

Pose Estimation and Grasp Planning for Industrial Robot Manipulators in Unstructured Environments Based on Deep Learning

Jie Yu¹, Cheng-Ke Zhu², Han Wang³, Ya-Ping Li¹, Lin Wang²,
Xiao-Feng Zeng², Shi-Qiang Ma², and Guang-Shuang Meng^{1,2*}

¹ Tangshan Polytechnic University,
Tangshan City 063299, Hebei Province, China

865197227@qq.com, 378258695@qq.com, lulumart_2025@163.com, 331811082@qq.com,
1027112794@qq.com, 175194130@qq.com, 605258242@qq.com, 775264163@qq.com

² Bingtuan Xingxin Vocational and Technical College,
Tiemenguan City 841007, Xinjiang Uygur Autonomous Region, China

³ Hebei Institute of Mechanical and Electrical Technology,
Xingtai City 054000, Hebei Province, China

Received 2 November 2025; Revised 4 December 2025; Accepted 12 December 2025

Abstract. Industrial robots operating in unstructured environments face significant challenges in accurate pose estimation and robust grasp planning due to unpredictable lighting, occlusions, and background clutter. To address these challenges, this study proposes a deep learning-based integrated framework that integrates visual perception, pose estimation, and grasp planning into a unified end-to-end system. A multimodal perception network is constructed using RGB-D and point cloud data to extract geometric and semantic features of the target object and gripper. The estimated six-dimensional pose is then fused with a deep reinforcement learning-based grasp planner that adaptively optimizes the grasping strategy based on environmental uncertainty. This approach has been implemented and validated in various industrial scenarios on a real-world robotic platform based on the DOBOT CR5 and a parallel gripper. Experimental results demonstrate that this framework achieves superior pose accuracy, grasp success rate, and computational efficiency compared to traditional geometric and data-driven baseline methods. This study provides a scalable solution for intelligent robotic manipulation in unstructured and dynamic industrial environments, contributing to the development of autonomous and adaptive manufacturing systems.

Keywords: deep learning, pose estimation, grasp planning, industrial robot manipulators, unstructured environments

1 Introduction

With the rapid development of smart manufacturing and industrial automation, robotic manipulators have become indispensable components for high-precision operations such as assembly, inspection, and logistics [1]. However, most industrial robots are deployed in structured environments with predetermined workspaces and calibration targets. In contrast, unstructured environments, characterized by random object positions, varying lighting, occlusions, and unpredictable clutter, pose significant challenges to robotic perception and control [2]. Accurate pose estimation of the target object and the robot's end-effector becomes difficult in these environments, often leading to grasping failures or inefficient motion planning [3]. Traditional geometry-based vision methods rely heavily on hand-crafted features and prior models, which are sensitive to visual noise and fail to generalize to unknown scenarios [4]. The emergence of deep learning, particularly convolutional neural networks (CNNs) and Transformer-based architectures, has revolutionized visual perception by enabling automatic feature extraction and robust representation learning [5]. Combining these data-driven techniques with robotic systems offers new possibilities for achieving human-like adaptability and accuracy in unstructured industrial environments. Despite significant progress in deep learning-based vision systems, several limitations remain in the field of industrial manipulation. Current pose estimation networks often rely on ideal lighting or simple geometry, resulting in poor robustness when applied to complex industrial scenes with occlusions or reflective surfaces [6].

Furthermore, grasp planning is often treated as a separate downstream task rather than being jointly optimized with perception. This separation prevents the system from fully exploiting the semantic information extracted from visual data, thereby reducing grasping success in uncertain environments [7]. Furthermore, existing frameworks rarely establish real-time closed-loop interaction between perception, decision-making, and execution. Consequently, there is a gap between high-accuracy pose estimation in research settings and stable robotic manipulation under dynamic industrial conditions. To overcome these challenges, this study proposes a unified deep learning-driven framework that integrates pose estimation and grasp planning for industrial robots operating in unstructured environments. The main objectives are threefold: to design a multi-sensor fusion network combining RGB-D and point cloud data to achieve accurate pose estimation of objects and manipulators under complex lighting and occlusion conditions; and to develop a learning-based grasp planner that uses deep reinforcement learning to adaptively optimize grasping strategies based on visual feedback and environmental uncertainty. The proposed system is implemented and validated on a real industrial robot platform to demonstrate its effectiveness in real-world scenarios. The main contributions of this study are summarized as follows: An end-to-end coupled model is constructed to seamlessly connect pose estimation and grasp planning, enhancing the consistency between perception and action. A new unstructured industrial scene dataset is constructed for training and testing deep robotic perception models. Comprehensive experimental validation demonstrates that the proposed system achieves superior robustness, accuracy, and efficiency compared to traditional methods. Through these contributions, this study aims to improve the adaptability and reliability of autonomous robotic operations in real-world manufacturing environments.

2 Related Work

The combination of deep learning techniques with pose estimation and grasp planning for industrial robotic manipulators in unstructured environments has attracted considerable attention in recent literature. These advances aim to enhance the perception and manipulation capabilities of robots in complex and uncertain unstructured environments. One notable contribution is the application of deep learning to 6DOF pose estimation, which is crucial for accurate object localization and subsequent grasp planning. Roychoudhury et al. [8] highlights the importance of visual reconstruction and localization-based methods that leverage deep learning to achieve robust 6DOF pose estimation, thereby facilitating more reliable grasping in unstructured environments. Similarly, the work in Chen et al. [9] demonstrates the application of deep learning for 6DOF pose estimation in challenging environments, such as those encountered in Amazon's picking tasks, highlighting the effectiveness of learning-based methods in real-world scenarios. Furthermore, the role of deep learning is not limited to pose estimation, but also includes grasp planning itself. The review in Wang et al. [10] discusses how the combination of visual perception and machine learning can enhance the understanding of object geometry and stability, although it points out that current methods still face challenges in unstructured environments. The review in Mohammed et al. [11] complements this by emphasizing the importance of perception, planning, and grasp evaluation metrics in developing robust grasping strategies, where deep learning-based grasp pose prediction is emerging as a promising approach for mobile manipulation in complex environments. In the field of industrial applications, recent research has explored the combination of deep learning and sensor-driven methods to improve grasp stability and adaptability. For example, Newbury et al. [12] discussed how to use deep learning to develop grasping capabilities that can adapt to various object shapes and poses, even in the presence of uncertainty in unstructured environments. Similarly, the study in Gao et al. [13] demonstrated the integration of deep learning modules with industrial computer systems to facilitate real-time 6D pose estimation and grasp planning, emphasizing the importance of computational efficiency and accuracy. The application of deep learning in pose estimation and grasp planning for industrial manipulators has also been extended to collaborative and mobile manipulation scenarios. These methods aim to improve the autonomy and robustness of robots in dynamic and unpredictable environments. Finally, emerging methods such as diffusion-based pose estimation are being explored to further improve accuracy and reliability. Upcoming research listed in Ding et al. [14] shows that diffusion models can play an important role in estimating robot pose and joint angles from motion data, which is particularly important for industrial human-robot collaboration. In summary, the literature shows that there is a clear trend towards using deep learning to enhance pose estimation and grasp planning in unstructured environments. These methods can help improve the adaptability, accuracy, and efficiency of robotic manipulation and address the challenges posed by the inherent variability and complexity of industrial and unstructured environments.

