

Robot Perception for Fully Automated Industrial Systems: A Review of Sensing Modalities, Learning Methods, and Open Challenges

Anirudha Gupta

Abstract—Complete automation of industrial processes, often described as “lights-out” manufacturing, requires robots that can sense, interpret, and act upon unstructured and changing environments without human supervision. Perception has therefore become the principal bottleneck separating today’s largely pre-programmed industrial robots from truly autonomous production cells. This paper reviews the state of the art in robot perception for industrial automation. We first survey the sensing modalities in use, including 2D machine vision, active and passive 3D sensing, force/torque and tactile sensing, and their fusion. We then examine the core perception tasks that enable autonomy: object detection and segmentation, 6D pose estimation, grasp detection and bin picking, visual inspection and anomaly detection, contact-rich assembly, perception for safe human–robot collaboration, calibration, and localization for mobile manipulators. We further discuss the learning paradigms that underpin recent progress, namely simulation-to-reality transfer, synthetic data, self-supervision, and vision–language foundation models, together with system integration through ROS 2 and digital twins. Finally, we identify open challenges, including transparent and reflective objects, deformable materials, data scarcity, uncertainty quantification, certification of learned components, and benchmark gaps, and outline research directions toward reliable, fully autonomous industrial robotic systems.

Index Terms—Robot perception, industrial robotics, machine vision, 3D sensing, pose estimation, grasp detection, anomaly detection, Industry 4.0, deep learning, complete automation.

I. INTRODUCTION

Industrial robots have been a cornerstone of manufacturing for more than five decades, yet the majority of deployed systems still operate in a narrowly defined manner: they repeat taught trajectories on parts that are presented at precisely known positions by fixtures, feeders, and conveyors. This paradigm is highly productive for high-volume, low-variety production, but it is expensive to reconfigure and fragile in the face of variation. The vision of Industry 4.0 [1], [2] calls for flexible, connected, and largely self-organizing production, in which robots adapt to new products, mixed batches, and unexpected situations with minimal human intervention. The International Federation of Robotics reports that annual installations of industrial robots have exceeded half a million units in recent years, with an operational stock of several million units worldwide [3]; the share of these robots that can genuinely perceive and adapt, however, remains comparatively small.

The gap between a programmed robot and an autonomous one is, to a large extent, a perception gap. A robot that must pick randomly oriented parts from a bin, inspect a surface

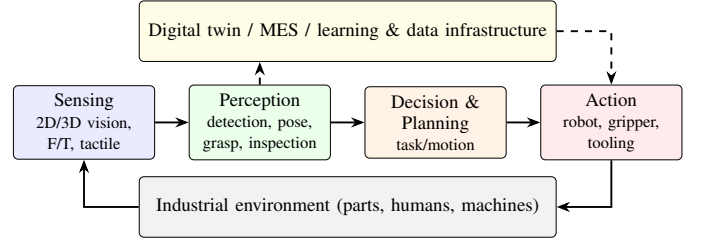


Fig. 1. The perception–action loop of an autonomous industrial robotic cell. Perception converts raw sensor data into task-relevant state estimates that drive planning and execution; feedback from the environment closes the loop.

for millimeter-scale defects, insert a connector with sub-millimeter tolerance, or share a workspace safely with people must build and continually update an internal model of its surroundings from noisy sensor data. Classical machine vision, based on hand-crafted features and carefully controlled lighting [4], [5], solved many constrained problems but generalizes poorly. Since 2012, deep learning [6], [7] has transformed what is achievable, while new commodity 3D sensors [8] and high-resolution tactile sensors [9], [10] have widened the range of signals a robot can exploit. More recently, large-scale foundation models [11], [12] and generalist robot policies [13]–[15] have begun to shift the field from task-specific perception modules toward broadly capable systems.

Several surveys cover individual parts of this landscape, for example grasp synthesis [16], [17], visual defect detection [18], human–robot collaboration [19], and the general limits of deep learning in robotics [20]. This review aims to give a single, connected picture of how perception enables *complete* automation in industrial settings. Its contributions are:

- a structured overview of industrial sensing modalities and their trade-offs (Section III);
- a task-oriented review of the perception capabilities required for autonomous industrial cells, from detection and pose estimation to inspection, assembly, and collaborative safety (Section IV);
- a discussion of the learning paradigms and data strategies that drive current progress (Section V) and of system-level integration (Section VI);
- a synthesis of open challenges and future research directions (Section VII).

