Terreran et al. 2021). To minimize processing time, DL enables the direct recognition of the human skeleton from 2-D camera inputs. This direct-from 2-D approach addresses challenges such as self-occlusion and partial occlusion of the human body. It can be achieved by defining a skeleton descriptor to capture both static and dynamic features of 3-D skeletons over short time intervals and by classifying multiple types of human motions or actions (Zhang et al. 2020).

It is important to optimise the operational speed of robotic systems while ensuring they respect human safety within the working environment. The robot must adapt its speed by detecting humans in its vicinity, enabling it to perform a safe stop when necessary. Accurate 3-D recognition of humans, their activities, and precise localization within the robot's workspace is critical in achieving this balance. This is combined with a penalty function

Fig. 11 Real-virtual data framework for cobots interaction at Robotics lab in IAI-KIT. Top: photos of the setup with a gripper close-up; the right block denotes Synthetic Data. Middle band: Real Data Pipeline. Bottom band: Virtual Data Pipeline, where recorded joint trajectories drive the simulated arm. Outputs from both pipelines are brought together for Compare Data and Validation. In particular, when real training data is scarce, predicted (virtual) human motion is generated and checked against observed motion, validating what the cobot has understood before it acts



to address the constrained nonlinear optimization problem of robot human interaction which adjusts the robot's handling speed (Chan and Tsai 2020).

Visual perception in object handling is crucial in cobotics. YOLOv5 and CNN-SVM (see Sect. 2.2) enable object detection, localization, and quality control of parts. These algorithms can be implemented in both simulated environments and real-world simulations, such as those using Gazebo a free robotic simulation environment (Yang, Zhou et al. 2023a, b).

Computer vision in collaborative robotics is not limited to human–robot or robot-to-robot interactions but also extends to robot-to-machine applications, such as additive manufacturing (AM). Spatio-temporal 3-D reconstruction (4D reconstruction) is used to monitor and enhance robotic AM processes. The system captures numerous images from various sensing sources during the collaborative process, enabling segmentation of ROIs. This allows for real-time tracking and modelling of dynamic, evolving objects, facilitating early detection and correction of deviations from the intended design (Chew et al. 2024).

In HRI, multimodal structures and data fusion are crucial to ensure the system operates effectively. This involves recognizing the human skeleton and pose to localize its position relative to the robot, and sensing the tactile interaction between the object and the human hand (Wong et al. 2024) or another robot hand (Shauri et al. 2012) for accurate handover or assembly tasks. Additionally, integrating gaze recognition helps detect human intention (Belcamino et al. 2024), thereby enhancing safety and smoothness of human–robot interaction (Bian et al. 2025). This combined approach offers a reliable and efficient solution which improves human–robot collaboration (Wong et al. 2024). Developing a model that can predict the presence of an obstacle and enable a robot to avoid it and complete a task without halting has been achieved using multi depth streams providing multiple views of the human skeleton (Forlini et al. 2024). In another case, deferens sensors such as RGB-D and binocular cameras were deployed at various positions around the workspace (Yang et al. 2018). Also, thermal imaging could help by using thermal multisensory fusion to enhance human target tracking, by combining omnidirectional infrared (IR) sensors with stereo IR vision using the thermal signature of the human body to achieve wide field-of-view tracking and precise distance measurement. This fusion allows improved accuracy in target localization by utilizing the maximum a posteriori estimation method to assess the reliability of each sensor rather than relying on visible-band-based tracking (Benli et al. 2019).

3.2 Vision-language approach

Vision can be applied as part of a language model in which the robot gets its instructions from a human (Zhao et al. 2025) or from another robot (Mandi, Jain et al. 2024) by text or voice and then proceeds accordingly. This approach is important in both human–robot interaction and robot/robot interaction using a guided vision and language system (Zitkovich et al. 2023). The system would also align command instructions issued via text or voice (Huang, Jiang et al. 2023), or as multimodal prompts combining language and visual inputs (Jiang et al. 2023), with a robot's perception which make learning better and faster (Wang et al. 2025). In this case, a learning framework translates natural language instructions into actionable robotic behaviour which enables intuitive cobot interactions (Du et al. 2023). It could be built based on a (i) Transformer model for executing complex/steps tasks, (ii) a Diffusion model for generating the trajectories of robots to enhance the handling of complex tasks with data efficiency, with (iii) a multimodal fusion network to align visual inputs with verbal instructions in order to ensure a comprehensive understanding of the environment, and (iv) a sophisticated policy model generating optimal actions for efficient and safe task execution (Ju, Wang et al. 2024a, b, c, d, e). LfD is a paradigm in which robots acquire task policies by observing and reproducing human demonstrations rather than through explicit programming. A LfD system can simplify the integration of cobots into manufacturing by addressing re-programming challenges which can occur due to the changes in product features. LfD can also utilize multimodal demonstration interfaces, including natural language instructions, visual observations, and hand-guiding, to enable robots to learn both the mechanics and semantic contexts of tasks to enable them to adapt to varying initial conditions and align task performance with specific customer requirements, streamlining customization and

operational efficiency (Lu et al. 2022). An improved vision-language model for dual robot systems or bimanual robotic manipulation is Vision-Language-Action (VLA), which integrates vision for scene understanding with language comprehension, translating human instructions into executable code which generates physical actions (Gbagbe, Cabrera et al. 2024).

In a vision-based cobot, the integration of YOLOv4 with point cloud data, combined with a Natural Language Processing (NLP) module for interpreting voice commands and converting them into robot instructions, enhances the robot's control and interaction capabilities, enabling more efficient and intuitive task execution (Hsu et al. 2023). Features such as MobileNet and ViT can be used to enhance robotic perception. MobileNet provides a lightweight and efficient framework for real-time perception, ensuring smooth operation on resource-constrained robotic systems (Cheng et al. 2024), while ViT excels at capturing global contextual information, making both ideal for tasks requiring a detailed understanding of complex visual scenes for multi-robots collaboration (Li et al. 2024) and/or human–robot collaboration (Subašić et al. 2025).

NLP enhances cobot interaction by enabling robots to better understand and execute tasks based on human commands, thus fostering a more seamless HRC (Rickert et al. 2007). This capability can be extended to support offline programming through large language models (LLMs) to enable automated code generation for various robotic tasks (Luo et al. 2024a, b, c). Implemented within a closed user-in-the-loop structure, LLMs allow for continuous operator input and feedback, ensuring adaptability and precision in collaborative task execution (Konstantinou et al. 2024). The use of Bidirectional LSTM-SRF and Bidirectional Encoder Representation (BERT) enables a high level of collaboration with robots through voice or text commands, addressing the limitations of programming skills among operators and facilitating more intuitive interactions (Park, Wang et al. 2024a, b, c, d, e).

The integration of a control system employing a Diffusion Policy with GPT-4 as the NLP model and OWL-ViT as the visual perception model enables precise robotic control when using the Robot Operating System (ROS) (Choi et al. 2024). This structure effectively converts verbal commands into actionable instructions, ensuring the robot accurately understands user intent and engages with its environment. GPT-4 processes and interprets natural language commands by utilizing the Chain of Thought technique to enhance intent recognition and contextual accuracy. These commands are then aligned with the vision system where the OWL-ViT model identifies and verifies the presence of necessary objects. Upon confirmation, the robot executes tasks guided by the Diffusion Policy (Ju, Wang et al. 2024a, b, c, d, e), delivering efficient task execution in cobots applications.

The utilization of OpenAI's GPT-3.5 Turbo better enables natural language communication between humans and robots in industrial settings. By addressing the limitations of pre-programmed commands, this integration facilitates dynamic, conversational exchanges, making robotic systems more intuitive and accessible (Dimitropoulos et al. 2024).

Autonomous robots that understand human instructions can significantly enhance the efficiency of human–robot operations, particularly where robot assistance is required to cope with unknown objects and/or provide on-demand assistance. 3-D object reconstruction and pose determination via neural object field modelling, facilitates object manipulation using a visual-based servo-driven mobile robotic system with navigation guidance. The LLM model provides text-based logic reasoning and high-level control command generation for natural HRI (Liu, Zhang et al. 2024a, b). It also, integrates language and vision into a unified framework that enables cobots to better interpret and act upon visual and textual inputs improving their ability to understand complex, multi-modal instructions and improving HRC in dynamic environments (Fan and Zheng 2024) and to better interact with environment.

To enhance interaction with the environment using vision-language guidance, a Reinforcement Learning (RL) motion control policy can be used for mobile manipulators. The RL is a machine learning method in which an agent learns a control policy by interacting with the environment and maximizing cumulative reward; Deep RL uses neural networks to approximate the policy and/or value functions (Yun et al. 2025). The RL motion control policy can ensure successful task completion in varied unstructured manufacturing scenarios. Such a RL framework significantly improves task execution and communication efficiency between human operators and

robots by enabling mixed reality head-mounted displays (MR-HMDs) for data collection, communication, and state representation (Niu et al. 2025). Such integration significantly improves collaboration, enabling robots to dynamically adapt to complex tasks while enhancing the user experience for operators (Zheng et al. 2024).

Such an advanced multi-modal foundation model, built around an LLM core and visual encoders, enhances embodied AI capabilities by enabling robots to plan and execute complex tasks such as the Franka kitchen (Gupta et al. 2020) in physical environments through integrated visual and language understanding. One of most important is the development of large data sets for this task such as the EgoCOT dataset that pairs videos from the Ego4D dataset with high-quality language instructions providing a richer training milieu. Such models, employing prefix tuning around the Mistral-7B LLM, could offer an efficient training approach that supports embodied planning. By integrating high-level planning with low-level control, the system forms a strong closed-loop framework for executing tasks with minimal error in dynamic environments. In this context EmbodiedGPT has demonstrated significant improvements in various embodied tasks, including planning, control, generating descriptions of an image, with the capacity to answer questions regarding the image (Mu et al. 2024).

An LLM can also access robot logs to gain a deeper understanding of the scene and context. The design of the Vision-Language Model (VLM) can allow the robot to answer questions with improved comprehension, combining both visual and contextual data. This capability enables more informed and context-aware responses, further improving the robot's interactive and operational efficiency (Sobrin-Hidalgo et al. 2024). Vision and language capabilities can provide the capacity of learning discrete concepts from sensory data and ground language (n-grams), and induce a grammar for the language describing the perceptual world. This is achieved incrementally and in a loosely-supervised manner, without storing all the previous data. Additionally, the robot can be guided through automatic sensing and interpretation of human actions, further facilitating collaboration. This capability allows the robot to anticipate and adapt to human intentions in real-time, enhancing efficiency and safety in shared working environments (Guo et al. 2023).

Embodied Question Answering, a transformer-based framework for robot navigation and question answering can align vision and language information via three key components: (1) an information fusion model using object tags as reference points to align vision and language, (2) a dedicated transformer block to capture extensive dependencies within visual scenes, and (3) integration of these components into a cohesive system (Luo et al. 2024a, b, c). Additionally, the implementation of large-scale VLMs can help robots map observations to actions while demonstrating emergent semantic reasoning capabilities. This approach expands task generalization, enabling robots to handle new objects, better comprehend complex commands, and perform basic reasoning tasks. By bridging visual perception and language understanding, VLMs significantly improve adaptability and efficiency in varied and dynamic environments (Brohan et al. 2023). A recent model in this area is the Visual Language Aligned Information which enables an agent to develop a more comprehensive global perception by using previously saved visual observations to enrich the semantic map around frontiers. For long-term goal prediction, the retrieved visual data is processed using the Chat Descriptor, a novel Visual-to-Language tool combining a LLM and a VLM with visual question-answering capabilities. This tool generates detailed frontier object descriptions and can calculate semantic similarity between target and frontier object categories. By integrating these semantic similarities with the frontier descriptions, the agent achieves more accurate long-term goal predictions, improving navigation and task efficiency (Luo et al. 2024a, b, c).

Multimodal planning and problem specification in robotics can be improved by combining language instructions with scene observations. This enables symbolic planners to operate within a language-guided framework. For example, the Vision-Language Interpreter framework combines LLMs and vision-language models to generate problem descriptions that incorporate a refinement mechanism. This mechanism can process anomalies in input messages (Cao et al. 2024), scenes (Jiao et al. 2023), or inspections (Li et al. 2025) providing feedback from the symbolic planner to iteratively refine the generated problem descriptions. As a result, both the accuracy and robustness of planning are improved (Shirai et al. 2024).

LLMs are not restricted to human–robot interaction, they can be extended to communications between multiple robots (Chen et al. 2024) which can be further extended to independent task assignment in distributed robots

using fuzzy logic control (Khelifa et al. 2024) or multi-agent reinforcement learning (Dong et al. 2024). The performance of LLMs can be substantially improved using prompting techniques such as in-context learning, which places greater emphasis on the token budget for the context window (the maximum number of tokens, i.e., sub-word units, that the model can consider at once) of multi-robot task planners (Chen et al. 2024). This approach has been shown to be applicable to augmented guide vehicles (Wang et al. 2024a, b, c, d, e). Additionally, robots equipped with LLMs have been shown to be capable of collaborative reasoning to develop task strategies (Liu et al. 2025). They can generate sub-task plans and task space waypoint paths, which are then used to optimise trajectory planning by, say, a multi-arm motion planner. The process is refined by incorporating environmental feedback, such as collision detection. This feedback prompts the LLM agents to adjust their plans and waypoints, resulting in improved task execution and adaptability to dynamic environments (Mandi, Jain et al. 2024).

Figure 12 presents the key concepts underpinning the current state of the art in natural-language-to-robot development in unified closed-loop architecture. A single pipeline links (i) natural-language intent capture from a human supervisor or peer robot, (ii) an LLM-powered code-synthesis module that translates those intentions into robot-executable C++ scripts, (iii) a vision backbone that continuously reconstructs scene state, and (iv) a

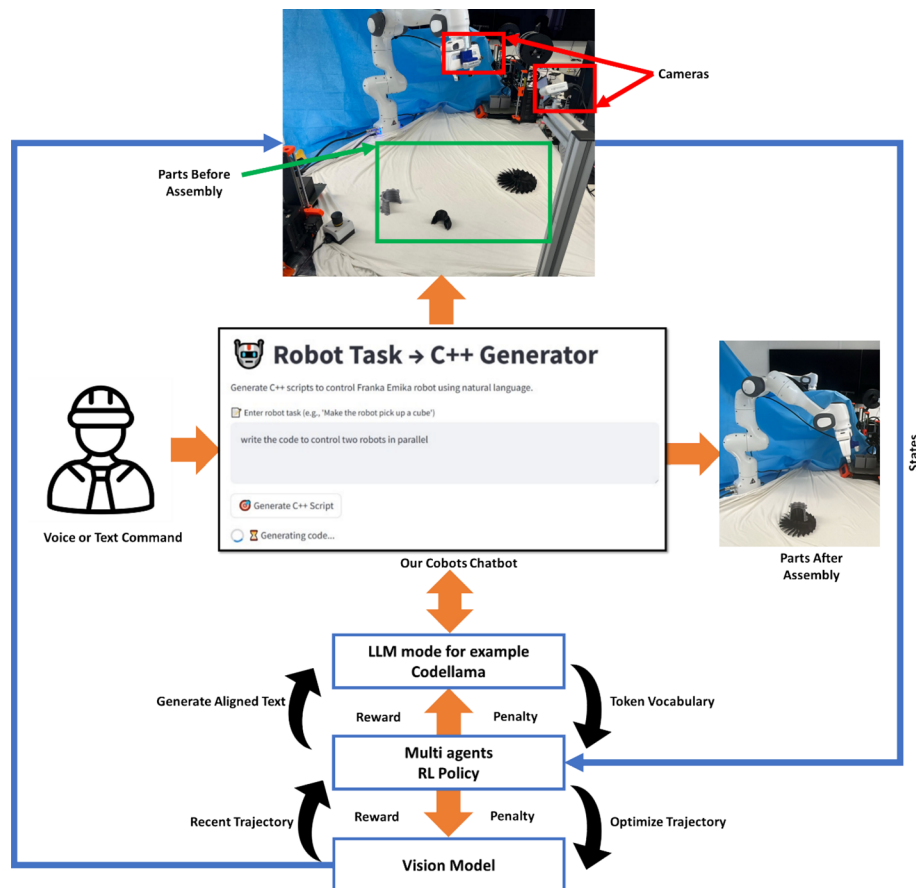


Fig. 12 Natural-language-to-robot pipeline combining vision, reinforcement learning, and a code-generating large language model in robotics lab at IAI-KIT. A human operator (left) specifies a high-level task in plain language through a “Robot Task to C++ Generator” interface (e.g., “write the code to control two robots in parallel”). The generator uses “LLM mode for example Codellama”, consults “Multi agents RL Policy” for feasible motion plans, and takes scene state from the “Vision Model.” In the workspace image, red boxes indicate the cameras and the green box marks the task workspace; in the execution image, the yellow box highlights the target. The generated C++ code is deployed to the robot, linking scene understanding, policy guidance, code synthesis, and physical execution in a closed loop. Overall, the framework integrates natural-language intent, LLM-based code synthesis, visual perception, and online reinforcement learning within a single closed loop, allowing commands to flow to the robots while perception and reward signals feed back into planning

multi-agent reinforcement-learning layer that refines trajectories online. High-level commands flow downward, while dense perceptual feedback and reward signals flow upward, allowing the LLM and the policy to co-adapt. Vision corrects language ambiguities, and policy steers the LLM toward feasible action sequences, and both components close the loop with the physical robot. In this way the diagram captures the current emphasis, from isolated perception or language modules toward tightly integrated, multimodal architectures that can be re-tasked in minutes rather than hours, while highlighting open research frontiers such as safe LLM grounding, vision-language reward shaping, and scalable multi-robot coordination. This architecture is modular and upgradeable, so newer LLMs, vision models, or policies can be dropped in as they emerge.

3.3 Vision-VR/AR approach

In the area of cognition-augmented human-centric robotics and collaborative robots, where human flexibility is paired with robotic assistance for mass personalization, computer vision can be used to address cognitive challenges (Li, Xie et al. 2024a, b, c, d). This subsection emphasizes that frequent changes in this procedure present cognitive hurdles, particularly in small and highly customized tasks (Pang et al. 2025).

