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

Pseudo-Labeling and Contextual Curriculum Learning for Online Grasp Learning in Robotic Bin Picking

Huy Le[1,2], Philipp Schillinger[2], Miroslav Gabriel[2], Alexander Qualmann[2], Ngo Anh Vien[2]

***Abstract*— The prevailing grasp prediction methods predominantly rely on offline learning, overlooking the dynamic grasp learning that occurs during real-time adaptation to novel picking scenarios. These scenarios may involve previously unseen objects, variations in camera perspectives, and bin configurations, among other factors. In this paper, we introduce a novel approach, SSL-ConvSAC, that combines semi-supervised learning and reinforcement learning for online grasp learning. By treating pixels with reward feedback as labeled data and others as unlabeled, it efficiently exploits unlabeled data to enhance learning. In addition, we address the imbalance between labeled and unlabeled data by proposing a contextual curriculum-based method. We ablate the proposed approach on real-world evaluation data and demonstrate promise for improving online grasp learning on bin picking tasks using a physical 7-DoF Franka Emika robot arm with a suction gripper. Video: https://youtu.be/OAro5pg8I9U**

I. Introduction

The core task in robotic manipulation is grasping, a fundamental skill that opens doors to more complex actions like pick and place or bin picking [1]. In bin picking, the goal is to take objects out of a container and put them in specific places, which has wide applications. However, bin picking is challenging due to issues like noisy perception, object obstructions, and collisions in planning. Thus, there is a need for a robust approach to handle this task effectively [2]. To address these challenges, modern grasping techniques have harnessed advanced deep learning methods. These techniques empower the model to predict grasping actions without relying on predefined models, thereby making them applicable to a broad spectrum of objects in unstructured environments [2], [3]. However, it is worth noting that most of the approaches discussed in the literature depend on supervised learning and offline training, potentially limiting their ability to adapt to unseen objects or new environmental conditions [4], [5].

To address these challenges, this paper's primary emphasis lies in addressing the issue of online grasp learning, which is often framed as a reinforcement learning problem [6], [7], [8], [9]. These works have commonly used a fully convolutional network (FCN) to learn dense pixel-wise grasp quality predictions, i.e. the critic. The pixel-wise parameterization is also used for the grasp primitive, i.e. the actor [10]. However during online learning, the agent receives sparse feedback of grasp success or failure at only one pixel location selected by the policy. The networks get updated accordingly through backpropagation via the loss at this only pixel point.

[1] Technical University Dortmund
[2] Bosch Center for Artificial Intelligence (BCAI), Germany. Email: firstname.lastname@de.bosch.com

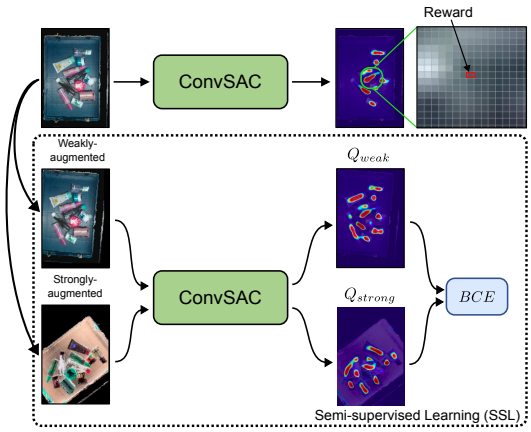

Fig. 1. SSL-ConvSAC: A combined SSL and ConvSAC grasp learning approach to address the sparse reward feedback problem in online grasp learning. During online learning, only one pixel point gets feedback, hence the loss is sparsely backpropagated. In contrast, SSL-ConvSAC will learn from both ground-truth and pseudo-labeled reward feedback.

In this paper, we propose an approach that is able to take advantages of backpropagation via the whole pixel points at each training step. In particular, we combine the advantages of semi-supervised learning (SSL) and RL-based online grasp learning. We cast the pixel point with reward feedback as labeled data, while the remaining pixels without reward feedback as unlabeled data. SSL allows us to enable for exploitation of unlabeled data to improve training progress and overall performance [11], [12]. To this end, we propose SSL-based fully Convolutional Soft-Actor Critic (SSL-ConvSAC), that combines both true rewards and pseudo-labeled rewards for grasp policy learning as depicted in Fig. 1.

To summarize, our main contributions are as follows:

- We address an unexplored problem in online grasp learning, *sparse reward feedback*. We propose a principled approach SSL-ConvSAC, that take advantages of unlabeled data to improve the learning efficiency.
- We show that different SSL methods can be integrated into SSL-ConvSAC. This integration also enables for curriculum learning-based SSL methods.
- We propose a contextual curriculum-based SSL-ConvSAC method to mitigate the data imbalance issue. We observe an extreme imbalance issue between the amount of labeled and unlabeled data. This problem could cause the training to diverge.
- We evaluate and ablate the proposed methods on real-world evaluation data and online grasp learning on bin picking tasks using a physical 7-DoF Franka Emika

robot arm with a suction gripper.