Fig. 1 illustrates the perception–action loop around which this review is organized.

II. SCOPE AND REVIEW METHODOLOGY

This review focuses on perception for *industrial* manipulation and production tasks, including manufacturing, assembly, machine tending, inspection, and logistics (order picking and depalletizing). Literature was identified through IEEE Xplore, ScienceDirect, the ACM Digital Library, and arXiv using combinations of the terms *robot perception*, *industrial robot*, *machine vision*, *bin picking*, *6D pose estimation*, *grasp detection*, *anomaly detection*, *tactile sensing*, and *human-robot collaboration*. We prioritized (i) peer-reviewed journal and major-conference papers, (ii) works that introduced widely adopted methods, datasets, or benchmarks, and (iii) recent work (2018 onward) that reflects the transition to learning-based and foundation-model approaches. Foundational classical works are included where they remain in industrial use. Perception for autonomous driving and purely domestic service robots is outside the scope except where methods transfer directly to industrial settings.

III. SENSING MODALITIES

A. 2D Machine Vision

Monochrome and color area-scan and line-scan cameras remain the workhorse of industrial perception owing to their low cost, high resolution, and mature tooling. Early robot guidance relied on template matching, blob analysis, and local feature descriptors such as SIFT [4]; a comparative review of these techniques for industrial robot guidance is given in [5]. Accurate intrinsic calibration [21] and hand-eye calibration [22] are prerequisites for converting image measurements into robot coordinates. Closed-loop visual servoing [23], [24] uses image features directly as feedback, reducing sensitivity to calibration errors. The principal limitation of 2D vision is the loss of depth, which makes pose estimation of arbitrarily oriented parts ambiguous.

B. 3D Sensing

Three families of 3D sensors dominate industrial use. *Structured-light* systems project coded patterns and triangulate depth [25]; they achieve sub-millimeter accuracy at short range and are widely used in bin-picking cells, but are sensitive to ambient light and specular surfaces. *Time-of-flight* (ToF) cameras measure the phase or delay of modulated light [26]; they are compact and fast but suffer from multipath interference and lower resolution. *Active and passive stereo*, exemplified by the Intel RealSense family [8], offers an inexpensive compromise and has made RGB-D sensing ubiquitous in research. Laser line scanners and LiDAR complement these for large workpieces and mobile platforms. Point clouds from these sensors are processed with classical registration such as ICP [27] or with learned point-set networks such as PointNet and PointNet++ [28], [29].

C. Force/Torque and Tactile Sensing

Vision alone cannot resolve the contact state during insertion, polishing, or grasping of slippery parts. Six-axis

TABLE I
COMPARISON OF INDUSTRIAL SENSING MODALITIES

Modality	Strengths	Limitations
2D camera	Low cost, high resolution, mature tooling	No depth; lighting-sensitive
Structured light	Sub-mm accuracy at short range	Ambient light, specular and dark surfaces
Time-of-flight	Compact, high frame rate	Multipath, lower resolution
Stereo (active/passive)	Inexpensive, works outdoors (passive)	Texture dependence, depth noise at range
LiDAR / laser line	Long range, large parts	Cost, sparse data (LiDAR)
Force/torque	Direct contact forces	Global only; no spatial contact map
Tactile	Local contact geometry, slip	Wear, integration, limited coverage

force/torque (F/T) sensors at the wrist are standard for force-controlled assembly. Tactile sensing [30] provides spatially distributed contact information at the fingertip. Vision-based tactile sensors such as GelSight [9] and the low-cost DIGIT [10] image the deformation of an elastomer membrane, yielding high-resolution contact geometry and shear estimates, and are increasingly being evaluated for industrial grasp and assembly tasks.

D. Multimodal Fusion

Robust autonomy usually requires combining modalities: RGB for appearance, depth for geometry, and force/tactile for contact. Learned fusion architectures, for example DenseFusion for RGB-D pose estimation [31] and self-supervised vision-touch representations for contact-rich manipulation [32], consistently outperform single-modality baselines. Table I summarizes the trade-offs among modalities.