Human–robot collaboration utilizing digital twin technology can enhance safe, flexible, and efficient collaboration in shared workspaces via the construction of a digital twin scene, mapping between real and virtual spaces, and planning and executing collaborative strategies. This could be built using a robust human mesh recovery algorithm to improve the reconstruction of occluded human bodies, and employ uncertainty estimation to enhance the action recognition algorithm while preserving risk controllability (Zhang et al. 2024a, b). The digital twin system could be extended to the remote operation process to guide cobots systems via augmented reality (AR) (Loy et al. 2025) and metaverse (Soliman et al. 2024). It maps the digital twins of industrial robots to their physical counterparts and visualized them through AR glasses, enabling intuitive control and interaction by human operators to improve multi-robot coordination. A key component of the system is the design and implementation of a multi-robot, multi-client communication mechanism that ensures synchronization across different operational states and clients (Li et al. 2022a, b). Additionally, it should not be restricted to HRI but should also enable better visualization and analysis of layout designs, task scenarios, robot programming, cycle time estimations, and safety analyses. Such an immersive interface is crucial for improving the setup and operational phases of collaborative robots generally used in applications with low payload limits (Dianatfar et al. 2021).

In the area of robot programming by demonstration, integrating AR interfaces into the manufacturing process can streamline and enhance the efficiency of programming industrial robots. Within the context of manufacturing tasks such as pointing and line tracing this could be done via three AR-based input methods: hand ray, head gaze, and eye gaze (Chu and Weng 2024). AR can also be used to assess which type of sensory cue (visual or tactile) is more effective in conveying the robot's intentions to a human operator, thereby improving safety, trust, and productivity in collaborative tasks (Chu and Liu 2023).

In the area of implementation head-mounted AR (HMD) displays, such as Microsoft's HoloLens Mixed Reality, integrated with the open source Robot Operating System (ROS) (Chovet et al. 2025) provided a good basis for interaction between operators and robotic arms to extend the number of workers able to use robotics in manufacturing by allowing operators to control robots through intuitive gestures without the need for extensive training or deep knowledge of robots, reducing the complexity traditionally associated with robotic programming (Lotsaris et al. 2021a, b). The application of AR enables better communication and interaction between human operators and robots by providing intuitive tools for direct instruction and real-time information exchange about the robot's status and the broader system's state. This is achieved through a ROS-based system that ensures good integration and communication between the AR headset and the mobile robotic platform (Lotsaris et al. 2021a, b). The integration of Brain-Computer Interface (BCI) technology with robotic systems through AR could improve control of brain-controlled robots. BCI technology allows direct control of devices using brain signals, notably electroencephalograms (EEG), and has two primary control methods: direct and shared. Direct control without HMD suffers from low efficiency and high user fatigue while shared control uses AR to enhance system

interaction and reduce cognitive load on the user (Hu 2024). AR can also be integrated with deep reinforcement learning to enhance the flexibility and intelligence of industrial robots in manufacturing environments (Liu, Tang et al. 2024a, b) and installed in multi-agent robotic systems (Timofeev 2001). This could be enhanced by integration of multi sensor structures which could introduce dynamic environments to cobots (Khatib et al. 2021).

Digital twinning of the real and virtual spaces has been proposed to facilitate planning, mapping and executing collaborative strategies (Wang et al. 2025a, b, c, d, e). A bespoke human mesh recovery algorithm employing uncertainty estimation was used to enhance reconstruction of occluded human bodies, while maintaining risk controllability (Zhang et al. 2024a, b). Integration with LLM (see Sect. 3.2) facilitated clear and intuitive communication between human operators and robots, integrated seamlessly with Digital Twins technology for real-time data processing and employment of behaviour-based control mechanisms (Gkournelos et al. 2024). A multimodal DL model to support human–robot collaboration provides user-centric manufacturing information in a mixed reality environment, correlating visual data with workers’ queries. The system offers real-time, relevant information to enhance productivity and adaptability in changing work environments. The implementation of digital twinning can be done by integrating vision and text dual encoders to analyse the visual input from smart glasses and text queries from workers. The information is displayed on the smart glasses, enabling a smooth interaction between workers, robots, and the manufacturing process (Choi et al. 2025).

4 Recent challenges, potential solutions, and future approaches

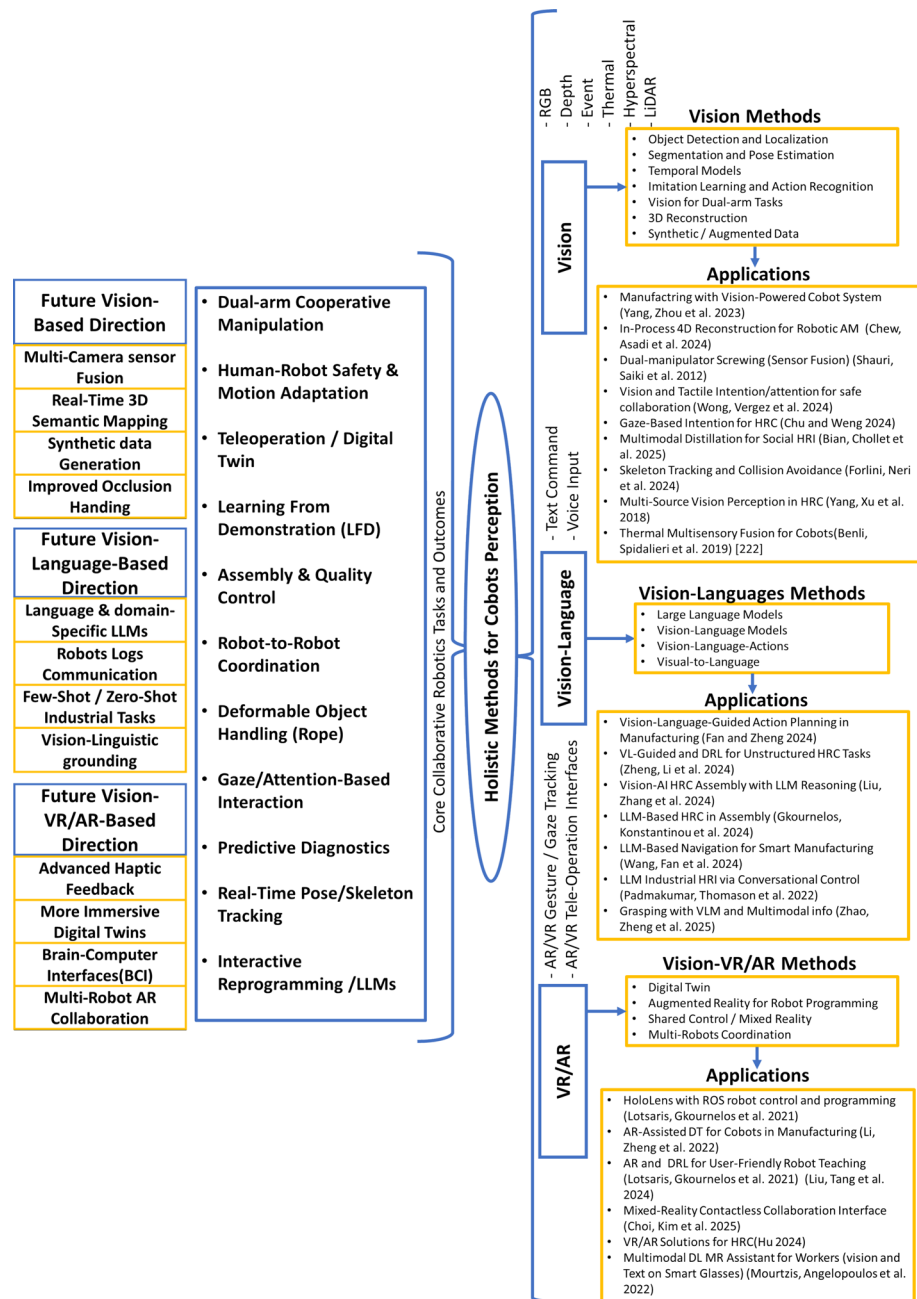
Figure 13 illustrates a state of the art holistic perception framework for cobots and points to emerging future directions in this field. The figure shows the integration of the three main components: Vision, Vision-Language, and VR/AR. Starting with Vision, we see object localization, human pose estimation, object pose estimation, and various cooperative operations (for instance, rope manipulation) that rely on accurate perception and tracking of scene elements. The Vision-Language pathway emphasizes multimodal fusion, combining visual input with linguistic descriptions to enhance contextual understanding for tasks such as command interpretation or complex scene analysis. Finally, VR/AR systems provide an immersive layer, enabling interactive visualization and real-time feedback loops that aid cobot operation and training. Collectively, these interconnected modules enable a more robust, flexible, and efficient cobot perception pipeline. For legibility, the individual vision, vision-language and VR/AR methods populating each pathway are not reproduced in the figure; they are listed with their supporting references in Table A.1. Collectively, these interconnected modules constitute a more robust, flexible and efficient cobot perception pipeline.

4.1 Safe collaboration in cobots system

Safe collaboration between humans and robots in shared workspaces progressed substantially with the integration of advanced vision, sensor fusion, and real-time motion control. Currently, many cobot systems employ depth or multi-camera setups to detect and track human presence and behaviour (Terreran Barcellona et al. 2021) using, for example, skeleton-based tracking or gaze detection in real time to avoid danger zones (Terreran Barcellona et al. 2021). By accurately determining a person’s position and motion, the robot can adjust its speed or perform a safe stop, ensuring worker safety without sacrificing efficiency (Lacevic et al. 2023). Additional methods, such as facial recognition or motion-speed analysis help predict potential collision scenarios and prevent hazardous movements (Sadrfaridpour and Wang 2018).

Despite these advances, several obstacles remain. High-volume, high-resolution vision data can be demanding computationally, especially in dynamic or cluttered backgrounds, where occlusions and changes in lighting or environment layout are frequent (Terreran Barcellona et al. 2021). Moreover, few-shot or zero-shot object detection techniques, key to rapidly handling a novel object or person, remain challenging in industrial settings. The defined algorithms for sensor fusion such as combining data from RGB cameras, depth sensors, IR sensors, etc.

Fig. 13 Holistic perception framework for collaborative robots. As shown, Vision, Vision-Language, and VR/AR are not separate topics but complementary modules of a single, more robust cobot perception pipeline



can also be slow or computationally heavy if not optimised, affecting real-time responsiveness. In addition, effectively combining a multi-camera system, where different robots or workstations each have their own cameras, introduces synchronization, calibration, and data overload issues.

Beyond these technical obstacles, it is important to recognise that safe collaboration is not only a perception and control problem but also a human machine coordination problem. A core requirement for safe coordination is that the human can anticipate what the robot will do next. A recent human automation research formalises this as the ability to answer the question "what will the automation do next?" and shows that it depends on how the joint system is designed and made observable to the operator, not on perception accuracy alone (Post et al. 2025). This requirement becomes critical because highly autonomous, high authority systems can act confidently on an internal model of the manufactures that no longer matches reality, behaving in a "wrong, strong, and silent" manner

when their model and the actual situation diverge (Dekker and Woods 2024). Such failures are manifestations of brittleness which is the sudden collapse in performance when events push a system beyond the boundaries of its design envelope. Resilience engineering argues that safety in open, changing environments depends not only on ever-more-accurate sensing, but also on graceful extensibility. This refers to the capacity to anticipate developing problems and adapt at the boundaries of competence (Woods 2018; Woods and Alderson 2021). Although these insights were developed largely in aviation and critical-infrastructure settings, they transfer directly to cobots. Therefore, future safety benchmarks should assess not only detection accuracy, but also the legibility and predictability of robot behaviour from the human's perspective. They should also evaluate the system's ability to degrade gracefully when its perceptual model is challenged.

Looking ahead, more powerful DNNs will cope better with occlusions, partial visibility, and changes in lighting (Zhang et al. 2020). Few-shot and zero-shot detection methods are expected to advance, making cobots more suited to unstructured environments. Real-time modelling of human motion using advanced filters, such as Kalman, and penalty-function-based speed optimization will enable robots to continually adjust their movement based on proximity or predicted human trajectories (Chan and Tsai 2020). Regarding data-processing, LLMs could integrate recordings from robots and machines which would assist the real time co-ordination of vision systems with multi-camera data streams and better interpret such contextual cues as warning signals. By reading or summarizing these recordings, each robot could, for example, be fed up-to-date information about potential hazards, thus increasing safety. Lastly, advanced multimodal frameworks that combine visual and speech cues with operator motion or attention analysis, could improve human–robot trust and reduce accidents, as the robot could modulate speed or behaviour when it detects stress or fatigue in the human operator.

4.2 Multi-robot and robot-machine interaction

Currently, multi-robot systems and robot-machine interactions are gaining traction for tasks (Fang et al. 2016) such as assembly (Chang et al. 2021) and manipulation of deformable objects (Suzuki et al. 2021). Using vision-driven coordination, often powered by CNNs (Hong, Dong et al. 2024), auto-encoders (Pei, Chen et al. 2024a, b) or LSTM (Tang, Dong et al. 2024) networks, each robot or machine obtains real-time information about object location, part geometries, or fault conditions (Fang et al. 2016, Chew et al. 2024). This enables specialized task division, reducing redundant movements and increasing throughput. In additive manufacturing processes, vision-based 4D reconstruction can even track an evolving 3-D-printed structure and detect deviations early for corrective action (Chew et al. 2024).

An important bottleneck is the shortage of high-quality, annotated data for less common tasks or unusual object interactions, though synthetic data generation may mitigate this data scarcity (Kim, Choi et al. 2024). Sensor fusion has to be implemented with care or the merging of different streams, such as depth, RGB, etc., can become burdensome with slower processing which inhibits real-time performance (Fan et al. 2025). Additionally, a multi-camera system across different robots or integrated with machines can generate synchronization and calibration problems for high precision assembly. It is noted that even small calibration errors can grow into large misalignments or even collisions (Suzuki et al. 2021).

Solutions to such problems usually include the adoption of advanced AI-driven techniques for real-time communication and dynamic synchronization of robots. To cope with large influxes of data without bottlenecks, sensor fusion is expected to develop algorithms which are more computationally efficient including, for example, lightweight neural networks and/or optimised inference pipelines. LLMs that integrate robot and machine log sheets would help orchestrate a network of robotic cameras, one for each mobile or stationary unit would unify the system's perception and enable more robust fault diagnosis and location tracking. Over time, these LLM-based frameworks might facilitate natural language queries such as “Identify the part that failed in Machine A”, aligning multiple camera feeds for rapid diagnosis. This approach unlocks higher-level planning, more efficient handovers between robots, and better fault tolerance, ultimately streamlining robot-machine integration.

4.3 Multi-robot and human interaction

When multiple robots share tasks with one or more human operators, current best practices frequently centre on LfD or DfL, where a human demonstrates how to perform a complex task such as knotting a rope, and dual-arm or multi-robot setups learn by observing hand motions and object trajectories (Kim et al. 2024a, b). Gaze-based attention can filter out irrelevant background details and narrow the learning scope (Kim et al. 2024a, b). Such methods significantly reduce training overheads and improve success rates.

Generalizing each user's demonstration to different robot configurations or tasks remains difficult, particularly when the environment is dynamic or contains complex backgrounds. Combining sensor data from multiple robots, each with its own camera and different vantage point, can also pose synchronization issues. Real-time analysis becomes more demanding if each robot and operator rely on different data fusion pipelines. However, LfD is frequently constrained by both the quantity and quality of available demonstrations, especially in high-mix, low-volume production environments (Kim, Choi et al. 2024).

Progress in advanced neural architectures offers the possibility of improved data fusion from cameras, depth sensors, event camera (Guan et al. 2024) and inertial measurements (Kang 2025). For real-time pose estimation and action prediction, Kalman filters (Xiao et al. 2025) or similar methods help minimize latency and reduce the noise in noisy data streams. LLMs integrated into imitation-learning pipelines could combine not only the operator's motions but also textual or verbal instructions, machine logs, and other robots' operational states. This multi-modal approach allows for a more complete understanding of tasks and workspace conditions, thereby enabling robots to better adapt to sudden changes such as new product variants or modifications in operator intent. Real-time analysis of operator emotion and attention signals could further boost safety and efficiency; if a human appears stressed, the system might slow robot speeds, adjust force thresholds, or alert other robots to proceed with caution. Having each robot use an LLM to access and interpret shared logs increases safety margins because all agents become aware of potential hazards or changes in real time.

4.4 Cobots in highly dynamic and changeable environments

In industries characterized by frequent design changes or custom orders, currently available cobots struggle with repeated reprogramming and labour-intensive data gathering. Vision systems must cope with dynamic scenes with unstructured layouts, moving objects, cluttered scene (Back et al. 2025) and unpredictable workflows while facing background complexity and partial occlusions (Guo, Zhang et al. 2024a, b) that necessitate sophisticated object detection (Terreran et al. 2021). There are few-shot (Kasaei and Xiong 2024) or zero-shot (Debnath et al. 2025) models, capable of recognizing novel items or tasks on-the-fly.

A significant limitation inherent in reliance on large, high-quality datasets is that they often quickly become obsolete (Kim, Choi et al. 2024). Additionally, real-time data processing can be hindered by complex sensor-fusion pipelines, merging RGB, thermal, and depth streams that can slow inference unless algorithms are continuously optimised. Multi-robot or multi-camera setups aggravate these performance concerns, as each new data source must be calibrated and synchronized. Ensuring safety while maintaining agility under these dynamic conditions remains an ongoing challenge.

Approaches that blend synthetic data generation for object pose estimation (Konstantinou et al. 2025) and human actions recognition (Knapp and Bohacek 2025), few-shot learning, and continual learning, provide the opportunity for lowering the overheads when retraining models for new tasks such as object pose estimation (Felice et al. 2025), human pose estimation (Ying et al. 2025), human action recognition (Wang et al. 2024a, b, c, d, e). While multi-modal perception that integrates thermal imaging, depth sensing, and standard RGB inputs can enhance scene understanding even in cluttered or poorly lit environments (Castilla-Arquillo et al. 2025), it also requires precise calibration, time-synchronization, and real-time fusion to be reliable on the shop floor. Notably, LLMs would be pivotal in fusing inputs from robots, machines, and sensor streams to coordinate large fleets of cobots under rapidly shifting conditions. By reading the logs of other robots, such as completed tasks,

encountered errors, and environment changes, each cobot could update its internal model and anticipate new product variants. It could also adopt different motion strategies to improve overall system adaptability and safety. Real-time operator feedback, including emotion and attention recognition, could trigger dynamic speed adjustments or force limits, ensuring safer and more efficient workflows. In the future, as LLMs become more deeply integrated with IL frameworks, cobots will be able to interpret not only visual and sensor cues but also high-level language instructions or constraints, drastically speeding up the reconfiguration process for new tasks or product lines.

4.5 Potential enhancements of cobots systems

4.5.1 Cobot perception fusion

Future cobots will coexist with humans, other robots, and constantly shifting clutter. Their perception core must therefore fuse heterogeneous depth cameras, LiDAR and neuromorphic event streams without flooding the shop-floor network or overloading edge GPUs. An information-aware fusion pipeline, where every sensor first broadcasts a self-estimated credibility map and a cross-view attention graph selects only the most informative voxels within a strict bandwidth-and-latency budget, offers a compelling way forward. By feeding the pruned data into a square-root information filter, sub-millimetre object-pose estimates become possible even as arms occlude one another. The open challenge, and opportunity, is to prove that such on-the-fly weighting outperforms today's brute-force "fuse everything" approach by an order of magnitude in compute and traffic, turning high-precision collaboration into a commodity feature rather than a premium add-on.

