2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

# Uncertainty-driven Exploration Strategies for Online Grasp Learning

Yitian Shi[1,2], Philipp Schillinger[2], Miroslav Gabriel[2], Alexander Qualmann[2], Zohar Feldman[2], Hanna Ziesche[2], Ngo Anh Vien[2]

***Abstract*— Existing grasp prediction approaches are mostly based on offline learning, while, ignoring the exploratory grasp learning during online adaptation to new picking scenarios, i.e., objects that are unseen or out-of-domain (OOD), camera and bin settings, etc. In this paper, we present an uncertainty-based approach for online learning of grasp predictions for robotic bin picking. Specifically, the online learning algorithm with an effective exploration strategy can significantly improve its adaptation performance to unseen environment settings. To this end, we first propose to formulate online grasp learning as an RL problem that will allow us to adapt both grasp reward prediction and grasp poses. We propose various uncertainty estimation schemes based on *Bayesian uncertainty quantification* and *distributional ensembles*. We carry out evaluations on real-world bin picking scenes of varying difficulty. The objects in the bin have various challenging physical and perceptual characteristics that can be characterized by semi- or total transparency, and irregular or curved surfaces. The results of our experiments demonstrate a notable improvement of grasp performance in comparison to conventional online learning methods which incorporate only naive exploration strategies. Video: https://youtu.be/fPKOrjC2QrU**

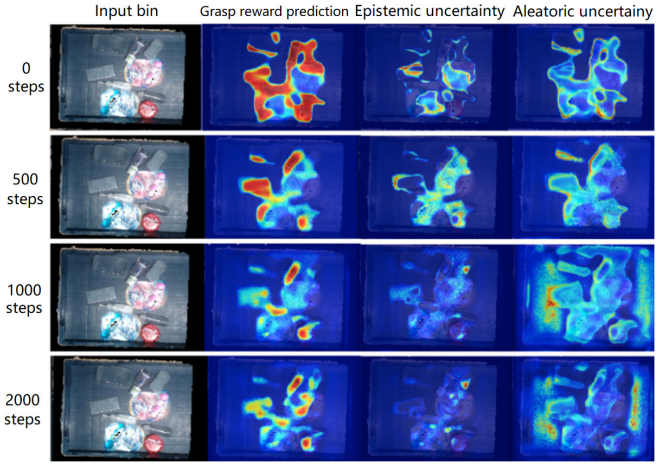



Fig. 1: Given an image of a picking scene (1st column), every 500 training steps our MV-ConvSACs predicts a grasp reward map (2nd column) and a normalized uncertainty map (3rd/4th columns).

## I. Introduction

Robotic picking or object grasping is a challenging task in robotics that requires the agent to select and execute the optimal grasp poses for various objects from observations of complex scenes, where the agent must search for an optimal grasping strategy in an exponentially large state space due to the inherent variability in object shapes, opaqueness, and materials, and in certain cases also due to the various characteristics of the robot's sensors and actuators. These challenges represent significant obstacles to designing hand-engineered algorithms as done in traditional approaches [1]. To overcome those challenges, modern grasping methods have successfully applied advanced deep learning techniques that enable model-free grasp predictions for a wide variety of objects in unstructured environments [2], [3]. However, methods [4]–[6] are based on supervised learning and offline training, and can not generalize well to OOD objects or new environment settings.

To address these challenges, this paper focuses on tackling the problem of online grasp learning. It does so by leveraging exploration capabilities to systematically search the space of grasp configurations to find the best grasps for picking objects from an OOD bin. Those objects are characterized by unknown physical and perceptual features such as semi- or total transparency, and irregular or curved surfaces.

In particular, we propose a study with a realistic assumption that online grasp learning with exploration capabilities can be used to fine-tune or adapt a pretrained grasping network to a set of OOD objects. We focus on data-efficient offline-to-online learning [7], aiming to enable the agent to quickly learn from a limited amount of real-world interactions given a grasping network that was pretrained using small-scale offline learning. To this end, we propose to formulate online grasp learning as a reinforcement learning (RL) problem and to leverage the Convolutional Soft Actor-Critic (ConvSAC) algorithm [8] to further update the policy. We propose various exploration strategies that are based on uncertainty estimation to enhance the data efficiency and the domain generalization of the online learning process. To achieve this, we utilize Bayesian uncertainty estimation [9] and quantile regression [10], [11] to compute a pixel-wise uncertainty map for a given input image. The uncertainty map allows the agent to actively choose the next picks that reduce ambiguity or uncertainty on the grasping scene while improving the grasp success in the long run, as depicted in Fig. 1. To summarize, our main contributions are as follows:

- The formulation of the exploration problem for online grasp learning allows us to integrate various exploration strategies into existing RL-based grasp learning methods to improve grasp performance online.
- New architectures for uncertainty estimation and princi-

[1]Karlsruhe Institute of Technology, Karlsruhe, Germany. Email: yitian.shi@kit.edu. This work is completed during the author's affiliation with the University of Stuttgart, Stuttgart, Germany.
[2]Bosch Center for Artificial Intelligence, Renningen, Germany. Email: firstname.lastname@de.bosch.com

pled exploration strategies specifically for bin picking. These exploration strategies are based on pixel-wise uncertainty maps that can be computed using Bayesian uncertainty estimation or distributional regression.
- Experiments and ablation studies on a real-world bin picking setup to i) demonstrate the proposed approach, and ii) understand the role of different types of uncertainties, i.e. epistemic and aleatoric.

