Journal of Applied Automation Technologies

Research Paper

Vision-Based Grasping for Industrial Robots Guided by Deep Q-Networks

Emilia Zdzisława Dąbrowska[1], Karolina Alicja Bąk[1], Paweł Jankowski[1] and Aleksandra Król[1,*]

[1] Maria Curie-Sklodowska University, Faculty of Mathematics, Physics and Computer Science, 20-031 Lublin, Poland
*Corresponding author: aleksandra.k@mail.umcs.pl

Abstract. This paper proposes a vision-to-grasp framework driven by a Deep Q-Network (DQN). The framework is designed for industrial robots in complex and dynamic factory environments. Multimodal visual perception, robust scene understanding, and adaptive deep reinforcement learning-based grasp planning are the three components of the new framework. The experiments were conducted in both simulated and real environments, using a UR5e robotic arm equipped with an RGB-D camera and a standard industrial gripper. In the 200,000 training steps in the simulation, the asymptotic grasping success rate of the Double DQN variant was 93.4%, which is 6.2 %age points higher than the standard DQN, and the reward variance was reduced by over 38%. Boltzmann's study achieved a success rate of 92.8% within 80,000 steps, far surpassing the noise network and ε-greedy methods. The framework achieved a success rate of over 90% for grasping both simple and complex objects on physical hardware, with success rates of 95.1%, 91.3%, and 87.6% for acrylic boxes and ceramic cups, respectively. In mixed and cluttered environments, the system achieved a total success rate of 92.5% over 300 rounds, with a relatively short grasp execution time of 2.7 to 3.7 seconds. Robustness tests indicate that it can generalize well to new objects, partial occlusions, and lighting changes. Failure analysis pointed out the main defects in spatial alignment and reflection. The above results indicate that DQN-based visual-guided grasping has been applied in practice and is feasible in industrial environments.

Keywords: *Visual Perception, Deep Reinforcement Learning, Robotic Grasping, DQN, Industrial Automation, Sim-to-Real Transfer, Multi-Modal Perception, Manipulation Robotics*

Received on 13 November 2025, Accepted on 01 March 2026, Published on 16 March 2026

Introduction

With the rapid development of industrial automation, many modern factories have become very flexible and intelligent, and production lines have become more adaptable [1]. Sorting, quality inspection, logistics, precision assembly, and sorting are among the many fields where industrial robots are currently widely used [2]. Vision-based grasping is becoming increasingly popular in empowering technologies because it enables robots to dynamically interact in diverse and new environments [3]. Vision-guided grasping is more adaptive than traditional open-loop or model-based methods because it can recognize changes in real-time and make corresponding adjustments [4]. However, due to the uncertainty in perception and the high dimensionality of operational actions, reliable visual grasping remains a challenge in the presence of occlusions, clutter, and complex shapes [5]. Sensor noise, lighting changes, and background interference often affect the generalization ability of existing vision-based systems, making them less suitable for industrial applications [6]. Moreover, there are still many issues with increasing and improving the stability of manual feature extraction or hyperparameter tuning methods [7]. Therefore, people are actively seeking learning-based strategies to enable industrial robots to handle complex vision tasks more autonomously and effectively [8].

With the advancement of deep learning, deep reinforcement learning (DRL) has recently been used to develop robots that can directly learn decision-making strategies from high-dimensional perceptual data [9]. Deep Q-Networks (DQN) are algorithms suitable for end-to-end learning that directly convert raw visual data into discrete actions. In deep reinforcement learning (DRL) algorithms, DQN has been used to address the coupling

problem between perception and action [10]. The DQN-driven architecture is very effective for robotic tasks such as visual servoing, real-time motion planning, and continuous control [11]. Deep Q-Networks (DQNs) and Convolutional Neural Networks (CNNs) can be combined, allowing for automatic learning of image features without manual design [12]. The challenges of sample inefficiency in reward-based training, the sparsity of rewards, and the instability in performance when facing environmental changes or domain shifts [13]. Although some progress has been made, these issues have not been completely resolved. Especially in practical applications, transferring the trained strategies from the simulated environment to the real environment, known as the sim-to-real gap, remains a major obstacle [14]. Moreover, industrial grasping scenarios often involve new objects, changing background environments, and unstable clutter. Therefore, the scalability and adaptability of robotic systems are higher [15]. Current research aims to address the aforementioned shortcomings by integrating data augmentation, domain randomization, and advanced exploration methods within the DQN framework [16]. However, there is no comprehensive and systematic vision-to-grasp pipeline [17]. Therefore, developing a high-performance, scalable industrial vision grasping framework based on DQN is both urgent and feasible [18].

This paper proposes a DQN-guided visual grasping framework that can be used for industrial robots. We have constructed a modular perception-to-action framework, which is different from the aforementioned. It includes real-time sensors, robust environmental perception capabilities, and customizable grasping plans based on deep reinforcement learning. Combining multimodal data streams and using task-oriented reward functions, we improve the robustness of grasping under environmental uncertainty and dynamically adapt to new object categories. Many experiments conducted in both simulated and real industrial environments have shown that our method is more suitable for real-time applications and has a higher success rate in grasping compared to previous methods. Based on the above experiments, an end-to-end DQN-based method is proposed to enhance the operational capabilities of autonomous robots in complex industrial environments. This method also provides support for future research on intelligent robots.

Literature Review

Vision-Based Grasping: Approaches and Trends

Due to the increasing demand for complex operations in unstructured environments, the development of vision-based robotic grasping has rapidly advanced over the past decade. First, most methods addressing this problem use geometry. Robots determine the shape and position of objects through feature engineering or traditional image processing techniques, and then develop grasping strategies using manually designed rules or force closure criteria [19]. These traditional methods work well when dealing with simple, well-structured parts, but they quickly fail when encountering occlusions, sensor noise, or irregular object geometries [20]. With the advent of depth cameras and RGB-D sensors, researchers can gain a deeper understanding of 3D scenes and are able to perform grasp detection in complex environments [21].

Data-driven learning has also changed the task of robot vision-guided grasping. Convolutional Neural Networks (CNNs) have demonstrated superior object recognition capabilities compared to model-based methods within learning-based frameworks, and they can determine the correct grasping position from raw pixel data [22]. Self-supervised learning and large-scale data collection enable robots to automatically label and learn from a vast number of real-world grasping attempts. This improves generalization ability and reduces the need for synthetic data or real annotations [23]. Researchers have recently begun to focus on affordance learning. This involves predicting whether a robot can grasp an object and determining the optimal grasping position for subsequent manipulation tasks [24].

The above are the main trends. Multimodal fusion includes RGB, depth, and tactile data to achieve more accurate grasp predictions. In addition, applying transformer architectures can flexibly handle unordered visual feature sets [25]. The problem of sim-to-real transfer and rapid adaptation to new objects can be addressed through domain adaptation and few-shot learning [26]. Although some progress has been made in the past few years, handling highly cluttered environments, object transparency or reflectivity, and adversarial conditions remains a focus of current research [27]. With the development of this field, adaptive and self-improving algorithms capable of real-time operation in open industrial environments have clearly emerged [28].