II. Related Work

A. Modern Robot Grasping Methods

Modern robot grasping methods are known for applications of deep learning techniques and training on a large amount of available data. Mahler et al. [13] and Zeng et al. [14] propose to predict grasp maps for both suction and parallel-jaw grasps through the employment of supervised datasets with RGB-D or depth. In a similar context, Morrison et al. [15] and Satish et al. [16] have adopted a comparable approach, focusing on predicting pixel-wise grasp quality maps and 4-DoF parallel-jaw grasp configurations. On the other hand, alternative strategies involve the utilization of point clouds [17], [18] aiming to infer dense pixel-wise grasp qualities and gripper configurations. Recent advancements have also introduced the concept of pixel-wise grasp maps and predictions for grasp configurations tailored to single-suction grippers [19] and multi-suction cup grippers [20].

B. Deep RL-based Grasping Methods

Vision-based Reinforcement Learning (RL) has emerged as a promising approach for enhancing robotic grasping. This technique involves harnessing visual information, such as RGB-D or point cloud data, to guide the actions of robots using RL networks [21]. RL-based methods [22], [23] typically adopt an end-to-end approach, wherein they optimize closed-loop policies for grasp planning directly from raw visual inputs. However, a significant challenge lies in the necessity for large quantities of high-quality training data, primarily due to the high dimensionality of visual [24] or depth [25] inputs. In addition to closed-loop RL methods, open-loop RL approaches are also applied in scenarios involving 6-DoF bin picking [6], [8], [9]. Several studies have explored the development of advanced online grasp learning techniques [5], [4], [26], [27]. However, their applications have primarily been demonstrated in scenarios involving single, isolated objects. One line of research [28] focuses on online exploratory grasp learning for the new scenes. Another line of research applying Equivariant Neural Network [29] archived good grasping performance.

C. Semi-Supervised Learning Methods

Consistency regularization and pseudo-labeling are two popular methods from Semi-supervised learning (SSL). FixMatch [11], a recently introduced novel approach, achieves competitive results by incorporating both weak and strong data augmentations, and the cross-entropy loss as the criterion for consistency regularization. Follow-up works propose improvements by introducing curriculum learning to FixMatch such as FlexMatch [12], FreeMatch [30], SoftMatch [31]. An alternative direction to improve FixMatch is to use pseudo labels to train a gentle teaching assistant (GTA) network. The student network only exploits knowledge from the GTA's feature extractor.

There have been applications of SSL, which are similar to our use-case, for object detection [32], [33], [34], and for problems with sparsely annotated image data [35], [36], [37], [38]. However we observe that the online grasp learning problem has far more extremely sparse annotation.

III. Problem Statement

This paper considers the online grasp learning problem for bin picking application. Given an RGB-D image of the scene $I \in R^{H\times W\times 4}$, we aim to online learn a grasping policy π that is a mapping from I to output map $\in R^{H\times W\times 4}$ that maximizes the long-term total grasp success rate. The policy output is a multi-channel map of pixel-wise 1-dim grasp quality Q and pixel-wise 3-dim grasp configurations A representing gripper rotation via Euler angles, which is commonly used in previous work [14], [20]. The action having the best grasp quality is selected to execute, specifically the selected pixel location is $(h^*, w^*) = \arg\max_{h',w'} Q[h', w']$, and the grasp configuration is extracted from the action map as $A[h^*, w^*]$. After each grasp attempt, the reward r_t is 1 if it succeeds in picking an object, otherwise 0. As a result, the policy $\pi(I)$ is optimized to maximize the total grasp success return $\sum_t r_t$. The reward feedback r is given to only the selected grasp, i.e. at pixel (h^*, w^*), where other pixel locations $\{h, w\}_{(h,w)\in H\times W,(h,w)\#(h^*,w^*)}$ do not obtain reward feedback. We frame this problem as **sparse reward feedback** with respect to the nature of extremely unbalanced ratio between the two type of labeled and unlabeled data. To improve data efficiency, our approach directly exploits both data with reward feedback at (h^*, w^*) and without reward feedback at $\{h, w\}_{(h,w)\in H\times W,(h,w)\#(h^*,w^*)}$. The proposed method is based on the synergy between pseudo-labeling for semi-supervised learning and reinforcement learning.

IV. Methodology

A. Problem Formulation

In this paper, we propose to formulate this online grasp learning problem as a Markov decision process (MDP), $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R})$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P}$ is the transition probability function, and $\mathcal{R}$ is the reward function. We adopt the fully convolutional network (FCN) which is the same architecture used by ConvSAC [9] and HACMan [10] to infer the dense grasp configuration map $A_\phi(s)$ via an actor network and approximate the dense grasp quality map via a critic network $Q_\theta(s, A_\phi(s))$. The network infers an embedding for each pixel location using a **Pixel Encoder** network. The **Actor** module views each pixel as a distinct state and convolves over encoded pixels and infers a Gaussian action for each of them. Actions are concatenated with their corresponding pixel embedding and evaluated by the **Critic** module resulting in a Q-value map Q_θ.