IV. PERCEPTION TASKS FOR COMPLETE AUTOMATION

A. Object Detection and Segmentation

Locating and identifying parts is the entry point of most perception pipelines. Two-stage detectors such as Faster R-CNN [33] and single-stage detectors such as YOLO [34] provide real-time bounding boxes, while Mask R-CNN [35] adds instance masks that are essential when parts overlap in clutter. These architectures, built on deep backbones [7], have largely replaced hand-crafted pipelines in new installations. Their weakness in industry is the need for labeled data for every new part; this has motivated promptable segmentation models such as the Segment Anything Model (SAM) [12] and open-vocabulary recognition built on vision-language pre-training such as CLIP [11], which can recognize or segment previously unseen objects with little or no task-specific annotation.

B. 6D Object Pose Estimation

Many industrial operations, such as machine tending, kitting, and assembly, require the full 6D pose (3D position and orientation) of a known part, typically available as a CAD

model. Classical approaches include template matching for texture-less objects (LINEMOD) [36] and point-pair features (PPF) [37], refined by ICP [27]. The texture-less, symmetric, and often metallic nature of industrial parts makes this problem difficult; the T-LESS dataset [38] was designed specifically to capture these properties, and the BOP benchmark [39] established a common evaluation protocol across datasets. Notably, early BOP results showed well-tuned PPF variants to be highly competitive with learned methods, a finding relevant to practitioners who value robustness.

Learning-based methods have since closed and surpassed this gap. PoseCNN [40] regresses pose directly from RGB; PVNet [41] votes for keypoints to achieve robustness to occlusion; DenseFusion [31] fuses color and depth per pixel; and DOPE [42] demonstrated that networks trained solely on synthetic data can achieve accuracy sufficient for robotic grasping. FoundationPose [43] represents a recent shift toward model-agnostic pose estimation and tracking of novel objects from a CAD model or a few reference images without object-specific training, an important step toward reducing the engineering effort required per new product variant.

C. Grasp Detection and Bin Picking

Random bin picking, i.e., picking parts directly from unstructured bins, is a canonical benchmark for industrial autonomy because it removes the need for part-specific feeders. Grasp synthesis approaches are commonly divided into analytic and data-driven methods [16]; [17] gives a review of vision-based grasping for parallel grippers. Deep learning for grasp detection was popularized by Lenz *et al.* [44] and by real-time CNN-based rectangle detection [45], [46]. GG-CNN [47] produces pixel-wise grasp quality maps fast enough for closed-loop reactive grasping.

Two influential strategies emerged for obtaining training data at scale. The Dex-Net line of work [48] trained grasp-quality networks on millions of synthetic point clouds labeled with analytic grasp metrics, and was later extended to ambidextrous policies that select between suction and parallel-jaw grippers, achieving high picking throughput on diverse objects [49]. In contrast, Levine *et al.* [50] collected hundreds of thousands of real grasp attempts across multiple robots, and QT-Opt [51] used large-scale off-policy reinforcement learning to achieve high success rates on unseen objects with emergent re-grasping behavior. Moving from planar to full 6-DoF grasps, GraspNet-1Billion [52] provides a large-scale benchmark, and Contact-GraspNet [53] generates 6-DoF grasps directly from scene point clouds in clutter.

The Amazon Picking Challenge exposed the gap between laboratory and warehouse conditions; analysis of the first challenge [54] highlighted perception as a key failure source, and the winning MIT-Princeton system [55] combined multi-affordance grasping with cross-domain image matching to pick novel objects without object-specific training. Table II summarizes representative grasping methods.