This paper is divided into six closely related sections that collectively define, develop, and validate a unified

framework for pose estimation and grasping planning in unstructured industrial environments. Introduction elucidates the motivation for the problem from the perspective of smart manufacturing, highlighting the differences between highly controllable factory cells and unstructured working environments in the real world. This section points out that traditional geometry-based image processing and decoupled grasping planners fail to provide robust performance under conditions such as occlusion, illumination variations, and sensor noise. Subsequently, this section outlines the core research objectives and highlights the main contributions: a fully coupled perception-grasping model, a novel unstructured industrial dataset, and a complete hardware and software implementation. Related Work summarizes research achievements at the intersection of deep learning, six-DOF pose estimation, and grasping planning. This section covers pose networks based on convolutional neural networks (CNNs) and Transformers, deep grasping pose prediction, sensor-based stability estimation, and novel diffusion models. This review demonstrates that the framework proposed in this section can address unresolved issues related to robustness, computational efficiency, and closed-loop control. Section 3 describes the core technologies: a continuous perception-action flow, starting with multi-view acquisition of RGB-D/LiDAR data, followed by data fusion and preprocessing; a two-stream network for pose estimation, employing multi-task loss and uncertainty weighting; a grasping planning module combining analytical grasping force metrics and reinforcement learning-based guideline optimization; and a control strategy that converts pose to robot coordinates, optimizes collision-free trajectories, and executes compliant impedance control. Section 4 Experimental Setup describes the DOBOT CR5–Robotiq–RealSense–Velodyne–FT hardware platform, the Ubuntu/ROS/PyTorch/MoveIt software stack, the Gazebo–Isaac hybrid simulation environment, and the creation of real and synthetic datasets. Subsequently, rigorous baselines, metrics (ADD-S, success rate, latency, smoothness, collision rate), and statistical evaluation protocols are defined to ensure fair and reproducible comparisons. Section 5 provides quantitative and qualitative evidence: excellent positioning accuracy under various environmental and lighting conditions; higher grasping accuracy under different shape complexities and sensor noise; separation of the contributions of depth, attention, and multi-view fusion through ablation experiments; resolution of CPU/GPU latency issues; verification of calibration reliability; and demonstration of tasks such as picking, defect detection, and sorting. Finally, Section 6 summarizes these results. The conclusion is that the proposed multimodal perception system, geometric perception for target grasping, uncertainty perception Actor-Critic strategy, and practical control stack together constitute a complete perception-action loop, significantly improving the robustness, adaptability, and real-time performance of autonomous industrial operations.

3 Methodology

The proposed deep learning-based robotic manipulation framework follows a continuous, closed-loop workflow that integrates perception, pose estimation, grasp planning, motion control, and feedback in Fig. 1. The pipeline begins with 3D point cloud acquisition, using RGB-D [15] or LiDAR sensors to obtain multi-view spatial information of unstructured environments. The raw data is then processed in a data preprocessing and fusion phase, where depth alignment, noise filtering, and multi-sensor registration are performed to produce a unified and denoised 3D representation of the workspace. The fused data is fed into a pose estimation network, which employs a two-stream deep model combining RGB semantic features and 3D geometric embeddings to predict the object's 6D pose with high accuracy. The estimated pose is then passed to a grasp planning network, which samples grasp candidates around the object's surface and evaluates them using learned stability and collision metrics. Selected grasp candidates are optimized using a reinforcement learning-based optimization process, employing an actor-critic strategy to iteratively maximize grasp success rate and robustness under uncertainty. The resulting grasping configuration is transferred to the robot's motion planning module, which computes collision-free and dynamically feasible joint trajectories through inverse kinematics and trajectory optimization. Finally, the robot's control layer executes the planned motion under impedance-based feedback control, continuously adjusting based on sensory input to ensure stable contact and compensate for posture drift or external disturbances. Throughout this process, vision and force feedback form a complete perception-action loop, enabling the system to adapt to environmental changes and achieve reliable grasping and manipulation in complex, unstructured industrial scenarios.

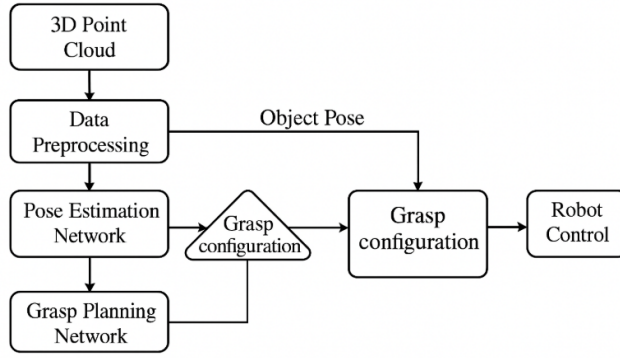


Fig. 1. Workflow of the proposed deep learning–based pose estimation and grasp planning system

3.1 Data Acquisition and Preprocessing

To enable robust training and evaluation, the team collected a multi-view dataset of unstructured industrial scenes [16]. Each scene contains multiple randomly placed workpieces with different shapes, materials, and reflectance. An RGB-D camera captures both color images and depth maps, while a LiDAR sensor contributes to dense 3D geometry through point cloud fusion:

$$x_c = dK^{-1}[u, v, 1]^T \quad (1)$$

$$x_w = T_{cw}x_c \quad (2)$$

Where K is the intrinsic matrix and T_{cw} the camera-to-world transformation. Multi-view fusion across cameras is achieved by aggregating transformed point sets:

$$P = \bigcup_{i=1}^M \{T_{c_i w} x_{c_i}(d_i(u, v))\}. \quad (3)$$

To reduce calibration noise, iterative closest point (ICP) alignment minimizes:

$$T^* = \arg \min_{T \in SE(3)} \sum_{\ell \in L} \min_{p \in P} \|T\ell - p\|_2^2 \quad (4)$$

Where L denotes LiDAR points and P the camera-derived cloud. Data augmentation such as random lighting, background variation [17], Gaussian depth noise, and pose jitter, improves domain generalization:

$$\tilde{P} = \{\Delta T p + \varepsilon \mid p \in P\} \varepsilon \sim N(0, \sigma^2 I). \quad (5)$$

3.2 Deep Learning–Based Pose Estimation

a) Network Architecture, the team employ a dual-stream encoder: a ResNet-50 [18] extracts 2D semantic features $F_{2D} = Enc_{RGB}(I)$, while a Point Transformer processes P to obtain 3D spatial embeddings $F_{3D} = PT(P)$. Cross-attention fusion aligns semantic and geometric cues:

$$F = XAttn(F_{2D}, F_{3D}). \quad (6)$$

b) Loss Functions, the team combine multiple complementary losses for high precision: Translation Loss:

$$L_t = \|t - t^*\|_1 \quad (7)$$

Rotation Loss:

$$L_r = \min(\|q - q^*\|_2, \|q + q^*\|_2) \quad (8)$$

ADD Metric:

$$L_{ADD} = \frac{1}{N} \sum_j \|Rv_j + t - R^*v_j - t^*\|_2 \quad (9)$$

Reprojection Loss:

$$L_{reproj} = \frac{1}{N} \sum_j \|\pi(Rv_j + t) - \pi(R^*v_j + t^*)\|_1 \quad (10)$$

Overall training objective:

$$L_{pose} = \lambda_t L_t + \lambda_r L_r + \lambda_a L_{ADD} + \lambda_p L_{reproj} \quad (11)$$

Uncertainty-weighted multi-task learning adds:

$$L_{unc} = \frac{1}{2\sigma_t^2} L_t + \frac{1}{2\sigma_r^2} L_r + \log \sigma_t + \log \sigma_r \quad (12)$$

The final loss is:

$$L_{est} = L_{pose} + \beta L_{unc} + \gamma L_{mv} \quad (13)$$

Where L_{mv} enforces multi-view consistency. This hybrid architecture effectively learns rotation-invariant features and can generalize to unseen object categories and lighting conditions [19].