## II. Related Works

### A. Modern Robot Grasping Methods

Modern robot grasping methods often employ deep learning techniques trained on extensive datasets to predict grasps [12], [13]. Most strategies depend on generating a grasp reward map where each pixel represents the likelihood of a successful grasp at this location. For instance, Mahler et al. [12] suggest the prediction of a grasp map for suction and parallel-jaw grasps through supervised datasets using RGB-D or depth as input. Morrison et al. [14] and Satish et al. [15] adopt this approach to predict both pixel-wise grasp reward maps and 4-DoF parallel-jaw grasp configurations. Alternative approaches utilize point clouds [16]–[21] or voxels [22] as input and predict dense grasp qualities and gripper configurations. In a subsequent study, grasp prediction is trained jointly with object shape reconstruction [23]. Recent works introduce the concept of pixel-wise grasp maps and grasp configuration predictions for single-suction grippers [24] and multi-suction cup grippers [25].

### B. Deep RL-based Grasping Methods

Vision-based RL has emerged as a promising approach to robotic grasping, which involves using visual information to guide robot actions through RL networks [26]–[28] and are often end-to-end and optimize close-loop policies for grasp planning from raw visual inputs. One of the primary obstacles lies in the requirement of substantial quantities of training data of exceptional quality, owing to the high dimensionality of visual [29] or depth [30] input. Additionally, the algorithm must effectively extrapolate the acquired knowledge to unknown scenarios while maintaining its efficient performance. Open-loop RL methods are also used for 6-DoF bin picking scenarios [8], [31]–[33] which need a substantial number of updates to achieve ideal performance. However, these works often resort to training from scratch and use a standard exploration policy, i.e. a policy entropy bonus, hence they require an extensive amount of online samples.

Several works have proposed more principled exploration strategies for online grasp learning [4], [5], [34]–[39], but so far singulated object scenes were addressed and their extensions to bin picking is non-trivial. Alternatively, meta-learning or few-shot learning have also been applied for online learning for bin picking [40]–[42]. These works show that the learned grasp can be quickly adapted to OOD objects. However, they require a few shots of context grasps provided either by an oracle or by executing passive actions. In contrast, our approach gathers these context grasp points via an actively exploring policy.

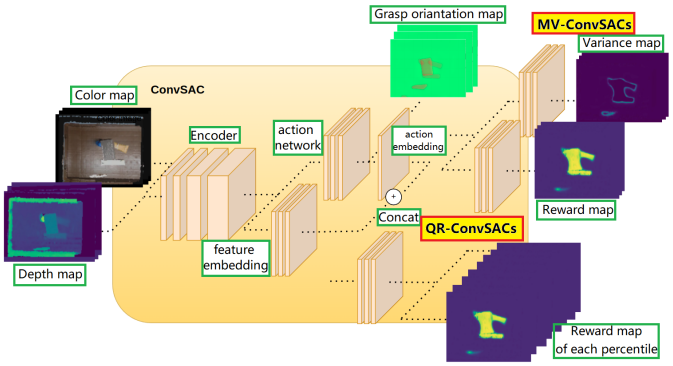



Fig. 2: Ensemble of probabilistic ConvSAC network architecture: **MV-ConvSACs** and **QR-ConvSACs**.

### C. Uncertainty Estimation in Deep RL

Uncertainty estimation is a research focus in various machine learning and deep learning subfields. Kendall et al. classify uncertainties into epistemic and aleatoric uncertainties [43]. Epistemic uncertainty arises from the model's lack of knowledge or information. Aleatoric uncertainty, on the other hand, is related to the inherent stochasticity (or noise) in the data. In the RL domain, Charpentier et al. [44] introduced desiderata influenced by supervised uncertainty estimation, encompassing both aleatoric and epistemic uncertainties. Lee et al. [45] propose deep ensemble techniques for the Soft Actor-Critic (SAC) algorithm [46], utilizing uncertainty estimates to re-weight sample transitions during policy updates. By incorporating uncertainty measures, the agents gain insights into the reliability of the collected data. Clements et al. [11] focused on estimating and disentangling uncertainties in distributional RL agents by estimating uncertainty at various quantiles [10], enabling more precise and reliable estimates of uncertainty in the RL setting. Additionally, methods based on Kalman filters as discussed in [47], have also demonstrated their success in Bayesian uncertainty estimation.

## III. Uncertainty-driven Offline-to-Online Robotic Grasp Learning

We consider the following online grasp learning problem setting. Given an RGB-D image of the scene, our goal is to predict multi-channel maps for a suction gripper: a *pixel-wise grasp reward map*, a *pixel-wise grasp orientation (rotation) map* and a *pixel-wise uncertainty (variance) map*. These prediction maps can be used to derive an exploration grasping action. After each grasp, the network receives sparse reward feedback (success or failure) and is updated accordingly.