This proposal sits within an active line of uncertainty- and resource-aware fusion research, and it is worth contrasting it with the established alternatives. Classical perception stacks usually fuse sensors with deterministic estimator methods such as the Kalman filter, or by simply concatenating all modalities in a single network. Both of these methods perform well in structured settings but degrade under real industrial conditions such as sensor noise, occlusion, and modality dropout, and the brute-force variant scales poorly in compute and bandwidth. Recent work has therefore moved toward learned, reliability-weighted fusion that mirrors the credibility-map idea advocated here, such as the Cocoon study (Cho et al. 2025), which performs object-level and feature-level uncertainty quantification so that each modality is weighted by its estimated reliability, outperforming both static and adaptive fusion under natural and artificial corruption. A parallel line targets precisely the bandwidth and latency budget emphasized here by fusing selectively rather than exhaustively, such as the study of HydraFusion (Malawade et al. 2022), which performs context-aware selective sensor fusion, the study of EcoFusion (Malawade et al. 2022), which makes that selection energy-aware, and SAMFusion (Palladin et al. 2025), which adapts the fused sensor set to operating conditions. The pipeline proposed here can be read as a synthesis of these two threads, combining per-sensor reliability weighting with resource-aware voxel selection and a square-root information filter for numerical stability, with its distinctive element being the explicit bandwidth-and-latency budget. The open question, as the surrounding text notes, is empirical. It is whether such credibility-gated, selective fusion can demonstrably outperform exhaustive fusion by a large margin in compute and traffic without sacrificing pose accuracy under heavy mutual occlusion.

4.5.2 Cobot task planning and smooth motion

Most industrial stacks still bolt a symbolic planner onto a separate motion controller, a division that can add unwanted delays (latency) and brittle hand-tuning. A single neural architecture that encodes plain-language commands, reasons about an object-centric scene graph, and simultaneously denoises a diffusion trajectory could erase that seam. In such a model, cross-modal reasoning layers emit discrete sub-goal tokens, e.g., "grasp," "insert," "tighten", which condition a diffusion decoder to generate jerk-bounded, continuous joint paths. Early evidence suggests that co-training planning tokens with trajectory smoothness lets the network "think and glide"

in one pass, meeting the decision quality of top symbolic planners while rivalling the polish of handcrafted diffusion controllers. For cobots, that means spoken or text instructions could translate directly into collision-free, human-friendly motion with latencies low enough for shared workspaces.

This unified planner controller proposal should likewise be positioned against the established alternatives. The modular pattern it critiques remains the classical task and motion planning paradigm. In this paradigm, a symbolic task planner is bolted onto a separate motion controller. This approach offers strong guarantees, but it suffers from the latency and brittle hand tuning noted here. At the other extreme, end-to-end behaviour cloning struggles with multi-modal action distributions. This is precisely why diffusion based visuomotor policies (Chi et al. 2025, Cho et al.) have become the dominant learned-controller approach. These policies usually use a generative model to capture the full distribution of expert behaviour. The specific idea of emitting discrete sub-goal tokens that condition a diffusion trajectory decoder is closely related to recent hierarchical diffusion methods. Hierarchical Diffusion Policy (Ma et al. 2024) chains a high-level, language-guided agent with a low-level, kinematics-aware diffusion policy. The high-level agent predicts the next-best end-effector pose, while the low-level policy respects joint constraints. SkillDiffuser (Liang, Mu et al. 2024) and ChainedDiffuser (Xian et al. 2023) similarly couple discrete skill or keypose prediction with trajectory diffusion. The proposal's distinctive ambition, relative to these methods, is to collapse the two levels into a single jointly trained pass rather than using separate high- and low-level agents. Its principal risk is also well documented. The inference latency of diffusion sampling is a recognised bottleneck for real-time control. Therefore, the claim of sufficiently low latency for shared workspaces is exactly the point that would require empirical demonstration.

4.5.3 Reinforcement learning utilization in enhancing cobot vision-language model

Reward engineering remains the bottleneck that keeps reinforcement learning out of most cobot deployments. A practical alternative is to freeze a large vision-language backbone such as Contrastive Language-Image Pretraining and attach a tiny projection head that contains the robot's proprioception. Trained on a modest "universal progress" corpus, the head outputs a scalar that rises monotonically as the live scene better satisfies a single English instruction. Because only the final layer is tuned, the module generalizes: swapping in a new sentence "align the two gears," "wipe the table". Off-the-shelf RL algorithms such as Soft Actor-critic or Proximal Policy Optimization can begin learning immediately, no bespoke metrics, no human clicks (typing inputs). If this approach proves suitable, it could widen substantially skill acquisition for small and medium-size enterprises that lack large annotation budgets.

This formulation is not isolated where it belongs to a rapidly growing family of foundation-model reward methods. It is therefore worth positioning it against the established alternatives for avoiding hand-crafted rewards. The seminal example is VLM-RM (Rocamonde et al. 2024), which uses the cosine similarity between a CLIP embedding of the current state and a natural-language prompt as the reward. It can also project the representation onto the direction between a baseline prompt and a target prompt. This lightweight projection is essentially equivalent to the mechanism proposed here. Work in 2025–2026 has scaled this idea beyond a single similarity score. It has moved toward dense and temporally grounded reward signals. The Vision-Language-Action-Critic Model (VLAC) (Zhai, Zhang et al. 2025) adapts a large vision-language model into a process reward model. Given paired observations and a language goal, it outputs a dense progress signal and a completion flag. It also rejects irrelevant prompts. Another model which, termed as ReWiND (Zhang et al. 2025), learns a language-conditioned reward that generalises to unseen task variations. It does so from only a small grounding dataset and without per-task demonstrations. Frozen foundation models have also been used as online, frame-level reward engines for closed-loop policy refinement. They extract supervision directly from unlabelled video trajectories (Wu et al. 2026). The same recipe has been applied to safety-critical control, for example VLM-RL for autonomous driving (Huang et al. 2025). The shared appeal is that goals are specified in natural language, the backbone's internet-scale priors transfer zero-shot to new tasks, and only a small head is trained, making this the cheapest of the three paradigms to deploy.

This contrasts with the two classical routes to avoiding manual rewards. Inverse reinforcement learning (Ng and Russell 2000, Ziebart et al. 2008) recovers a reward from expert demonstrations, but the problem is inherently under-determined, often requires an expensive inner reinforcement-learning loop, and scales poorly to high-dimensional visual states. Preference-based reinforcement learning (Christiano et al. 2017, Lee et al. 2021) instead learns a reward from human comparisons of trajectory segments. It aligns well with human intent, but is limited by the cost, noise, and ambiguity of human queries, which recent work such as PRMT (Wang et al. 2026) addresses through multimodal feedback and foundation-model trajectory synthesis. The frozen vision-language reward sidesteps both the demonstration requirement of inverse reinforcement learning and the human-query burden of preference learning by reading the reward from a pretrained model. However, it has two known weaknesses. First, its reward granularity can be coarse because vision-language models have limited spatial and fine-grained reasoning, making the signal less reliable for contact-rich manipulation. Second, it can be susceptible to reward hacking when the policy exploits the projection head. Importantly, the paradigms are complementary rather than mutually exclusive. A vision-language model can supply the preference labels used in preference-based reinforcement learning, as shown by (Wang, Sun et al. 2024) and (Liu et al. 2025). Therefore, a frozen-model reward can serve as a cheap, generalisable prior that is later refined by a small amount of human- or model-generated preference feedback. The frozen vision-language reward is best positioned not as a replacement for inverse reinforcement learning or preference learning, but as the most deployment-efficient starting point, with preference-based refinement as a natural second stage.

4.5.4 Towards an end-to-end perception–action loop

Taken together, these three directions sketch a future cobot stack with almost no handcrafted glue. Adaptive fusion supplies trustworthy geometry; the hybrid VLA network turns natural-language goals into smooth, safe trajectories; and the zero-shot reward head closes the loop by eliminating manual tuning. Such an ecosystem would let integrators add new tasks via a sentence and a few demonstrations rather than weeks of sensor calibration and controller coding, moving collaborative robots closer to true plug-and-play versatility on the factory floor.

Figure 14 highlights how three emerging research threads, information-aware sensor fusion, language-conditioned planning plus control, and zero-shot reward estimation, can be combined into a single cobot stack that all but eliminates bespoke engineering. By pushing credibility assessment to the sensor edge, the fusion layer delivers trustworthy geometry without saturating factory networks. The hybrid VLA model then transforms high-level, human-readable intents into smooth, safe joint commands in one GPU pass, avoiding the latency and brittleness of traditional planner-controller seams. Finally, the zero-shot reward head acts as an automatic supervisor, using the same fused perception to measure task progress and drive continual policy refinement. Together these components form a self-contained perception–action loop that lets integrators deploy new tasks with little more than sentence-level description and a handful of demonstration runs, an essential step toward truly flexible, plug-and-play collaborative robotics on the factory floor.

In addition to what is described in Fig. 14, two further considerations are worth raising, and they relate to the loop in deliberately different ways. The first, cybersecurity, is best understood not as a general concern bolted onto the system but as a direct extension of safe perception itself: adversarial perturbation of visual input, sensor spoofing, and tampering along the perception-to-action chain attack the very signals the loop depends on, and such risks grow increasingly unavoidable as cobots are networked (Balasubramanian et al. 2025). Recent security surveys of robotic systems catalogue these threats, from LiDAR and camera spoofing to perception manipulation along the operator-robot channel, and report that many have already been demonstrated on deployed platforms (Cao et al. 2019, Sabouri 2026). Next-generation cobots should therefore be shipped with an IEC 62443-aligned "security spine" in which every depth camera or joint drive is assigned its credibility map and streams across only those micro-segmented, mutually authenticated channels negotiated with handshakes; the per-sensor credibility maps already computed for fusion then double as an integrity check, so that spoofed or tampered inputs

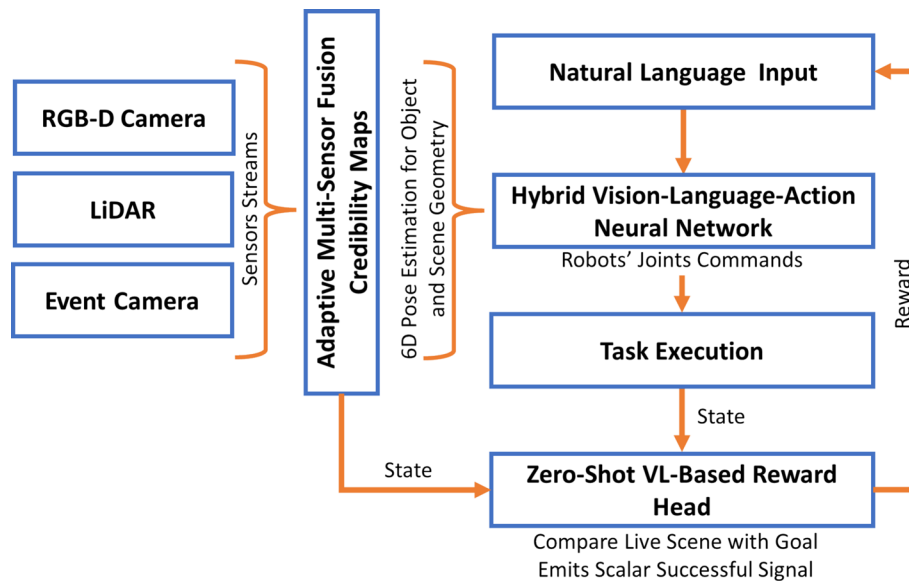


Fig. 14 Conceptual end-to-end perception-to-action loop for plug-and-play collaborative robots: RGB-D cameras, spinning or solid-state LiDARs and ultrafast event cameras (left column) stream heterogeneous data into an Adaptive Multi-Sensor Fusion block that produces per-pixel/beam credibility maps and a sub-millimetre 6-DoF scene estimate. Combining that geometry with a Natural-Language Input command (top right), conditions a Hybrid Vision-Language-Action Neural Network which simultaneously plans discrete sub-goals and denoises a jerk-limited trajectory. Executed motions propagate to the Zero-Shot Vision-Language-based Reward Head (bottom right), which converts live camera views plus the original goal sentence into a dense success signal. Two feedback pathways close the loop: (i) a blue arrow supplies fused geometry to the reward head for more precise scoring, and (ii) a yellow arrow feeds the scalar reward back to the hybrid network, enabling real-time reinforcement or adaptive fine-tuning without manual intervention

are down-weighted before they can reach the planner. This would close the attack chains already demonstrated on cobots, replace today's flat VLANs with per-task enclaves, and let integrators treat cybersecurity as a built-in service rather than an after-market patch (Hollerer et al. 2021). Securing the perception pipeline is, in this sense, a genuine next step for the loop itself, which is why it is shown as a cross-cutting layer in Fig. 10. The second consideration, quantum AI, is of a different character: it is a longer-horizon, general-purpose enabler rather than an immediate next step for the loop, and its relevance to cobot perception remains largely speculative at present (Devadas and T 2025, Sun, Sun et al. 2026). In principle, Quantum Machine Learning (QML) superposition could evaluate many candidate grasps, poses or motion primitives in parallel, shrinking trajectory-planning time, while entanglement could let a variational quantum circuit model the tight, non-linear correlation across heterogeneous sensors such as depth, LiDAR and event streams, so the fusion block converges on a sharper 6-DoF scene estimate from fewer frames. Early quantum reinforcement learning, in which a parameterised quantum circuit serves as the function approximator within an otherwise classical loop, has begun to demonstrate quantum agents for robot navigation and control, though so far only in simulation and on small-scale problems (Hohenfeld et al. 2024, Kaldari et al. 2025). Such approaches could in turn reduce jerk, torque and collision risk and accelerate reinforcement-learning updates that would otherwise stall on large state spaces, with global optimisation tasks, such as policy search, reward shaping and rare-failure simulation, off-loaded to a cloud Quantum Processing Unit (QPU) via a hybrid algorithm such as the variational quantum eigensolver, while the low-latency control loop stays on classical edge GPUs. These gains, however, remain contingent on a level of quantum hardware maturity that is not yet available; we therefore present quantum AI as a broader, forward-looking direction rather than as a component of the loop proposed here.

5 Conclusion

This review delivers a forward-looking, state of the art appreciation of advanced visual-perception algorithms for collaborative robots (cobots). It traces the field's evolution from single-camera, task-specific vision to heterogeneous pipelines that fuse RGB-D, LiDAR, neuromorphic event streams, large vision-language models and AR/VR feedback, showing how this multimodal stack supports object localization, human-action recognition, language-conditioned planning, diffusion-based motion generation and zero-shot reward shaping. Four research topics frame the discussion, (i) the current state of the art, (ii) enabling algorithms and datasets, (iii) open problems and (iv) prospective solutions for cobot perception.

Although multimodal perception and language-conditioned control have advanced rapidly, seven critical shortcomings still stand between today's prototypes and true plug-and-play cobots on the factory floor:

1. Large-scale sensor fusion has yet to show that credibility-weighted pruning of RGB-D, LiDAR and event streams can keep pose errors below a millimetre when cutting bandwidth and edge-GPU load by a factor of ten.
2. Safety remains reactive; no benchmarked pipelines fuse frame-based and neuromorphic data to predict human motion early enough to prevent collisions in shared HRC assembly bays.
3. Few- and zero-shot perception break downs when parts arrive deformed, damaged or heavily occluded, emphasising the need for richer synthetic-plus-real datasets and tougher domain-adaptation techniques.
4. Cross-robot perception sharing is still confined to the laboratory because lightweight, trustworthy algorithms for auto-synchronizing multiple cameras and arms in cluttered, reconfigurable cells are not production-ready.
5. Reinforcement learning is blocked by reward engineering; frozen VLM and tiny proprio head reward heads have not been proven stable or sample-efficient beyond toy demonstrations.
6. Open, highly-granular multimodal datasets are missing, specifically a body of pair time-synchronized visual logs with language, proprioception and operator state, plus affect-aware benchmarks to test how well cobots adjust speed or force when humans appear stressed.
7. Quantum-accelerated solvers remain unintegrated; routing data over a hardened QPU link combining fusion and trajectory networks in trusted hardware roots which protects every packet must be achieved without introducing the latency, bandwidth overheads, or attack surface that would cancel out the promised speed-ups.

This review finds that multimodal, credibility-weighted perception, coupled with vision-language models, enhance high-precision object localization, reliable human-action recognition and language-conditioned task planning in collaborative workspaces. This review integrated stack enables cobots to translate plain-language goals into smooth, high speed, collision-aware trajectories while simultaneously generating self-supervised reward signals that drive online policy refinement. At the same time, this review is scoped to visual perception, and several studies surveyed here show that vision alone can be insufficient for safety-critical and contact-rich human-robot collaboration. Occlusion and physical contact are precisely where camera-based perception is weakest (Lee et al. 2019, Liu et al. 2024, Jiang et al. 2025). Accordingly, several works have combined vision with haptic, force/torque, tactile, and audio sensing to support safer physical interaction (Bdiwi 2014; Tannous et al. 2020; Guo et al. 2023; Wong et al. 2024). Therefore, multimodal perception is increasingly regarded as a prerequisite for safe physical interaction, rather than an optional extension.

Future research should converge on an end-to-end perception-action architecture that delivers plug-and-play deployment without bespoke tuning. At the sensing layer, credibility-aware fusion must be implemented on embedded GPUs, with head-to-head benchmarks that quantify bandwidth, latency and sub-millimetres without posing accuracy problems under heavy occlusion. Building on this foundation, predictive-safety modules that combine RGB-D, LiDAR and event streams should learn to forecast human motion and pre-empt collisions in cluttered, reconfigurable cells. Perception robustness demands synthetic-plus-real datasets and domain-adaptation schemes that can generalize to include deformed, damaged or novel parts with few- and/or zero-shots, while dynamic-calibration algorithms must auto-synchronize multi-camera, multi-robot setups on-the-fly. A hybrid

VLA network, co-training scene-graph reasoning, language-conditioned sub-goal tokens and diffusion-generated trajectories, can then translate sentence-level instructions into smooth, jerk-bounded motion within 50 ms decision cycles. Reinforcement learning should be driven by lightweight, proprio-aware reward heads grafted onto frozen vision-language backbones to accelerate skill acquisition without hand-crafted metrics. Finally, publicly available, high-granularity multimodal datasets, complete with time-synchronized visual logs, operator affect labels and robot states, are essential for benchmarking cross-robot perception sharing and affect-aware speed modulation. Together, these directions would yield agile, human-centred cobots capable of commissioning new tasks in minutes rather than weeks across FMS assembly/disassembly cells and inspection lines.