TABLE II
REPRESENTATIVE LEARNING-BASED GRASPING APPROACHES

Method	Input	Data source	Key idea
Lenz <i>et al.</i> [44]	RGB-D	Real, labeled	Two-stage cascaded detection
Dex-Net 2.0 [48]	Depth	Synthetic	Analytic labels at scale
Levine <i>et al.</i> [50]	RGB	Real, self-collected	Servoing via success prediction
QT-Opt [51]	RGB	Real, self-collected	Off-policy deep RL
GG-CNN [47]	Depth	Real, labeled	Pixel-wise, closed-loop
Contact-GraspNet [53]	Point cloud	Synthetic	6-DoF grasps in clutter

D. Visual Inspection and Anomaly Detection

Automated quality inspection is a prerequisite for closing the production loop without human checkers. A survey of visual defect detection in industry [18] distinguishes statistical, structural, filter-based, and learning-based approaches. A key practical difficulty is that defects are rare and diverse, so supervised training sets are hard to collect. Unsupervised anomaly detection, which learns only from defect-free samples, has therefore become the dominant paradigm, catalyzed by the MVTec AD dataset [56]. Methods based on pretrained feature embeddings, such as PaDiM [57] and PatchCore [58], model the distribution of normal patch features and achieve near-saturated image-level detection performance on MVTec AD, although localization of subtle and logical anomalies remains challenging. Mounting inspection cameras on robots enables viewpoint planning for complex parts and integration of inspection directly into manufacturing cells.

E. Contact-Rich Manipulation and Assembly

Assembly tasks such as peg-in-hole insertion, connector mating, and screw driving involve tolerances below the accuracy of typical vision-based pose estimates. Successful systems therefore combine vision for coarse alignment with force/torque or tactile feedback for fine alignment. Lee *et al.* [32] showed that a self-supervised, multimodal representation of vision and touch substantially improves sample efficiency of policy learning for peg insertion. High-resolution tactile sensors [9], [10] enable in-hand pose estimation and slip detection, which are needed when the grasped part shifts during transport.

F. Perception for Safe Human-Robot Collaboration

Complete automation does not imply the complete absence of people; maintenance, changeovers, and hybrid workflows still bring humans into robot cells. Safety requirements are defined by ISO 10218 [59] and, for collaborative operation, ISO/TS 15066 [60], which specifies speed-and-separation monitoring and power-and-force limiting. Speed-and-separation monitoring depends on reliable human detection and distance estimation. Villani *et al.* [19] survey safety, interfaces, and applications of industrial collaboration, and

Halme *et al.* [61] review vision-based safety systems. Human pose estimators such as OpenPose [62] enable fine-grained tracking of body parts and intention inference, but their use in safety functions is constrained by the need for certifiable, deterministic performance, which deep networks do not yet provide.

G. Calibration and Localization

Every perception output must ultimately be expressed in the robot frame. Camera calibration [21] and hand-eye calibration [22] remain foundational, and drift in these parameters is a frequent cause of failure in long-running cells, motivating on-line self-calibration [63]. For autonomous mobile manipulators performing machine tending or intralogistics, simultaneous localization and mapping (SLAM) is required; Cadena *et al.* [64] review the field, and ORB-SLAM2 [65] is a widely used visual SLAM baseline. The accuracy of mobile-base localization is typically insufficient for manipulation, so a local visual re-registration step against fiducials or workpiece features is commonly added at each workstation.

V. LEARNING PARADIGMS AND DATA STRATEGIES

A. Simulation-to-Reality Transfer and Synthetic Data

Labeled real data is the most expensive resource in industrial perception. Domain randomization [66] trains models on simulated images with randomized textures, lighting, and camera poses so that the real world appears as another variation. DOPE [42] and Dex-Net [48] demonstrate that purely synthetic training can produce deployable pose estimators and grasp planners. Because industrial parts usually come with CAD models, synthetic data generation is particularly attractive in manufacturing, and it integrates naturally with digital twins [67].

B. Self-Supervised and Large-Scale Robot Learning

Self-supervised approaches let robots label their own data, for example by recording grasp success [50], [51] or by predicting contact from vision [32]. This shifts the cost from human annotation to robot time, which can be scheduled during idle shifts in a factory.