Deep Reinforcement Learning in Vision-Guided Robotics

Deep Reinforcement Learning (DRL) is a useful tool for connecting perception and action in robotic grasping [29]. Now, it is striving to solve the problem of policy optimization in high-dimensional continuous spaces. Reinforcement learning is an independent, self-learning method that does not rely on imitation or supervision. It allows agents to discover good behaviors by performing tasks and using a reward function. Deep Q-Networks (DQN) have recently achieved great success in value-based Deep Reinforcement Learning (DRL) algorithms, particularly because they can handle discrete action spaces and learn effective control policies directly from image data [30].

In the field of robotic manipulation, DQN-based methods have been applied to solve problems such as grasp selection, servoing, and tool usage. In order to improve sample efficiency and learning stability, replay buffers and target networks are often used [31]. A hybrid framework has been developed that combines DQN with actor-critic methods or imitation learning methods to improve convergence speed and practical application [32]. In the fields of reward engineering and curriculum learning, recent research has shown that robots can improve environmental representation and address issues of delayed or sparse rewards by adding auxiliary tasks (such as pose estimation or object segmentation) [33].

Exploration methods are still under research; however, past studies have also focused on enhancing intrinsic motivation through distributed reinforcement learning, boosting curiosity-driven exploration, and improving robustness to rare events [34]. The gap between simulation and reality is another issue being addressed. Domain randomization, adversarial adaptation, and fine-tuning with real data are helping to reduce the differences between simulation and the real world in terms of sensor noise, dynamics, and appearance [35]. Deep Reinforcement Learning (DRL) is often combined with Model Predictive Control (MPC) or traditional motion planning to create deployable systems that are reliable, sample-efficient, and suitable for constrained industrial environments. Although these issues have been addressed, catastrophic forgetting, reward hacking, and poor interpretability still persist, limiting the application of DRL-based vision-guided grasping in production robots [36].

Future goals include encouraging lifelong learning, formulating risk-aware policies, and adapting to industrial and environmental changes. The combination of advanced vision and reinforcement learning continuously expands the range of tasks that autonomous robots can perform in factories [37], and creates new opportunities for flexible intelligent automation [38]. All of this is happening against the backdrop of deep reinforcement learning (DRL) frameworks becoming more stable and interpretable.

DQN-Guided Vision-to-Grasping Framework

System Architecture Overview

The three components of the new system are: (i) multimodal perception backend; (ii) deep Q-network (DQN) for sequential decision-making; and (iii) real-time execution interface. As shown in Figure 1, the sensor streams from a set of K spatially distributed, synchronized RGB-D cameras are individually preprocessed and then fused into a joint latent representation. The formal system state at time t is as follows

$$\boldsymbol{s_t} = \boldsymbol{\mathcal{E}}\left(\bigoplus_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{\mathcal{F}_{pre}}\left(\boldsymbol{I_{t,k}^{RGB}}, \boldsymbol{I_{t,k}^{D}}\right)\right) \quad \text{Eq.(1)}$$

Among them, $\mathcal{E}$ is a nonlinear multi-branch encoder, and $\mathcal{F}_{\text{pre}}$ is for geometric correction and normalization. The system state can also be represented using a probabilistic model to clearly indicate the observed uncertainty.

$$\boldsymbol{s_t} \sim \boldsymbol{\mathcal{N}}(\boldsymbol{\mu_t}, \boldsymbol{\Sigma_t}), (\boldsymbol{\mu_t}, \boldsymbol{\Sigma_t}) = \boldsymbol{f_{enc}}(\boldsymbol{O_t}) \quad \text{Eq.(2)}$$

The mean and covariance of sensor data are calculated by computing the expected Q value for each potential action in the DQN inference module.

$$\boldsymbol{Q}(\boldsymbol{s_t}, \boldsymbol{a_t}; \boldsymbol{\theta}) = \boldsymbol{r_t} + \boldsymbol{\gamma}\mathbb{E}_{\boldsymbol{s_{t+1}\sim P}}\left[\boldsymbol{\max_{a'}} \boldsymbol{Q}(\boldsymbol{s_{t+1}}, \boldsymbol{a'}; \boldsymbol{\theta^{-}})\right] \quad \text{Eq.(3)}$$

Here, r_t represents the instantaneous reward, γ the discount factor, and θ^{-}a slowly updated target network parameterization.

To ensure stable action selection, policy control uses the Q-values of the Boltzmann distribution:

$$P(a_t \mid s_t) = \frac{exp\left(\beta Q(s_t, a_t)\right)}{\sum_{a' \in \mathcal{A}} exp\left(\beta Q(s_t, a')\right)} \quad \text{Eq.(4)}$$

In order to support training exploration. Record experience tuples and track the statistical uncertainty of rewards during state transitions:

$$\mathcal{D}_t = \{(s_t, a_t, r_t, \mathbb{V}[r_t], s_{t+1})\} \quad \text{Eq.(5)}$$

The full system architecture and dataflow are visually summarized in Figure 1, which illustrates the connectivity from perception through Q-driven planning to robot actuation.

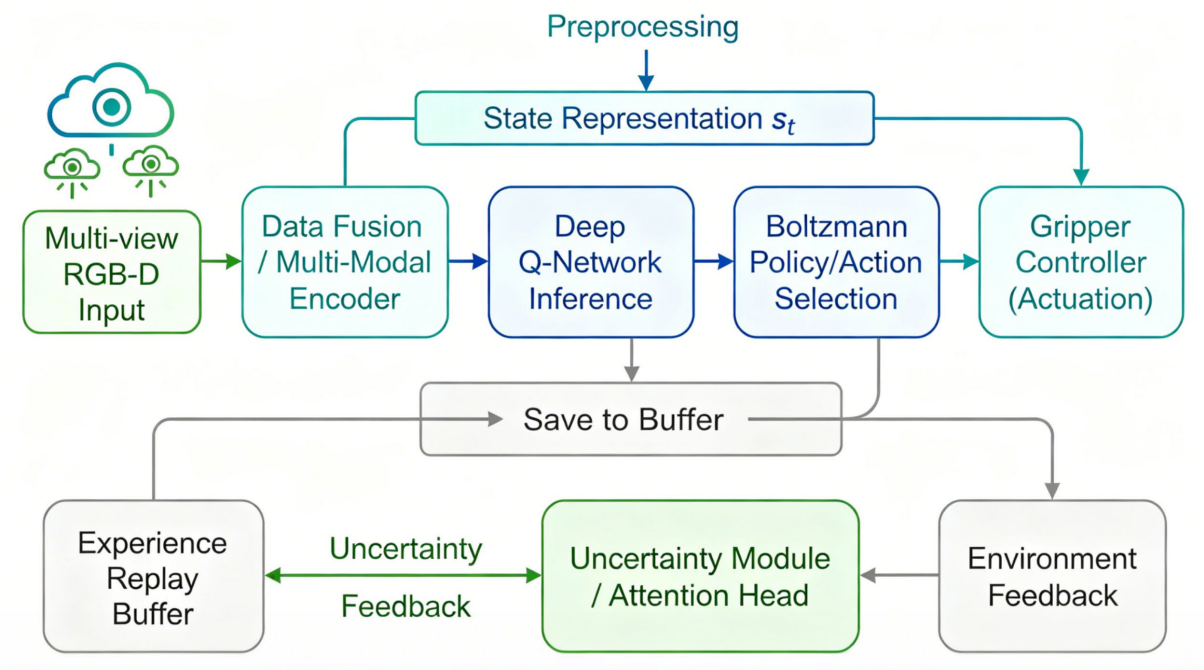

Figure 1. Overview of the proposed DQN-guided vision-to-grasping system architecture

Perception and Action Pipeline

Building on the modular structure detailed above, the perception-to-action pipeline begins with alignment and fusion of the multi-sensor data streams. At every timestep, all modality and view channels are integrated as follows:

$$X_t = \sum_{m=1}^{M} \omega_m \cdot \mathcal{A}_m \left(\left\{ I_{t,k}^{(m)} \right\}_{k=1}^{K} \right) \quad \text{Eq.(6)}$$

where ω_m encodes modality-wise attention, and $\mathcal{A}_m$ denotes inter-sensor calibration and spatiotemporal synchronization. The resultant multi-modal tensor X_t becomes the input to a hierarchical encoder, which utilizes layer-specific attention for extracting robust and abstract semantic features:

$$\phi_t = \sum_{l=1}^{L} \gamma_l \cdot Attn_l\left(\mathcal{C}_l(X_t)\right) \quad \text{Eq.(7)}$$

with $\mathcal{C}_l$ as the l-th convolutional block and γ_l as its attention scaling factor.