6 Appendix A: Detailed graphical summaries of visual perception foundation studies in Sect. 2

See Tables 2, 3, 4 and 5.

Table 2 Object detection methods reviewed in Sect. 2.2.2

Method	Year	Type	Input	Key contribution	References
R-CNN	2014	2-D, two-stage	RGB	Selective Search ROIs $\rightarrow$ CNN features $\rightarrow$ SVM $\rightarrow$ box regression; established CNN-based detection	(Girshick et al. 2014)
Fast R-CNN	2015	2-D, two-stage	RGB	ROI pooling from a shared conv feature map; removes redundant computation	(Girshick 2015)
Faster R-CNN	2015	2-D, two-stage	RGB	RPN unifies proposal, extraction, classification and regression; near real-time	(Ren et al. 2016)
R-FCN	2016	2-D, two-stage	RGB	Fully convolutional; position-sensitive score maps; fewer parameters	(Dai et al. 2016)
Mask R-CNN	2017	2-D, two-stage	RGB	Adds a pixel-level mask branch to Faster R-CNN	(He et al. 2017) (Rettenberger et al. 2023) (Kijdech and Vongbun- yong 2024)
U-Net	2015	2-D, encoder-decoder	RGB	Contracting/expanding paths with skip connections; precise boundaries from limited data	(Ronneberger et al. 2015) (Rasheed et al. 2024)
Attention U-Net	2018	2-D, encoder-decoder	RGB	Adds attention gates to U-Net to focus on relevant regions	(Lukács et al. 2025)
SSD	2016	2-D, one-stage	RGB	Anchor-based class + box-offset prediction; multi-scale, fully convolutional	(Liu et al. 2016)
Feature Fusion SSD	2017	2-D, one-stage	RGB	Fuses fine + contextual features across layers	(Li et al. 2017)
YOLO (series, v1 $\rightarrow$)	2016	2-D, one-stage	RGB/point cloud	Grid-based simultaneous box + class prediction; anchors and clustering in later versions	(Sapkota et al. 2025) (Redmon et al. 2016)
YOLOv11	2024	2-D, one-stage (anchor-free)	RGB/2-D, 3-D, point cloud	Anchor-free, decoupled head, multi-scale FPN/PAN neck, task-aligned assignment; unified detection/seg/pose/OBB	Khanam (Khanam and Hussain 2024) (Terven et al. 2023)
RetinaNet	2017	2-D, one-stage	RGB	Focal loss emphasizes hard samples; addresses class imbalance	(Lin et al. 2018)
Libra R-CNN	2019	2-D, two-stage	RGB	IoU-balanced sampling, Balanced Feature Pyramid, balanced L1 loss	Pang et al. 2019
TridentNet	2019	2-D, two-stage (multi-branch)	RGB	Parallel branches with scale-specific receptive fields	Li et al. 2019a, b, c
DETR	2020	2-D, transformer	RGB	Set-based detection; models global object relations; no NMS	Carion et al. 2020 (Zhang et al. 2021)
Deformable DETR	2020	2-D, transformer	RGB	Sparse deformable sampling on relevant regions; faster convergence	Zhu et al. 2020 (Zhu et al. 2020)
ViT	2020	2-D, transformer	RGB	Patch-based encoding captures global context (classification)	(Dosovitskiy 2020)
DeiT	2021	2-D, transformer	RGB	Distillation-based training; effective on small datasets	(Touvron et al. 2021)
Mono3-D	2016	3-D, monocular	RGB (mono)	Ground-plane prior; 3-D box proposals via energy minimization + CNN refinement	(Chen et al. 2016)
MonoDIS	2019	3-D, monocular	RGB (mono)	Loss disentanglement per parameter; signed-IoU loss for training stability	(Simonelli et al. 2019)
MonoGRNet	2019	3-D, monocular	RGB (mono)	Four-stage pipeline with instance-depth module; faster inference	(Qin et al. 2019)
MonoRCNN	2021	3-D, monocular	RGB (mono)	Distance decomposed into physical + projection heights	Shi et al. 2021
Local surface embeddings	2020	3-D, CAD-matching	RGB + CAD	Aligns 2-D points to 3-D CAD points; Mask R-CNN + RANSAC for 6D pose	(Pitteri et al. 2020)
SurfEmb (dense corr.)	2022	3-D, CAD-matching	RGB + CAD	Unsupervised dense 2-D/3-D matches; contrastive embeddings; handles occlusion	(Haugaard and Buch 2022)
Relative-pose embeddings	2024	3-D, CAD-matching	RGB + CAD	U-Net + attention discriminative embeddings; generalizes to new objects	(Nguyen et al. 2024)

Table 2 (continued)

Method	Year	Type	Input	Key contribution	References
CAD/depth template matching	2022	3-D, CAD-matching	RGB/depth + CAD	Local CAD/depth representations for unseen or occluded objects	(Nguyen et al. 2022)
Siamese network	2024	3-D, CAD-matching	RGB (scene + reference)	Compares observed scene against a reference; also high-speed tracking	(Cabrera et al. 2024)
BEVDet	2021	3-D, BEV/multi-camera	Multi-cam RGB	Projects 2-D features to BEV; augmentation + improved NMS	(Huang, Huang et al. 2021)
BEVFormer	2022	3-D, BEV/multi-camera	Multi-cam RGB	Deformable sparse attention; BEV queries; reduces calibration reliance	(Li et al. 2022)
BEVFormer v2	2023	3-D, BEV/multi-camera	Multi-cam RGB	Temporal fusion of historical + current BEV features	(Yang et al. 2023a, b)
					(Yang, Chen et al. 2023a, b)
BEVerse	2022	3-D, BEV/multi-camera	Multi-cam RGB	Joint detection, mapping and motion prediction; spatio-temporal BEV + multi-task	(Zhang et al. 2022)
BEVDepth	2023	3-D, BEV/multi-camera	Multi-cam RGB	Depth supervision + spatial alignment for accuracy	(Li et al. 2023)
BEVStereo	2023	3-D, BEV/multi-camera	Multi-cam RGB	Temporal stereo over multi-view sequences	(Li et al. 2023)
Stereo R-CNN	2019	3-D, stereo	Stereo RGB	Extends Faster R-CNN; key point branch + photometric alignment	(Li, Chen et al. 2019a, b)
DSGN/DSGN++	2020/2022	3-D, stereo	Stereo RGB	3-D volume from stereo; single-stage depth + detection; ++ adds plane sweeping + dual-view volume	(Chen et al. 2022)
YOLOStereo3-D	2021	3-D, stereo	Stereo RGB	Pixel-level correlations, ghost module, hierarchical fusion; no disparity/LiDAR	(Liu, Wang et al. 2021)
Continuous Disparity Network	2020	3-D, stereo	Stereo RGB	Wasserstein-distance loss stabilizes depth gradients	(Garg et al. 2020)
Voxel R-CNN	2021	3-D, point/voxel (voxel)	LiDAR/point cloud	Coarse voxels suffice for high accuracy at lower cost	(Deng, Shi et al. 2021)
PointRCNN	2019	3-D, point/voxel (point)	LiDAR/point cloud	Two-stage on raw points; foreground/background segmentation	(Shi et al. 2019)
PV-RCNN/PV-RCNN++	2020/2023	3-D, point/voxel (hybrid)	LiDAR/point cloud	Combines voxel grids with point keypoint features	(Shi, Guo et al. 2020, Shi et al. 2023)
PointPillars	2019	3-D, point/voxel	LiDAR/point cloud	Pillar encoding + PointNet + 2-D conv; fast and accurate	(Lang et al. 2019)
Point-cloud Transformer (reg.)	2024	3-D, point/voxel	Point cloud	Transformer features improve point-cloud accuracy	(Wang et al. 2024a, b, c, d, e)
Fusion ViT	2023	3-D, fusion	RGB + LiDAR	Transformer-only hierarchical fusion of images + point clouds	(Xiang and Zhang 2023)
Cross Modal Transformer	2023	3-D, fusion	RGB + LiDAR	Image + point tokens → 3-D boxes; no explicit view transform	(Yan, Liu et al. 2023a, b)
Voxel/Image enhancement fusion	2024	3-D, fusion	RGB + LiDAR	Mutual Voxel/Image enhancement modules bridge the modality gap	(Song and Wang 2024)
Multi-layer fusion network	2024	3-D, fusion	RGB + LiDAR	Dense + adaptive fusion; boosts small-object detection	(Guo and Hu 2024)

Table 3 Instance segmentation methods reviewed in Sect. 2.2.3

Method	Year	Approach	Key idea	References
Mask R-CNN	2017	2-D, detection-based	Mask branch alongside box head; high-quality per-instance masks (~5 fps); extensible to pose	(He et al. 2017)
MaskLab	2018	2-D, detection-based	Faster R-CNN + semantic + direction features; separates instances of the same class	(Chen et al. 2018)
Hybrid Task Cascade	2019	2-D, cascade	Exploits the reciprocal detection-segmentation relationship in a cascade	(Chen et al. 2019)
PointRend	2020	2-D, rendering-based	Adaptive point-wise predictions via iterative subdivision	(Kirillov et al. 2020)
TensorMask	2019	2-D, dense sliding-window	4-D tensor output per location with custom operators	(Chen et al. 2019)
YOLOACT/YOLACT++	2019	2-D, prototype-based (real-time)	Prototype masks + per-instance coefficients; no repooling	(Bolya et al. 2019a, b)
SOLO	2020	2-D, location-based	Assigns “instance categories” by location/size; no box or post-processing	(Wang et al. 2020)
BlendMask	2020	2-D, top-down + bottom-up	Blends instance-level context with pixel-level detail	(Chen et al. 2020)
Shape-Aware Mono 3-D	2023	3-D, monocular	Instance-seg head + 3-D box projection; ADS depth metric; robust to occlusion	(Chen et al. 2023a, b)
MonoCInIS	2021	3-D, monocular	Camera-independent geometric reasoning; intrinsic-agnostic	(Heylen et al. 2021)
CD3IS	2024	3-D, monocular	Double-Snake depth + instance seg; adaptive point-cloud filtering	(Tang et al. 2024a, b)
CNOS	2023	3-D, CAD-based	FastSAM proposals + DINOv2 descriptors + matching; unseen objects, no retraining	(Nguyen et al. 2023)
Mask2CAD	2020	3-D, CAD-based	Shared-embedding CAD-template retrieval + pose refinement	(Kuo et al. 2020)
BEVSegFormer	2023	3-D, BEV/multi-camera	Deformable transformer BEV semantic segmentation; arbitrary camera setups	(Peng et al. 2023a, b)
ODIN	2024	3-D, multi-view	Alternates within-view 2-D and cross-view 3-D fusion; joint 2-D + 3-D	(Jain et al. 2024)
Instant NeRF (seg)	2023	3-D, NeRF-based	Learns an instance field; 3-D masks supervised by multi-view 2-D	(Liu et al. 2023)
Adaptive context-aware net	2023	3-D, RGB-D	Lightweight encoder-decoder; channel/spatial fusion attention	(Zhang et al. 2023)
DiPFormer	2024	3-D, RGB-D	Depth SAO + Depth LCA; multi-scale MLP decoder; real-time	(Chen et al. 2024)
ScanNet200 expansion	2022	3-D, point/voxel	200-class vocabulary; language-driven pre-training for class imbalance	(Rozenberszki et al. 2022)
DyCo3-D	2021	3-D, point/voxel	Proposal-free dynamic kernels; Transformer for long-range context	(He et al. 2021)
JSNet	2020	3-D, point/voxel	Joint instance + semantic; mean-shift clustering; no proposals	(Zhao and Tao 2020)
PolarNet	2020	3-D, point/voxel (LiDAR)	Polar BEV grid balances point distribution; near real-time	(Zhang et al. 2020)
SoftGroup	2022	3-D, point/voxel	Soft grouping + top-down refinement; reduces error propagation	(Vu et al. 2022)
Stratified Transformer	2022	3-D, point/voxel (transformer)	Dense-local/sparse-distant sampling; relative position encoding	(Lai et al. 2022)
Cylinder3-D	2020	3-D, point/voxel (LiDAR)	Cylindrical partitioning avoids 2-D projection loss	(Zhou et al. 2020)
TransFusion	2022	3-D, fusion (RGB+LiDAR)	Transformer soft association; resilient to calibration errors	(Bai et al. 2022)
BEVFusion	2023	3-D, fusion (RGB+LiDAR)	Shared BEV; optimised pooling; low latency, multi-task	(Liu et al. 2023)
HFA-Net	2025	3-D, fusion (geom.+colour)	HFEC + HBAC + ternary distance attention pooling	(Wen et al. 2025)

Table 4 Object pose estimation methods reviewed in Sect. 2.2.4

Method	Year	Level	Input	Key idea	References
GDR-Net	2021	Instance	RGB	Direct regression + dense 2-D-3-D geometry; end-to-end, no PnP/RANSAC	(Wang et al. 2021)
CDPN	2019	Instance	RGB	Disentangles rotation/translation; robust to texture-less/occluded objects	(Li et al. 2019)
CosyPose	2020	Instance	Multi-view RGB	Single-view hypotheses + matching + global bundle adjustment	(Labbé et al. 2020)
CAPTRA	2021	Category	Point-cloud seq	9-DoF tracking; canonicalisation + RotationNet + CoordinateNet	(Weng et al. 2021)
FS-Net	2021	Category	RGB-D	Orientation-aware autoencoder; 3-D graph conv; decoupled rotation	(Chen et al. 2021)
One stage 3-D recon + pose	2022	Category	RGB-D (single view)	Box-free pixel-wise centres; shape + 6D + size at 40 fps	(Irshad et al. 2022)
MegaPose	2022	Unseen	RGB	Render-and-compare + coarse estimator; large synthetic training set	(Labbé et al. 2022)
GigaPose	2024	Unseen	RGB	2-DoF template sampling; fast nearest-neighbour matching; CAD-based	(Nguyen et al. 2024)
ZS6D	2024	Unseen (zero shot)	RGB	ViT descriptors + patch matches + RANSAC-PnP	(Ausserlechner et al. 2024)
FoundPose	2023	Unseen	RGB	DINOv2 patch descriptors + template retrieval + feature-metric alignment	(Örnek et al. 2025)

Table 5 Human activity recognition approaches for HRC reviewed in Sect. 2.2.5

Method/Approach	Sensing	Model	HRC application	References
Multi camera 3-D pose + activity	Multi-camera 2-D	Particle filter (2-D-joint fusion) + LSTM	Recognizes 10 HRC activities (incl. construction sites)	(Jang et al. 2023)
Operator task action recognition	RGB	ResNet-34 + LSTM	Interprets implicit operator cues in assembly	(Moutinho et al. 2023)
Human robot teaming (HrT)	RGB	ML-based action recognition	Monitors human actions in HRC assembly	(Mehak et al. 2024)
Knowledge-encoded recognition	RGB	RCC + PDDL probabilistic model	Cuts training time via domain-knowledge encoding	(Lee et al. 2019)
Multi scale gesture control	RGB	Static/dynamic gesture + YOLOv5	70-command library for flexible production lines	(Wang et al. 2025a, b, c, d, e)
Multimodal gesture recognition	RGB + infrared	Gesture fusion model	Robust hand-gesture recognition under high illumination	(Wang et al. 2024a, b, c, d, e)
Safety motion detection	RGB	DL/CNN + distance algorithm	Real-time safe distance monitoring	(Chiurco et al. 2022)
Low cost pose recognition	RGB	Transfer-learning pose model	Path planning and collision avoidance	(Liu and Wang 2021)
Motion trajectory prediction	Depth (Kinect v2)	GMM + Expectation–Maximization	Predicts trajectories to avoid collisions	(Kang et al. 2018)
Proactive action prediction	RGB	RNN + Monte-Carlo dropout	Pre-emptive action; fewer mis-triggers	(Zhang et al. 2020a, b, c, d)
Lightweight motion prediction	RGB	MSLP network	Smooth, efficient human motion prediction	(Zou 2024)
Variable motion prediction	RGB	Memory graph NN + diffusion model	Predicts missing/uncertain human motion	(Gui, Li et al. 2024a, b, c, d)
INTERACT dataset	3-D skeleton, RGB-D, eye gaze, robot joints	Multimodal benchmark dataset	Benchmark for human cobot interaction	(Yasar et al. 2024)
Virtual HRC data generation	Digital/simulated	Musculoskeletal model + admittance controller	Generates virtual HRC training data	(Yao et al. 2024)
Deformable object understanding	RGB (time series)	Convolutional autoencoder + LSTM	In-air rope knotting; rope/human state tracking	(Suzuki et al. 2021)

Author contributions M.S.: Conceptualization, Methodology, Investigation, Visualization, Writing Original Draft. M.R.: Supervision, Validation, Proofreading, Reviewing and Editing. V.H.: Supervision, Validation, Proofreading, Reviewing and Editing, Funding Acquisition. K.N.: Proofreading, Project Administration, Funding Acquisition, Resources Acquisition. A.E.: Conceptualization, Methodology, Supervision, Validation, Proofreading, Reviewing and Editing. All authors read and approved the final manuscript.