Similar to the setting of ConvSAC [9], the state s is represented by a 7-dim input data, which is composed of a color image I_c, a normal surface map I_n, and a height map I_c, $s_t = (I_c, I_n, I_d)_t$ 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. As a RL-based grasp learning approach, we maintain a replay buffer of samples $\{s_t, a_t, r_t\}$, where $a_t = A_t[h_t, w_t]$ is an

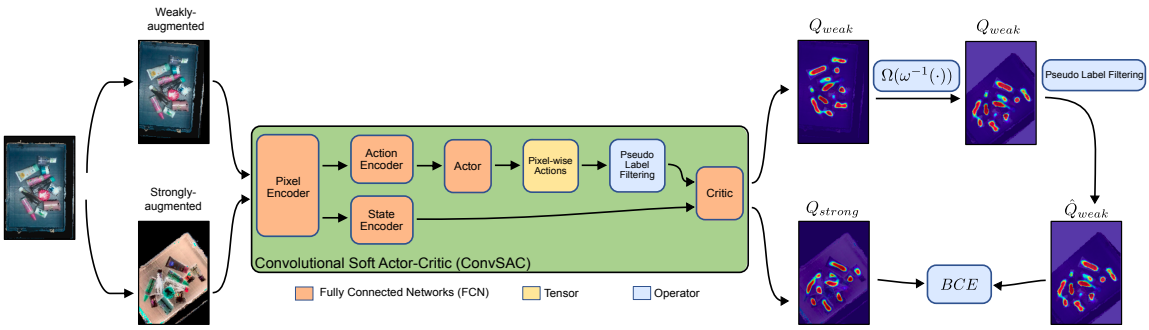

Fig. 2. SSL-ConvSAC Pipeline: Both strongly and weakly augmented images are fed to a ConvSAC network. The pixel-wise prediction of a weakly augmented input is used to provide pseudo labels if their confidence are above a threshold. These pixel-wise pseudo-labeled rewards are then used to compute consistency regularization to update both the Critic and Actor receiving the strongly augmented image as input.

action at the only selected or *labeled* pixel (h_t, w_t). The critic and actor networks are updated similar to ConvSAC [9], where the critic loss is formulated as a classification task with reward labels $r \in \{0, 1\}$ denoting grasp failure or success. The critic uses a BCE loss and the episode horizon terminates after each grasp attempt. Specifically, we define the critic and actor losses for the *labeled* pixels as follows:

$$\begin{aligned} \mathcal{L}^l_{\text{critic}} &= \text{BCE}\left(Q_t\left(s_t, a_t\right), r_t\right) \\ \mathcal{L}^l_{\text{actor}} &= \alpha \log \pi\left(a_t | s_t\right) - Q_t\left(s_t, a_t\right), \end{aligned} \tag{1}$$

where α is an entropy regularization coefficient. Note that this update backpropagates the loss through only one pixel (h_t, w_t) at both the grasp quality map Q and action map A.

We propose to tackle the problem of sparse reward feedback in RL-based online grasp learning through semi-supervised learning. In our particular problem, we are facing an extremely unbalanced data issue. Specifically, for each input image the amount of labeled data is only $N_l = 1$ while the amount of unlabeled data is $N_u = (H \times W) - 1$ is the remaining state pixels. This setting is due to the fact that rearranging the scene to the previous state in order to collect grasp samples at other pixel locations can result in a different state in a real-world setup. In addition, we tackle a realistic setting where online learning operates on an industrial picking cell without interruption.

B. General Approach

Our main contribution is to i) first leverage state-of-the-art SSL techniques for the online grasp learning problem, such as FixMatch [11] and curriculum learning-based SSL FlexMatch [12] and FreeMatch [30], and ii) propose a new contextual curriculum learning-based SSL. We introduce a general approach that enables the integration of different SSL methods, called SSL-ConvSAC, as depicted in Fig. 2.