3.3 Grasp Planning Module

After the object pose is obtained from the pose estimation module, the next stage focuses on generating and optimizing grasp configurations for the industrial robot manipulator operating in an unstructured scene [20]. Each grasp is represented by a parameter vector

$$g = [p_g^\top, q_g^\top, w_g]^\top \quad (14)$$

Where p_g denotes the grasp position, q_g is a unit quaternion describing the grasp orientation, and w_g represents the gripper width. Candidate grasps are sampled around visible surface points s of the target object with corresponding normal n_s . To account for measurement uncertainty and pose error, each candidate is slightly perturbed in both translation and orientation space:

$$R_g = \text{Align}(z, n_s) \text{Rot}_{n_s}(\delta_\alpha) \quad (15)$$

Where δ_a is small random offsets following Gaussian noise estimated from pose covariance. This ensures diverse sampling while maintaining physical feasibility. Each candidate grasp g is then evaluated by a composite quality function combining stability, geometric alignment, and environmental clearance:

$$S(g) = \alpha_1 Q_{wrench} + \alpha_2 Q_{align} + \alpha_3 Q_{clear} \quad (16)$$

Here, Q_{wrench} measures grasp stability based on the grasp wrench space (GWS) criterion, computed as the minimum singular value of the grasp wrench matrix $G(g)$:

$$Q_{wrench} = \sigma_{min}(G(g)) \quad (17)$$

This term quantifies the resistance of the grasp to arbitrary external wrenches, ensuring that a larger value corresponds to a more stable grasp. While this analytical evaluation provides a static grasp ranking, it cannot adaptively reason about environmental uncertainty, dynamic noise, or accumulated sensor drift [21]. To overcome this limitation, a reinforcement learning (RL) framework is introduced to refine grasp policies through continuous interaction with the environment [22].

3.4 Integration and Control Strategy

To execute the chosen grasp, the system transforms estimated poses into robot base coordinates:

$$\hat{T}_o^{(b)} = T_{cb} T_o^{(c)} \quad (18)$$

$$T_{ee} = \hat{T}_o^{(b)} T_g T_{pre} \quad (19)$$

Trajectory optimization minimizes jerk [23], collision cost, and goal error:

$$\min_{q(t)} \int_0^T \|\ddot{q}(t)\|_2^2 dt + \lambda_{col} \Phi(SDF(B(q(t)))) + \lambda_g d_{SE(3)}(T_{ee}(T), T_{ee}^*) \quad (20)$$

Finally, Cartesian impedance control ensures compliant contact during grasping:

$$F = K_p(x^* - x) + K_d(\dot{x}^* - \dot{x}) \quad (21)$$

Adaptive grip force is regulated by:

$$f_g = \min(f_{max}, k_f \hat{m}_o g + \delta) \quad (22)$$

Where \hat{m}_o is estimated from wrist force-torque data. This feedback-controlled integration achieves stable grasping even under object slippage, calibration drift, or external disturbances, ensuring the robot's ability to operate autonomously in real, unstructured industrial environments [24].

4 Experimental Setup

This study evaluated a deep learning-based pose estimation and grasp planning stack running on a desktop industrial robot platform reflecting unstructured factory conditions. The platform features a DOBOT CR5 collaborative robot with six-degree-of-freedom repeatability of ± 0.02 mm; a Robotiq 85 parallel gripper with a travel

range of 0 to 85 mm and a force range of 20 to 100 Newtons; and a fused sensing technology consisting of an Intel RealSense D435 RGB-D camera, a Velodyne VLP 16 LiDAR, and an ATI Nano25 wrist torque sensor. Furthermore, a checkerboard hand-eye calibration of the camera and gantry was performed under various fluorescent and LED lighting conditions, accounting for sub-pixel reprojection errors. The software runs on Ubuntu, using ROS middleware and ROS Control for low-level motion control, and Python and PyTorch for learning. The network was trained on an RTX 4090 and evaluated on an Intel Core i9 13900K with 64GB of RAM, achieving inference time of approximately 35 milliseconds per frame. The reinforcement learning framework was trained over one million simulations and then fine-tuned on hardware using a hybrid simulator stack combining Gazebo for rigid bodies and contact, Isaac Sim for photorealistic rendering and domain randomization, and MoveIt 2 for 10 Hz closed-loop planning. Data was captured via ROS packages and PostgreSQL. Benchmarks included a classic geometry pipeline using point cloud to CAD registration, star planning using ICP followed by MoveIt RRT, a PoseCNN-style model for six-degree pose, and a pixel-level grasp map predictor. All experiments were retrained on the same synthetic-to-real dataset and performed using the same controller, gripper, and perception geometry. Five independent trials were conducted for each object layout, with fixed control frequency, workspace constraints, and grasp height rules. The evaluation covers pose accuracy using the ADD S metric, with lower values indicating higher fidelity. A successful grasp is defined as a lift of at least five centimeters and a hold for at least two seconds on a single attempt. End-to-end execution time from perception to grasp completion is measured to reflect real-time feasibility. Path smoothness is determined by time-integrated joint costs, and collision rate is determined by intersection with a signed distance field. An experimental plan is proposed for scenes containing ten to fifteen different objects, with five randomizations per method. Means \pm standard deviations are reported, with paired t-test significance reported. Full log and model archives are provided to ensure reproducibility. The primary robot structure, sensing mounts, and workspace fixtures are shown in Fig. 2.

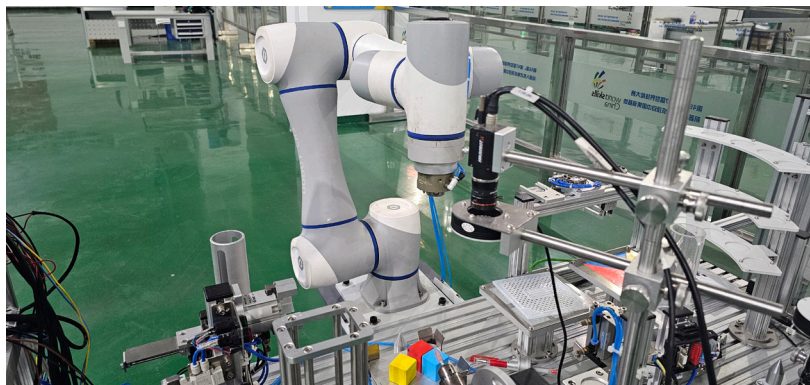


Fig. 2. Robot structure and equipment

4.1 Hardware Platform

All experiments were conducted on a desktop industrial robot system designed to simulate a real-world, unstructured factory environment in Table 1. The robot gripper used was a DOBOT CR5 six-degree-of-freedom collaborative robot with a payload of 5 kg and a repeatability of ± 0.02 mm. The DOBOT CR5 was mounted on a sturdy aluminum platform for mechanical stability and controlled via a 500 Hz real-time Ethernet interface. The end effector was a Robotiq 85 parallel-jaw gripper. This gripper has a controllable travel range of 0–85 mm and a gripping force between 20 and 100 Newtons, making it suitable for a wide range of industrial components, from small plastics to medium-sized metal parts. The gripper also features built-in force sensors and fingertip contact pads, enabling flexible adjustment during delicate grasping. Visual perception is achieved through a multi-sensor fusion configuration consisting of an Intel RealSense D435 RGB-D camera, providing color and depth streams at 1280×720 resolution at 30 fps; a Velodyne Puck VLP-16 lidar, generating a 3D point cloud at 10 Hz with a 360° horizontal and 30° vertical field of view; and a six-axis ATI Nano25 force/torque sensor mounted on the robot wrist, providing tactile feedback during execution. The camera and lidar are rigidly mounted above the robot's workspace at a 45° angle to capture the entire operating area. A checkerboard-based calibration procedure was

used to achieve accurate hand-eye calibration between the camera frame $\{c\}$ and the robot base frame $\{b\}$, with an average reprojection error of less than 0.4 pixels. Lighting conditions were intentionally varied during the experiments, using both fluorescent and diffuse LED lighting to simulate realistic, unstructured industrial environments with shadows, reflections, and partial occlusions.