Funding This work was carried out with the financial support of the Karlsruhe Nano Micro Facility (KNMFi, www.knmf.kit.edu), a Helmholtz Research Infrastructure at Karlsruhe Institute of Technology (KIT, www.kit.edu) and under the Helmholtz Research Programme MSE (Materials Systems Engineering) at KIT. Also, the authors would like to acknowledge the support provided by the KIT-Publication Fund of Karlsruhe Institute of Technology.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest There is no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Apps AZ, Wang V, Perejogin TL Molloy and Mahony R (2025) Asynchronous Multi-Object Tracking with an Event Camera. 2025 IEEE International Conference on Robotics and Automation (ICRA), IEEE.
- Armeni I, Sener O, Zamir AR, Jiang H, Brilakis I, Fischer M, Savarese S (2016). 3D Semantic Parsing of Large-Scale Indoor Spaces. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
- Auddy S, Paolillo A, Piater J, Saveriano M (2025) Imitation learning-based direct visual servoing using the large projection formulation. *Robot Auton Syst* 190:104971
- Ausserlechner P, Habberger D, Thalhammer S, Weibel JB, Vincze M (2024). Zs6d: Zero-shot 6d object pose estimation using vision transformers. 2024 IEEE International Conference on Robotics and Automation (ICRA), IEEE.
- Back S, Lee J, Kim K, Rho H, Lee G, Kang R, Lee S, Noh S, Lee Y, Lee T, Lee K (2025) GraspClutter6D: A Large-scale Real-world Dataset for Robust Perception and Grasping in Cluttered Scenes. *arXiv preprint [arXiv:2504.06866](https://arxiv.org/abs/2504.06866)*.
- Bai X, Hu Z, Zhu X, Huang Q, Chen Y, Fu H, Tai CL (2022) Transfusion: robust lidar-camera fusion for 3d object detection with transformers. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.
- Balasubramanian P, Liyana S, Sankaran H, Sivaramakrishnan S, Pusuluri S, Pirttikangas S, Peltonen E (2025) Generative AI for cyber threat intelligence: applications, challenges, and analysis of real-world case studies. *Artif Intell Rev* 58(11):336
- Bascetta L, Ferretti G, Rocco P, Ardö H, Bruyninckx H, Demeester E, Di Lello E (2011) Towards safe human-robot interaction in robotic cells: An approach based on visual tracking and intention estimation. *IEEE International Conference on Intelligent Robots and Systems*.
- Bdiwi M (2014) Integrated sensors system for human safety during cooperating with industrial robots for handing-over and assembling tasks. *Procedia CIRP* 23:65–70
- Belcamino V, Takase M, Kilina M, Carfi A, Mastrogiovanni F, Shimada A, Shimizu S (2024) Gaze-Based Intention Recognition for Human-Robot Collaboration. *Proceedings of the 2024 International Conference on Advanced Visual Interfaces*. Arenzano, Genoa, Italy, Association for Computing Machinery: Article 36.
- Bellandi P, Sansoni G, Vertuan A (2012) Development and characterization of a multi-camera 2D-vision system for enhanced performance of a drink serving robotic cell. *Robot Comput Integr Manuf* 28(1):35–49
- Benli E, Spidalieri RL, Motai Y (2019) Thermal multisensor fusion for collaborative robotics. *IEEE Trans Ind Inform* 15(7):3784–3795
- Berens F, Koschinski Y, Badami MK, Geimer M, Elser S, Reischl M (2024) Adaptive training for robust object detection in autonomous driving environments. *IEEE Trans Intell Vehicles* 10:1–15

- Bhargava A, Sachdeva A, Sharma K, Alsharif MH, Uthansakul P, Uthansakul M (2024) Hyperspectral imaging and its applications: a review. *Heliyon* 10(12):e33208
- Bian T, Chollet M, Guha T (2025) Robust understanding of human-robot social interactions through multimodal distillation. arXiv preprint [arXiv:2505.06278](https://arxiv.org/abs/2505.06278).
- Bolya D, Zhou C, Xiao F, Lee YJ (2019) Yolact: real-time instance segmentation. *Proceedings of the IEEE/CVF international conference on computer vision*.
- Bolya D, Zhou C, Xiao F, Lee YJ (2019b) YOLACT++ better real-time instance segmentation. *IEEE Trans Pattern Anal Mach Intell* 44:1108–1121
- Borboni A, Reddy KVV, Elamvazuthi I, AL-Quraishi MS, Natarajan E, Azhar Ali SS (2023) The expanding role of artificial intelligence in collaborative robots for industrial applications: a systematic review of recent works. *Machines* 11(1):111
- Boschetti G, Faccio M, Granata I, Minto R (2023) 3D collision avoidance strategy and performance evaluation for human–robot collaborative systems. *Comput Ind Eng* 179:109225
- Brachmann E, Krull A, Michel F, Gumhold S, Shotton J, Rother C (2014) Learning 6D object pose estimation using 3D object coordinates. *Computer Vision – ECCV 2014*, Cham, Springer International Publishing.
- Brohan A, Brown N, Carbajal J, Chebotar Y, Chen X, Choromanski K, Ding T, Driess D, Dubey A, Finn C (2023) Rt-2: vision-language-action models transfer web knowledge to robotic control. arXiv preprint [arXiv:2307.15818](https://arxiv.org/abs/2307.15818).
- Cabrera JJ, Román V, Gil A, Reinoso O, Payá L (2024) An experimental evaluation of Siamese Neural Networks for robot localization using omnidirectional imaging in indoor environments. *Artif Intell Rev* 57(8):198
- Caesar H, Bankiti V, Lang AH, Vora S, Liong VE, Xu Q, Krishnan A, Pan Y, Baldan G, Beijbom O (2020) nuscenes: a multimodal dataset for autonomous driving. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.
- Cai Z, Vasconcelos N (2018) Cascade r-cnn: Delving into high quality object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*.
- Cao Y, Ma Y, Zhu Y, Ting KM (2024) Revisiting streaming anomaly detection: benchmark and evaluation. *Artif Intell Rev* 58(1):8
- Cao Y, Xiao C, Cyr B, Zhou Y, Park W, Rampazzi S, Chen QA, Fu K, Mao ZM (2019) Adversarial Sensor Attack on LiDAR-based Perception in Autonomous Driving. *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. London, United Kingdom, Association for Computing Machinery: 2267–2281.
- Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S (2020) End-to-end object detection with transformers. *European conference on computer vision*, Springer.
- Castilla-Arquillo R, Pérez-del-Pulgar CJ, Gerdes L, Garcia-Cerezo A, Olivares-Mendez M (2025) OmniUnet: A Multimodal Network for Unstructured Terrain Segmentation on Planetary Rovers Using RGB, Depth, and Thermal Imagery. arXiv preprint [arXiv:2508.00580](https://arxiv.org/abs/2508.00580).
- Cazzato D, Cimorelli C, Sanchez-Lopez JL, Voos H, Leo M (2020) A survey of computer vision methods for 2D object detection from unmanned aerial vehicles. *J Imaging*. <https://doi.org/10.3390/jimaging6080078>
- Cervantes J, Garcia-Lamont F, Rodríguez-Mazahua L, Lopez A (2020) A comprehensive survey on support vector machine classification: applications, challenges and trends. *Neurocomputing* 408:189–215
- Chai J, Zeng H, Li A, Ngai EWT (2021) Deep learning in computer vision: a critical review of emerging techniques and application scenarios. *Mach Learn Appl* 6:100134
- Chan CC, Tsai CC (2020) Collision-free speed alteration strategy for human safety in human-robot coexistence environments. *IEEE Access* 8:80120–80133
- Chang P, Luo R, Dorostian M, Padir T (2021) A shared control method for collaborative human-robot plug task. *IEEE Robot Autom Lett* 6(4):7429–7436
- Chen H, Sun K, Tian Z, Shen C, Huang Y, Yan Y (2020) BlendMask: top-down meets bottom-up for instance segmentation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Chen J, Er MJ (2024) Dynamic YOLO for small underwater object detection. *Artif Intell Rev* 57(7):165
- Chen K, Pang J, Wang J, Xiong Y, Li X, Sun S, Feng W, Liu Z, Shi J, Ouyang W, Loy CC (2019) Hybrid task cascade for instance segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.
- Chen LC, Hermans A, Papandreou G, Schroff F, Wang P, Adam H (2018) Masklab: Instance segmentation by refining object detection with semantic and direction features. *Proceedings of the IEEE conference on computer vision and pattern recognition*.
- Chen L, Wang Y, Miao Z, Mo Y, Feng M, Zhou Z, Wang H (2023a) Transformer-based imitative reinforcement learning for multirobot path planning. *IEEE Trans Ind Inform* 19(10):10233–10243
- Chen P, Lu W (2021) Deep reinforcement learning based moving object grasping. *Inf Sci* 565:62–76
- Chen S, Han T, Zhang C, Liu W, Su J, Wang Z, Cai G (2024) Depth Matters: Exploring Deep Interactions of RGB-D for Semantic Segmentation in Traffic Scenes. arXiv preprint [arXiv:2409.07995](https://arxiv.org/abs/2409.07995).
- Chen W, Jia X, Chang HJ, Duan J, Shen L, Leonardis A (2021) Fs-net: Fast shape-based network for category-level 6d object pose estimation with decoupled rotation mechanism. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

- Chen W, Zhao J, Zhao WL, Wu SY (2023b) Shape-aware monocular 3D object detection. *IEEE Trans Intell Transp Syst* 24(6):6416–6424
- Chen X, Girshick R, He K, Dollár P (2019) Tensormask: a foundation for dense object segmentation. *Proceedings of the IEEE/CVF international conference on computer vision*.
- Chen X, Kundu K, Zhang Z, Ma H, Fidler S, Urtasun R (2016) Monocular 3d object detection for autonomous driving. *Proceedings of the IEEE conference on computer vision and pattern recognition*.
- Chen Y, Arkin J, Zhang Y, Roy N, Fan C (2024) Scalable multi-robot collaboration with large language models: centralized or decentralized systems? 2024 IEEE International Conference on Robotics and Automation (ICRA).
- Chen Y, Huang S, Liu S, Yu B, Jia J (2022) Dsgn++: exploiting visual-spatial relation for stereo-based 3d detectors. *IEEE Trans Pattern Anal Mach Intell* 45(4):4416–4429
- Cheng X, Lu F, Liu Y (2024) Lightweight hybrid model based on MobileNet-v2 and Vision Transformer for human–robot interaction. *Eng Appl Artif Intell* 127:107288
- Chew SY, Asadi E, Vargas-Uscategui A, King P, Gautam S, Bab-Hadiashar A, Cole I (2024) In-process 4D reconstruction in robotic additive manufacturing. *Robot Comput Integr Manuf* 89:102784
- Chi C, Xu Z, Feng S, Cousineau E, Du Y, Burchfiel B, Tedrake R, Song S (2025) Diffusion policy: visuomotor policy learning via action diffusion. *Int J Robot Res* 44(10–11):1684–1704
- Chi M, Chang S, Guo Z, Huang S, Li Z, Li J, Xia Z, Zheng Z, Ren Q (2024) Research on target recognition and grasping of dual-arm cooperative mobile robot based on vision. 2024 International Symposium on Intelligent Robotics and Systems (ISoIRS).
- Chirurgo A, Frangella J, Longo F, Nicoletti L, Padovano A, Solina V, Mirabelli G, Citraro C (2022) Real-time detection of worker's emotions for advanced human-robot interaction during collaborative tasks in smart factories. *Procedia Comput Sci* 200:1875–1884
- Chmielowiec A (2025) Utilization of Neural Networks and U-Net Architecture for Cladding Area Detection for Collaborative Robots. *Innovations in Mechatronics Engineering IV*. Springer Nature Switzerland, Cham
- Cho H, Kang JY, Kim Y, Yoon KJ (2025) Ev-3dod: pushing the temporal boundaries of 3d object detection with event cameras. *Proceedings of the Computer Vision and Pattern Recognition Conference*.
- Cho M, Cao Y, Sun J, Zhang Q, Pavone M, Park JJ, Yang H, Mao Z (2025) Cocoon: Robust multi-modal perception with uncertainty-aware sensor fusion. *International Conference on Learning Representations*.
- Choi S, Kim D, Ahn M, Choi D (2024) Large language model based collaborative robot system for daily task assistance. *JMST Advances* 6(3):315–327
- Choi SH, Kim M, Lee JY (2025) Smart and user-centric manufacturing information recommendation using multimodal learning to support human-robot collaboration in mixed reality environments. *Robot Comput-Integr Manuf* 91:102836
- Chovet LP, Garcia GM, Bera A, Richard A, Yoshida K, Olivares-Mendez MA (2025) Performance Comparison of ROS2 Middlewares for Multi-robot Mesh Networks in Planetary Exploration. *J Intell Rob Syst* 111(1):18
- Christiano PF, Leike J, Brown T, Martic M, Legg S, Amodei D (2017) Deep reinforcement learning from human preferences. *Adv Neural Information Processing Systems* 30:120
- Chu C-H, Liu Y-L (2023) Augmented reality user interface design and experimental evaluation for human-robot collaborative assembly. *J Manuf Syst* 68:313–324
- Chu C-H, Weng C-Y (2024) Experimental analysis of augmented reality interfaces for robot programming by demonstration in manufacturing. *J Manuf Syst* 74:463–476
- Cong Y, Chen R, Ma B, Liu H, Hou D, Yang C (2023) A comprehensive study of 3-D vision-based robot manipulation. *IEEE Trans Cybern* 53(3):1682–1698
- Contreras M, Jain A, Bhatt NP, Banerjee A, Hashemi E (2024) A survey on 3D object detection in real time for autonomous driving. *Front Robot AI* 11:120
- Cordts, M., M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth and B. Schiele (2016). The Cityscapes Dataset for Semantic Urban Scene Understanding. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
- D'Emilia G, Di Gasbarro D (2017) Review of techniques for 2D camera calibration suitable for industrial vision systems. *J Phys Conf Ser* 841(1):012030
- Dai A, Chang AX, Savva M, Halber M, Funkhouser T, Nießner M (2017) Scannet: Richly-annotated 3d reconstructions of indoor scenes. *Proceedings of the IEEE conference on computer vision and pattern recognition*.
- Dai J, He K, Li Y, Ren S, Sun J (2016) Instance-sensitive fully convolutional networks. *Computer vision – ECCV 2016*. Springer International Publishing, Cham
- Debnath A, Stein GJ, Kosecka J (2025). SemNav: a model-based planner for zero-shot object goal navigation using vision-foundation models. *arXiv preprint [arXiv:2506.03516](https://arxiv.org/abs/2506.03516)*.
- Dekker SWA, Woods DD (2024) Wrong, strong, and silent: what happens when automated systems with high autonomy and high authority misbehave? *J Cogn Eng Decis Mak* 18(4):339–345
- Deng J, Shi S, Li P, Zhou Y, Zhang Y, Li H (2021) Voxel r-cnn: Towards high performance voxel-based 3d object detection. *Proceedings of the AAAI conference on artificial intelligence*.

- Devadas RM, T S (2025) Quantum machine learning: a comprehensive review of integrating AI with quantum computing for computational advancements. *MethodsX* 14:103318
- Dianatfar M, Latokartano J, Lanz M (2021) Review on existing VR/AR solutions in human–robot collaboration. *Procedia CIRP* 97:407–411
- Dimitropoulos N, Papalexis P, Michalos G, Makris S (2024) Advancing human-robot interaction using AI – a large language model (LLM) approach. *Advances in Artificial Intelligence in Manufacturing*. Springer Nature Switzerland, Cham
- Dong L, He Z, Song C, Yuan X, Zhang H (2024) Multi-robot social-aware cooperative planning in pedestrian environments using attention-based actor-critic. *Artif Intell Rev* 57(4):108
- Dosovitskiy A (2020) An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint [arXiv:2010.11929](https://arxiv.org/abs/2010.11929)*.
- Du G, Wang K, Lian S, Zhao K (2021) Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: a review. *Artif Intell Rev* 54(3):1677–1734
- Du Y, Yang S, Dai B, Dai H, Nachum O, Tenenbaum J, Schuurmans D, Abbeel P (2023) Learning universal policies via text-guided video generation. *Adv Neural Inf Process Syst* 36:9156–9172
- Duarte L, Polito M, Gastaldi L, Neto P, Pastorelli S (2024) Demonstration of real-time event camera to collaborative robot communication. *Advances in Italian mechanism science*. Springer Nature Switzerland, Cham
- Edozie E, Shuaibu AN, John UK, Sadiq BO (2025) Comprehensive review of recent developments in visual object detection based on deep learning. *Artif Intell Rev* 58(9):277
- Fan J, Zheng P (2024) A vision-language-guided robotic action planning approach for ambiguity mitigation in human–robot collaborative manufacturing. *J Manuf Syst* 74:1009–1018
- Fan Z, Zhang L, Wang X, Shen Y, Deng F (2025) LiDAR, IMU, and camera fusion for simultaneous localization and mapping: a systematic review. *Artif Intell Rev* 58(6):174
- Fang S, Huang X, Chen H, Xi N (2016) Dual-arm robot assembly system for 3C product based on vision guidance. 2016 IEEE International Conference on Robotics and Biomimetics (ROBIO).
- Felice D, Remus A, Gasperini S, Busam B, Ott L, Thalhammer S, Tombari F, Avizzano CA (2025) InstantPose: zero-shot instance-level 6D pose estimation from a single view. *IEEE Robot Autom Lett* 10(6):6023–6030
- Feng Z, Hu G, Sun Y, Soon J (2020) An overview of collaborative robotic manipulation in multi-robot systems. *Annu Rev Control* 49:113–127
- Flacco F, Kröger T, De Luca A, Khatib O (2012) A depth space approach to human-robot collision avoidance. *Proceedings - IEEE International Conference on Robotics and Automation*.
- Forlini M, Neri F, Ciccarelli M, Palmieri G, Callegari M (2024) Experimental implementation of skeleton tracking for collision avoidance in collaborative robotics. *Int J Adv Manuf Technol* 134(1):57–73
- Fu C, Lu K, Zheng G, Ye J, Cao Z, Li B, Lu G (2023) Siamese object tracking for unmanned aerial vehicle: a review and comprehensive analysis. *Artif Intell Rev* 56(1):1417–1477
- Gallego G, Delbrück T, Orchard G, Bartolozzi C, Tabá B, Censi A, Leutenegger S, Davison AJ, Conradt J, Daniilidis K, Scaramuzza D (2022) Event-based vision: a survey. *IEEE Trans Pattern Anal Mach Intell* 44(1):154–180
- Gao Z (2023) MMTSA: multi-modal temporal segment attention network for efficient human activity recognition. *Proc ACM Interact Mob Wearable Ubiquitous Technol* 7(3):96
- Garg D, Wang Y, Hariharan B, Campbell M, Weinberger KQ, Chao W-L (2020) Wasserstein distances for stereo disparity estimation. *Adv Neural Inf Process Syst* 33:22517–22529
- Gavenski N, Meneguzzi F, Luck M, Rodrigues O (2024) A survey of imitation learning methods, environments and metrics. *arXiv preprint [arXiv:2404.19456](https://arxiv.org/abs/2404.19456)*.
- Gbagbe KF, Cabrera MA, Alabbas A, Alyunes O, Lykov A, Tsetserukou D (2024) Bi-VLA: Vision-Language-Action Model-Based System for Bimanual Robotic Dexterous Manipulations. *arXiv preprint [arXiv:2405.06039](https://arxiv.org/abs/2405.06039)*.
- Geiger A, Lenz P, Stiller C (2013) Vision meets robotics: the KITTI dataset. *Int J Robot Res* 32(11):1231–1237
- Girshick R (2015) Fast r-cnn. *Proceedings of the IEEE international conference on computer vision*.
- Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*.
- Gkournelos C, Konstantinou C, Makris S (2024) An LLM-based approach for enabling seamless human-robot collaboration in assembly. *CIRP Ann* 73(1):9–12
- Guan W, Chen P, Zhao H, Wang Y, Lu P (2024) EVI-SAM: robust, real-time, tightly-coupled event–visual–inertial state estimation and 3D dense mapping. *Adv Intell Syst* 6(12):2400243
- Gui H, Li M, Yuan Z (2024) A behavioral conditional diffusion probabilistic model for human motion modeling in multi-action mixed human-robot collaboration. *Adv Eng Inform* 62:102742
- Guo B, Zhang H, Wang H, Li X, Jin L (2024) Adaptive occlusion object detection algorithm based on OL-IoU. *Sci Rep* 14(1):27644
- Guo D, Liu H, Sun F (2023) Audio–visual language instruction understanding for robotic sorting. *Robot Auton Syst* 159:104271