We adopt consistency regularization for SSL to rewrite the losses of the actor A_ϕ and critic Q_θ. The critic and actor models are updated using a joint objective based on labeled and unlabeled data. The updates using labeled data are defined in Eq. 1. The updates using unlabeled data are written as in Eq. 2, given a data sample (s, a, r) where action a encodes labeled pixel (h, w) with reward r while unlabeled pixels are $U = \{h', w'\}_{(h',w') \in H \times W, (h',w') \# (h,w)}$.

$$\begin{aligned} \mathcal{L}^u_{\text{critic}} &= \frac{1}{N_u} \lambda(\hat{Q}; U)\, \text{BCE}\left(\hat{Q}, Q(\hat{s}, \pi(\hat{s}))\right) \\ \mathcal{L}^u_{\text{actor}} &= \frac{1}{N_u} \lambda(\hat{Q}; U)\left(\alpha \log \pi\left(A | \hat{s}\right) - Q\left(\hat{s}, \pi(\hat{s})\right)\right) \end{aligned} \tag{2}$$

where $\hat{s} = \Omega(s)$ denotes a strongly-augmented data given input s. $\hat{Q} = [Q(\omega(s), \pi(\omega(s)) > 0.5]$ computes pixel-wise pseudo labels, i.e. $\{0, 1\}$, where $\omega(s)$ is a weakly-augmented data. $Q(\Omega(s), \pi(\Omega(s)))$ and $\pi(\Omega(s))$ compute a grasp quality map and an action map of the strongly-augmented data. The weight $\lambda(\hat{Q}; U) \in R^{H*W}$ is a pixel-wise weighting function that can be defined differently according to a different choice of an SSL method. The conditioning on U means that only unlabeled pixels matter in this operation, i.e. $\lambda(\hat{Q}; U)$ has a zero value at labeled pixel (h, w) and values in range $[0, 1]$ at unlabeled pixels. Please note that our suggested SSL objective computes the pixel-wise loss (including BCE with no reduction), enabling it to utilize parallel computations in fully convolutional networks to process the loss for all N_u unlabeled data points simultaneously.

As a result, the joint objectives of the actor and critic are $\mathcal{L}_{\text{critic}} = \mathcal{L}^l_{\text{critic}} + \mathcal{L}^u_{\text{critic}}$ and $\mathcal{L}_{\text{actor}} = \mathcal{L}^l_{\text{actor}} + \mathcal{L}^u_{\text{actor}}$, respectively. The SSL objective for unlabeled data is pixel-wise computed, therefore the final loss is then a sum of losses over all pixels.

Note that whenever there are $\arg\max$ or $\max$ operations on grasp quality map $Q \in R^{H \times W}$, we implicitly assume $Q \in R^{H \times W \times 2}$ for binary classes, specifically, $Q[:, :, 1] = Q$ which is Q-value for class *success*, $Q[:, :, 0] = 1.0 - Q$ for class *failure*. And $\arg\max$ or $\max$ operations are applied across the last axis.

1) FixMatch-based SSL-ConvSAC: We first propose to leverage FixMatch [11] for SSL-ConvSAC. We define a constant threshold τ based on which pseudo-labels with high confidence will be retained. In particular, the weighting function is computed as follows

$$\lambda(\hat{Q}_t; U_t) = \mathbb{I}\left(\max(\hat{Q}_t) \geq \tau\right) \tag{3}$$

where $\mathbb{I}$ is an identity function.

2) Curriculum-based SSL-ConvSAC: We leverage two curriculum-based SSL frameworks FlexMatch [12] and FreeMatch [30]. Instead of using a fixed constant threshold τ, FlexMatch and FreeMatch introduce curriculum learning to tune τ in order to control the way pseudo labels from indi-

vidual class are retained. The following proposals computes an adaptive threshold that can be used for re-computing the weighting function specifically for each class c as follows.

$$\lambda_t(\hat{Q}_t; U_t) = \mathbb{I}\left(\max(\hat{Q}_t) \geq \tau_t(\arg\max \hat{Q}_t)\right) \quad (4)$$

where τ_t will be adapted according to curriculum learning.

a) FlexMatch-based SSL-ConvSAC: Similar to original FlexMatch, we propose to model learning effect $\sigma_t(c)$, $c \in \{0, 1\}$ at each training step t, where class with fewer samples having their prediction confidence reach the threshold is considered to have a greater learning difficulty or a worse learning status. Assuming that the size of the replay buffer is $|B|$, then the total number of unlabeled pixels is $N_u \times |B|$. The learning effect is computed as follows

$$\begin{aligned}\sigma_t(c) = \sum_{s\in B}\sum_{n=1}^{N_u} \mathbb{I}\left(\max Q(\omega(s), \pi(\omega(s)) > \tau\right) \\ \cdot\mathbb{I}\left(\arg\max Q(\omega(s), \pi(\omega(s)) = c\right).\end{aligned} \quad (5)$$

The operation inside the identity function is pixel-wise as usual. The summation takes sum over all unlabeled pixels n and across samples in the replay buffer. As a result, the adaptive threshold $\tau_t(c)$ can now be computed by normalizing $\sigma_t(c)$ in the range $[0, 1]$ as follows.

$$\beta_t(c) = \frac{\sigma_t(c)}{\max_c \sigma_t}, \quad \tau_t(c) = \beta_t(c) \cdot \tau, \quad (6)$$

where normalized learning effect $\beta_t(c)$ is equal to 1 for the best-learned class and lower for the hard classes. We also use a similar warm-up process and a non-linear mapping function from FlexMatch to enable the thresholds to have a non-linear increasing curve in the range from 0 to 1.