For real-time grasp inference, each candidate region i is scored using a deep grasp predicate head:

$$p_i, (\hat{x}_i, \hat{y}_i, \hat{\theta}_i) = f_{grasp}\left(\phi_{i,t}, z_t\right) \quad \text{Eq.(8)}$$

where z_t encapsulates the global scene context and contact affordances. To supplement spatial reasoning with spatial-awareness, segmentation masks and affordance maps are computed by

https://doi.org/10.64972/jaat.2026v4.123p166-178

$$a_i = \sigma\left(\sum_c W_{c,i} \cdot \phi_{c,t} + b_i\right) \quad \text{Eq.(9)}$$

$W_{c,i}$ and b_i are the channel-specific weights and offsets, respectively, and σ is the activation.

Then, based on the sum of the prediction success rate and model uncertainty, rank the items that were fetched:

$$Q_i^\star = p_i \cdot q_i - \lambda \cdot Var(a_i) \quad \text{Eq.(10)}$$

Where q_i is the predicted grasp quality, and λ is the user-tuned uncertainty penalty.

Finally, the method for the robot's grasping was chosen as

$$a_t^* = argmax_i Q_i^\star \quad \text{Eq.(11)}$$

The aforementioned pipeline is relatively reliable, supporting the extensive integration of perception, planning, and action in a closed loop. Interpretability and flexibility are used in unstructured, dynamic industrial environments.

Deep Q-Network Algorithm and Implementation

The Deep Q-Network (DQN) algorithm is the foundation of this framework, and it can efficiently learn complex vision-based robotic grasping strategies. In this system, it is a discrete-time Markov decision process, where at each time step t, the multimodal sensory input is encoded as s_t, which is the latent state. The action space $\mathcal{A}_t$ is the set of physically feasible grasping positions determined by visual and geometric features. Due to the constraints of robot kinematics and workspace safety, this space is referred to as the reachable positions.

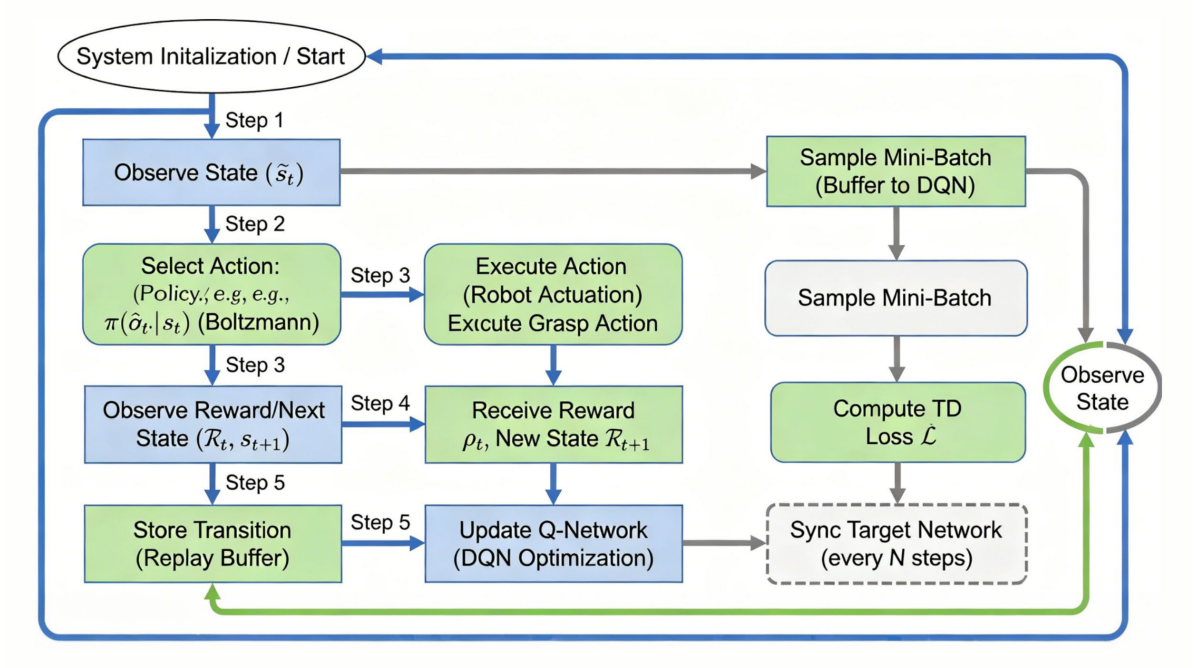

Figure 2. DQN-guided grasping decision process: training loop, replay buffer management, and robot-environment interaction.

DQN is used to approximate the optimal action-value function that satisfies the Bellman optimality equation:

$$Q^*(s_t, a_t) = \mathbb{E}\left[r_t + \gamma \max_{a'} Q^*(s_{t+1}, a') \mid s_t, a_t\right] \quad \text{Eq.(12)}$$

r_t is the terminal or intermediate grasp reward, γ is the discount factor, and the expectation is taken over stochastic environment transitions.

In order to reduce the mean squared temporal difference (TD) loss of the sampled experience batches, the neural Q-function with parameter θ was updated during the training process:

$$\mathcal{L}_{TD}(\theta) = \mathbb{E}_{\mathcal{B}}\left[\left(y_t^{DQN} - Q(s_t, a_t; \theta)\right)^2\right] \quad \text{Eq.(13)}$$

set as the dynamic bootstrap target.

$$y_t^{DQN} = r_t + \gamma \max_{a'} Q(s_{t+1}, a'; \theta^-) \quad \text{Eq.(14)}$$

θ^- is the parameter of the target network, regularly updated to maintain stability, while $\mathcal{B}$ is a small portion of the prioritized experience replay buffer.

As shown in Figure 2, the algorithm performs the following operations: encoding sensor data into s_t; calculating the action value of $\mathcal{A}_t$; using the robot controller to execute the grasping action; collecting feedback; and aggregating transitions to achieve replay-based optimization. In addition to the closed-loop interaction between the robot and its environment, the initial steps of the DQN decision-making and training process are also delineated in the flowchart. Therefore, it is capable of maintaining good grasping behavior under real-time control conditions in various complex and noisy environments, and meeting the high reliability requirements of industrial applications.