- Guo M-H, Cai J-X, Liu Z-N, Mu T-J, Martin RR, Hu S-M (2021) PCT: point cloud transformer. *Comput vis Media* 7(2):187–199
- Guo Y, Hu H (2024) Multi-Layer Fusion 3D Object Detection via Lidar Point Cloud and Camera Image. *Appl Sci*. <https://doi.org/10.3390/app14041348>
- Gupta A, Kumar V, Lynch C, Levine S (2020) Relay Policy Learning: Solving Long-Horizon Tasks via Imitation and Reinforcement Learning. *Proceedings of the Conference on Robot Learning*.
- Halme R-J, Lanz M, Kämäräinen J, Pieters R, Latokartano J, Hietanen A (2018) Review of vision-based safety systems for human-robot collaboration. *Procedia CIRP* 72:111–116
- Hameed A, Ordys A, Możaryn J, Sibilska-Mroziewicz A (2023) Control system design and methods for collaborative robots: review. *Appl Sci*. <https://doi.org/10.3390/app13010675>
- Haugaard RL, Buch AG (2022) SurfEmb: Dense and Continuous Correspondence Distributions for Object Pose Estimation with Learnt Surface Embeddings. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- He K, Gkioxari G, Dollár P, Girshick R (2017) Mask r-cnn. *Proceedings of the IEEE international conference on computer vision*.
- He, T., C. Shen and A. Van Den Hengel (2021). Dyco3d: Robust instance segmentation of 3d point clouds through dynamic convolution. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.
- He Y, Chen S (2019) Recent advances in 3D data acquisition and processing by time-of-flight camera. *IEEE Access* 7:12495–12510
- He Y, Wu M, Liu S (2020) An optimisation-based distributed cooperative control for multi-robot manipulation with obstacle avoidance. *IFAC-PapersOnLine* 53(2):9859–9864
- Hetto T, Schulz S, Reischl M (2023) Categorization and evaluation methods for control strategies of bilateral tasks in arm prosthetics. *IEEE Access* 11:52501–52508
- Heylen, J., M. De Wolf, B. Dawagne, M. Proesmans, L. Van Gool, W. Abbeloos, H. Abdelkawy and D. O. Reino (2021). Monocinis: Camera independent monocular 3d object detection using instance segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- Hidayatullah, P., N. Syakrani, M. R. Sholahuddin, T. Gelar and R. Tubagus (2025) YOLOv8 to YOLO11: A Comprehensive Architecture In-depth Comparative Review. *arXiv preprint arXiv:2501.13400*.
- 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. *Computer Vision – ACCV 2012, Berlin, Heidelberg, Springer Berlin Heidelberg*.
- Hodan 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. *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE.
- Hodaň T, Michel F, Brachmann E, Kehl W, Buch AG, Kraft D, Drost B, Vidal J, Ihrke S, Zabulis X, Sahin C, Manhardt F, Tombari F, Kim T-K, Matas J, Rother C (2018) BOP: benchmark for 6D object pose estimation. *Computer vision – ECCV 2018*. Springer International Publishing, Cham
- Hohenfeld H, Heimann D, Wiebe F, Kirchner F (2024) Quantum deep reinforcement learning for robot navigation tasks. *IEEE Access* 12:87217–87236
- Hollerer S, Fischer C, Brenner B, Papa M, Schlund S, Kastner W, Fabini J, Zseby T (2021) Cobot attack: a security assessment exemplified by a specific collaborative robot. *Procedia Manuf* 54:191–196
- Hong Q, Dong H, Deng W, Ping Y (2024) Education robot object detection with a brain-inspired approach integrating Faster R-CNN, YOLOv3, and semi-supervised learning. *Front Neurorob* 17:120
- Horaud R, Hansard M, Evangelidis G, Ménier C (2016) An overview of depth cameras and range scanners based on time-of-flight technologies. *Mach Vis Appl* 27(7):1005–1020
- Hosang J, Benenson R, Schiele B (2017) Learning non-maximum suppression. *Proceedings of the IEEE conference on computer vision and pattern recognition*.
- Hsu CCJ, Hwang PJ, Wang WY, Wang YT, Lu CK (2023) Vision-based mobile collaborative robot incorporating a multi-camera localization system. *IEEE Sens J* 23(18):21853–21861
- Hu J (2024) Augmented-reality based brain-computer interface of robot control. *Heliyon* 10(5):e26255
- Huang J, Huang G, Zhu Z, Ye Y, Du D (2021). Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*.
- Huang S, Jiang Z, Dong H, Qiao Y, Gao P, Li H (2023) Instruct2Act: mapping multi-modality instructions to robotic actions with large language model. *ArXiv abs/2305.11176*.
- Huang Z, Sheng Z, Qu Y, You J, Chen S (2025) VLM-RL: A unified vision language models and reinforcement learning framework for safe autonomous driving. *Transp Res C Emerg Technol* 180:105321
- Hwang P-J, Hsu C-C, Chou P-Y, Wang W-Y, Lin C-H (2022) Vision-based learning from demonstration system for robot arms. *Sensors (Basel)*. <https://doi.org/10.3390/s22072678>
- Hwang PJ, Hsu CC, Wang WY (2020) Development of a mimic robot—learning from demonstration incorporating object detection and multi-action recognition. *IEEE Consum Electron Mag* 9(3):79–87

- Irshad MZ, Kollar T, Laskey M, Stone K, Kira Z (2022) Centersnap: Single-shot multi-object 3d shape reconstruction and categorical 6d pose and size estimation. 2022 International Conference on Robotics and Automation (ICRA), IEEE.
- Jain A, Katara P, Gkanatsios N, Harley AW, Sarch G, Aggarwal K, Chaudhary V (2024) Odin: a single model for 2d and 3d perception. arXiv preprint [arXiv:2401.02416](https://arxiv.org/abs/2401.02416).
- Jang Y, Jeong I, Younesi Heravi M, Sarkar S, Shin H, Ahn Y (2023) Multi-camera-based human activity recognition for human-robot collaboration in construction. *Sensors*. <https://doi.org/10.3390/s23156997>
- Javaid M, Haleem A, Singh RP, Rab S, Suman R (2022a) Exploring impact and features of machine vision for progressive industry 4.0 culture. *Sens Int* 3:100132
- Javaid M, Haleem A, Singh RP, Rab S, Suman R (2022b) Significant applications of cobots in the field of manufacturing. *Cognitive Robotics* 2:222–233
- Jiang S, Zhao S, Fan Y, Yin P (2025). GelFusion: Enhancing robotic manipulation under visual constraints via visuotactile fusion. arXiv preprint [arXiv:2505.07455](https://arxiv.org/abs/2505.07455).
- Jiang, Y., A. Gupta, Z. Zhang, G. Wang, Y. Dou, Y. Chen, L. Fei-Fei, A. Anandkumar, Y. Zhu and L. Fan (2023). VIMA: robot manipulation with multimodal prompts. Proceedings of the 40th International Conference on Machine Learning. Honolulu, Hawaii, USA, JMLR.org. **202**: Article 611.
- Jiao R, Wan Y, Poiesi F, Wang Y (2023) Survey on video anomaly detection in dynamic scenes with moving cameras. *Artif Intell Rev* 56(3):3515–3570
- Jin P, Huang J, Liu F, Wu X, Ge S, Song G, Clifton D, Chen J (2022) Expectation-maximization contrastive learning for compact video-and-language representations. *Adv Neural Inf Process Syst* 35:30291–30306
- Ju Z, Wang H, Luo J, Sun F (2024) Enhancing human–robot communication with a comprehensive language-conditioned imitation policy for embodied robots in smart cities. *Comput Commun* 222:177–187
- K Kas (2024) Faster region based convolution neural network with context iterative refinement for object detection. *Meas Sens* 31:101025
- Kaldari, J., S. Tariq, S. Al-Kuwari, S. Y.-C. Chen, S. Chatzinotas and H. Shin (2025). Quantum reinforcement learning: recent advances and future directions. arXiv preprint [arXiv:2510.14595](https://arxiv.org/abs/2510.14595).
- Kang J (2025) Deep depth-flow odometry with inertial sensor fusion. *Electron Lett* 61(1):e70409
- Kang, J., K. Jia, F. Xu, F. Zou, Y. Zhang and H. Ren (2018). Real-Time Human Motion Estimation for Human Robot Collaboration. 2018 IEEE 8th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER).
- Kasaei H, Xiong S (2024) Lifelong ensemble learning based on multiple representations for few-shot object recognition. *Robot Auton Syst* 174:104615
- Kerssies T, Cavagnero N, Hermans A, Norouzi N, Averta G, Leibe B, Dubbelman G, De Geus D (2025) Your vit is secretly an image segmentation model. Proceedings of the Computer Vision and Pattern Recognition Conference.
- Keshvarparast A, Battini D, Battaia O, Pirayesh A (2024) Collaborative robots in manufacturing and assembly systems: literature review and future research agenda. *J Intell Manuf* 35(5):2065–2118
- Khanam R and M Hussain (2024) Yolov11: An overview of the key architectural enhancements. arXiv preprint [arXiv:2410.17725](https://arxiv.org/abs/2410.17725).
- Khatib M, Al Khudir K, De Luca A (2021) Human-robot contactless collaboration with mixed reality interface. *Robot Comput Integr Manuf* 67:102030
- Khelifa R, Hamza T, Fatma B (2024) Fuzzy task assignment in heterogeneous distributed multi-robot system. *Artif Intell Rev* 58(1):14
- Kijdech D, Vongbunyong S (2024) Manipulation of a complex object using dual-arm robot with Mask R-CNN and grasping strategy. *J Intell Robot Syst* 110(3):103
- Kim D, Choi M, Um J (2024a) Digital twin for autonomous collaborative robot by using synthetic data and reinforcement learning. *Robot Comput-Integr Manuf* 85:13
- Kim H, Ohmura Y, Kuniyoshi Y (2024b) Goal-conditioned dual-action imitation learning for dexterous dual-arm robot manipulation. *IEEE Trans Robot* 40:2287–2305
- Kimhi M, Vainshtein D, Baskin C, Di Castro D (2025) Robot instance segmentation with few annotations for grasping. 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE.
- Kirillov A, Wu Y, He K, Girshick R (2020) Pointrend: Image segmentation as rendering. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.
- Knapp V, Bohacek M (2025) Synthetic human action video data generation with pose transfer. *Synthetic Data for Computer Vision Workshop@ CVPR* 2025.
- Kobayashi S, Wan W, Kiyokawa T, Koyama K, Harada K (2023) Obtaining an object's 3D model using dual-arm robotic manipulation and stationary depth sensing. *IEEE Trans Autom Sci Eng* 20(3):2075–2087
- Konstantinou C, Antonarakos D, Angelakis P, Gkournelos C, Michalos G, Makris S (2024) Leveraging generative AI prompt programming for human-robot collaborative assembly. *Procedia CIRP* 128:621–626

- Konstantinou C, Kampouroglou N, Theodoris N, Basamakakis F, Gkournelos C, Makris S (2025) Leveraging generative AI for synthetic data generation: improving 6-DOF pose estimation in assembly systems. *Advances in Artificial Intelligence in Manufacturing II*, Cham, Springer Nature Switzerland.
- Krüger J, Katschinski V, Surdilovic D, Schreck G (2010) Flexible assembly systems through workplace-sharing and time-sharing human-machine cooperation (PISA). *ISR 2010 (41st International Symposium on Robotics) and ROBOTIK 2010 (6th German Conference on Robotics)*, VDE.
- Kuo, W., A. Angelova, T.-Y. Lin and A. Dai (2020). Mask2CAD: 3D Shape Prediction by Learning to Segment and Retrieve. *Computer Vision – ECCV 2020*, Cham, Springer International Publishing.
- Labbé Y, Carpentier J, Aubry M, Sivic J (2020) Cosypose: Consistent multi-view multi-object 6d pose estimation. *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, Springer.
- Labbé Y, Manuelli L, Mousavian A, Tyree S, Birchfield S, Tremblay J, Carpentier J, Aubry M, Fox D, Sivic J (2022) Megapose: 6d pose estimation of novel objects via render & compare. *arXiv preprint [arXiv:2212.06870](https://arxiv.org/abs/2212.06870)*.
- Lacevic B, Zanchettin AM, Rocco P (2023) Safe human-robot collaboration via collision checking and explicit representation of danger zones. *IEEE Trans Autom Sci Eng* 20(2):846–861
- Lai, X., J. Liu, L. Jiang, L. Wang, H. Zhao, S. Liu, X. Qi and J. Jia (2022). Stratified transformer for 3d point cloud segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.
- Lang, A. H., S. Vora, H. Caesar, L. Zhou, J. Yang and O. Beijbom (2019). Pointpillars: Fast encoders for object detection from point clouds. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.
- Langås EF, Zafar MH, Sanfilippo F (2025) Exploring the synergy of human-robot teaming, digital twins, and machine learning in Industry 5.0: a step towards sustainable manufacturing. *J Intell Manuf.* <https://doi.org/10.1007/s10845-025-02580-x>
- Lee K, Smith L, Abbeel P (2021) PEBBLE: Feedback-Efficient Interactive Reinforcement Learning via Relabeling Experience and Unsupervised Pre-training. *International Conference on Machine Learning*.
- Lee MA, Zhu Y, Srinivasan K, Shah P, Savarese S, Fei-Fei L, Garg A, Bohg J (2019) Making Sense of Vision and Touch: Self-Supervised Learning of Multimodal Representations for Contact-Rich Tasks. *2019 International Conference on Robotics and Automation (ICRA)*.
- Lee SU, Hofmann A, Williams B (2019) A Model-Based Human Activity Recognition for Human–Robot Collaboration. *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.
- Li C, Zheng P, Li S, Pang Y, Lee CKM (2022a) AR-assisted digital twin-enabled robot collaborative manufacturing system with human-in-the-loop. *Robot Comput Integr Manuf* 76:102321
- Li J, Wu J, Zhao W, Bai S, Bai X (2024) Partglee: A foundation model for recognizing and parsing any objects. *European Conference on Computer Vision*, Springer.
- Li J, Xu X, Jiang Z, Jiang B (2024b) Adaptive Kalman filter for real-time visual object tracking based on autocovariance least square estimation. *Appl Sci.* <https://doi.org/10.3390/app14031045>
- Li P, Chen X, Shen S (2019) Stereo r-cnn based 3d object detection for autonomous driving. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Li S, Xie H-L, Zheng P, Wang L (2024c) Industrial metaverse: a proactive human-robot collaboration perspective. *J Manuf Syst* 76:314–319
- Li Y, Bao H, Ge Z, Yang J, Sun J, Li Z (2023) Bevstereo: Enhancing depth estimation in multi-view 3d object detection with temporal stereo. *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Li Y, Chen Y, Wang N, Zhang Z (2019) Scale-aware trident networks for object detection. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*.
- Li Y, Ge Z, Yu G, Yang J, Wang Z, Shi Y, Sun J, Li Z (2023) Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Li Z, Ren L, Yang J, Zhao Y, Wu X, Xu Z, Bai X, Zhao H (2024) VIP: Vision Instructed Pre-training for Robotic Manipulation. *arXiv preprint [arXiv:2410.07169](https://arxiv.org/abs/2410.07169)*.
- Li Z, Wang J (2019) CDPN: Coordinates-Based Disentangled Pose Network for Real-Time RGB-Based 6-DoF Object Pose Estimation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*: 7677–7686.
- Li Z, Wang W, Li H, Xie E, Sima C, Lu T, Yu Q, Dai J (2022) Bevformer: Learning bird’s-eye-view representation from multi-camera images via spatiotemporal transformers. *European conference on computer vision*, Springer.
- Li Z, Yan Y, Wang X, Ge Y, Meng L (2025) A survey of deep learning for industrial visual anomaly detection. *Artif Intell Rev* 58(9):279
- Li Z, Yang L, Zhou F (2017) FSSD: feature fusion single shot multibox detector. *arXiv preprint [arXiv:1712.00960](https://arxiv.org/abs/1712.00960)*.
- Liang Z, Mu Y, Ma H, Tomizuka M, Ding M, Luo P (2024) Skilldiffuser: Interpretable hierarchical planning via skill abstractions in diffusion-based task execution. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Liao Y, Xie J, Geiger A (2023) KITTI-360: a novel dataset and benchmarks for urban scene understanding in 2D and 3D. *IEEE Trans Pattern Anal Mach Intell* 45(03):3292–3310
- Lin T-Y, Goyal P, Girshick R, He K, Dollar P (2018) Focal loss for dense object detection. *IEEE Trans Pattern Anal Mach Intell.* <https://doi.org/10.1109/TPAMI.2018.2858826>