### A. Problem Formulation

Here we suppose to model the online grasp learning problem on a bin picking setting as an MDP $(\mathcal{S}, \mathcal{A}, \mathcal{T}, r)$ with state space $\mathcal{S}$, action space $\mathcal{A}$, transition function $\mathcal{T}$ and reward function $r$. In each step, the system observes a state $s_t \in S$, an action $a_t \in \mathcal{A}$ is taken by the policy $\pi(a_t|s_t)$

and the reward is received from the environment $r(s_t, a_t)$. A new state $s_{t+1}$ follows upon based on the transition $\mathcal{T}$ from the environment. In the setting of [8], the state $s_t$ is represented by a set that is composed of one color image, normal map, and height map $s_t = (I_c, I_n, I_d)_t$ with $I_c \in \mathbb{R}^{H*W*3}$, $I_n \in \mathbb{R}^{H*W*3}$ and $I_d \in \mathbb{R}^{H*W*1}$. In our experiments, the states are captured by a stereo sensor with a top-down view of the object bin. The action $a_t$ corresponds to a three-dimensional orientation represented by Euler angles $(\alpha_t, \beta_t, \gamma_t)$ and Cartesian coordinates $(x_t, y_t, z_t)$. Due to the rotational symmetry of our suction gripper ($\gamma_t$ ignored), we define the grasp action as $a_t = (x_t, y_t, \alpha_t, \beta_t) \in \mathcal{A}$ since $z_t$ is directly extracted from the height map. The reward $r_t$ is 1 when a successful grasp is executed or 0 otherwise. We aim to optimize a policy $\pi(s) : \mathcal{S} \mapsto \mathcal{A}$ that maximizes the total grasp success return $\sum_t r_t$.

We follow a similar framework as proposed by ConvSAC [8] and HACMan [48] to create network architectures for continuous actions, as well as $Q$-learning for discrete actions. In particular, the **Actor** module $\pi(s)$ infers pixel-wise Gaussian actions, resulting in an action map denoted as $A$. These actions are then concatenated with the embedding of the input image and evaluated by a **Critic** module, resulting in a $Q$-value map denoted as $Q$, namely *grasp reward map*.

### B. Ensemble of Probabilistic ConvSAC

Here we propose an extension of ConvSAC to model the critic's uncertainty using two strategies. The *first strategy* is to model heteroscedastic aleatoric uncertainty following the Gaussian likelihood [43] of each pixel of the Q-value map. The *second strategy* is to predict the Q-value for each pre-defined distributional quantile [11] individually. Figure 2 shows our proposed network architectures. The training of the actors is similar to the ConvSAC's actor training procedure, where we implement a probabilistic action as $\pi_\theta(s) \sim \mathcal{N}(A_\mu(s), A^2_\sigma(s))$.

*1) Gaussian-based Uncertainty Estimation:* Suppose we have the input data $s$, the network will encode it and generate the grasp orientation map $A_\mu(s) \in \mathbb{R}^{H*W*3}$ for each pixel $D = \{a_i(s)\}_{i<H*W, i\in\mathbb{N}}$. To criticize the pixel-wise action $a_i(s)$, the critic network will take the embedding from both the input feature as well as the action map and finally output the reward map or $Q$-value map as well as the variance map, denoted as $Q(s, A_\mu(s)) \in \mathbb{R}^{H*W*1}$ and $\mathrm{Var}(s, A_\mu(s)) \in \mathbb{R}^{H*W*1}$. We denote $i$ as the pixel index of a specific map for all the notations mentioned above. For simplicity, we call this architecture as **Mean-Variance Convolutional Soft Actor-Critics (MV-ConvSACs)**. Specifically, we build another head of the ConvSAC's critic network to predict the variance map $\mathrm{Var}(s, A_\mu(s))$ parallel with the reward map $Q$. Inspired by [49], our critic network is trained in a supervised manner with negative log-likelihood (NLL) loss [43].

We further propose to improve the MV-ConvSACs framework using ensemble learning [50]. Following [7], using multiple agents of Q-functions can achieve a higher resolution of pessimism for OOD data. Suppose $N$ MV-ConvSACs agents are considered with an ensemble of $Q$-functions and their variance $\{Q_j(s, A_\mu(s)), \mathrm{Var}_j(s, A_\mu(s))\}_{j=1}^N$. As a result, the epistemic and aleatoric uncertainties, and the total uncertainty can be computed as follows:

$$\bar{V}_{ale}(s, A_\mu(s)) = \frac{1}{N}\sum_{j=1}^{N} \mathrm{Var}_j(s, A_\mu(s))$$

$$\bar{V}_{epi}(s, A_\mu(s)) = \frac{1}{N}\sum_{j=1}^{N} \Big(Q_j(s, A_\mu(s)) - \bar{Q}\Big)^2$$

$$\bar{V}_{all}(s, A_\mu(s)) = \bar{V}_{ale}(s, A_\mu(s)) + \bar{V}_{epi}(s, A_\mu(s))$$

where $\bar{Q} = \sum_j^N Q_j(s, \pi_\mu(s))$.

*2) Quantile-based Distributional Learning:* While **MV-ConvSACs** assume the critic's reward map $Q$ being a pixel-wise Gaussian distribution, quantile-based estimation allows us to capture the full range of possible outcomes and their associated probabilities. By using quantile estimations, we can easily estimate the value at a specific percentile of the reward distribution represented by individual percentiles.

Here we aim to construct a discrete quantile-regressed critic estimator. Suppose the target reward map is a random variable $Z(s, \pi(s)) \in \mathbb{R}^{H*W*1}$, we estimate the percentile corresponding to the $k$'th quantile $\tau_k$ by $Z_{\tau_k}(s, A(s))$ with $k \in [1, K]$ for $K$ quantiles in total. To realize this, we simply construct $K$ output heads for each quantile regression member. For notational simplicity, we denote the $k$'th quantile reward estimate $Z(s, A(s))$ as $\hat{Q}_k \in \mathbb{R}^{H*W*1}$. Given this architecture, the $Q$-value map is estimated as $Q(s, A(s)) = \mathbb{E}_{k\sim\mathcal{U}[1,K]}[\hat{Q}_k]$. Similar to before, we call this architecture as **Quantile Regression Convolutional Soft Actor-Critics (QR-ConvSACs)**.