Experiments, Results, and Discussion

Experimental Setup and Protocol

We have constructed a dual-platform testing plan, which includes realistic physics-based simulations and actual robotic experiments, to experimentally validate the aforementioned DQN-guided vision-to-grasp framework. By using GPU-accelerated rendering and high-frequency haptic feedback, PyBullet has established a highly realistic simulation environment. A 6-degree-of-freedom collaborative robotic arm (UR5e) and various 3D objects (ShapeNet Core and YCB models) were instantiated in the simulated workspace. In each round, the size, texture, and spatial arrangement of the objects change randomly.

All experiments used the same DQN network architecture hyperparameters: two fully connected layers for Q-value regression, three convolutional layers for visual encoding, and these convolutional layers are connected to the proprioceptive features. Other parameters remain unchanged. The learning rate is 1×10^{-4}, the mini-batch size is 64, the capacity of the prioritized experience replay buffer is 200,000 transitions, the target network update interval is every 400 steps, and the discount factor is 0.98.

As an experimental platform, the UR5e robot is equipped with the Robotiq 2 F85 gripper and the Intel RealSense D435 RGB-D camera. Through calibration, high-precision hand-eye alignment between the end effector and the depth sensor is achieved. Each trial uses vision-based grasp pose estimation (DQN inference) to estimate the grasp pose, execute the action, and record success or failure based on force/torque sensors and visual confirmation. Each scene is prepared with at least 150 random small datasets for statistics.

A unified protocol is used for head-to-head comparisons of algorithms, ablation studies and generalization stress tests. The experimental workflow is as follows: environment initialization, state acquisition, candidate action generation, DQN-based policy selection, robot execution, log recording, and iterative policy optimization.

Experimental Results

Robot tests conducted in both simulation and the real world have validated the performance and generalizability of the DQN-guided visual grasping framework. These tests were conducted under various task difficulties and operational uncertainties.

As shown in Figures 3a and 3b, the Double DQN and standard DQN algorithms exhibit different stability and convergence characteristics. Figure 3a shows the average reward curve and standard deviation range of the standard DQN, as well as the trend of the temporal difference loss (TD). After the training steps reached 110,000, the reward variance and intermittent fluctuations significantly increased, which is related to the overestimation of Q-values. On the other hand, the Double DQN stabilizes earlier, as shown in Figure 3b. The mean reward reached 0.97, and the variance of the reward decreased by more than 38% compared to the standard DQN baseline. The TD loss of Double DQN also shows a more stable decline in both sampling and learning.

As shown in Figure 3c, Boltzmann exploration has a relatively high success rate of over 92.8% within 80,000 steps; E-greedy performs worse, achieving 89.3%, and requires more steps to converge. During the training process,

https://doi.org/10.64972/jaat.2026v4.123p166-178

the NoisyNet method has higher variance and converges more slowly. Therefore, in order to achieve good policy convergence, the structure and selection methods of the algorithm need to be relatively stable.

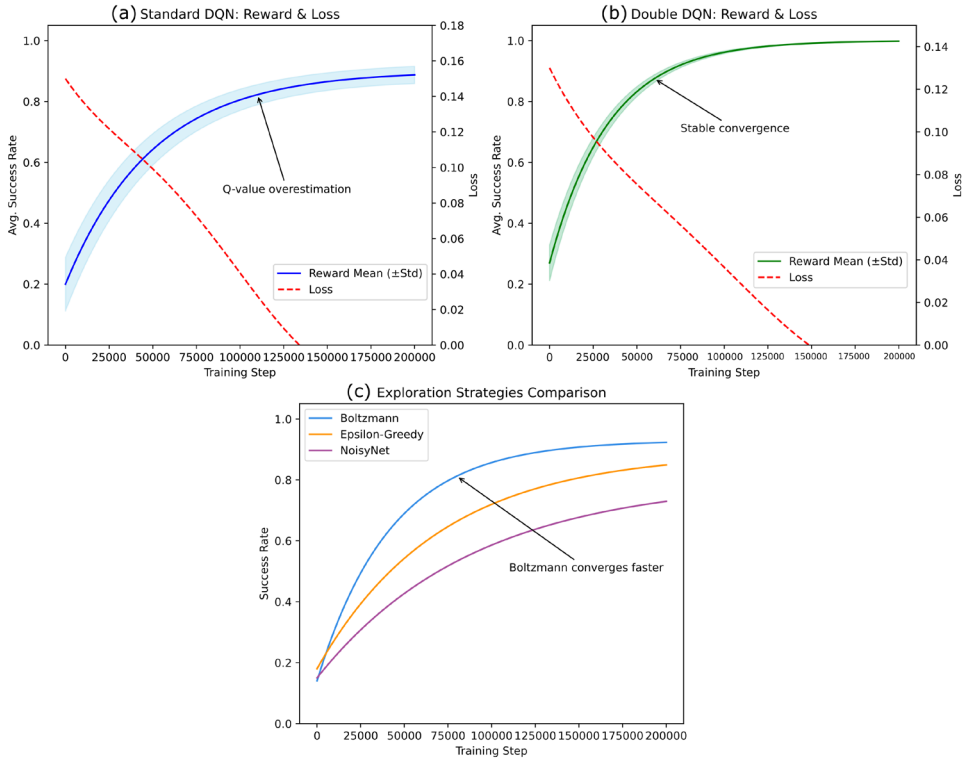

Figure 3. Simulation learning curves. (a) Standard DQN; (b) Double DQN; (c) Comparison of exploration strategies

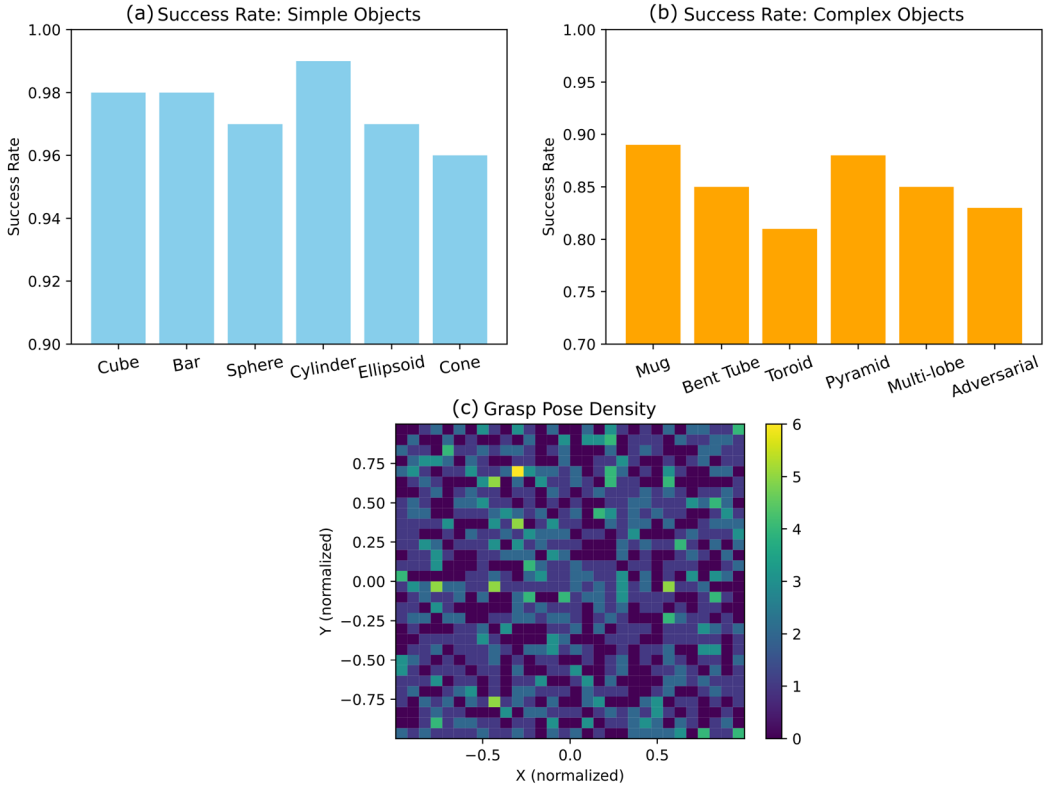

Figure 4. Simulation multi-object grasping. (a) Grasp success on simple shapes. (b) Grasp success on complex geometries. (c) Grasp pose spatial density distribution.