- Lin, T.-Y., M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár and C. L. Zitnick (2014). Microsoft coco: Common objects in context. *Computer vision—ECCV 2014: 13th European conference, zurich, Switzerland, September 6–12, 2014, proceedings, part v 13*, Springer.
- Lintvedt N (2023) Thermal imaging in robotics as a privacy-enhancing or privacy-invasive measure? Misconceptions of privacy when using thermal cameras in robots. *Digital Society* 2(3):33
- Liu C, Tang D, Zhu H, Nie Q, Chen W, Zhao Z (2024a) An augmented reality-assisted interaction approach using deep reinforcement learning and cloud-edge orchestration for user-friendly robot teaching. *Robot Comput-Integr Manuf* 85:102638
- Liu CY, Liang JJ, Li THS, Chang KC (2019) Motion imitation and augmentation system for a six degrees of freedom dual-arm robot. *IEEE Access* 7:153986–153998
- Liu H, Wang L (2018) Gesture recognition for human-robot collaboration: a review. *Int J Ind Ergon* 68:355–367
- Liu H, Wang L (2021) Collision-free human-robot collaboration based on context awareness. *Robot Comput-Integr Manuf* 67:101997
- Liu, K., Z. Tang, D. Wang, Z. Wang, X. Li and B. Zhao (2025). Coherent: Collaboration of heterogeneous multi-robot system with large language models. 2025 IEEE International Conference on Robotics and Automation (ICRA), IEEE.
- Liu L, Guo F, Zou Z, Duffy VG (2024b) Application, development and future opportunities of collaborative robots (cobots) in manufacturing: a literature review. *Int J Hum Comput Interact* 40(4):915–932
- Liu R, Bai C, Lyu J, Sun S, Du Y, Li X (2025) Vlp: Vision-language preference learning for embodied manipulation. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.
- Liu S, Zhang J, Wang L, Gao RX (2024c) Vision AI-based human-robot collaborative assembly driven by autonomous robots. *CIRP Ann Manuf Technol* 73(1):13–16
- Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC (2016) Ssd: Single shot multibox detector. *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, Springer.
- Liu X, Zheng Y, Gao S, Song R, Gao R, Li Y (2024) A Trajectory Planning Method Based on Imitation Learning for Hydraulic Robotic Manipulators. 2024 IEEE International Conference on Robotics and Biomimetics (ROBIO).
- Liu Y, Hu B, Huang J, Tai YW, Tang CK (2023) Instance neural radiance field. *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- Liu Y, Wang L, Liu M (2021) Yolostereo3d: A step back to 2d for efficient stereo 3d detection. 2021 IEEE international conference on Robotics and automation (ICRA), IEEE.
- Liu Z, Chi C, Cousineau E, Kuppuswamy N, Burchfiel B, Song S (2024) Maniwav: Learning robot manipulation from in-the-wild audio-visual data. *arXiv preprint [arXiv:2406.19464](https://arxiv.org/abs/2406.19464)*.
- Liu, Z., H. Tang, A. Amini, X. Yang, H. Mao, D. L. Rus and S. Han (2023). Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. 2023 IEEE international conference on robotics and automation (ICRA), IEEE.
- Longhini A, Wang Y, Garcia-Camacho I, Blanco-Mulero D, Moletta M, Welle M, Alenyà G, Yin H, Erickson Z, Held D (2024) Unfolding the literature: a review of robotic cloth manipulation. *Annu Rev Control Robot Auton Syst*. <https://doi.org/10.1146/annurev-control-022723-033252>
- Lotsaris K, Fousekis N, Koukas S, Aivaliotis S, Kousi N, Michalos G, Makris S (2021a) Augmented reality (AR) based framework for supporting human workers in flexible manufacturing. *Procedia CIRP* 96:301–306
- Lotsaris K, Gkourmelos C, Fousekis N, Kousi N, Makris S (2021b) AR based robot programming using teaching by demonstration techniques. *Procedia CIRP* 97:459–463
- Loy WW, Donovan J, Rittenbruch M, Belek Fialho Teixeira M (2025) Exploring AR-enabled human–robot collaboration (HRC) system for exploratory collaborative assembly tasks. *Construction Robotics* 9(2):19
- Lu S, Berger J, Schilp J (2022) System of robot learning from multi-modal demonstration and natural language instruction. *Procedia CIRP* 107:914–919
- Lukács HI, Beregi BZ, Porteleki B, Fischl T, Botzheim J (2025) Attention U-Net-based semantic segmentation for welding line detection. *Sci Rep* 15(1):15276
- Luo H, Guo Z, Wu Z, Teng F, Li T (2024a) Transformer-based vision-language alignment for robot navigation and question answering. *Information Fusion* 108:102351
- Luo H, Wu J, Liu J, Antwi-Afari MF (2024b) Large language model-based code generation for the control of construction assembly robots: a hierarchical generation approach. *Dev Built Environ* 19:100488
- Luo H, Zeng Y, Yang L, Chen K, Shen Z, Lv F (2024c) VLAI: exploration and exploitation based on visual-language aligned information for robotic object goal navigation. *Image vis Comput* 151:105259
- Ma, X., S. Patidar, I. Haughton and S. James (2024). Hierarchical diffusion policy for kinematics-aware multi-task robotic manipulation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Malawade, A. V., T. Mortlock and M. A. A. Faruque (2022). EcoFusion: energy-aware adaptive sensor fusion for efficient autonomous vehicle perception. *Proceedings of the 59th ACM/IEEE Design Automation Conference*. San Francisco, California, Association for Computing Machinery: 481–486.

- Malawade, A. V., T. Mortlock and M. A. A. Faruque (2022). HydraFusion: Context-Aware Selective Sensor Fusion for Robust and Efficient Autonomous Vehicle Perception. 2022 ACM/IEEE 13th International Conference on Cyber-Physical Systems (ICCPs).
- Male J, Martinez-Hernandez U (2023) Deep learning based robot cognitive architecture for collaborative assembly tasks. *Robot Comput Integr Manuf* 83:102572
- Mandi, Z., S. Jain and S. Song (2024). RoCo: Dialectic Multi-Robot Collaboration with Large Language Models. 2024 IEEE International Conference on Robotics and Automation (ICRA).
- Manoharan, M., A. N. Shridhar, V. Y. Vinod and S. Kumaraguru (2020). "A Novel Volume Decomposition Methodology for Multi-Robots Collaborative Additive Manufacturing." 2020 IEEE 4th Conference on Information & Communication Technology (CICT): 1–6.
- Martini E, Boldo M, Bombieri N (2024) FLK: a filter with learned kinematics for real-time 3D human pose estimation. *Signal Process* 224:109598
- Matin A, Islam MR, Zhu Y, Wang X, Huo H, Xu G(2024) Hybrid Deep Learning for Assembly Action Recognition in Smart Manufacturing.
- Maxim V, Ligus J, Ligusova J, Kovaluk D, Sabol M (2023) Processing of collected data using an industrial 2D visual camera system. 2023 IEEE 21st World Symposium on Applied Machine Intelligence and Informatics (SAMI).
- Mehak S, Kelleher JD, Guilfoyle M, Leva MC (2024) Action recognition for human-robot teaming: exploring mutual performance monitoring possibilities. *Machines*. <https://doi.org/10.3390/machines12010045>
- Minz PS, Saini CS (2021) RGB camera-based image technique for color measurement of flavored milk. *Meas Food* 4:100012
- Mohamed A, Culverhouse PF, De Azambuja R, Cangelosi A, Yang C (2017) Automating active stereo vision calibration process with cobots. *IFAC-PapersOnLine* 50(2):163–168
- Mohsen S, Elkaseer A, Scholz SG (2021) Industry 4.0-oriented deep learning models for human activity recognition. *IEEE Access* 9:150508–150521
- Moutinho D, Rocha LF, Costa CM, Teixeira LF, Veiga G (2023) Deep learning-based human action recognition to leverage context awareness in collaborative assembly. *Robot Comput-Integr Manuf* 80:102449
- Mu Y, Zhang Q, Hu M, Wang W, Ding M, Jin J, Wang B, Dai J, Qiao Y, Luo P (2024) Embodiedgpt: vision-language pre-training via embodied chain of thought. *Adv Neural Inf Process Syst*. <https://doi.org/10.52202/075280-1090>
- Ng AY, Russell SJ (2000) Algorithms for Inverse Reinforcement Learning. *Proceedings of the Seventeenth International Conference on Machine Learning*, Morgan Kaufmann Publishers Inc. pp. 663–670.
- Nguyen VN, Groueix T, Ponimatkin G, Hu Y, Marlet R, Salzmann M, Lepetit V (2024) NOPE: Novel Object Pose Estimation from a Single Image. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
- Nguyen VN, Groueix T, Ponimatkin G, Lepetit V, Hodan T (2023) Cnos: a strong baseline for cad-based novel object segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- Nguyen VN, Groueix T, Salzmann M, Lepetit V (2024) Gigapose: Fast and robust novel object pose estimation via one correspondence. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Nguyen VN, Hu Y, Xiao Y, Salzmann M, Lepetit V (2022) Templates for 3D Object Pose Estimation Revisited: Generalization to New Objects and Robustness to Oclusions. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
- Nguyen VN, Tyree S, Guo A, Fourmy M, Gouda A, Lee T, Moon S, Son H, Ranftl L, Tremblay J, Brachmann E (2025) BOP Challenge 2024 on Model-Based and Model-Free 6D Object Pose Estimation. *arXiv preprint* [arXiv:2504.02812](https://arxiv.org/abs/2504.02812).
- Niedermaier T, Berens F, Reischl M, Elser S (2025) A novel metric for 6D pose estimation. *Addressing Errors and False Detections for More Reliable Evaluation* 73(2):125–135
- Niu B, Luo B, Cui Y, Xu X, Zhao Y, Feng Y (2025) HIAT: human-in-the-loop reinforcement learning with auxiliary task. *Artif Intell Rev* 58(9):273
- Örnek EP, Labbé Y, Tekin B, Ma L, Keskin C, Forster C, Hodan T (2023) Foundpose: unseen object pose estimation with foundation features. *arXiv preprint* [arXiv:2311.18809](https://arxiv.org/abs/2311.18809).
- Örnek EP, Labbé Y, Tekin B, Ma L, Keskin C, Forster C, Hodan T (2025). Foundpose: Unseen object pose estimation with foundation features. *European Conference on Computer Vision*, Springer.
- Padmakumar A, Thomason J, Shrivastava A, Lange P, Narayan-Chen A, Gella S, Piramuthu R, Tur G, Hakkani-Tur D (2022) TEACH: task-driven embodied agents that chat. *Proc AAAI Conf Artif Intell* 36(2):2017–2025
- Pagire V, Chavali M, Kale A (2025) A comprehensive review of object detection with traditional and deep learning methods. *Signal Process* 237:110075
- Palladin, E., R. Dietze, P. Narayanan, M. Bijelic and F. Heide (2025). SAMFusion: Sensor-Adaptive Multimodal Fusion for 3D Object Detection in Adverse Weather. *Computer Vision – ECCV 2024*, Cham, Springer Nature Switzerland.
- Pang J, Chen K, Shi J, Feng H, Ouyang W, Lin D (2019) Libra R-CNN: Towards Balanced Learning for Object Detection. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
- Pang J, Zheng P, Fan J, Liu T (2025) Towards cognition-augmented human-centric assembly: a visual computation perspective. *Robot Comput-Integr Manuf* 91:102852

- Park S, Wang X, Menassa CC, Kamat VR, Chai JY (2024) Natural language instructions for intuitive human interaction with robotic assistants in field construction work. *Autom Constr* 161:105345
- Pei G, Chen T, Jiang X, Liu H, Sun Z, Yao Y (2024) Videomac: Video masked autoencoders meet convnets. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Peng L, Chen Z, Fu Z, Liang P, Cheng E (2023) Bevsegformer: Bird's eye view semantic segmentation from arbitrary camera rigs. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*.
- Perrett T, Masullo A, Burghardt T, Mirmehdi M, Damen D (2021) Temporal-relational crosstransformers for few-shot action recognition. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.
- Pitteri G, Bugeau A, Ilic S, Lepetit V (2020) 3D Object Detection and Pose Estimation of Unseen Objects in Color Images with Local Surface Embeddings. *ArXiv abs/2010.04075*.
- Post AR, Nijveldt M, Ijtsma and D. Woods (2025). Determining what's next: visual analytics for evaluating human-automation
- Qi CR, Su K, Mo LJ Guibas (2017) Pointnet: Deep learning on point sets for 3d classification and segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*.
- Qin, Z., J. Wang and Y. Lu (2019). Monognet: A geometric reasoning network for monocular 3d object localization. *Proceedings of the AAAI conference on artificial intelligence*.
- Qu J, Zhang F, Wang Y, Fu Y (2019) Human-like coordination motion learning for a redundant dual-arm robot. *Robot Comput Integr Manuf* 57:379–390
- Rasheed M, Jasim WM, Farhan R (2024) Enhancing robotic grasping with attention mechanism and advanced UNet architectures in generative grasping convolutional neural networks. *Alex Eng J* 102:149–158
- Redmon J, Divvala S, Girshick R, Farhadi A (2016) You Only Look Once: Unified, Real-Time Object Detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
- Ren S, He K, Girshick R, Sun J (2016) Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans Pattern Anal Mach Intell* 39(6):1137–1149
- Rettenberger L, Münke FR, Bruch R, Reischl M (2023) Mask R-CNN outperforms U-Net in instance segmentation for overlapping cells. *Curr Dir Biomed Eng* 9(1):335–338
- Reynolds D (2009) Gaussian mixture models. In: Li SZ, Jain A (eds) *Encyclopedia of biometrics*. Springer US, Boston, MA, pp 659–663
- Rickert M, Foster ME, Giuliani M, By T, Panin G, Knoll A (2007) Integrating Language, Vision and Action for Human Robot Dialog Systems. *Universal Access in Human-Computer Interaction. Ambient Interaction*. Springer Berlin Heidelberg, Berlin, Heidelberg
- Rocamonde, J., V. Montesinos, E. Nava, E. Perez and D. Lindner (2024). Vision-language models are zero-shot reward models for reinforcement learning. *International Conference on Learning Representations*.
- Ronneberger, O., P. Fischer and T. Brox (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *ArXiv abs/1505.04597*.
- Rozenberszki DO Litany A (2022) Language-grounded indoor 3d semantic segmentation in the wild. *European Conference on Computer Vision*, Springer.
- Sabouri, M. H. (2026). Cybersecurity of teleoperated quadruped robots: a systematic survey of vulnerabilities, Threats, and Open Defense Gaps. *ArXiv abs/2602.23404*.
- Sadrifaridpour B, Wang Y (2018) Collaborative assembly in hybrid manufacturing cells: an integrated framework for human-robot interaction. *IEEE Trans Autom Sci Eng* 15(3):1178–1192
- Saikia A, Di Vece C, Bonilla S, He C, Magbagbeola M, Mennillo L, Czempiel T, Bano S, Stoyanov D (2025) Robotic arm platform for multi-view image acquisition and 3D reconstruction in minimally invasive surgery. *IEEE Robot Autom Lett*. <https://doi.org/10.1109/lra.2025.3540529>
- Salem M, Wael O, Attallah MM, Elkaseer A (2026) Tailoring patient-specific cranial implants for bone reconstruction via end-to-end deep learning image-to-print approach. *IEEE Access* 14:25677–25690
- Sampieri A, di Melendugno GM, Avogaro A, Cunico F, Setti F, Skenderi G, Cristani M, Galasso F (2022). Pose forecasting in industrial human-robot collaboration. *European Conference on Computer Vision*, Springer.
- Sapkota R, Flores-Calero M, Qureshi R, Badgujar C, Nepal U, Poulouse A, Zeno P, Vaddevolu UBP, Khan S, Shoman M, Yan H, Karkee M (2025) YOLO advances to its genesis: a decadal and comprehensive review of the You Only Look Once (YOLO) series. *Artif Intell Rev* 58(9):274
- Shafizadegan F, Naghsh-Nilchi AR, Shabaninia E (2024) Multimodal vision-based human action recognition using deep learning: a review. *Artif Intell Rev* 57(7):178
- Shah R, Doss ASA, Lakshmaiya N (2025) Advancements in AI-enhanced collaborative robotics: towards safer, smarter, and human-centric industrial automation. *Results Eng* 27:105704
- Shauri RLA, Saiki K, Toritani S, Nonami K (2012) Sensor integration and fusion for autonomous screwing task by dual-manipulator hand robot. *Procedia Eng* 41:1412–1420
- Shi, C., Y. Zhang, B. Yang, J. Tang, Y. Ma and S. Yang (2025). Part2Object: Hierarchical Unsupervised 3D Instance Segmentation. *Computer Vision – ECCV 2024*, Cham, Springer Nature Switzerland.

- Shi, S., C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang and H. Li (2020). Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.
- Shi S, Jiang L, Deng J, Wang Z, Guo C, Shi J, Wang X, Li H (2023) PV-RCNN++: point-voxel feature set abstraction with local vector representation for 3D object detection. *Int J Comput vis* 131(2):531–551
- Shi S, Wang X, Li H (2019) Pointtrcn: 3d object proposal generation and detection from point cloud. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.
- Shi X, Ye Q, Chen X, Chen C, Chen Z, Kim TK (2021) Geometry-based distance decomposition for monocular 3d object detection. *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- Shirai K, Beltran-Hernandez CC, Hamaya M, Hashimoto A, Tanaka S, Kawaharazuka K, Tanaka K, Ushiku Y, Mori S (2024) Vision-Language Interpreter for Robot Task Planning. 2024 IEEE International Conference on Robotics and Automation (ICRA).
- Silberman N, Hoiem D, Kohli P, Fergus R (2012) Indoor Segmentation and Support Inference from RGBD Images. *Computer Vision – ECCV 2012, Berlin, Heidelberg, Springer Berlin Heidelberg*.
- Simonelli A, Buló SR, Porzi L, López-Antequera M (2019). Disentangling monocular 3d object detection. *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- Singh R, Liu J, Van Wyk K, Chao YW, Lafleche JF, Shkurti F, Ratliff N, Handa A (2024) Synthetica: Large Scale Synthetic Data for Robot Perception. *arXiv preprint [arXiv:2410.21153](https://arxiv.org/abs/2410.21153)*.
- Sobrin-Hidalgo, D., M. Á. González-Santamarta, Á. M. Guerrero-Higueras, F. J. Rodríguez-Lera and V. Matellán-Olivera (2024). "Enhancing Robot Explanation Capabilities through Vision-Language Models: a Preliminary Study by Interpreting Visual Inputs for Improved Human-Robot Interaction." *arXiv preprint [arXiv:2404.09705](https://arxiv.org/abs/2404.09705)*.
- Soliman MM, Ahmed E, Darwish A, Hassanien AE (2024) Artificial intelligence powered metaverse: analysis, challenges and future perspectives. *Artif Intell Rev* 57(2):36
- Song, S., S. P. Lichtenberg and J. Xiao (2015). SUN RGB-D: A RGB-D scene understanding benchmark suite. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
- Song, Y. and L. Wang (2024). "BiCo-Fusion: Bidirectional Complementary LiDAR-Camera Fusion for Semantic-and Spatial-Aware 3D Object Detection." *arXiv preprint [arXiv:2406.19048](https://arxiv.org/abs/2406.19048)*.
- Srinivasan M, Chakrabarty A, Quirynen R, Yoshikawa N, Mariyama T, Cairano SD (2021) Fast multi-robot motion planning via imitation learning of mixed-integer programs. *IFAC-PapersOnLine* 54(20):598–604
- Su X, Wang Y, Mao J, Chen Y, Yin A, Zhao B, Zhang H, Liu M (2022) A review of pharmaceutical robot based on hyperspectral technology. *J Intell Robot Syst* 105(4):75
- Subašić, S., L. Banjanović-Mehmedović, H. Subašić, I. Karabegović and E. Husak (2025). Vision Transformer-Based Multimodal Fusion of Gesture and Object Classification in Human-Robot Collaboration. *Advances in Service and Industrial Robotics, Cham, Springer Nature Switzerland*.
- Sun, X., J. Wu, X. Zhang, Z. Zhang, C. Zhang, T. Xue, J. B. Tenenbaum and W. T. Freeman (2018). Pix3D: Dataset and Methods for Single-Image 3D Shape Modeling. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.
- Sun, Y., B. Sun, J. Tian and X. Zhang (2026). Research progress on quantum neural networks and quantum machine learning.
- Sun Z, Zhou Y, Tang S, Luo J, Zhao B (2024) Noise suppression zeroing neural network for online solving the time-varying inverse kinematics problem of four-wheel mobile manipulators with external disturbances. *Artif Intell Rev* 57(8):211
- Suzuki K, Kanamura M, Suga Y, Mori H, Ogata T (2021) In-air Knotting of Rope using Dual-Arm Robot based on Deep Learning. 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).
- Tang Y, Dong P, Tang Z, Chu X, Liang J (2024) Vmrnn: Integrating vision mamba and lstm for efficient and accurate spatiotemporal forecasting. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Tang Z, Chen G, Wang R, Miao Z, Dai M, Ma Y, Liao X (2024b) CD3IS: cross dimensional 3D instance segmentation network for production workshop. *J Intell Manuf* 35(7):3273–3289
- Tannous M, Miraglia M, Inglese F, Giorgini L, Ricciardi F, Pelliccia R, Milazzo M, Stefanini C (2020) Haptic-based touch detection for collaborative robots in welding applications. *Robot Comput Integr Manuf* 64:101952
- Terreran, M., L. Barcellona, D. Evangelista and S. Ghidoni (2021) Multi-view Human Parsing for Human-Robot Collaboration. 2021 20th International Conference on Advanced Robotics (ICAR).
- Terven J, D.-M. Córdova-Esparza D-M, J.-A. Romero-González J-A (2023) A comprehensive review of yolo architectures in computer vision: from yolov1 to yolov8 and yolo-nas. *Mach Learn Knowl Extr* 5(4):1680–1716
- Timofeev AV (2001) Neural control, multi-agent navigation and virtual reality of robots. *IFAC Proceed Vol* 34(6):1543–1548
- Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H (2021) Training data-efficient image transformers & distillation through attention. *International conference on machine learning, PMLR*.
- Vicentini F (2020) Collaborative robotics: a survey. *J Mech Des*. <https://doi.org/10.1115/1.4046238>
- Vu, T., K. Kim, T. M. Luu, T. Nguyen and C. D. Yoo (2022). Softgroup for 3d instance segmentation on point clouds. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Wächter M, Ovchinnikova E, Wittenbeck V, Kaiser P, Szedmak S, Mustafa W, Kraft D, Krüger N, Piater J, Asfour T (2018) Integrating multi-purpose natural language understanding, robot's memory, and symbolic planning for task execution in humanoid robots. *Robot Auton Syst* 99:148–165