Inspired by this work [51] that proposes an application of ensemble learning for distributional Q-learning, we also use $N$ QR-ConvSACs agents with an ensemble of quantile heads $\{\hat{Q}_{k,j}(s, A(s))\}_{j\in[1,N], k\in[1,K]}$. Given this, the estimate $Q$-value map is computed as:

$$Q(s, A(s)) = \frac{1}{K \times N}\sum_{k=1}^{K}\sum_{j=1}^{N} \hat{Q}_{k,j}(s, A(s))$$

The training of QR-ConvSACs' critics relies on quantile regression. In particular, following [11] we propose to train the prediction heads according to the Huber loss [10] respectively and model the estimated uncertainty as:

$$\bar{V}_{epi}(s) = \frac{1}{K \times N}\sum_{k=1}^{K}\sum_{j=1}^{N} \left(\hat{Q}_{k,j} - \frac{1}{N}\sum_{j=1}^{N}\hat{Q}_{k,j}\right)^2$$

$$\bar{V}_{ale}(s) = \frac{1}{K}\sum_{k=1}^{K} \left(\frac{1}{N}\sum_{j=1}^{j}\hat{Q}_{k,j} - Q(s, A(s))\right)^2$$

### C. Data-efficient Online Training

*1) Multi-processed Online Learning:* We build a pipeline for multi-processed online learning inspired by Ape-X [52], which is designed to efficiently collect valuable online data and train the agents in a scalable manner. We parallelize

the training of individual agents from the ensemble on a single robotic picking cell. This distributed training pipeline achieves the ratio of **6:1** between *the number of training steps vs. one online collected data sample* (one robot grasp), with an ensemble of $N = 3$ agents. In addition, our pipeline can be theoretically extended to multi-robot data collection for better efficiency as well.

*2) UCB-based Exploration Policy:* Suppose in the online inference process, our stereo sensor captures the state $s$, and our ensemble agents give the correspondent reward map $Q(s, A_\mu(s))$, variance map $V(s, A_\mu(s))$ as well as the action map $A_\mu(s) = \frac{1}{N}\sum_{j=1}^{N} A_{\mu,j}(s)$ as the output action during inference. We design an exploration policy based using the UCB strategy [53], [54] as:

$$Q_{\text{UCB}}(s, A_\mu(s)) = Q(s, A_\mu(s)) + \delta \cdot V(s, A_\mu(s)), \quad (1)$$

where $\delta \in \mathbb{R}_+$ is the UCB ratio, which is a hyperparameter that indicates the degree of exploration on uncertain pixel regions. In the experiment section, we will further investigate the effect of different uncertainty types, $\bar{V}_{ale}$, $\bar{V}_{epi}$ or the total uncertainty $\bar{V}_{all}$.

*3) Online Training with Data Buffer:* We are supposed to update our networks with the online collected data from the shared data buffer. Online learning utilizes similar objectives as offline training, while only on the selected pixel indexed by $i_{best} = \{h^*, w^*\}$ that is selected according to the probabilistic policy in the inference process, where $(h^*, w^*) = \arg\max_{h',w'} Q_{\text{UCB}}[h', w']$, and the best action parameter is extracted from the action map as $A[h^*, w^*]$. Once the updates are completed, the new network parameters are sent to the parameter buffer periodically.

## IV. Experiments

### A. Experiment Setup

Our study was conducted on the Franka Emika Robot equipped with a Schmalz suction gripper. We aim to enable the robot to accurately detect and manipulate objects by mounting a Realsense d415 camera to capture a clear top-down view of the bin. We prepared two sets of objects, each with specific characteristics and experimental purposes as in Fig. 3 (b). The first object set contains opaque objects with rigid and regular shapes in Fig. 3 (b) (top) including several rectangular boxes and cylindrical bottles with various sizes and textures. The second set contains "hard" objects with various characteristics that increase the difficulty of grasping in Fig. 3 (b) (bottom) that are characterized by semi- or total transparency, irregular or curved surfaces, which are used to evaluate the performance of our proposed algorithm.

### B. Technical Details

Here we report the background configurations and hyper-parameter settings of our experiments. We use an ensemble of $N = 3$ QR- or MV-ConvSAC agents. During both offline and online training, we trained both QR- or MV-ConvSAC and the baseline ConvSAC for 4000 steps with a mini-batch size of 12 and a learning rate of $1e-4$ for both the actor and the critic. We apply data augmentation through proper affine