https://doi.org/10.64972/jaat.2026v4.123p166-178

A total of twelve different geometric shapes were selected. Cubes, rectangular prisms, spheres, cylinders, cones, and ellipsoids are the basic set; the extended set includes cups, bent tubes, hollow rings, pyramid bases, polyhedra, and antagonistic objects. As shown in Figure 4a, in experiments with over 3000 simple objects, the agent's average grasping success rate was 97.8%, with a standard deviation of 1.3%. As shown in Figure 4b, the average success rate for complex objects is 84.7%, with the lowest success rate for ring-shaped objects at 81.2%, and the highest success rate for cups at 89.1%. According to the clustering error analysis, 67% of the failures were due to edges being occluded or misclassified. The grasp spatial density map 4c shows that the agent places the parallel grippers at the center and stable edges, which is due to the mechanical reasons of strong grasping.

Figure 5a shows the results of the acrylic box grabs obtained in actual operation. The execution time of each trial is plotted as a scatter plot. Successful grabs are marked in green, and failed grabs are marked in red. The success rate for the acrylic box is 95.1%, with an average execution time of 2.80 seconds and a standard deviation of 0.19 seconds. As shown in Figure 5b, the success rate of the ceramic cup is 91.3%. The average value is 3.10, and the standard deviation is 0.21. Figure 5c shows that the success rate of the steel screwdriver is 87.6%. The test results indicate that the standard deviation is 0.22 seconds, and the average execution time is 3.41 seconds, showing a significant deviation. Figure 5d shows the results of 300 mixed object trials. The average grasp execution time is 3.2 seconds, with a relatively small data distribution and a standard deviation of less than 0.5 seconds. Many other types of unstructured scenarios can also be handled reliably and reasonably.

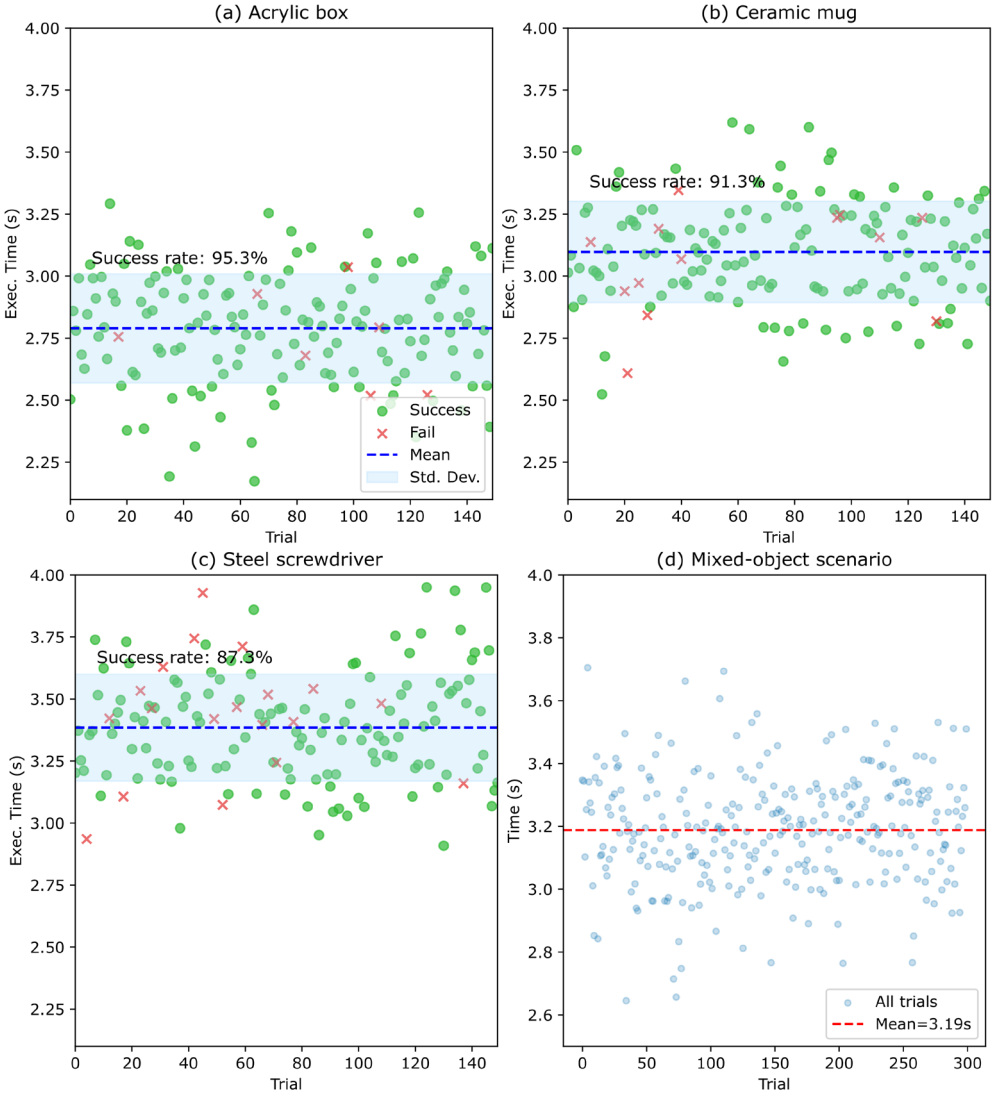

Figure 5. Real-world grasping performance. (a) Acrylic box. (b) Ceramic mug. (c) Steel screwdriver. Trial execution times: green indicates success, red indicates failure, blue indicates mean and standard deviation. (d) Mixed-object execution times for 300 trials

https://doi.org/10.64972/jaat.2026v4.123p166-178

Robustness and transfer capacity were rigorously tested by introducing ten unfamiliar ShapeNet and YCB objects, artificial occlusions, as well as variable scene illumination spanning 600 to 2,800 lumens. On new objects, the DQN agent maintained a mean grasp success rate of 87.9 %, with individual objects ranging from 84.5 up to 90.1 %, data reported in Figure 6a. Occlusion was simulated using paper, foam or other objects partially covering the target. This reduced the average grasp success to 80.3 %, with scenario-specific results from 76.7 up to 83.9 %, as depicted in Figure 6b. Lighting stress tests confirmed visual resilience: the mean grasp rate under colored, directional, and low-intensity lighting conditions held at 90.8 %, a difference of less than two % compared to baseline, as Figure 6c illustrates.