C. Foundation Models and Generalist Policies

Since 2021, foundation models pretrained on internet-scale data have reshaped perception. CLIP [11] aligns images and text, enabling open-vocabulary recognition; SAM [12] provides zero-shot segmentation; and self-supervised visual features such as DINOv2 [68] transfer well to dense tasks. In robotics, RT-1 [13] showed that a transformer trained on a large real-world dataset can execute hundreds of instructions, RT-2 [14] transferred web-scale vision-language knowledge to robotic control, and the Open X-Embodiment collaboration [15] pooled data across many robot platforms to train cross-embodiment policies. Diffusion Policy [69] offers an expressive visuomotor policy class well suited to learning from demonstrations. These approaches blur the boundary between perception and control. Their industrial relevance is significant,

as they promise rapid onboarding of new tasks, but inference latency, precision, and verifiability currently limit deployment in high-throughput production.

VI. SYSTEM INTEGRATION

Perception modules deliver value only when integrated into reliable cells. The Robot Operating System (ROS) [70] became the de facto research middleware; ROS 2 [71], built on DDS, adds real-time, security, and multi-robot features that make it suitable for industrial deployment. Digital twins [67] provide a virtual replica of the cell for offline programming, synthetic data generation, and validation of perception changes before rollout. Edge computing with embedded GPUs allows deep models to run beside the robot with deterministic latency, while integration with manufacturing execution systems (MES) lets perception outputs, such as inspection results, feed traceability and process control, closing the loop envisioned by Industry 4.0 [2].

VII. OPEN CHALLENGES AND FUTURE DIRECTIONS

Despite rapid progress, several challenges stand between current research and reliable, fully autonomous industrial cells.

1) *Difficult materials*: Transparent, reflective, and dark objects corrupt the depth measurements of common 3D sensors. ClearGrasp [72] showed that learned depth completion can recover usable geometry for transparent objects, but metallic and shiny parts, which are ubiquitous in manufacturing, remain a persistent failure mode. Deformable objects such as cables, textiles, and food require state representations far richer than a 6D pose [73].

2) *Data efficiency and changeover*: High-mix production demands that a new part be handled within hours rather than weeks. Model-free pose estimation [43], open-vocabulary models [11], [12], and CAD-driven synthetic data [66] are promising, but few studies measure end-to-end changeover time as a metric.

3) *Reliability and uncertainty*: Industrial cells target availability well above that demonstrated in research. Perception systems should know when they are wrong. Bayesian and ensemble-based uncertainty estimation [74] and the broader agenda outlined by Sünderhauf *et al.* [20] remain underused in industrial perception pipelines, where uncertainty could trigger re-sensing, alternative grasps, or human assistance.

4) *Certification and safety*: Safety standards [59], [60] were written for deterministic systems. Methods for verifying and validating learned perception components, and for establishing their safety integrity, are an open research and standardization problem.

5) *Benchmark gap*: Benchmarks such as BOP [39], GraspNet-1Billion [52], and MVTec AD [56] have driven progress, but they rarely capture long-horizon operation, cycle-time constraints, sensor drift, or recovery from failure. Benchmarks that measure throughput, mean picks between failures, and changeover effort would better reflect industrial needs [54].

6) *Foundation models at the edge*: Vision–language–action models [14], [15] offer generalization but require substantial computation and do not yet reach the precision of specialized pipelines. Hybrid architectures, in which foundation models handle task understanding and novelty while specialized, verifiable modules handle precise localization and control, are a likely near-term path.

7) *Multimodal and active perception*: Few deployed systems exploit tactile sensing or actively choose viewpoints. Combining vision, force, and touch [30], [32] with active viewpoint selection could resolve occlusions and ambiguities that static sensing cannot.

VIII. CONCLUSION

Perception is the enabling technology for the transition from programmed industrial robots to fully autonomous production systems. This review traced the field from sensing hardware, through core perception tasks of detection, pose estimation, grasping, inspection, assembly, collaborative safety, and localization, to the learning paradigms and system-integration practices that support them. The evidence shows a clear trajectory: from hand-engineered, part-specific pipelines toward data-driven, increasingly general models trained with synthetic data, self-supervision, and internet-scale pretraining. Complete automation, however, demands more than accuracy on benchmarks. It demands reliability over months of operation, rapid adaptation to new products, robust handling of difficult materials, calibrated awareness of uncertainty, and verifiable safety. Addressing these requirements, through better multimodal sensing, uncertainty-aware and certifiable learning, industry-relevant benchmarks, and hybrid systems that combine foundation models with precise specialized modules, is the central research agenda for the coming years.