b) FreeMatch-based SSL-ConvSAC: Instead of adjusting the confidence threshold according to only the current step's information as in FlexMatch, FreeMatch proposes to self-adapt this value according to the model learning progress. In particular, a self-adaptive global threshold is computed as in Eq. 7 to track the overall learning status globally across all classes among unlabeled data.

$$\tau_t^{\text{global}} = \alpha\tau_{t-1}^{\text{global}} + (1-\alpha)\frac{1}{N_u|B|}\sum_{s\in B}\sum_{n=1}^{N_u}\max q_s, \quad (7)$$

with $t > 0$ and $\tau_0^{\text{global}} = 1/2$ (as the number of classes is 2, i.e. success or failure), where we denote $q_s = Q(\omega(s), \pi(\omega(s))$ and $\alpha \in (0, 1)$ is the momentum decay of the exponential moving average of the confidence. A self-adaptive local threshold to adjust the global threshold in a class-specific fashion is computed as follows.

$$\tilde{p}_t(c) = \alpha\tilde{p}_{t-1}(c) + (1-\alpha)\frac{1}{N_u|B|}\sum_{s\in B}\sum_{n=1}^{N_u} q_s(c), \quad (8)$$

with $t > 0$ and $\tilde{p}_0 = 1/2$. As a result, the final adaptive threshold for each individual class is computed as

$$\tau_t(c) = \frac{\tilde{p}_t(c)}{\max_{c\in\{\text{failure,success}\}}\{\tilde{p}_t(c)\}} \cdot \tau_t^{\text{global}} \quad (9)$$

The fairness regularization introduced in FreeMatch has been excluded from our experiments because it did not demonstrate any advantages.

3) Contextual Curriculum-based SSL-ConvSAC: We observe that the main challenge in our setting compared to standard SSL is at the extreme imbalance between labeled and unlabeled data. This will quickly lead to the confirmation bias problem [39]. The authors show that most SSL methods can suffer from this problem if the mini-batch contains a ratio of 1:100 between labeled and unlabeled data. Inspired by [39], we propose two main technical and one fundamental contributions that can improve generalization and reduce confirmation bias. The technical contributions aim to reduce the confidence of the network as suggested by [39].

- **Lower-bounded confidence threshold**: This helps to filter pseudo labels with low-confidence for curriculum-based methods. In particular, we lower-bound the adaptive threshold as $\tau_t = \max\{\tau_t, \tau_{\text{lb}}\}$, τ_{lb} is a predefined lower-bound confidence threshold.
- **Soft-weighting function**: The current hard-weighting λ_t in Eq. 4 treats pseudo labels of both low and high confidence equally as long as their confidence is above the threshold. We propose using soft-weighting via a soft-max function: $\lambda_t(\hat{Q}_t; U_t) \propto \exp\left(\hat{Q}_t[\mathbb{I}_t]\right)$, where $\mathbb{I}_t = \mathbb{I}\left(\max(\hat{Q}_t) \geq \tau_t(\arg\max \hat{Q}_t)\right)$.
- **Contextual curriculum-based learning**: All previous SSL-ConvSAC variants compute thresholds adaptively to each class, i.e. in FlexMatch and FreeMatch-based SSL-ConvSAC and these values $\sigma_t, \beta_t, \tau_t, \tilde{p}_t$ are 2-dim. However, we observe that different pixel locations in an input image though having the same class, e.g. success, their grasp quality values are not necessarily identical. We propose to represent $\sigma_t, \beta_t, \tau_t, \tilde{p}_t \in R^{H\times W\times 2}$ to depend on pixel contexts. As a result, we can rewrite Eqs. 5, 7, 8 as pixel-wise versions as follows,

$$\begin{aligned}\sigma_t(c) &= \sum_{s\in B}\mathbb{I}\left(\max Q(\omega(s), \pi(\omega(s)) > \tau\right) \\ &\quad\cdot\mathbb{I}\left(\arg\max Q(\omega(s), \pi(\omega(s)) = c\right). \\ \tau_t^{\text{global}} &= \alpha\tau_{t-1}^{\text{global}} + (1-\alpha)\frac{1}{|B|}\sum_{s\in B}\max q_s \\ \tilde{p}_t(c) &= \alpha\tilde{p}_{t-1}(c) + (1-\alpha)\frac{1}{|B|}\sum_{s\in B} q_s(c),\end{aligned} \quad (10)$$

The final equations Eq. 6, 9 are now pixel-wise computed, too. As a result, the weighting function in Eq. 4 involves fully pixel-wise terms.

V. Experiment Setting

We carry out two set of experiments: i) a large scale evaluation task using pre-collected data with a setup described in Section V-B.3; ii) online learning directly on a physical robot system with a setup described in Section V-B.4.