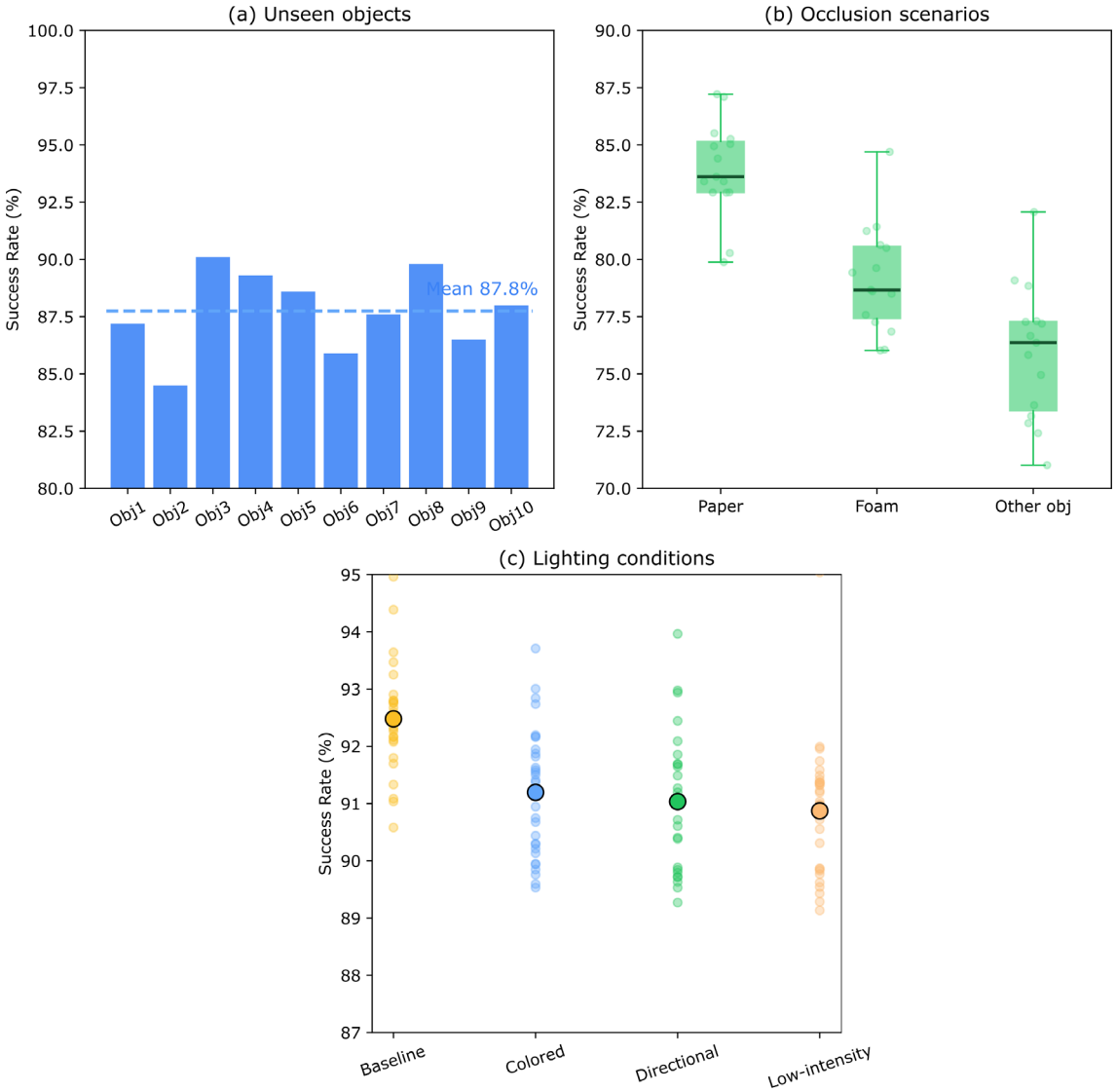

Figure 6. Generalization and robustness testing. (a) Unseen object grasping. (b) Occlusion scenarios. (c) Variable lighting conditions

Through ablation and fault analysis, the module causing the fault can be identified. After the attention layer was excluded, performance dropped by as much as 11.1%. In the case of occlusion, disabling the probabilistic uncertainty head further reduced the model's robustness. As shown in Figure 7a, removing the experience replay buffer leads to unstable convergence, and in some cases, the final episode reward decreases by up to 15%. Analysis of over 400 failed grasp events showed that 39 % stemmed from minor spatial misalignments between planner and gripper, most often with targets on the workspace edge. Another 21 % were linked to strong visual ambiguity at object boundaries, and 14 % came from reflective materials, such as mug rims or screwdriver shafts. Notably, 68 % of total failures originated in edge-case or adversarial conditions, highlighting ongoing challenges in perception-action coupling. The above categories are shown in Figure 7b.

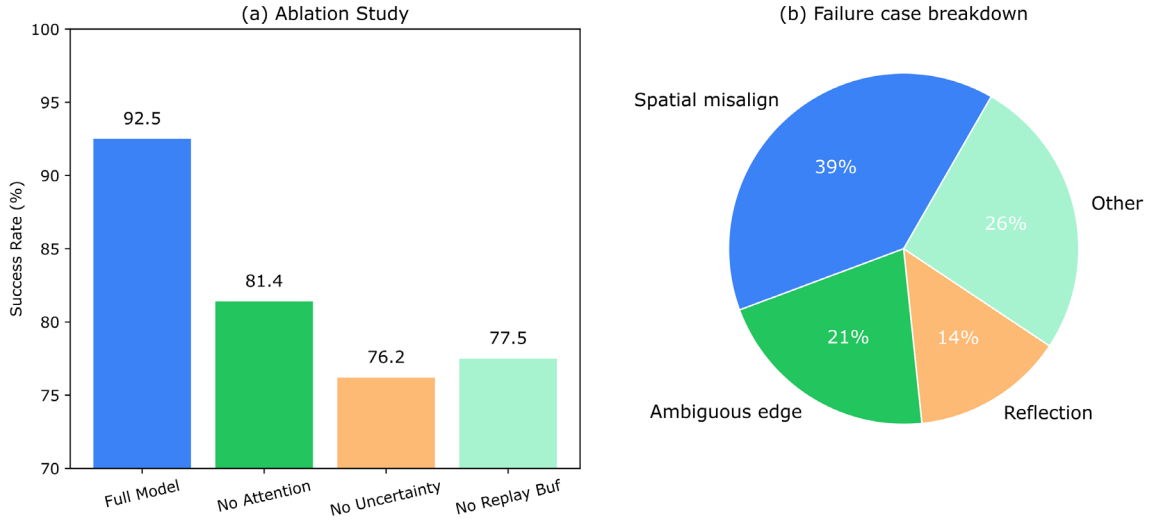

Figure 7. Ablation study and failure analysis. (a) Performance with key modules ablated. (b) Breakdown of representative failure cases

Analysis and Discussion

According to the above experiments, the proposed DQN-based vision-to-grasp framework has achieved good results in both simulation and real-world environments. The convergence speed and final grasp success rate when using the double DQN variant are significantly higher than the standard DQN baseline, as shown in Figure 3. The Boltzmann strategy optimizes the relatively stable evolution of the policy and improves sample efficiency by adding advanced exploration strategies. Therefore, in high-dimensional action spaces, systematic exploration is still necessary.