(a) Robot configuration (b) Bin picking objects

Fig. 3: Grasping experiment setup

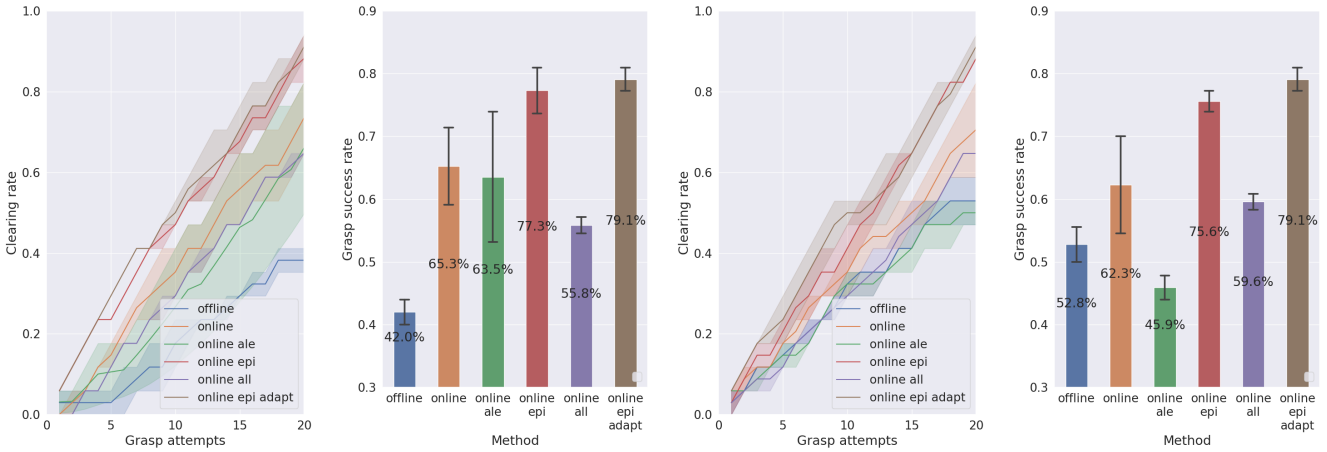
(rotation and translation) and color jitter operations. For a fair comparison, all the models with different exploration strategies are trained for another 3000 steps in our ablation studies. During online learning, our robot will grasp from the bin with 10-17 randomly chosen objects from the online object set. We set the UCB ratio $\delta = 1$ in Eq. (1). The period to transfer the networks from the training process to the inference process is set to be after every 10 steps of updates.

*a) Offline training:* We first train all methods offline. The offline dataset consists of 300 scenes of random 5-10 objects. It can be represented by $\mathbb{D}_{offline} = \{I_c, I_n, I_d, A, Q\}_j$, where $I_c, I_n, I_d$ are defined in Section III-A, and $Q$ is the pixel-wise approximated ground truth $Q$-reward map. The approximated ground truth $Q$ map is computed simply using background subtraction to detect object regions, and then assign values proportional to the inverse of the standard deviation of the object surface normals [25]. The action map for a suction gripper at each pixel is assigned to its respective negative normal vector.

*b) Baselines:* We evaluate and compare our methods to the two ensembles ConvSAC versions: offline-trained MV- and QR-ConvSAC (called *offline* version, correspondingly), and these versions with online fine-tuning (called *online* version) using exploration strategies from the standard SAC algorithm.

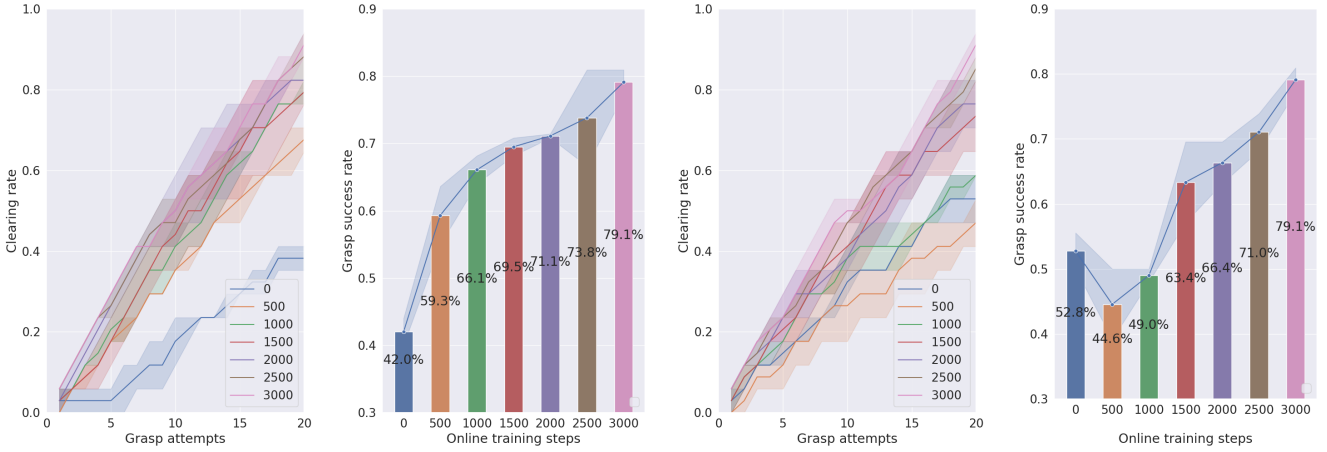
*c) Metrics:* Our evaluation and comparison use two metrics: i) *Grasp success rate*: The percentage of successful grasps over grasp attempts, and ii) *Clearing rate*: The percentage of objects removed over the total objects in a bin.

*d) Evaluation scenes:* To achieve a fair comparison, we predefined a set of 2 evaluation scenes of 17 objects, which contains a selected subset from our online objects. We rearranged the scenes for each algorithm in a consistent manner, so that any observed differences in performance could be attributed solely to the algorithmic approach rather




(a) Success rate (MV-ConvSACs) (b) Final success rate (MV-ConvSACs) (c) Success rate (QR-ConvSACs) (d) Final success rate (QR-ConvSACs)

Fig. 4: Ablation on uncertainty exploration strategies for MV-ConvSACs and QR-ConvSACs




(a) Clearing rate (MV-ConvSACs) (b) Grasp success (MV-ConvSACs) (c) Clearing rate (QR-ConvSACs) (d) Grasp success (QR-ConvSACs)

Fig. 5: Online training steps for MV- and QR-ConvSACs

than to any differences in the scene arrangement.