REFERENCES

- [1] H. Lasi, P. Fettke, H.-G. Kemper, T. Feld, and M. Hoffmann, “Industry 4.0,” *Business & Information Systems Engineering*, vol. 6, no. 4, pp. 239–242, 2014.
- [2] L. D. Xu, E. L. Xu, and L. Li, “Industry 4.0: State of the art and future trends,” *International Journal of Production Research*, vol. 56, no. 8, pp. 2941–2962, 2018.
- [3] International Federation of Robotics, “World robotics 2024 – industrial robots,” IFR Statistical Department, Frankfurt, Germany, 2024.
- [4] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [5] L. Pérez, Í. Rodríguez, N. Rodríguez, R. Usamentiaga, and D. F. García, “Robot guidance using machine vision techniques in industrial environments: A comparative review,” *Sensors*, vol. 16, no. 3, p. 335, 2016.
- [6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2012, pp. 1097–1105.
- [7] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [8] L. Keselman, J. I. Woodfill, A. Grunnet-Jepsen, and A. Bhowmik, “Intel RealSense stereoscopic depth cameras,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 1–10.
- [9] W. Yuan, S. Dong, and E. H. Adelson, “GelSight: High-resolution robot tactile sensors for estimating geometry and force,” *Sensors*, vol. 17, no. 12, p. 2762, 2017.
- [10] M. Lambeta, P.-W. Chou, S. Tian, B. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, G. Kammerer, D. Jayaraman, and R. Calandra, “DIGIT: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation,” *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 3838–3845, 2020.
- [11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning transferable visual models from natural language supervision,” in *Proc. Int. Conf. Machine Learning (ICML)*, 2021, pp. 8748–8763.
- [12] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, “Segment anything,” in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2023, pp. 4015–4026.
- [13] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar *et al.*, “RT-1: Robotics transformer for real-world control at scale,” in *Proc. Robotics: Science and Systems (RSS)*, 2023.
- [14] B. Zitkovich, T. Yu, S. Xu, P. Xu *et al.*, “RT-2: Vision-language-action models transfer web knowledge to robotic control,” in *Proc. Conf. Robot Learning (CoRL)*, 2023, pp. 2165–2183.
- [15] Open X-Embodiment Collaboration, “Open X-Embodiment: Robotic learning datasets and RT-X models,” in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2024, pp. 6892–6903.
- [16] J. Bohg, A. Morales, T. Asfour, and D. Kragic, “Data-driven grasp synthesis—a survey,” *IEEE Transactions on Robotics*, vol. 30, no. 2, pp. 289–309, 2014.
- [17] G. Du, K. Wang, S. Lian, and K. Zhao, “Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: A review,” *Artificial Intelligence Review*, vol. 54, no. 3, pp. 1677–1734, 2021.
- [18] T. Czimmermann, G. Ciuti, M. Milazzo, M. Chiurazzi, S. Roccella, C. M. Oddo, and P. Dario, “Visual-based defect detection and classification approaches for industrial applications—a survey,” *Sensors*, vol. 20, no. 5, p. 1459, 2020.
- [19] V. Villani, F. Pini, F. Leali, and C. Secchi, “Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications,” *Mechatronics*, vol. 55, pp. 248–266, 2018.
- [20] N. Sünderhauf, O. Brock, W. Scheirer, R. Hadsell, D. Fox, J. Leitner, B. Upcroft, P. Abbeel, W. Burgard, M. Milford, and P. Corke, “The limits and potentials of deep learning for robotics,” *International Journal of Robotics Research*, vol. 37, no. 4–5, pp. 405–420, 2018.
- [21] Z. Zhang, “A flexible new technique for camera calibration,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330–1334, 2000.
- [22] R. Y. Tsai and R. K. Lenz, “A new technique for fully autonomous and efficient 3D robotics hand/eye calibration,” *IEEE Transactions on Robotics and Automation*, vol. 5, no. 3, pp. 345–358, 1989.
- [23] S. Hutchinson, G. D. Hager, and P. I. Corke, “A tutorial on visual servo control,” *IEEE Transactions on Robotics and Automation*, vol. 12, no. 5, pp. 651–670, 1996.
- [24] F. Chaumette and S. Hutchinson, “Visual servo control. I. Basic approaches,” *IEEE Robotics & Automation Magazine*, vol. 13, no. 4, pp. 82–90, 2006.
- [25] J. Geng, “Structured-light 3D surface imaging: A tutorial,” *Advances in Optics and Photonics*, vol. 3, no. 2, pp. 128–160, 2011.
- [26] S. Foix, G. Alenyà, and C. Torras, “Lock-in time-of-flight (ToF) cameras: A survey,” *IEEE Sensors Journal*, vol. 11, no. 9, pp. 1917–1926, 2011.
- [27] P. J. Besl and N. D. McKay, “A method for registration of 3-D shapes,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, 1992.
- [28] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, “PointNet: Deep learning on point sets for 3D classification and segmentation,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 652–660.
- [29] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, “PointNet++: Deep hierarchical feature learning on point sets in a metric space,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
- [30] R. S. Dahiya, G. Metta, M. Valle, and G. Sandini, “Tactile sensing—from humans to humanoid,” *IEEE Transactions on Robotics*, vol. 26, no. 1, pp. 1–20, 2010.
- [31] C. Wang, D. Xu, Y. Zhu, R. Martín-Martín, C. Lu, L. Fei-Fei, and S. Savarese, “DenseFusion: 6D object pose estimation by iterative dense fusion,” in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3343–3352.