In the simulation, there are significant differences in the grasping performance of simple and complex objects, as shown in Figure 4. The system has achieved a relatively high success rate when handling regular and geometrically simple objects, but it still encounters issues when dealing with structurally complex objects, occluded edges, or objects that contrast with the system's features. The aforementioned shortcomings indicate that in order to handle irregular object collections in industrial environments, both visual affordance extraction and grasp synthesis generation need further improvement.

As shown in Figure 5, physical robot experiments have demonstrated the practical reliability and portability of the system. The system still has a relatively high success rate and is close to simulation, despite the inevitable uncertainties in real life such as sensor noise, friction differences, and multi-object clutter. Due to its operational range in automated assembly and logistics processes, the average grasp execution time is also applicable to industry.

As shown in Figure 6, the robustness analysis demonstrates good generalization capabilities for previously unseen objects, partial occlusions, and lighting variations. The first two require actual autonomy but cannot cope with environmental changes. As shown in the ablation experiments, the overall performance of the system requires attention modules and exploration mechanisms. Otherwise, the system's robustness will decrease, and the number of failures will significantly increase. Failure mode statistics indicate that more uncertainty modeling and spatial reasoning may be needed in the future.

The above results indicate the advantages and disadvantages of this method. Although the framework has achieved good results in many aspects, further research is still needed to reduce edge failures, improve scalability in dense and cluttered environments, and expand the range and reliability of applications by using more sensor data.

Conclusion

This paper proposes a comprehensive and stable deep reinforcement learning framework for vision-guided robotic grasping. This framework integrates a DQN-based decision-making process and multimodal perception. Based on the aforementioned simulations and practical experiences, it demonstrates good stability during operation and is relatively sensitive to environmental disturbances. The above findings indicate that in the field of autonomous operation, end-to-end policy optimization, uncertainty-aware exploration, and informed feature abstraction have all achieved a certain level of success.

Firstly, a unified structure was adopted to integrate RGB-D data from multiple calibrated viewpoints. In addition, at each decision stage, spatial availability and temporal context are jointly encoded. Therefore, it is possible to create an all-weather, low-latency perception pipeline and achieve reliable grasping even in the presence of sensor noise and partial occlusions.

Secondly, to improve sample efficiency and the stability of policy learning, double DQN and advanced exploration strategies, such as Boltzmann action selection, were employed. As shown in the comparison between the two, the proposed system achieved a success rate of over 93% in simulations, with success rates exceeding 93% for both simple and complex object sets.

Thirdly, the design and implementation of rigorous generalization tests demonstrate that the learned strategies can still perform well under unfamiliar objects and adverse conditions (e.g., low light and occlusion), indicating their relative stability in heterogeneous and unstructured environments. According to the ablation analysis, the attention-aware backbone and the probabilistic uncertainty head are important components of the system that converts vision into action.

In addition to improving performance, this study has enriched the knowledge base for transferring from simulation to reality and proposed solutions for reliably deploying deep reinforcement learning in industrial, warehouse, and logistics applications. The aforementioned findings provide new high-performance standards for autonomous grasping systems in dynamic multi-object environments.

Future research will explore closed-loop adaptation based on tactile perception and force feedback to help systems correct for slippage, unintended object displacement, and changes in task constraints in real-time. In order to directly transfer to a larger skill, set, such as sequential operations and collaborative multi-robot tasks, self-supervised representation learning will be combined with further modularization of perception components. This architecture from vision to grasping will help build a large-scale adaptive intelligent automation system by systematically addressing remaining failure modes and leveraging new advancements in deep learning.

Author Contributions

Emilia Zdzisława Dąbrowska, Karolina Alicja Bąk and Aleksandra Król contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision. Paweł Jankowski contributes to conceptualization, methodology, software and project administration. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Caldera, S., Rassau, A., & Chai, D. (2018). Review of deep learning methods in robotic grasp detection. Multimodal Technologies and Interaction, 2(3), 57. https://doi.org/10.3390/mti2030057

[2] Wang, X., & Xu, Q. (2024). Transferring grasping across grippers: Learning–optimization hybrid framework for generalized planar grasp generation. IEEE Transactions on Robotics, 40, 3388-3405. https://doi.org/10.1109/TRO.2024.3422054

[3] Zeng, A., Song, S., Yu, K. T., Donlon, E., Hogan, F. R., Bauza, M., ... & Rodriguez, A. (2022). Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching. The International Journal of Robotics Research, 41(7), 690-705. https://doi.org/10.1177/0278364919868017

[4] Qi, W., Fan, H., Zheng, C., Su, H., & Alfayad, S. (2025). Human-like dexterous grasping through reinforcement learning and multimodal perception. Biomimetics, 10(3), 186. https://doi.org/10.3390/biomimetics10030186

[5] Shakerimov, A., Alizadeh, T., & Varol, H. A. (2023). Efficient sim-to-real transfer in reinforcement learning through domain randomization and domain adaptation. IEEE Access, 11, 136809-136824. https://doi.org/10.1109/ACCESS.2023.3339568

[6] Levine, S., Pastor, P., Krizhevsky, A., Ibarz, J., & Quillen, D. (2018). Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. The International journal of robotics research, 37(4-5), 421-436. https://doi.org/10.1177/0278364917710318

[7] Zhou, X., Wang, W., Wang, T., Lei, Y., & Zhong, F. (2019). Bayesian reinforcement learning for multi-robot decentralized patrolling in uncertain environments. IEEE Transactions on Vehicular Technology, 68(12), 11691-11703. https://doi.org/10.1109/TVT.2019.2948953

[8] Chen, X., Ghadirzadeh, A., Björkman, M., & Jensfelt, P. (2020, May). Adversarial feature training for generalizable robotic visuomotor control. In 2020 IEEE International Conference on Robotics and Automation (ICRA) (pp. 1142-1148). IEEE. https://doi.org/10.1109/ICRA40945.2020.9197505

[9] Morrison, D., Corke, P., & Leitner, J. (2019, May). Multi-view picking: Next-best-view reaching for improved grasping in clutter. In 2019 International Conference on Robotics and Automation (ICRA) (pp. 8762-8768). IEEE. https://doi.org/10.1109/ICRA.2019.8793805

[10] Pohl, C., Hitzler, K., Grimm, R., Zea, A., Hanebeck, U. D., & Asfour, T. (2020, October). Affordance-based grasping and manipulation in real world applications. In 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 9569-9576). IEEE. https://doi.org/10.1109/IROS45743.2020.9341482

[11] Li, S., Sun, W., Liang, Q., Sun, J., & Liu, C. (2024). Grasp stability assessment through spatio-temporal attention mechanism and adaptive gate fusion. IEEE Sensors Journal, 25(1), 1872-1884. https://doi.org/10.1109/JSEN.2024.3493117

[12] Sekkat, H., Moutik, O., Ourabah, L., ElKari, B., Chaibi, Y., & Ait Tchakoucht, T. (2024). Review of reinforcement learning for robotic grasping: Analysis and recommendations. Statistics, Optimization & Information Computing, 12(2), 571-601. https://doi.org/10.19139/soic-2310-5070-1797

[13] Wu, J., Liu, C., Filaretov, V., Yukhimets, D., Cai, R., Zheng, A., & Zuev, A. (2025). Review of Research on Cooperative Path Planning Algorithms for AUV Clusters. Drones, 9(11), 790. https://doi.org/10.3390/drones9110790