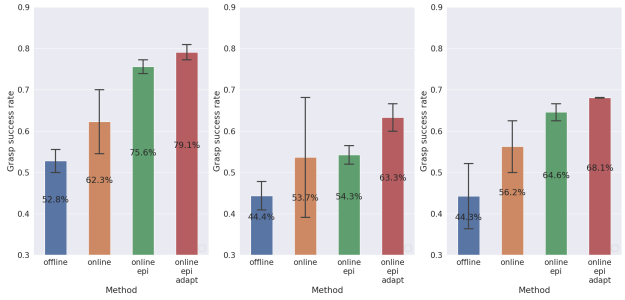
### C. Online Learning Results

*1) Ablation on the Online Exploration Strategies:* In this section, we mainly focus on the uncertainty type for Gaussian-UCB exploration as in Eq. (1). Figure 4 shows our results averaged over the evaluation scenes, where we evaluate the networks by setting an episode of 25 grasp attempts per scene. The episode can terminate earlier the all objects are cleared. The shading regions and error bars represent the mean estimate's first standard deviation. The results show that our online learning approaches MV- and QR-ConvSACs achieve a significant improvement over both ConvSAC ensemble baselines offline and online.

Figure 4 (a, b) shows that for MV-ConvSACs the exploration strategy using epistemic uncertainty gives the highest final clearing rate of above 90%, while also showing a relatively high grasp success. QR-ConvSACs show a similar tendency, they also achieve comparable clearing rates and grasp success. Overall, epistemic uncertainty is shown to be particularly effective in guiding the agent's exploration towards promising actions and strategies, leading to more efficient and targeted experiences and resulting in faster and more effective learning. On the other hand, these results consistently prove that explorations using aleatoric uncertainty or total uncertainty can mislead the learning process, hence converge to a poorer policy than the baselines.

We further propose to adaptively tune the UCB ratio from 1 to 0 (marked as "online epi adaptive" in Fig. 4) that follows a cosine decay. These results demonstrate that the adaptive exploration strategy performs better than the fixed UCB ratio in terms of clearing rates and grasp success. Specifically, we observe that a higher level of exploration is needed in the initial stage of learning, while a lower level of exploration is more appropriate in later stages to consolidate the learned knowledge.

*2) Ablation on the Number of Online Data:* We further conducted an ablation study on the amount of collected




(a) 10 heads (b) 20 heads (c) 100 heads

Fig. 6: QR-ConvSACs: Ablation of quantile head numbers

online data (proportional to the training steps) on the learning efficiency and consistency of our approaches. During the online learning of the exploration of epistemic uncertainty with adaptive UCB ratio, we recorded the checkpoints after every 500 training steps and evaluated on our evaluation scenes. The results are presented in Fig. 5, which shows an overall increase in both metrics during online learning.

To gain a deeper understanding of the performance of our framework, we visualize the evolution of reward maps and uncertainty maps during the training process, as shown in Fig. 1. Our observations show that the high reward regions tend to concentrate on parts of objects that can be grasped safely with a high probability, such as the center region of a homogeneous surface. On the other hand, false positives on hard objects are identified as the main exploration targets during the online training. Exploring these areas can help the agent learn more effective grasping strategies in the presence of uncertainty. Similarly, transparent objects can lead to false negatives due to the difficulty of accurately sensing the object's shape and position. Exploring these regions can help the agent to improve its understanding of the environment and reduce the number of false negatives.

*3) Ablation on the Number of Quantile Heads of QR-ConvSACs:* In this final experiment in Fig. 6, we report our ablation study on the number of quantile heads of QR-ConvSACs, where the improvement of the success rate does not increase after 20. The final grasp success rate of adaptive epistemic exploration achieves the best performance if using the same number of heads, while the best performance is around 79 % with 10 quantile heads. For each quantile head choice, the adaptive setting also has the highest clearing rates, with above 90%, 70%, and 75% for 10, 20, and 100 quantiles, respectively.

## V. Conclusions

In this work, we have proposed uncertainty-based exploration for online grasp learning. The proposed strategies are based on deep ensemble learning, uncertainty estimation, and UCB exploration. Specifically, we investigate and employ two ways to integrate uncertainty estimation, including mean-variance estimation (MV-ConvSACs) and quantile regression (QR-ConvSACs). During training, we online train the offline initialized agents on real-world bin picking scenes. Based on a view that integrating uncertainty estimation into the algorithm improved its performance by providing more potential regions to be explored, we studied different exploration strategies based on the UCB algorithm by utilizing different types of uncertainties and found that epistemic uncertainty is much more informative than aleatoric uncertainty. Epistemic uncertainty refers to uncertainty due to a lack of knowledge, which can be completed by further grasp trials. While aleatoric uncertainty is due to randomness or noise in the data that is generally unavoidable. To sum up, our proposed methods show a promising overall success rate, generalization ability, and efficiency in the bin picking of unseen objects. Our future work will look at an extension into online learning of non-prehensile manipulation. In addition, improved exploration strategies on a fused space between visual and proprioceptive modalities will open more applications to other manipulation tasks.

## References


[1] J. Bohg, A. Morales, T. Asfour, and D. Kragic, "Data-driven grasp synthesis—a survey," *IEEE Trans. on robotics*, vol. 30, no. 2, pp. 289–309, 2013.

[2] K. Kleeberger, R. Bormann, W. Kraus, and M. F. Huber, "A survey on learning-based robotic grasping," *Current Robotics Reports*, vol. 1, no. 4, pp. 239–249, 2020.

[3] R. Newbury, M. Gu, L. Chumbley, A. Mousavian, C. Eppner, J. Leitner, J. Bohg, A. Morales, T. Asfour, D. Kragic *et al.*, "Deep learning approaches to grasp synthesis: A review," *IEEE Trans. on Robotics*, 2023.