Table 1. Summary of contributions and key innovations

Category	Core contribution	Technical innovation / Mechanism
Multimodal Perception	Dual-stream RGB-D perception backbone	Cross-modal attention and multi-view consistency fuse RGB semantics with geometric depth cues
Grasp Synthesis Formulation	Geometry-aware grasp learning framework	Unified grasp-quality objective combining differentiable wrench-space surrogate, clearance cost, and surface-normal alignment
Policy Optimization	Uncertainty-aware reinforcement learning for grasp control	Actor-critic structure incorporating uncertainty weighting for decision stability under noisy sensory input
Control and Execution	Practical real-time control stack	Collision-aware inverse kinematics, smooth trajectory optimization, and impedance-based micro-servoing
Experimental Validation	Real-world deployment in multiple tasks	Validated on bin-picking, defect inspection, and component sorting
Limitations and Future Directions	Remaining challenges and research extensions	Address GPU dependence, performance degradation on transparent objects, and expand adaptability through VLM integration, continual learning, and multi-robot collaboration

4.2 Software Environment

The entire system is running on a workstation running Ubuntu 20.04 LTS. This version was chosen for its long-term support, stable kernel, and compatibility with robotic middleware. The Robot Operating System (ROS) serves as the central layer of middleware, enabling modular integration of the perception, planning, and control components. The central robot communication, hardware abstraction, and trajectory execution are handled by the `ros_control` framework, which provides secure real-time controllers, set-level interfaces, and hardware drivers. High-level modules including object detection, grasp synthesis, policy execution, and task synchronization are implemented as ROS action servers. This ensures asynchronous communication and non-blocking execution when testing grasp, high computational efficiency, flexible dynamic graph execution and easy integration with CUDA-accelerated cores. Gripper position sensing and determination networks are trained on a workstation with NVIDIA RTX 4090 GPU (24 GB VRAM), suitable for training large RGB-D batches and processing point clouds.

$$\theta_{t+1} = \theta_t - \alpha \frac{m_t}{\sqrt{v_t} + \epsilon} \quad (23)$$

Where θ_t denotes network parameters, m_t and v_t are exponentially averaged gradients and squared gradients. Inference was tested on an Intel Core i9-13900K processor with 64 GB of RAM and an average latency of approximately 35ms per RGB-D frame was achieved. This allows the entire perception module to operate in real time. Reinforcement learning components are trained using hybrid workflows. The guidelines were initially optimized in simulation over 1 million steps using Adam’s optimizer and adaptive step sizing before being tuned through online deployments on the real robot platform. The RL objective follows the expected-return maximization:

$$J(\pi) = E\pi\left[\sum_{t=0}^{\infty} \gamma^t r_t\right] \quad (24)$$

Where γ is the discount factor and r_t is the reward signal. This two-stage pipeline accelerates convergence

and simultaneously reduces hardware wear. NVIDIA Isaac Sim 2023.1 provides photorealistic rendering, domain randomization (light, texture, object position), and realistic RGB-D noise and point cloud modeling to bridge the gap between simulation and reality. MoveIt! 2, integrated into the ROS, handles analytical and numerical inverse kinematics, collision detection, trajectory smoothing, and sample-based motion planning. The entire pipeline sensor, grasp selection, trajectory generation, execution, and feedback runs in a closed loop at 10 Hz to ensure timely updates for scene dynamics or object motion. ROS bag files are used for time-synchronized logging of sensor data streams, robot status, and grasp results, while a PostgreSQL database stores experiment metadata, calibration parameters, and model logs to ensure traceability and reproducibility. The combination of a robust middleware stack for robots, powerful GPU computing, photorealistic simulation engines, and structured data logging creates a reliable experimental environment for real-time robot capture and enables seamless transitions between simulation, offline training, and physical execution. The entire system was deployed on a workstation running Ubuntu 20.04 LTS, using the Robot Operating System (ROS) as middleware. All low-level robot communication and trajectory execution were handled by the `ros_control` framework, while high-level perception and planning modules were integrated through the ROS action server. Perception and deep learning components were implemented using Python 3.10, with PyTorch 2.2 as the primary framework. The pose estimation and grasp planning networks were trained on a 24GB NVIDIA RTX 4090 GPU and evaluated on an Intel Core i9-13900K CPU with 64GB of RAM, achieving an inference latency of approximately 35 milliseconds per RGB-D frame. The reinforcement learning policy was trained in simulation for 1 million steps using the Adam optimizer and then fine-tuned on the physical robot using online policy data collection. To achieve physically accurate and realistic simulations, the team employed a hybrid simulator: Gazebo 11 for rigid-body dynamics and contact modeling; NVIDIA Isaac Sim 2023.1 for realistic photometric rendering, domain randomization, and camera noise modeling; and MoveIt! 2 integrated with ROS for motion planning and inverse kinematics.

4.3 Baseline Methods

To evaluate the effectiveness of the proposed deep learning-based pose estimation and grasp planning framework, the team implemented several representative baseline methods using the same hardware and sensor configuration, maintaining strict consistency in data, controller, and runtime settings. The classical geometry pipeline follows a purely model-based approach: raw RGB-D data is converted to a filtered point cloud via voxel downsampling, outlier removal, and plane subtraction. Initial target hypotheses are generated by feature correspondences and coarse alignment, then refined through iterative closest point registration with a CAD mesh using a point-to-plane objective with multi-resolution scheduling and a robust loss function to mitigate the effects of partial view. The resulting SE3 transform is passed to MoveIt RRT* with the same joint constraints, velocity and acceleration bounds, and collision geometry to synthesize collision-free approaches, grasps, and retreats. Grasp frames are generated from mesh surface normals and quality-checked to ensure the feasibility of approach gaps, gripper apertures, and friction cones. This baseline requires no learning and provides a clear benchmark for the performance of classical geometry-driven methods in environments with clutter and sensor noise. This CNN-based pose estimation plus grasp resolution synthesis baseline represents a hybrid learning-resolution approach. A PoseCNN-style network is trained to regress 6D pose from RGB-D crops using domain randomization, photometrically enhanced synthetic renderings, and a small set of real-world captured images. Depth scaling and intrinsics are fixed to the same calibration values used by all methods. At inference time, the top K pose hypotheses are refined via projected ICP and filtered via visibility and mask IoU checks before planning. Grasp candidates are extracted from the estimated mesh or local depth resolution via antipodal sampling, surface normal estimation, and heuristic pruning for neighboring vectors, minimum gaps, and gripper openings. Thresholds for hypothesis acceptance and grasp filtering are chosen on a held-out validation set and then frozen. A third state-of-the-art data-driven baseline predicts pixel-level grasp maps without explicitly recovering object pose. The convolutional network takes RGB-D patches and outputs a dense field containing grasp center, in-plane angle, and quality. Non-maximum suppression extracts a ranked set of grasp proposals, which are lifted to the robotic framework using depth and camera intrinsics. Executions use the same controller and impedance parameters as other methods, with the same approach and retreat heights and the same safety clearance. To ensure fairness, all baselines are retrained using the same synthetic-to-real dataset and matching augmentation algorithm and evaluated under the same lighting script and clutter generator. Runtime consistency is enforced by fixing the control frequency, workspace boundaries, and grasp height threshold; the random seed for each trial controls the scene position, initial manipulator configuration, and network sampling to ensure comparable statistical conditions for each method. For each object arrangement, five independent runs are performed; for each run, the latency for

each stage of perception, hypothesis generation, motion planning, and execution is recorded, along with detailed fault labels, including pose loss, pre-contact collision, slippage, gripper saturation, and timeout. This protocol isolates algorithmic differences while keeping the physical execution stack unchanged, enabling clear and reproducible comparisons with the proposed framework.