- Wang B, Song C, Li X, Zhou H, Yang H, Wang L (2025a) A deep learning-enabled visual-inertial fusion method for human pose estimation in occluded human-robot collaborative assembly scenarios. *Robot Comput Integr Manuf* 93:102906
- Wang B, Yu L, Zhang B (2024a) AL-MobileNet: a novel model for 2D gesture recognition in intelligent cockpit based on multi-modal data. *Artif Intell Rev* 57(10):282
- Wang, G., F. Manhardt, F. Tombari and X. Ji (2021). Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Wang G, Zhang C, Liu S, Zhao Y, Zhang Y, Wang L (2025b) Multi-robot collaborative manufacturing driven by digital twins: advancements, challenges, and future directions. *J Manuf Syst* 82:333–361
- Wang K, Cheng L, Chen W, Zhang P, Lin L, Zhou F, Li G (2025c) MarvelOVD: marrying object recognition and vision-language models for robust open-vocabulary object detection. *Computer vision – ECCV 2024*. Springer Nature Switzerland, Cham
- Wang R, Zhao D, Yuan Z, Shao T, Chen G, Kao D, Hong S, Min B-C (2026) Print: preference-based reinforcement learning with multimodal feedback and trajectory synthesis from foundation models. *Adv Neural Inf Process Syst* 38:153095–153135
- Wang T, Fan J, Zheng P (2024b) An LLM-based vision and language cobot navigation approach for human-centric smart manufacturing. *J Manuf Syst* 75:299–305
- Wang T, Liu Z, Wang L, Li M, Wang XV (2024c) Data-efficient multimodal human action recognition for proactive human–robot collaborative assembly: a cross-domain few-shot learning approach. *Robot Comput Integr Manuf* 89:102785
- Wang, X., T. Kong, C. Shen, Y. Jiang and L. Li (2020). Solo: Segmenting objects by locations. *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII* 16, Springer.
- Wang, Y. and E. Johns (2025) One-Shot Dual-Arm Imitation Learning.
- Wang, Y., Z. Sun, J. Zhang, Z. Xian, E. Biyik, D. Held and Z. Erickson (2024). RL-VLM-F: Reinforcement Learning from Vision Language Foundation Model Feedback. In: S. Ruslan, K. Zico, H. Katherine et al. *Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, PMLR*. 235: 51484–51501.
- Wang Y, Zhou P, Geng G, An L, Zhang Q (2024e) End-to-end point cloud registration with transformer. *Artif Intell Rev* 58(1):26
- Wang Z, Yan J, Yan G, Yu B (2025d) Multi-scale control and action recognition based human-robot collaboration framework facing new generation intelligent manufacturing. *Robot Comput Integr Manuf* 91:102847
- Wang Z, Zhou Z, Song J, Huang Y, Shu Z, Ma L (2025e) VLATest: testing and evaluating vision-language-action models for robotic manipulation. *Proc ACM Softw Eng*. <https://doi.org/10.1145/3729343>
- Wen C, Zhang L, Ren J, Hong R, Li C, Yang C, Lv Y, Chen H (2025) HFA-Net: hybrid feature-aware network for large-scale point cloud semantic segmentation. *Artif Intel Rev* 58(4):102
- Weng Y, Wang H, Zhou Q, Qin Y, Duan Y, Fan Q, Chen B, Su H, Guibas LJ (2021) Captra: Category-level pose tracking for rigid and articulated objects from point clouds. *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- Wilson AN, Gupta KA, Koduru BH, Kumar A, Jha A, Cenkeramaddi LR (2023) Recent advances in thermal imaging and its applications using machine learning: a review. *IEEE Sens J* 23(4):3395–3407
- Wong CY, Vergez L, Suleiman W (2024) Vision- and tactile-based continuous multimodal intention and attention recognition for safer physical human-robot interaction. *IEEE Trans Autom Sci Eng* 21(3):3205–3215
- Woods DD (2018) The theory of graceful extensibility: basic rules that govern adaptive systems. *Environ Syst Decis* 38(4):433–457
- Woods DD, Alderson DL (2021) Progress toward resilient infrastructures: are we falling behind the pace of events and changing threats? *J Crit Infrastruct Policy* 2(2):5–18
- Wu Y, Yuan W, Qi A, Guizilini V, Mao J, Wang Y (2026) Large reward models: Generalizable online robot reward generation with vision-language models. *arXiv preprint* [arXiv:2603.16065](https://arxiv.org/abs/2603.16065).
- Xia W, Zheng H, Xu W, Xu X (2025) Large vision-language models enabled novel objects 6D pose estimation for human-robot collaboration. *Robot Comput Integr Manuf* 95:103030
- Xian, Z., N. Gkanatsios, T. Gervet, T.-W. Ke and K. Fragkiadaki (2023). ChainedDiffuser: Unifying Trajectory Diffusion and Keypose Prediction for Robotic Manipulation. *Proceedings of The 7th Conference on Robot Learning*. T. Jie, T. Marc and D. Kourosh. *Proceedings of Machine Learning Research, PMLR*. 229: 2323–2339.
- Xiang X, Zhang J (2023) FusionViT: Hierarchical 3D Object Detection via LiDAR-Camera Vision Transformer Fusion. *arXiv preprint* [arXiv:2311.03620](https://arxiv.org/abs/2311.03620).
- Xiang Y, Schmidt T, Narayanan V, Fox D (2017) Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint* [arXiv:1711.00199](https://arxiv.org/abs/1711.00199).
- Xiao X, Tie J, Zhao J, Wang Z, Li Y, Dou Q, Wang L (2025) "Spike-Kal: A Spiking Neuron Network Assisted Kalman Filter." *arXiv preprint* [arXiv:2504.12703](https://arxiv.org/abs/2504.12703).
- Xu C, Nguyen TK, Dixon E, Rodriguez C, Miller P, Lee R, Shah P, Ambrus R, Nishimura H, Itkina M (2025) Can we detect failures without failure data? uncertainty-aware runtime failure detection for imitation learning policies. *arXiv preprint* [arXiv:2503.08558](https://arxiv.org/abs/2503.08558).

- Yahyaoui MA, Oujabour M, Letaifa LB, Bohi A (2025) Multi-face emotion detection for effective Human-Robot Interaction. arXiv preprint [arXiv:2501.07213](https://arxiv.org/abs/2501.07213).
- Yan, J., Y. Liu, J. Sun, F. Jia, S. Li, T. Wang and X. Zhang (2023). Cross modal transformer: Towards fast and robust 3d object detection. Proceedings of the IEEE/CVF International Conference on Computer Vision.
- Yan S, Tao X, Xu D (2021) High-precision robotic assembly system using three-dimensional vision. *Int J Adv Robot Syst* 18(3):17298814211027029
- Yan X, Gilani SZ, Feng M, Zhang L, Qin H, Mian A (2020) Self-supervised learning to detect key frames in videos. *Sensors*. <https://doi.org/10.3390/s20236941>
- Yang C, Chen Y, Tian H, Tao C, Zhu X, Zhang Z, Huang G, Li H, Qiao Y, Lu L, Zhou J (2023) Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
- Yang G, Wang Y, Shi D, Wang Y (2025) Golden Cudgel Network for Real-Time Semantic Segmentation. Proceedings of the Computer Vision and Pattern Recognition Conference.
- Yang S, Xu W, Liu Z, Zhou Z, Pham DT (2018) Multi-source vision perception for human-robot collaboration in manufacturing. 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC).
- Yang X, Zhou Z, Sørensen JH, Christensen CB, Ünalán M, Zhang X (2023b) Automation of SME production with a Cobot system powered by learning-based vision. *Robot Comput Integr Manuf* 83:102564
- Yao B, Yang B, Xu W, Ji Z, Zhou Z, Wang L (2024) Virtual data generation for human intention prediction based on digital modeling of human-robot collaboration. *Robot Comput Integr Manuf* 87:102714
- Yasar MS, Islam MM, Iqbal T (2024) PoseTron: Enabling Close-Proximity Human-Robot Collaboration Through Multi-human Motion Prediction. 2024 19th ACM/IEEE International Conference on Human-Robot Interaction (HRI).
- Ying JJ-C, Chen Y-H, Zhang J (2025) Few-shot learning-based human pose estimation model. *Inf Sci* 717:122320
- Yu X, Zhang X, Xu C, Ou L (2024) Human–robot collaborative interaction with human perception and action recognition. *Neurocomputing* 563:126827
- Yun H, Lo S, Diepenbrock CH, Bailey BN, Earles JM (2024) VisTA-SR: Improving the Accuracy and Resolution of Low-Cost Thermal Imaging Cameras for Agriculture. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
- Yun J-W, Na M, Hwang Y, Song J-B (2025) Similar assembly state discriminator for reinforcement learning-based robotic connector assembly. *Robot Comput Integr Manuf* 91:102842
- Zaki AMA, Carnevale M, Schlette C, Giberti H (2023) On the use of low-cost 3D stereo depth camera to drive robot trajectories in contact-based applications. *Int J Adv Manuf Technol* 128(9):3745–3759
- Zanuttigh P, Marin G, Dal Mutto C, Dominio F, Minto L, Cortelazzo GM et al (2016) Operating principles of time-of-flight depth cameras. In: Zanuttigh P, Marin G, Dal Mutto C (eds) Time-of-flight and structured light depth cameras: technology and applications. Springer International Publishing, Cham, pp 81–113
- Zare M, Kebria PM, Khosravi A, Nahavandi S (2024) A survey of imitation learning: algorithms, recent developments, and challenges. *IEEE Trans Cybern* 54(12):7173–7186
- Zenhari S, Möhring H-C, Torbati AV (2024) Comprehensive analysis of kinematic models based on the DH method and screw theory for a five-axis machine tool. *Procedia CIRP* 130:1745–1751
- Zhai S, Zhang Q, Zhang T, Huang F, Zhang H, Zhou M, Zhang S, Liu L, Lin S, Pang J (2025) A Vision-Language-Action-Critic Model for Robotic Real-World Reinforcement Learning. *ArXiv abs/2509.15937*.
- Zhang G, Luo Z, Cui K, Lu S (2021) Meta-DETR: Image-level few-shot object detection with inter-class correlation exploitation. arXiv preprint [arXiv:2103.11731](https://arxiv.org/abs/2103.11731).
- Zhang J, Liu H, Chang Q, Wang L, Gao RX (2020a) Recurrent neural network for motion trajectory prediction in human-robot collaborative assembly. *CIRP Ann* 69(1):9–12
- Zhang J, Luo Y, Anwar A, Sontakke SA, Lim JJ, Thomason J, Biyik E, Zhang J (2025) ReWiND: Language-Guided Rewards Teach Robot Policies without New Demonstrations. Proceedings of The 9th Conference on Robot Learning. L. Joseph, S. Shuran and P. Hae-Won. Proceedings of Machine Learning Research, PMLR. **305**: 460--488.
- Zhang K, Xu W, Yao B, Ji Z, Hu Y, Feng H (2020) Human motion recognition for industrial human-robot collaboration based on a novel skeleton descriptor. 2020 IEEE 16th International Conference on Automation Science and Engineering (CASE).
- Zhang X, Li M, Lim JH, Weng Y, Tay YWD, Pham H, Pham Q-C (2018) Large-scale 3D printing by a team of mobile robots. *Autom Constr* 95:98–106
- Zhang Y, Ding K, Hui J, Liu S, Guo W, Wang L (2024a) Skeleton-RGB integrated highly similar human action prediction in human–robot collaborative assembly. *Robot Comput-Integr Manuf* 86:102659
- Zhang Y, Li Y, Wang R, Lu J, Ma X, Qiu M (2020c) PSAC: proactive sequence-aware content caching via deep learning at the network edge. *IEEE Trans Netw Sci Eng* 7(4):2145–2154
- Zhang Y, Xiong C, Liu J, Ye X, Sun G (2023) Spatial-information guided adaptive context-aware network for efficient RGB-D semantic segmentation. *IEEE Sens J*. <https://doi.org/10.1109/jsen.2023.3304637>

- Zhang Y, Zhou Z, David P, Yue X, Xi Z, Gong B, Foroosh H (2020) Polarnet: An improved grid representation for online lidar point clouds semantic segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.
- Zhang Y, Zhu Z, Zheng W, Huang J, Huang G, Zhou J, Lu J (2022) Beverse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving. *arXiv preprint [arXiv:2205.09743](https://arxiv.org/abs/2205.09743)*.
- Zhang Z, Ji Y, Tang D, Chen J, Liu C (2024b) Enabling collaborative assembly between humans and robots using a digital twin system. *Robot Comput-Integr Manuf* 86:102691
- Zhao L, Tao W (2020) Jsnet: joint instance and semantic segmentation of 3d point clouds. *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Zhao Z, Zheng D, Chen Y, Luo J, Wang Y, Huang P, Yang C (2025) Integrating with multimodal information for enhancing robotic grasping with vision-language models. *IEEE Trans Autom Sci Eng* 22:13073–13086
- Zheng H, Xia W, Xu X (2025) A human-robot collaborative assembly framework with quality checking based on real-time dual-hand action segmentation. *Robot Comput Integr Manuf* 94:102976
- Zheng P, Li C, Fan J, Wang L (2024) A vision-language-guided and deep reinforcement learning-enabled approach for unstructured human-robot collaborative manufacturing task fulfilment. *CIRP Ann* 73(1):341–344
- Zhong H, Wang H, Wu Z, Zhang C, Zheng Y, Tang T (2021) A survey of LiDAR and camera fusion enhancement. *Procedia Comput Sci* 183:579–588
- Zhou H, Zhu X, Song X, Ma Y, Wang Z, Li H, Lin D (2020) Cylinder3D: An Effective 3D Framework for Driving-scene LiDAR Semantic Segmentation. *ArXiv abs/2008.01550*.
- Zhou Y, Gallego G, Rebecq H, Kneip L, Li H, Scaramuzza D (2018) Semi-dense 3D Reconstruction with a Stereo Event Camera. *Computer Vision – ECCV 2018*. Springer International Publishing, Cham
- Zhu X, Su W, Lu L, Li B, Wang X, Dai J (2020) Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint [arXiv:2010.04159](https://arxiv.org/abs/2010.04159)*.
- Ziebart BD, Maas AL, Bagnell JA, Dey AK (2008) Maximum entropy inverse reinforcement learning. *Proceedings of the 23rd national conference on Artificial intelligence - Volume 3*. Chicago, Illinois, AAAI Press: 1433–1438.
- Zitkovich, B., T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid, Q. Vuong, V. Vanhoucke, H. Tran, R. Soricut, A. Singh, J. Singh, P. Sermanet, P. R. Sanketi, G. Salazar, M. S. Ryoo, K. Reymann, K. Rao, K. Pertsch, I. Mordatch, H. Michalewski, Y. Lu, S. Levine, L. Lee, T.-W. E. Lee, I. Leal, Y. Kuang, D. Kalashnikov, R. Julian, N. J. Joshi, A. Irpan, B. Ichter, J. Hsu, A. Herzog, K. Hausman, K. Gopalakrishnan, C. Fu, P. Florence, C. Finn, K. A. Dubey, D. Driess, T. Ding, K. M. Choromanski, X. Chen, Y. Chebotar, J. Carbajal, N. Brown, A. Brohan, M. G. Arenas and K. Han (2023). RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. *Proceedings of The 7th Conference on Robot Learning*. T. Jie, T. Marc and D. Kourosh. *Proceedings of Machine Learning Research, PMLR*. **229**: 2165–2183.
- Zou J (2024) Simplified neural architecture for efficient human motion prediction in human-robot interaction. *Neurocomputing* 588:127683

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Mahmoud Salem¹  · Markus Reischl¹ · Veit Hagenmeyer¹ · Katja Nau¹ · Ahmed Elkaseer^{2,3}

✉ Mahmoud Salem
mahmoud.salem@kit.edu

¹ Institute for Automation and Applied Informatics (IAI), Karlsruhe Institute of Technology (KIT), 76344 Eggenstein-Leopoldshafen, Germany

² Faculty of Engineering, British University in Egypt (BUE), El Sherouk City, Suez Desert Road, Cairo 11837, Egypt

³ Department of Computer Science, Gulf University for Science and Technology, 32093 Hawally, Kuwait