A. Weak-Strong Augmentation Pipelines

Our set of weak and strong augmentation are: i) **Color transformation** in RandAugment [40], which includes AutoContrast, Brightness, Contrast, Equalize, Posterize, Sharpness, and Solarize. ii) **Geometric transformation**: Rotation and shift operations that alter the information about normal vectors in the seven-channel image, and matching operation [41]. iii) **Noise Operator**: Uniform noise and Binary noise.

We apply color transformation on RGB channels, Uniform noise on Depth channel, Binary noise on normal vectors channels, Geometric transformation on the whole seven channels and grasp action configurations. In the weakly augmentation pipeline, we set random Rotation in $[-10, 10]$ degree, and random shifting with $[-10, 10]$ pixels, color Jittering on RGB channel. Strong augmentation pipeline sets random rotation in the range $[-180, 180]$ degree, and shifting $[-30, 30]$ pixels on the whole seven channels. For the depth channel, we use uniform noise of 5mm range and 10% zero out in normal vectors.

B. Technical Details

1) Training Setting: We use the same ConvSAC network parameters as in [9], and slightly different hyperparameters in FixMatch [11]. Concretely, the optimizer for all experiments is Adam with $(\beta_1, \beta_2) = (0.9, 0.999)$, weight decay 0.0001, and learning rate 0.0001. The momentum decay α in FreeMatch is set to 0.95. We perform an exponential moving average model with momentum of 0.99.

2) Offline training: We first train a ConvSAC model offline for 100 epochs to use i) for data collection in the large-scale evaluation task, ii) as a warm-start policy network for online learning on real robots. The offline dataset consists of 72 simple scenes of random 4 objects from the set in Fig. 4 (right). It can be represented by $\{I, A, Q\}$, where input image I are defined in Section IV-A, and Q is the pixel-wise approximated ground truth Q-reward map as discussed in [20]. The action map A pertaining to a suction gripper is assigned to negative surface normal vectors. This pretrained policy is 64% grasp success as shown in Fig. 3 (Left) before any online learning starts.

3) Large-Scale Evaluation Setup:

a) Evaluation data: We collected in total 900 online data points using the pre-trained ConvSAC model above. The scene is prepared with randomly 10-12 objects from the online training objects. Each data point is a sample with sparse reward feedback $\{s, a, r\}$. We use a replay buffer of 500 points for the training set. We train all models in the evaluation task for 500 epochs on a *Nvidia A100* GPU with a batch size of 4. The evaluation set consists of the remaining 400 data points. Only for this set, for each image s we also generate negative samples from background areas using background subtraction given an empty bin image. The negative samples are (s, a random action, $r = 0$). Our evaluation metric is the mean squared error (MSE) between the grasp quality prediction and ground-truth reward 0/1.

b) Comparing methods: We benchmark our approach against the following configurations:

1) **Online (ON)**: ConvSAC without SSL. We additionally run ON with 3000 data points for benchmarking.
2) **FixMatch (FI)**: FixMatch SSL-ConvSAC with a confidence threshold set at $\tau = 0.95$.
3) **FlexMatch (FL)**: FlexMatch SSL-ConvSAC with different lower-bound confidence threshold $\tau_{\text{lb}} = \{0.5, 0.7, 0.9\}$, w./w.o soft-weighting function, and w./w.o contextual curriculum-based learning.

Methods	S	C	5	100	Full
ON 500					0.34 ± 0.13
ON 3000					0.11 ± 0.04
FI C0.95					0.07 ± 0.02
FR L0.9	N	N	0.39 ± 0.13	0.4 ± 0.2	0.08 ± 0.02
FR L0.9	Y	N	0.29 ± 0.15	0.34 ± 0.07	0.08 ± 0.01
FR L0.9	N	Y	0.38 ± 0.12	0.32 ± 0.02	0.08 ± 0.01
FR L0.9	Y	Y	0.26 ± 0.03	0.43 ± 0.19	0.07 ± 0.02
FR L0.7	N	N	0.3 ± 0.18	0.36 ± 0.14	0.11 ± 0.05
FR L0.7	Y	N	0.36 ± 0.15	0.57 ± 0.16	0.1 ± 0.05
FR L0.7	N	Y	0.42 ± 0.03	0.38 ± 0.1	0.09 ± 0.02
FR L0.7	Y	Y	0.45 ± 0.07	0.46 ± 0.08	0.08 ± 0.02
FR L0.5	N	N	0.35 ± 0.08	0.41 ± 0.14	191.81 ± 148.86
FR L0.5	Y	N	0.37 ± 0.2	0.58 ± 0.16	12.59 ± 11.89
FR L0.5	N	Y	0.39 ± 0.15	0.43 ± 0.15	191.78 ± 148.92
FR L0.5	Y	Y	0.32 ± 0.13	0.38 ± 0.06	192.51 ± 162.96
FL L0.9	N	N	0.48 ± 0.3	0.29 ± 0.07	0.07 ± 0.01
FL L0.9	Y	N	0.41 ± 0.11	0.47 ± 0.22	0.08 ± 0.02
FL L0.9	N	Y	0.33 ± 0.24	0.28 ± 0.07	0.08 ± 0.02
FL L0.9	Y	Y	0.27 ± 0.19	0.47 ± 0.31	$\mathbf{0.06 \pm 0.01}$
FL L0.7	N	N	0.29 ± 0.1	0.53 ± 0.28	0.11 ± 0.05
FL L0.7	Y	N	0.33 ± 0.08	0.63 ± 0.26	0.11 ± 0.05
FL L0.7	N	Y	0.46 ± 0.24	0.46 ± 0.23	0.07 ± 0.02
FL L0.7	Y	Y	0.32 ± 0.08	0.44 ± 0.12	$\mathbf{0.06 \pm 0.02}$
FL L0.5	N	N	0.35 ± 0.04	0.38 ± 0.07	191.76 ± 149.0
FL L0.5	Y	N	0.2 ± 0.03	0.46 ± 0.24	38.84 ± 33.39
FL L0.5	N	Y	0.42 ± 0.2	0.44 ± 0.26	191.7 ± 149.11
FL L0.5	Y	Y	0.38 ± 0.07	0.57 ± 0.31	192.46 ± 162.95