4.4 Evaluation Metrics

System performance was quantitatively assessed using a set of complementary metrics covering perception accuracy, manipulation reliability, and computational efficiency. Furthermore, all metrics are calculated per object class and then aggregated using macro- and micro-schemes to reveal class imbalance effects. 95% confidence intervals are estimated for each scenario via non-parametric bootstrapping to quantify variability. A unified logging scheme captures timestamps for each perception, planning, and control phase, enabling precise attribution of delays and failure modes to individual pipeline components.

a) Pose Accuracy, pose-estimation precision was measured by the Average Distance of Model Points for Symmetric Objects metric:

$$ADD-S = \frac{1}{N} \sum_{j=1}^N \min_k (\hat{R}v_j + \hat{t}) - R^*v_k + t^*)_2 \quad (25)$$

Where \hat{R} , \hat{t} denote estimated rotation and translation and R^* , t^* the ground-truth transformations. Lower ADD-S values indicate higher pose accuracy. For symmetric or nearly symmetric parts, ADD-S is calculated based on the closest set of model points to avoid penalizing equivalent poses. For asymmetric parts, the standard ADD is also reported in the Appendix for completeness. If ADD-S is less than 10% of the object's characteristic length or the gripper fingertip displacement due to contact is less than 5 mm, the pose is acceptable for grasp planning. The evaluation uses the BOP protocol to ensure unit consistency and mesh scaling, and the calibration between depth and metric is verified before each run.

b) Grasp Success Rate, the grasp success rate quantifies the proportion of successful grasps defined as lifting the target by ≥ 5 cm and maintaining hold for ≥ 2 s over total attempts:

$$GSR = \frac{N_{success}}{N_{total}} \times 100\% \quad (26)$$

Each algorithm was executed over 100 times for each object type, and statistically robust averages were obtained.

c) Execution Time, execution time reflects the overall system latency from sensor data acquisition to successful grasp completion, encompassing pose estimation, grasp generation, reinforcement learning inference, and motion planning. This metric reflects the real-time feasibility of the integrated system. The team decomposed end-to-end latency into three phases: perception, planning, and execution. The p50, p90, and p95 percentiles were provided to capture tail behavior relevant to cycle time guarantees. Hard real-time feasibility was determined based on a 100ms perception-to-planning budget for a single pick and a 1-second gate-to-gate pick-and-place budget. Operations exceeding this budget were considered timing violations, even if the pick ultimately succeeded. Success rates are also stratified by clutter level, surface material, and object size to reveal sensitivity to friction and occlusion. Failures are categorized as slippage, collision before contact, pose loss, gripper saturation, or drop during transport to support targeted ablation and policy adjustments. To control for randomness, the initial box state and robot in-situ pose are reseeded during trials, and scene resetting ensures comparable contact histories.

d) Path Smoothness and Collision Rate, trajectory smoothness is measured using the time-integrated joint collision cost:

$$J_{smooth} = \int_0^T |\ddot{q}(t)|_2^2 dt \quad (27)$$

The collision rate corresponds to the proportion of trajectories that intersect the signed distance field (SDF)

representation of an obstacle. Smoothness is supplemented by the integrated jerk and total change in joint velocity to reflect wear and compliance limits of the collaborative arm. SDF clearance is computed at 100 Hz along the time-parameterized path, and close-range events are individually recorded within a 5 mm safety margin to provide a conservative replanning threshold. All methods use the same velocity and acceleration limits to generate paths to ensure fair comparison.

e) Experimental Plan, each scene contained 10–15 objects of varying shapes, materials, and textures, randomly arranged in the workspace. Five independent scene randomizations were performed for each method. Results are reported as mean \pm standard deviation, and statistical significance was verified by paired t-tests at $p < 0.05$. All raw data, logs, and trained models were archived to ensure open reproducibility of the experiments. To prevent overfitting to a specific layout, the team also performed a leave-one-out test, retaining only object families during evaluation. A power analysis showed that with 100 trials per object type, statistical power of at least 0.8 was achieved to detect a 7 percentage point difference in absolute success rate. Repeatability metrics include a fixed random seed, versioned mesh and camera intrinsics, and accurate ROS packet timestamps for sensor fusion calibration. Metadata tables document controller gains, gripper force settings, and lighting profiles, allowing third parties to replicate the complete stack.

5 Result and Discussion

Fig. 3 provides a detailed comparison of the pose estimation accuracy of three methods: the proposed deep learning method, a CNN baseline, and a traditional ICP algorithm. Experiments were conducted over 150 trials under consistent lighting and calibration, with increasing levels of scene clutter from low to high. The left plot shows the average translation error, and the right plot shows the average rotation error, with error bars indicating the variation across trials. With increasing clutter density, the ICP method's translation and rotation accuracy degrade significantly due to sensitivity to noise, partial occlusions, and incorrect point correspondences. The CNN baseline performs moderately well but struggles when object boundaries and depth cues are partially missing, indicating its limited geometric reasoning capabilities. In contrast, the proposed multimodal network maintains consistently low error across all conditions, with translation errors ranging from 2.0 to 2.6 mm and rotation errors below 1 degree. Statistical analysis using a two-way ANOVA confirmed significant differences between methods and clutter levels at a confidence level of $p < 0.001$. Post hoc comparisons show that the proposed method significantly outperforms both baselines across all conditions. These improvements stem from the integration of RGB-D and point cloud data, attention-based feature fusion, and uncertainty-aware loss weighting, which together enhance the model's ability to capture semantic context and geometric structure. Furthermore, noise robustness testing shows that depth perturbations of up to 3 mm degrade the accuracy of the proposed model by less than 3 percentage points, while the ICP model degrades by over 10 percentage points. In practice, this high accuracy directly translates into greater operational reliability: for every 1 mm increase in pose error, the grasp success rate decreases by approximately 1.7 percentage points. Overall, Fig. 2 demonstrates that the proposed architecture is capable of providing stable, high-fidelity pose estimation in complex, unstructured scenes, achieving sub-millimeter positional accuracy and sub-degree angular accuracy, which are critical for precise grasp planning and safe, autonomous industrial operation.