- [32] M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2019, pp. 8943–8950.
- [33] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.
- [34] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.
- [35] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017, pp. 2961–2969.
- [36] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab, "Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes," in *Proc. Asian Conf. Computer Vision (ACCV)*, 2012, pp. 548–562.
- [37] B. Drost, M. Ulrich, N. Navab, and S. Ilic, "Model globally, match locally: Efficient and robust 3D object recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 998–1005.
- [38] T. Hodaň, P. Haluza, Š. Obdržálek, J. Matas, M. Lourakis, and X. Zabulis, "T-LESS: An RGB-D dataset for 6D pose estimation of texture-less objects," in *Proc. IEEE Winter Conf. Applications of Computer Vision (WACV)*, 2017, pp. 880–888.
- [39] T. Hodaň, F. Michel, E. Brachmann, W. Kehl *et al.*, "BOP: Benchmark for 6D object pose estimation," in *Proc. European Conf. Computer Vision (ECCV)*, 2018, pp. 19–34.
- [40] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes," in *Proc. Robotics: Science and Systems (RSS)*, 2018.
- [41] S. Peng, Y. Liu, Q. Huang, X. Zhou, and H. Bao, "PVNet: Pixel-wise voting network for 6DoF pose estimation," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4561–4570.
- [42] J. Tremblay, T. To, B. Sundaralingam, Y. Xiang, D. Fox, and S. Birchfield, "Deep object pose estimation for semantic robotic grasping of household objects," in *Proc. Conf. Robot Learning (CoRL)*, 2018, pp. 306–316.
- [43] B. Wen, W. Yang, J. Kautz, and S. Birchfield, "FoundationPose: Unified 6D pose estimation and tracking of novel objects," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 17 868–17 879.
- [44] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *International Journal of Robotics Research*, vol. 34, no. 4–5, pp. 705–724, 2015.
- [45] J. Redmon and A. Angelova, "Real-time grasp detection using convolutional neural networks," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2015, pp. 1316–1322.
- [46] S. Kumra and C. Kanan, "Robotic grasp detection using deep convolutional neural networks," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, 2017, pp. 769–776.
- [47] D. Morrison, P. Corke, and J. Leitner, "Learning robust, real-time, reactive robotic grasping," *International Journal of Robotics Research*, vol. 39, no. 2–3, pp. 183–201, 2020.
- [48] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. Aparicio Ojea, and K. Goldberg, "Dex-Net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," in *Proc. Robotics: Science and Systems (RSS)*, 2017.
- [49] J. Mahler, M. Matl, V. Satish, M. Danielczuk, B. DeRose, S. McKinley, and K. Goldberg, "Learning ambidextrous robot grasping policies," *Science Robotics*, vol. 4, no. 26, p. eaau4984, 2019.
- [50] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen, "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection," *International Journal of Robotics Research*, vol. 37, no. 4–5, pp. 421–436, 2018.
- [51] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, and S. Levine, "Scalable deep reinforcement learning for vision-based robotic manipulation," in *Proc. Conf. Robot Learning (CoRL)*, 2018, pp. 651–673.
- [52] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "GraspNet-1Billion: A large-scale benchmark for general object grasping," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 444–11 453.
- [53] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox, "Contact-GraspNet: Efficient 6-DoF grasp generation in cluttered scenes," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2021, pp. 13 438–13 444.