[4] M. Danielczuk, A. Balakrishna, D. S. Brown, S. Devgon, and K. Goldberg, "Exploratory grasping: Asymptotically optimal algorithms for grasping challenging polyhedral objects," *arXiv preprint arXiv:2011.05632*, 2020.

[5] L. Fu, M. Danielczuk, A. Balakrishna, D. S. Brown, J. Ichnowski, E. Solowjow, and K. Goldberg, "Legs: Learning efficient grasp sets for exploratory grasping," in *International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 8259–8265.

[6] M. Gilles and V. Rau, "Continual learning of vacuum grasps from grasp outcome for unsupervised domain adaption," in *2022 2nd International Conference on Robotics, Automation and Artificial Intelligence (RAAI)*. IEEE, 2022, pp. 164–171.

[7] S. Lee, Y. Seo, K. Lee, P. Abbeel, and J. Shin, "Offline-to-online reinforcement learning via balanced replay and pessimistic q-ensemble," in *Conference on Robot Learning (CoRL)*, 2021.

[8] Z. Feldman, H. Ziesche, N. A. Vien, and D. D. Castro, "A hybrid approach for learning to shift and grasp with elaborate motion primitives," in *International Conference on Robotics and Automation (ICRA)*. IEEE Press, 2022, p. 6365–6371.

[9] Y. Gal, "Uncertainty in deep learning," Ph.D. dissertation, University of Cambridge, 2016.

[10] W. Dabney, M. Rowland, M. G. Bellemare, and R. Munos, "Distributional reinforcement learning with quantile regression," in *Advancement of Artificial Intelligence (AAAI)*, 2018, pp. 2892–2901.

[11] W. R. Clements, B.-M. Robaglia, B. van Delft, R. B. Slaoui, and S. Toth, "Estimating risk and uncertainty in deep reinforcement learning," *ArXiv*, vol. abs/1905.09638, 2019.

[12] J. Mahler, F. T. Pokorny, B. Hou, M. Roderick, M. Laskey, M. Aubry, K. Kohlhoff, T. Kröger, J. Kuffner, and K. Goldberg, "Dex-net 1.0: A cloud-based network of 3d objects for robust grasp planning using a multi-armed bandit model with correlated rewards," in *International Conference on Robotics and Automation (ICRA)*, 2016, pp. 1957–1964.

[13] M. Gilles, Y. Chen, T. R. Winter, E. Z. Zeng, and A. Wong, "Metagraspnet: A large-scale benchmark dataset for scene-aware ambidextrous bin picking via physics-based metaverse synthesis," in *International Conference on Automation Science and Engineering (CASE)*. IEEE, 2022, pp. 220–227.

[14] D. Morrison, J. Leitner, and P. Corke, "Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach," in *RSS XIV, Pittsburgh, USA*, 2018.

[15] V. Satish, J. Mahler, and K. Goldberg, "On-policy dataset synthesis for learning robot grasping policies using fully convolutional deep networks," *Robotics and Automation Letters*, vol. 4, no. 2, pp. 1357–1364, 2019.

[16] D. Yang, T. Tosun, B. Eisner, V. Isler, and D. Lee, "Robotic grasping through combined image-based grasp proposal and 3d reconstruction," in *International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 6350–6356.

[17] K.-Y. Jeng, Y.-C. Liu, Z. Y. Liu, J.-W. Wang, Y.-L. Chang, H.-T. Su, and W. H. Hsu, "GDN: A coarse-to-fine (c2f) representation for end-to-end 6-dof grasp detection," *Conference on Robot Learning (CoRL)*, pp. 220–231, 2021.

[18] P. Ni, W. Zhang, X. Zhu, and Q. Cao, "Pointnet++ grasping: learning an end-to-end spatial grasp generation algorithm from sparse point clouds," in *International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 3619–3625.

[19] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *Computer Vision and Pattern Recognition Conference (CVPR)*, 2020, pp. 11 444–11 453.

[20] Y. Li, L. Schomaker, and S. H. Kasaei, "Learning to grasp 3d objects using deep residual u-nets," in *RO-MAN*. IEEE, 2020, pp. 781–787.

[21] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox, "Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes," in *International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 438–13 444.

[22] M. Breyer, J. J. Chung, L. Ott, R. Siegwart, and J. I. Nieto, "Volumetric grasping network: Real-time 6 DOF grasp detection in clutter," in *Conference on Robot Learning (CoRL)*, vol. 155. PMLR, 2020, pp. 1602–1611.

[23] Z. Jiang, Y. Zhu, M. Svetlik, K. Fang, and Y. Zhu, "Synergies between affordance and geometry: 6-dof grasp detection via implicit representations," in *RSS XVII, Virtual*, 2021.

[24] H. Cao, H.-S. Fang, W. Liu, and C. Lu, "Suctionnet-1billion: A large-scale benchmark for suction grasping," *Robotics and Automation Letters*, vol. 6, no. 4, pp. 8718–8725, 2021.

[25] P. Schillinger, M. Gabriel, A. Kuss, H. Ziesche, and N. A. Vien, "Model-free grasping with multi-suction cup grippers for robotic bin picking," *CoRR*, vol. abs/2307.16488, 2023.

[26] J. Kober, J. A. Bagnell, and J. Peters, "Reinforcement learning in robotics: A survey," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1238–1274, 2013.

[27] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, and S. Levine, "Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation." in *Conference on Robot Learning (CoRL)*, vol. abs/1806.10293, 2018.