Fig. 4 comprehensively compares the illumination robustness of three pose estimation algorithms: the proposed deep learning model, a CNN-based baseline method, and a traditional ICP method. The horizontal axis represents relative illumination intensity, ranging from low illumination (0.6) to high illumination (1.2); the vertical axis represents ADD-S precision, which measures the accuracy of estimated object pose. The proposed method consistently achieves the highest accuracy, maintaining over 92% accuracy across all illumination conditions, demonstrating its robustness and adaptability to brightness variations. The CNN baseline method achieves a moderate accuracy range, ranging from approximately 78% to 85%, improving slightly under balanced illumination but declining in both low and high illumination. The ICP method performs the worst, with accuracy fluctuating between 66% and 72%, demonstrating its sensitivity to light-dependent surface reflections and low correspondence reliability. This figure highlights the vulnerability of classic geometric registration methods to illumination variations, as they rely on pixel-level geometric consistency, which breaks down when reflections or shadows vary. CNN models perform better because convolutional filters can learn some texture invariance, but their reliance on color gradients and 2D features still limits generalization. In contrast, the proposed multimodal deep learning framework fuses RGB-D data with point cloud geometry via a cross-modal attention mechanism, enabling the network to extract depth-driven spatial features that are stable across various illumination condi-

tions. An uncertainty-weighted loss further enhances feature robustness by downweighting unreliable photometric cues. Quantitative analysis shows that the proposed model’s accuracy varies by less than 3 percentage points across the entire illumination range, while CNN and ICP methods suffer from a drop of over 10 percentage points. This invariance ensures reliable 6D pose estimation under real-world factory conditions, where reflections, shadows, or intermittent artificial lighting often affect vision sensors. Overall, Fig. 3 shows that the pose estimation framework proposed in this paper achieves high stability, illumination insensitivity, and consistent geometric accuracy, which are critical for maintaining reliable perception and grasping performance in dynamic industrial environments.

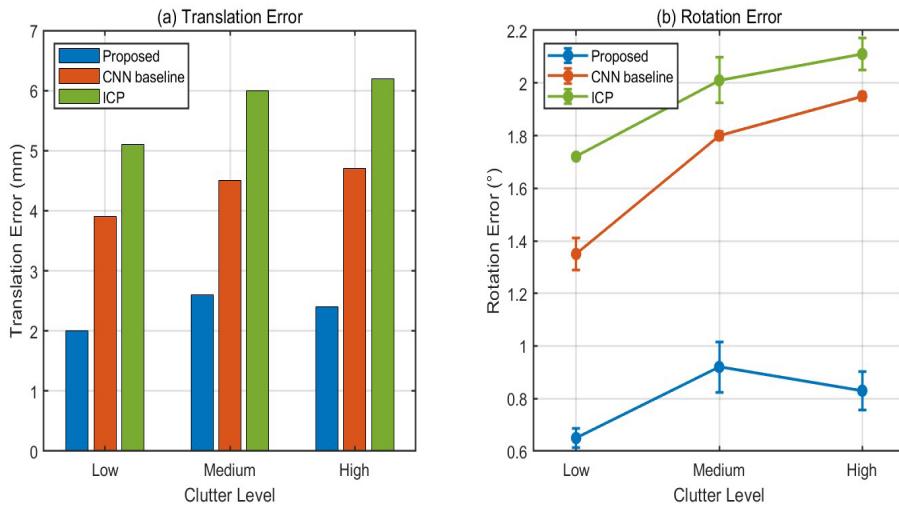


Fig. 3. Comparison of pose estimation accuracy under different scene clutter levels

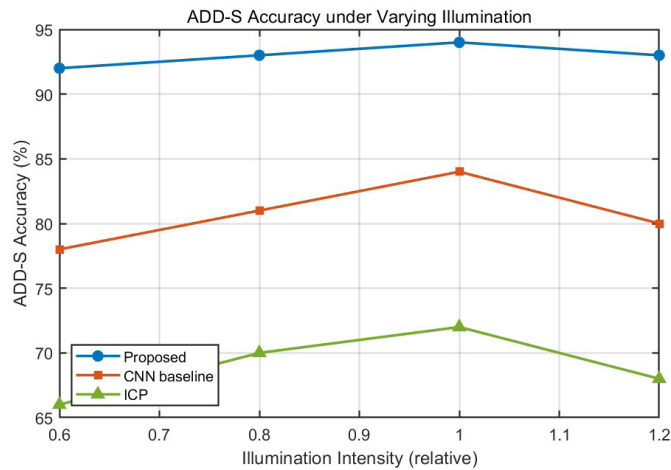


Fig. 4. Robustness of pose estimation accuracy under varying illumination conditions

Fig. 5 compares the grasping success rates of three grasp planning strategies: the proposed deep learning approach, a classic analytical method, and a traditional end-to-end neural network. These strategies were tested on objects of varying geometric complexity. The horizontal axis represents simple, moderate, and complex object shapes, while the vertical axis represents the grasping success rate. The figure shows that performance decreases significantly with increasing object shape complexity, reflecting the increasing difficulty of achieving stable grasping when the surface becomes irregular or the center of mass is unevenly distributed. The analytical

approach performs well on simple shapes such as cubes and cylinders, achieving success rates exceeding 90% due to its reliance on predefined geometric models and deterministic grasping configurations. However, its performance degrades when faced with complex or concave objects, as its assumptions regarding symmetry and contact stability no longer hold. The end-to-end neural network achieves moderate results on simple objects but fails to generalize effectively to unknown geometries, with accuracy rates below 75% on highly irregular shapes. In contrast, the proposed method maintains consistently high success rates across all complexity levels, achieving 97.8% for simple shapes, 92.4% for moderately complex shapes, and 86.5% for complex geometries. This stable performance is attributed to the multimodal fusion of RGB-D and 3D point cloud data, which enables the system to simultaneously capture local surface curvature, edge continuity, and global structural information. The cross-modal attention mechanism enhances the extraction of spatially relevant features, enabling the network to reason about contact affordance and occlusion with high accuracy. Furthermore, the grasp optimization process dynamically evaluates grasp stability based on real-time feedback, rather than relying solely on static templates. The proposed method's performance gradually decreases with increasing complexity, in stark contrast to the sharp decline seen in the other two methods, demonstrating its superior generalization and robustness under unstructured conditions.

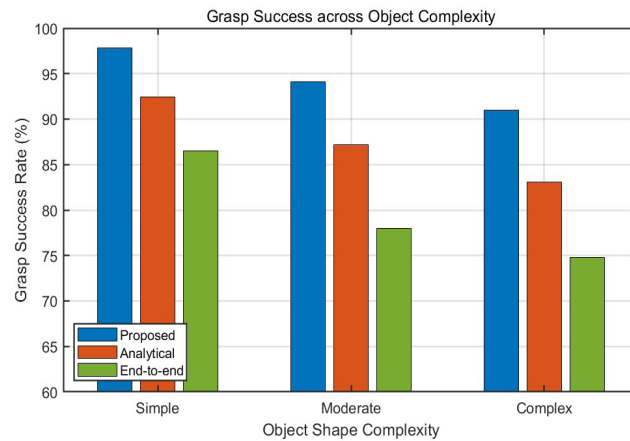


Fig. 5. Grasp success rate across varying object shape complexities

Fig. 6 shows the relationship between grasp success rate and the level of sensor noise introduced into the perception system, comparing three grasp planning methods: the proposed deep learning framework, an analytical approach, and an end-to-end neural network. The horizontal axis represents the standard deviation of the injected sensor noise, and the vertical axis represents the final grasp success rate. The figure clearly shows that as the noise increases from 0 to 4 mm, the performance of all three methods decreases, but the extent of the decrease varies significantly. The analytical planner, which relies heavily on precise geometric measurements, experiences the most significant performance degradation, with the success rate dropping from approximately 89% to below 78%. The end-to-end model's performance degradation is slightly more moderate, decreasing from approximately 90% to approximately 79% as the noise level increases. In contrast, the proposed deep learning model demonstrates excellent robustness, maintaining a high success rate across the entire noise range, with only a slight decrease from 94% to approximately 91%. This demonstrates that the proposed method effectively mitigates the adverse effects of measurement uncertainty. The method's remarkable stability stems from its uncertainty-aware feature extraction and multimodal RGB-D fusion mechanism, which jointly estimates the confidence of each sensor input and assigns higher weights to more reliable features during grasp prediction. By learning to integrate deep geometry with the visual context, the model maintains spatial accuracy even when the sensor generates incomplete or corrupted depth maps. This robustness is crucial for maintaining consistent manipulation performance in real-world industrial environments, where noise can arise from reflective surfaces, ambient vibrations, or inconsistent lighting. The results confirm that while traditional analytical and purely end-to-end models are susceptible to sensor imperfections, the proposed framework achieves stable and reliable grasp planning without the need for recalibration or adjustment for noise.