TABLE I

MSE ON EVALUATION SET ($\times 10^{-3}$): COLUMN S IS SHORT FOR SOFT-WEIGHTING FUNCTION, COLUMN C IS SHORT FOR CONTEXTURE, AND 5, 100, FULL INDICATE TOP k PSEUDO-LABELS SELECTION. PREFIX L INDICATES τ_{LB} VALUE. (Y/N ARE SHORT FOR YES/NO)

4) **FreeMatch (FR)**: FreeMatch SSL-ConvSAC with the same options as FlexMatch SSL-ConvSAC.

All results are averaged over 3 training trials per setting. We report the mean and its standard deviation.

4) Physical Robot Setup: The experiment was conducted on the *Franka Emika* Robot with *Schmalz* suction gripper and an over the shoulder Realsense d415 camera as seen in Fig. 4. The online training objects consist of 12 common objects depicted in Fig. 4 middle. Each online learning scene is prepared with randomly 10-12 objects.

a) Metrics: We measure performance based on the *Grasp Success Rate* (SR) and *Bin Completion Rate* (BCR), averaged across the latest 15 bins. Each episode finishes if the bin is clear or the robot has already attempted 15 grasps on this bin. SR is defined as the ratio of successful grasps, and BCR is the ratio of complete bin.

b) Online training setting: We online learn all comparing models with a learning rate of $(1e-5)$ for both the actor and critic from the pretrained ConvSAC. The model runs on Ubuntu 20.04 and is equipped with an Intel Xeon E3-1505M CPU and a Nvidia GeForce RTX 3090 GPU.

c) Comparing methods: We run these algorithms: i) Online, ii) FlixMatch SSL-ConvSAC with $\tau = 0.95$, iii) Flexmatch SSL-ConvSAC with contexture, softmax weighting, $\tau_{\text{lb}} = 0.9$, and full pseudo-label. All results are averaged over 2 training trials per setting. We report the mean and its standard deviation.

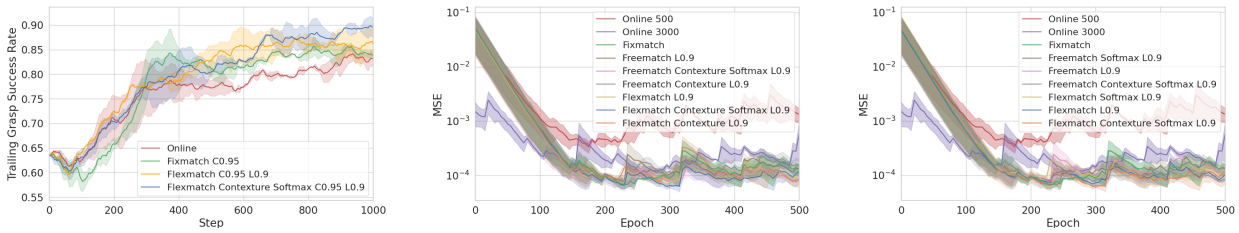

Fig. 3. (Left) Online training plots on physical robot. Trailing grasp success rate over the latest 15 bins vs. number of online sample steps. The asynchronous training has a ratio of 10:1 steps per grasp attempt. (Center) Comparisons of contextural models. (Right) Comparisons of soft-weighting models.