[28] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen, "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection," *The International Journal of Robotics Research*, vol. 37, no. 4-5, pp. 421–436, 2018.

[29] A. Bicchi and V. Kumar, "Robotic grasping and contact: a review," in *International Conference on Robotics and Automation (ICRA). Symposia Proceedings (Cat. No.00CH37065)*, vol. 1, 2000, pp. 348–353 vol.1.

[30] P. Schmidt, N. Vahrenkamp, M. Wächter, and T. Asfour, "Grasping of unknown objects using deep convolutional neural networks based on depth images," in *International Conference on Robotics and Automation (ICRA)*, 2018, pp. 6831–6838.

[31] A. Zeng, S. Song, S. Welker, J. Lee, A. Rodriguez, and T. Funkhouser, "Learning synergies between pushing and grasping with self-supervised deep reinforcement learning," 2018.

[32] L. Berscheid, P. Meißner, and T. Kröger, "Robot learning of shifting objects for grasping in cluttered environments," in *International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 612–618.

[33] L. Berscheid, C. Friedrich, and T. Kröger, "Robot learning of 6 dof grasping using model-based adaptive primitives," in *International Conference on Robotics and Automation (ICRA)*.

[34] Q. Lu, M. Van der Merwe, and T. Hermans, "Multi-fingered active grasp learning," in *International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 8415–8422.

[35] C. Eppner and O. Brock, "Visual detection of opportunities to exploit contact in grasping using contextual multi-armed bandits," in *International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 273–278.

[36] O. B. Kroemer, R. Detry, J. Piater, and J. Peters, "Combining active learning and reactive control for robot grasping," *Robotics and Autonomous systems*, vol. 58, no. 9, pp. 1105–1116, 2010.

[37] H. Y. Li, M. Danielczuk, A. Balakrishna, V. Satish, and K. Goldberg, "Accelerating grasp exploration by leveraging learned priors," in *International Conference on Automation Science and Engineering (CASE)*. IEEE, 2020, pp. 110–117.

[38] M. Laskey, J. Mahler, Z. McCarthy, F. T. Pokorny, S. Patil, J. Van Den Berg, D. Kragic, P. Abbeel, and K. Goldberg, "Multi-armed bandit models for 2d grasp planning with uncertainty," in *International Conference on Automation Science and Engineering (CASE)*. IEEE, 2015, pp. 572–579.

[39] J. Oberlin and S. Tellex, "Autonomously acquiring instance-based object models from experience," *Robotics Research: Volume 2*, pp. 73–90, 2018.

[40] W. Guo, W. Li, Z. Hu, and Z. Gan, "Few-shot instance grasping of novel objects in clutter," *Robotics and Automation Letters*, vol. 7, no. 3, pp. 6566–6573, 2022.

[41] R. Chen, N. Gao, N. A. Vien, H. Ziesche, and G. Neumann, "Meta-learning regrasping strategies for physical-agnostic objects," *arXiv preprint arXiv:2205.11110*, 2022.

[42] L. Barcellona, A. Bacchin, A. Gottardi, E. Menegatti, and S. Ghidoni, "Fsg-net: a deep learning model for semantic robot grasping through few-shot learning," in *International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 1793–1799.

[43] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *Annual Conference on Neural Information Processing Systems (NuerIPS)*, 2017, p. 5580–5590.

[44] B. Charpentier, R. Senanayake, M. J. Kochenderfer, and S. Günnemann, "Disentangling epistemic and aleatoric uncertainty in reinforcement learning," *ArXiv*, vol. abs/2206.01558, 2022.

[45] K. Lee, M. Laskin, A. Srinivas, and P. Abbeel, "Sunrise: A simple unified framework for ensemble learning in deep reinforcement learning," in *International Conference on Machine Learning (ICML)*, vol. abs/2007.04938, 2020.

[46] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International Conference on Machine Learning (ICML)*, vol. 80, 10–15 Jul 2018, pp. 1861–1870.

[47] P. Wagner, X. Wu, and M. F. Huber, "Kalman bayesian neural networks for closed-form online learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 8, 2023, pp. 10 069–10 077.

[48] W. Zhou, B. Jiang, F. Yang, C. Paxton, and D. Held, "Learning hybrid actor-critic maps for 6d non-prehensile manipulation," *CoRR*, vol. abs/2305.03942, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2305.03942

[49] B. Goodrich, A. Kuefler, and W. D. Richards, "Depth by poking: Learning to estimate depth from self-supervised grasping," *International Conference on Robotics and Automation (ICRA)*, pp. 10 466–10 472, 2020.

[50] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Annual Conference on Neural Information Processing Systems (NuerIPS)*. Curran Associates Inc., 2017, p. 6405–6416.

[51] Y. Jiang, J. Z. Kolter, and R. Raileanu, "Uncertainty-driven exploration for generalization in reinforcement learning," in *Deep Reinforcement Learning Workshop in Annual Conference on Neural Information Processing Systems (NuerIPS)*, 2022.

[52] D. Horgan, J. Quan, D. Budden, G. Barth-Maron, M. Hessel, H. van Hasselt, and D. Silver, "Distributed prioritized experience replay," in *ICLR*, 2018.

[53] R. Y. Chen, S. Sidor, P. Abbeel, and J. Schulman, "UCB exploration via q-ensemble," ICLR submission, OpenReview, 2018.

[54] X. Wu, M. El-Shamouty, C. Nitsche, and M. F. Huber, "Uncertainty-guided active reinforcement learning with bayesian neural networks," in *International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 5751–5757.