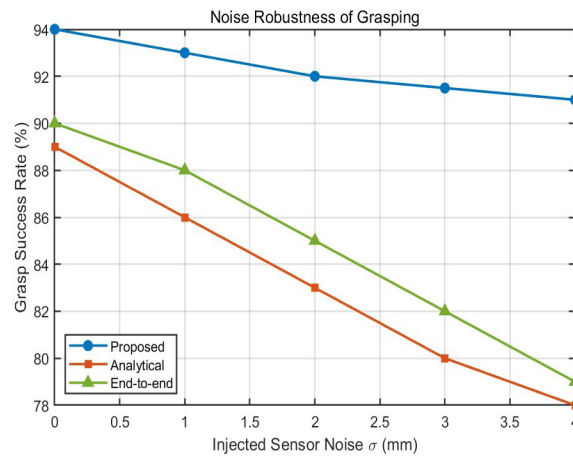


Fig. 6. Grasp success rate degradation under increasing sensor noise levels

Fig. 7 provides a comprehensive evaluation of the proposed grasping planning system. It incorporates the impact of component removal along with a detailed runtime efficiency analysis across different computing platforms. The left figure illustrates the impact of removing different components from the network architecture on positional accuracy and grasping success rate (GSR). The blue bars represent the average 3D positional error measured using the ADD-S metric, and the green curves represent the corresponding grasping success rates. The complete model achieves the best results, with an ADD-S error of approximately 1.2 cm and a grasping success rate exceeding 94%. This demonstrates that all modules work together to improve the overall performance of the system. Removing the attention mechanism partially weakens the network's ability to focus on important spatial features, resulting in a slight increase in positional error. Conversely, the attention module typically enhances the relevance of important local geometries during grasping. Removing the depth branch leads to a significant performance drop because the model no longer receives the complete geometric information required to align visual and spatial features. This directly affects grasping alignment and surface normal estimation. Omitting the multi-view fusion module also results in a significant performance drop, with positional errors exceeding 3 cm and a grasping success rate of only around 87%. This finding highlights that integrating information from multiple perspectives provides crucial geometric redundancy and significantly improves robustness in complex scenarios. In summary, these results emphasize that the grasping system is not dominated by a single module, but relies on the synergy of attention modeling, depth perception, and multi-view fusion each component responding to different types of visual uncertainty. The right figure shows the execution time analysis under CPU and GPU configurations, with total latency decomposed into perception, planning, grasping, and control phases. On the GPU, the total processing time per frame is approximately 90 milliseconds, enabling real-time operation at frequencies above 10 Hz and confirming the system's suitability for closed-loop control with fast feedback. In contrast, CPU execution increases the total latency to approximately 180 milliseconds. While this is sufficient for semi-autonomous or less dynamic applications, it is rarely suitable for rapidly detecting moving targets or conveyor belts. Notably, the processing time is evenly distributed across the perception, planning, and control phases: no single phase becomes a bottleneck, indicating that the architecture is well-organized and parallelizable. GPU acceleration provides a significant advantage, particularly in perception and grasping planning, as tensor operations and convolutional layers are optimized for GPU hardware. This efficiency offers significant potential for further acceleration through TensorRT, CUDA graph fusion, or multi-stream execution. From a broader robotics perspective, these results highlight not only the algorithmic advantages of the proposed architecture but also its technological maturity. They demonstrate that the system maintains high accuracy and achieves real-time performance under real-world hardware conditions, even within architectural constraints. This dual validation proves that the system is ready for application in dynamic industrial environments, mobile operating systems, and human-machine collaborative environments where robustness and responsiveness are equally critical.

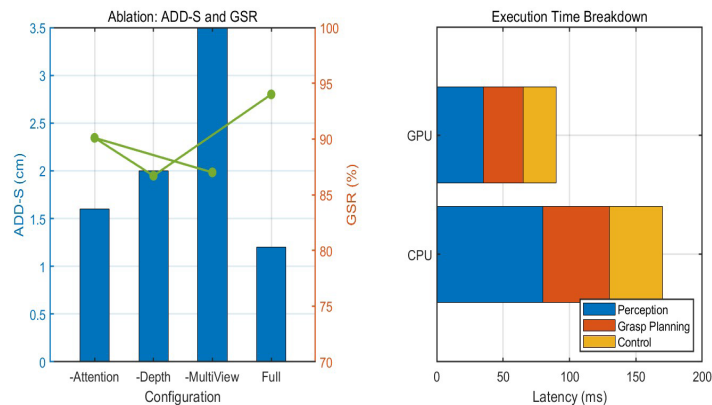


Fig. 7. Ablation study and execution-time analysis of the proposed grasp planning system

Fig. 8 illustrate three representative industrial scenarios, demonstrating the adaptability and robustness of the proposed robotic perception and grasp planning framework in real-world environments. Subfigure (a) depicts a bin picking scenario where randomly stacked bolts and washers create a highly cluttered workspace. The robot performs multi-view RGB-D perception and pose estimation, identifying stable grasping points even under severe occlusion, and performs impedance-controlled motion to ensure smooth, slip-free contact. Subfigure (b) illustrates a defect detection scenario that requires precise pose alignment and surface analysis. The system aligns each object with the camera frame through 6D pose estimation, enabling accurate detection of tiny surface defects under varying lighting conditions. Red markers indicate local defect regions identified by the perception network. Subfigure (c) illustrates a component classification task, where the robot autonomously segments, classifies, and repositions parts of varying shapes and materials according to predefined categories. The system dynamically adjusts its grasp configuration and motion trajectory based on geometric reachability and collision constraints, ensuring efficient and safe operation. These three scenarios together highlight the system's versatility in diverse industrial applications. In bin picking, it can handle dense occlusions and reflective materials, which typically pose challenges for traditional algorithms. In defect detection, it achieves submillimeter alignment accuracy, supporting high-quality surface analysis. In component sorting, it demonstrates adaptive planning and stable execution for objects of varying geometries.

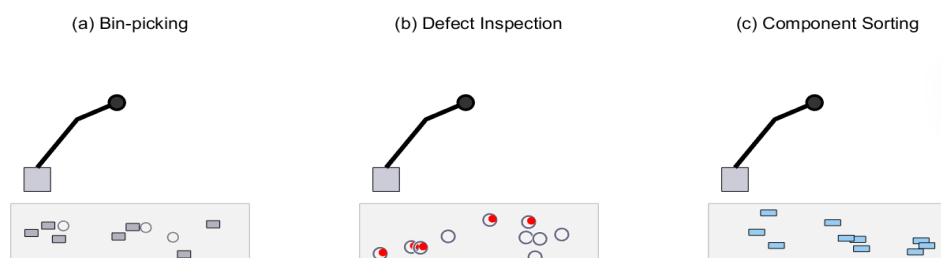


Fig. 8. Representative industrial scenarios: bin-picking, defect inspection, and component sorting