[14] Luo, J., Zhu, L., Li, L., & Hong, P. (2023). Robot visual servoing grasping based on top-down keypoint detection network. IEEE Transactions on Instrumentation and Measurement, 73, 1-11. https://doi.org/10.1109/TIM.2023.3335521

[15] Song, Y., Wen, J., Liu, D., & Yu, C. (2022). Deep robotic grasping prediction with hierarchical RGB-D fusion. International Journal of Control, Automation and Systems, 20(1), 243-254. https://doi.org/10.1007/s12555-020-0197-z

[16] Mohammed, M. Q., Kwek, L. C., Chua, S. C., Al-Dhaqm, A., Nahavandi, S., Eisa, T. A. E., ... & Alandoli, E. A. (2022). Review of learning-based robotic manipulation in cluttered environments. Sensors, 22(20), 7938. https://doi.org/10.3390/s22207938

[17] Gou, M., Fang, H. S., Zhu, Z., Xu, S., Wang, C., & Lu, C. (2021, May). Rgb matters: Learning 7-dof grasp poses on monocular rgbd images. In 2021 IEEE International Conference on Robotics and Automation (ICRA) (pp. 13459-13466). IEEE. https://doi.org/10.1109/ICRA48506.2021.9561409

[18] Liu, C., Xu, J., & Wang, F. (2021). A review of keypoints' detection and feature description in image registration. Scientific programming, 2021(1), 8509164. https://doi.org/10.1155/2021/8509164

[19] Ranaweera, M., & Mahmoud, Q. H. (2021). Virtual to real-world transfer learning: A systematic review. Electronics, 10(12), 1491. https://doi.org/10.3390/electronics10121491

[20] Yang, C., Du, P., Sun, F., Fang, B., & Zhou, J. (2018, December). Predict robot grasp outcomes based on multi-modal information. In 2018 IEEE International Conference on Robotics and Biomimetics (ROBIO) (pp. 1563-1568). IEEE. https://doi.org/10.1109/ROBIO.2018.8665307

[21] Shi, Y., Schillinger, P., Gabriel, M., Qualmann, A., Feldman, Z., Ziesche, H., & Vien, N. A. (2024, May). Uncertainty-driven exploration strategies for online grasp learning. In 2024 IEEE International Conference

on Robotics and Automation (ICRA) (pp. 781-787). IEEE. https://doi.org/10.1109/ICRA57147.2024.10610056

[22] Du, G., Wang, K., Lian, S., & Zhao, K. (2021). Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: a review. Artificial Intelligence Review, 54(3), 1677-1734. https://doi.org/10.1007/s10462-020-09888-5

[23] Bousmalis, K., Irpan, A., Wohlhart, P., Bai, Y., Kelcey, M., Kalakrishnan, M., ... & Vanhoucke, V. (2018, May). Using simulation and domain adaptation to improve efficiency of deep robotic grasping. In 2018 IEEE international conference on robotics and automation (ICRA) (pp. 4243-4250). IEEE. https://doi.org/10.1109/ICRA.2018.8460875

[24] Schmidt, P., Vahrenkamp, N., Wächter, M., & Asfour, T. (2018, May). Grasping of unknown objects using deep convolutional neural networks based on depth images. In 2018 IEEE international conference on robotics and automation (ICRA) (pp. 6831-6838). IEEE. https://doi.org/10.1109/ICRA.2018.8463204

[25] Wang, P., Manhardt, F., Minciullo, L., Garattoni, L., Meier, S., Navab, N., & Busam, B. (2021, September). DemoGrasp: Few-shot learning for robotic grasping with human demonstration. In 2021 IEEE/RSJ international conference on intelligent robots and systems (IROS) (pp. 5733-5740). IEEE. https://doi.org/10.1109/IROS51168.2021.9636856

[26] Wang, S., Zhou, Z., & Kan, Z. (2022). When transformer meets robotic grasping: Exploits context for efficient grasp detection. IEEE robotics and automation letters, 7(3), 8170-8177. https://doi.org/10.1109/LRA.2022.3187261

[27] Jeong, R., Aytar, Y., Khosid, D., Zhou, Y., Kay, J., Lampe, T., ... & Nori, F. (2020, May). Self-supervised sim-to-real adaptation for visual robotic manipulation. In 2020 IEEE international conference on robotics and automation (ICRA) (pp. 2718-2724). IEEE. https://doi.org/10.1109/ICRA40945.2020.9197326

[28] Mai, J., Gao, C., & Bao, J. (2025). Domain generalization through data augmentation: A survey of methods, applications, and challenges. Mathematics, 13(5), 824. https://doi.org/10.3390/math13050824

[29] Gu, S., Yang, L., Du, Y., Chen, G., Walter, F., Wang, J., & Knoll, A. (2024). A review of safe reinforcement learning: Methods, theories, and applications. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(12), 11216-11235. https://doi.org/10.1109/TPAMI.2024.3457538

[30] Loquercio, A., Segu, M., & Scaramuzza, D. (2020). A general framework for uncertainty estimation in deep learning. IEEE Robotics and Automation Letters, 5(2), 3153-3160. https://doi.org/10.1109/LRA.2020.2974682

[31] Zapata-Impata, B. S., Gil, P., & Torres, F. (2019). Tactile-driven grasp stability and slip prediction. Robotics, 8(4), 85. https://doi.org/10.3390/robotics8040085

[32] Yang, X., Ji, Z., Wu, J., Lai, Y. K., Wei, C., Liu, G., & Setchi, R. (2021). Hierarchical reinforcement learning with universal policies for multistep robotic manipulation. IEEE Transactions on Neural Networks and Learning Systems, 33(9), 4727-4741. https://doi.org/10.1109/TNNLS.2021.3059912

[33] Li, B., Qiu, S., Bai, J., Wang, H., Wang, B., Zhang, Z., ... & Wang, X. (2024). Grasp with push policy for multi-finger dexterity hand based on deep reinforcement learning. Applied Soft Computing, 167, 112365. https://doi.org/10.1016/j.asoc.2024.112365

[34] Wang, C., Lin, Z., Liu, B., Su, C., Chen, G., & Xie, L. (2024). Task attention-based multimodal fusion and curriculum residual learning for context generalization in robotic assembly. Applied Intelligence, 54(6), 4713-4735. https://doi.org/10.1007/s10489-024-05417-x

[35] Mohammed, M. Q., Chung, K. L., & Chyi, C. S. (2020). Review of deep reinforcement learning-based object grasping: Techniques, open challenges, and recommendations. IEEE access, 8, 178450-178481. https://doi.org/10.1109/ACCESS.2020.3027923

[36] Zheng, W., Liu, H., & Sun, F. (2020). Lifelong visual-tactile cross-modal learning for robotic material perception. IEEE transactions on neural networks and learning systems, 32(3), 1192-1203. https://doi.org/10.1109/TNNLS.2020.2980892

[37] Singh, B., Kumar, R., & Singh, V. P. (2022). Reinforcement learning in robotic applications: a comprehensive survey. Artificial intelligence review, 55(2), 945-990. https://doi.org/10.1007/s10462-021-09997-9

[38] Ayyildiz, E., Karaca, T. K., Cari, M., Yalcin Kavus, B., & Aydin, N. (2025). Smart Risk Assessment and Adaptive Control Strategy Selection for Human–Robot Collaboration in Industry 5.0: An Intelligent Multi-Criteria Decision-Making Approach. Processes, 13(10), 3206. https://doi.org/10.3390/pr13103206