- [54] N. Correll, K. E. Bekris, D. Berenson, O. Brock, A. Causo, K. Hauser, K. Okada, A. Rodriguez, J. M. Romano, and P. R. Wurman, "Analysis and observations from the first Amazon picking challenge," *IEEE Transactions on Automation Science and Engineering*, vol. 15, no. 1, pp. 172–188, 2018.
- [55] A. Zeng, S. Song, K.-T. Yu, E. Donlon, F. R. Hogan *et al.*, "Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2018, pp. 3750–3757.
- [56] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "MVTec AD—a comprehensive real-world dataset for unsupervised anomaly detection," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9592–9600.
- [57] T. Defard, A. Setkov, A. Loesch, and R. Audigier, "PaDiM: A patch distribution modeling framework for anomaly detection and localization," in *Proc. Int. Conf. Pattern Recognition (ICPR) Workshops*, 2021, pp. 475–489.
- [58] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 14 318–14 328.
- [59] International Organization for Standardization, "ISO 10218-1:2011 — robots and robotic devices — safety requirements for industrial robots — part 1: Robots," ISO, Geneva, Switzerland, 2011.
- [60] —, "ISO/TS 15066:2016 — robots and robotic devices — collaborative robots," ISO, Geneva, Switzerland, 2016.
- [61] R.-J. Halme, M. Lanz, J. Kämäräinen, R. Pieters, J. Latokartano, and A. Hietanen, "Review of vision-based safety systems for human-robot collaboration," *Procedia CIRP*, vol. 72, pp. 111–116, 2018.
- [62] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Real-time multi-person 2D pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172–186, 2021.
- [63] V. S. R. Karamreddy and L. Mitchell, "Validation and exploration of multimodal deep-learning camera–lidar calibration models," *Automotive Technical Papers* 381098, 2026.
- [64] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Transactions on Robotics*, vol. 32, no. 6, pp. 1309–1332, 2016.
- [65] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.
- [66] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, 2017, pp. 23–30.
- [67] F. Tao, H. Zhang, A. Liu, and A. Y. C. Nee, "Digital twin in industry: State-of-the-art," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 4, pp. 2405–2415, 2019.
- [68] M. Oquab, T. Darcet, T. Moutakanni *et al.*, "DINOv2: Learning robust visual features without supervision," *Transactions on Machine Learning Research*, 2024.
- [69] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," in *Proc. Robotics: Science and Systems (RSS)*, 2023.
- [70] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng, "ROS: An open-source robot operating system," in *Proc. ICRA Workshop on Open Source Software*, 2009.
- [71] S. Macenski, T. Foote, B. Gerkey, C. Lalancette, and W. Woodall, "Robot operating system 2: Design, architecture, and uses in the wild," *Science Robotics*, vol. 7, no. 66, p. eabm6074, 2022.
- [72] S. Sajjan, M. Moore, M. Pan, G. Nagaraja, J. Lee, A. Zeng, and S. Song, "ClearGrasp: 3D shape estimation of transparent objects for manipulation," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2020, pp. 3634–3642.
- [73] J. Sanchez, J.-A. Corrales, B.-C. Bouzgarrou, and Y. Mezouar, "Robotic manipulation and sensing of deformable objects in domestic and industrial applications: A survey," *International Journal of Robotics Research*, vol. 37, no. 7, pp. 688–716, 2018.

- [74] A. Kendall and Y. Gal, “What uncertainties do we need in Bayesian deep learning for computer vision?” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5574–5584.