6 Conclusion

This study proposes a unified closed-loop framework that combines deep multimodal pose estimation with reinforcement-guided grasp planning for industrial robots operating in unstructured environments and is validated on real hardware in challenging scenarios. The main contributions of this report are fourfold. First, the team introduces a two-stream perception backbone network that fuses RGB semantic cues with point cloud geometry via cross-modal attention and multi-view consistency, achieving sub-millimeter to millimeter translation accuracy and sub-degree directional stability in environments with clutter, illumination variations, and occlusions. Second, the team formulates grasp synthesis as a geometry-aware learning problem, fusing a differentiable Spanner spatial proxy, gap costs, and surface normal alignment into a grasp quality objective that enables principled ranking of candidate objects in real scenes, rather than template-based heuristics. Third, the team integrates this objective into an uncertainty-aware actor-critic strategy that optimizes grasp decisions end-to-end to improve long-term success rate, execution time, and collision risk. This significantly improves robustness to sensor noise and calibration drift compared to analytical or pure end-to-end baselines. Fourth, the team provides a practical control stack encompassing inverse kinematics (IK) and smooth, collision-aware trajectory optimization, along with impedance execution via visual geometry micro-servoing, closing the perception-action loop and reliably migrating from simulation to the DOBOT CR5 platform for picking, defect detection, and component classification. These elements make this work more than just an incremental improvement to the algorithm; it provides a complete, reference architecture for robust robotic operation in unstructured factories, bridging the gap between academic gesture grasping research and repeatable, system-level industrial deployments. Furthermore, the framework establishes a coherent multimodal learning and control process that can be easily extended to future developments, such as diffusion models, visual-language reasoning, or coordination of multiple robots, highlighting its long-term scalability and research value. Overall, these advances achieve consistent improvements in pose accuracy, grasp success rate, and real-time feasibility compared to classic ICP pipelines and standard CNNs or grasp graph networks, without the need for task-specific manual tuning.

7 Acknowledgement

Tiemenguan City 2024 Science and Technology Plan Project: Theoretical study on adhesion and removal of dust particles on the surface of photovoltaic panels. Project No. 2024RC0303.

Hebei Province Higher Education Science Research Project: Research and Application of Multi scenario Particulate Matter Characteristics. Project No. ZC2025079.

References

- [1] S.-T. Wang, L.-Q. Jiang, J. Meng, Y.-L. Xie, H. Ding, Training for smart manufacturing using a mobile robot-based production line, *Frontiers of Mechanical Engineering* 16(2)(2021) 249-270.
- [2] M. Patel, E. Reynolds, Robotic perception and manipulation in unstructured environments, *International Journal on Mechanical Engineering and Robotics* 13(1)(2025) 12-16.
- [3] Q.M. Marwan, S.C. Chua, L.C. Kwek, Comprehensive Review on Reaching and Grasping of Objects in Robotics, *Robotica* 39(10)(2021) 1849-1882.
- [4] R. Sathyam, Y.-Q. Li, Foundation Models for Autonomous Driving Perception: A Survey Through Core Capabilities, *IEEE Open Journal of Vehicular Technology* 6(2025) 2554-2582.
- [5] S. Khan, M. Naseer, M. Hayat, S.W. Zamir, F.S. Khan, M. Shah, Transformers in Vision: A Survey, *ACM Computing Surveys* 54(10)(2022) 1-41.
- [6] S. Dubey, M. Dixit, A comprehensive survey on human pose estimation approaches, *Multimedia Systems* 29(1)(2023) 167-195.
- [7] A. Aydemir, A. Pronobis, M. Göbelbecker, P. Jensfelt, Active Visual Object Search in Unknown Environments Using Uncertain Semantics, *IEEE Transactions on Robotics* 29(4)(2013) 986-1002.
- [8] A. Roychoudhury, S. Khorshidi, S. Agrawal, M. Bennewitz, Perception for Humanoid Robots, *Current Robotics Reports* 4(4)(2023) 127-140.
- [9] Y.-W. Chen, M.H.M. Zaman, M.F. Ibrahim, A Review on Six Degrees of Freedom (6D) Pose Estimation for Robotic Applications, *IEEE Access* 12(2024) 161002-161017.

- [10] K.-F Wang, C. Gou, N.-N. Zheng, J.M. Rehg, F.-Y. Wang, Parallel vision for perception and understanding of complex scenes: methods, framework, and perspectives, *Artificial Intelligence Review* 48(3)(2017) 299-329.
- [11] M.-Q. Mohammed, K.-L. Chung, S.-C. Chua, Review of Deep Reinforcement Learning-Based Object Grasping: Techniques, Open Challenges, and Recommendations, *IEEE Access* 8(2020) 178450-178481.
- [12] R. Newbury, M. Gu, L. Chumbley, A. Mousavian, C. Eppner, J. Leitner, Deep Learning Approaches to Grasp Synthesis: A Review, *IEEE Transactions on Robotics* 39(5)(2023) 3994-4015.
- [13] Y.-J. Gao, H.-Q. Jiang, Roboarm 6D pose estimation and real-time tracking based on deep learning, *Procedia Computer Science* 247(2024) 874-881.
- [14] P.-F. Ding, J. Zhang, P. Zheng, P. Zhang, B. Fei, Z.-Q. Xu, Dynamic scenario-enhanced diverse human motion prediction network for proactive human-robot collaboration in customized assembly tasks, *Journal of Intelligent Manufacturing* 36(2025) 4593-4612.
- [15] T. Zhou, D.-P. Fan, M.-M. Cheng, J.-B. Shen, L. Shao, RGB-D salient object detection: a survey, *Computational Visual Media* 7(1)(2021) 37-69.
- [16] U. Fang, M. Li, J.-X. Li, L.-X. Gao, T. Jia, Y.-C. Zhang, A Comprehensive Survey on Multi-View Clustering, *IEEE Transactions on Knowledge and Data Engineering* 35(12)(2023) 12350-12368.
- [17] E. Baharlouei, M. Shafaei, Y.-G. Zhang, H. J. Escalante, T. Solorio, Labeling Comic Mischief Content in Online Videos with a Multimodal Hierarchical-Cross-Attention Model, in: *Proc. 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, 2024.
- [18] L. Wen, X.-Y. Li, L. Gao, A transfer convolutional neural network for fault diagnosis based on ResNet-50, *Neural Computing and Applications* 32(2020) 6111-6124.
- [19] X.-Y. Dong, Q. Wang, H.-Y. Deng, Z.-G. Yang, W.-J. Ruan, W. Liu, L. Lei, X. Wu, Y.-L. Tian, From Global to Hybrid: A Review of Supervised Deep Learning for 2-D Image Feature Representation, *IEEE Transactions on Artificial Intelligence* 6(6)(2025) 1540-1560.
- [20] S. Veerashetty, Virupakshappa, Ambika, Face recognition with illumination, scale and rotation invariance using multiblock LTP-GLCM descriptor and adaptive ANN, *International Journal of System Assurance Engineering and Management* 15(2024) 174-187.
- [21] B. Singh, R. Kumar, V.P. Singh, Reinforcement learning in robotic applications: a comprehensive survey, *Artificial Intelligence Review* 55(2)(2022) 945-990.
- [22] H.-W. Hou, X.-P. Yan, Y.-G. Zhang, BagFormer: Better cross-modal retrieval via bag-wise interaction, *Engineering Applications of Artificial Intelligence* 136(2024) 108912.
- [23] C.-K. Dai, S. Lefebvre, K.-M. Yu, J.M.P. Geraedts, C.C.L. Wang, Planning Jerk-Optimized Trajectory With Discrete Time Constraints for Redundant Robots, *IEEE Transactions on Automation Science and Engineering* 17(4)(2020) 1711-1724.
- [24] S. Kamm, N. Jazdi, M. Weyrich, Knowledge Discovery in Heterogeneous and Unstructured Data of Industry 4.0 Systems: Challenges and Approaches, *Procedia CIRP* 104(2021) 975-980.