Fig. 4. Robot setup (left), online objects (middle), offline objects (right)

Method	Softmax	Contexture	BCR (%)	SR (%)
ON			60	82.5
FI C0.95			70	83
FL C0.95 L0.9	N	N	80	86
FL C0.95 L0.9	Y	Y	**93.3**	**90**

TABLE II

RESULTS ON PHYSICAL ROBOTS. (Y/N ARE SHORT FOR YES/NO)

VI. EXPERIMENT RESULTS

A. Large-Scale Evaluation

a) Extreme imbalanced data: We first show how extreme the imbalanced data issue is. For each curriculum-based variants, we allows a ratio of 5:1, 10:1 (though many of them with sufficient confidence), and all of pseudo labels accepted to training. The results in Table I show that if the ratio is as low as most settings in SSL literature, the training does not have issues, and performs only comparably as the baseline ON. With an extreme ratio, i.e. *Full*, vanilla curriculum-based methods can diverge if measures from our contextual curriculum-based SSL-ConvSAC are not applied, i.e. $\tau_{\text{lb}} = 0.5$ means no lower-bound threshold as the boundary between two classes is at confidence 0.5. The reason is they enroll too many erroneous pseudo-labels with low confidence at the beginning of the training.

b) Contextual curriculum-based SSL-ConvSAC: The settings with higher lower bound of 0.7 and 0.9 have consistently achieved good evaluation losses, especially on Full. This shows that this technical fix mitigates the divergence issue and also can take advantages of having more unlabeled data. Finally, the best performing setting is when all measures of contextual curriculum-based SSL-ConvSAC are activated, together with FixMatch SSL-ConvSAC. Fig. 3 (Center, Right) also shows the training plots of SSL-ConvSACs and baselines. These results tell that SSL methods are able to help training faster and converge to a lower loss compared to baselines Online 500 & 3000. Again, Contexture & softmax weighting SSL-ConvSAC achieved the best training progress and the final loss in all settings.

B. Online Learning on Physical Robots

In physical robot experiments, ours reached an 80% success rate after just 400 steps, whereas the online model required approximately 600 steps, as illustrated in Fig 3 (Left). Flexmatch SSL-ConvSAC with contexture, softmax weighting, $\tau_{\text{lb}} = 0.9$, and full pseudo-label achieved the best grasp success rate and stable around 90%. This observation is further supported by Table II, where the Flexmatch SSL-ConvSAC outperformed Online model with 93.3 % BCR, and 90% SR. Note that all training plots have a slight drop at an early stage due to the distribution shift from the pretrained model, i.e. different object portfolio and the offline training scenes have 4 objects instead of 10-12 objects in clutter of online scenes. The out-performance of Flexmatch SSL-ConvSAC compared to non-curriculum FixMatch SSL-ConvSAC, because this online learning setting is based on a pretrained model. This allows the curriculum to benefit from a high confidence model at the beginning, which might not be as high as $\tau = 0.95$ used by FixMatch.

VII. CONCLUSION

This paper introduces SSL-ConvSAC, a novel approach that combines advantages of SSL and RL for online grasp learning. We pose the sparse reward feedback problem for online grasp learning. We show that naively integrating SSL with RL is not enough to tackle this problem, as the amount of unlabeled data in this setting is overwhelming. To mitigate this issue, we propose a contextual curriculum learning approach. We show that the proposed approach is able to fully exploit unlabeled data in our application to improve the overall performance. We demonstrate it on a real-world bin picking setup with a grasp success rate of 90 %, and bin completion of 93%. As future work, the topics of pseudo-labeling for closed-loop grasping and online learning on flexible objects would be challenging but bring many interesting applications.

REFERENCES

[1] J. Bohg, A. Morales, T. Asfour, and D. Kragic, "Data-driven grasp synthesis—a survey," *IEEE Transactions on robotics*, vol. 30, no. 2, pp. 289–309, 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 Transactions 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 *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 8259–8265.

[6] 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.

[7] L. Berscheid, P. Meißner, and T. Kröger, "Robot learning of shifting objects for grasping in cluttered environments," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2019, Macau, SAR, China, November 3-8, 2019*. IEEE, 2019, pp. 612–618.

[8] L. Berscheid, C. Friedrich, and T. Kröger, "Robot learning of 6 dof grasping using model-based adaptive primitives," in *IEEE International Conference on Robotics and Automation, ICRA 2021, Xi'an, China, May 30 - June 5, 2021*. IEEE, 2021, pp. 4474–4480.

[9] 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.

[10] 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

[11] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in neural information processing systems*, vol. 33, pp. 596–608, 2020.

[12] B. Zhang, Y. Wang, W. Hou, H. Wu, J. Wang, M. Okumura, and T. Shinozaki, "Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18 408–18 419, 2021.

[13] 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 *IEEE International Conference on Robotics and Automation (ICRA)*, 2016, pp. 1957–1964.

[14] A. Zeng, S. Song, K.-T. Yu, E. Donlon, F. R. Hogan, M. Bauza, D. Ma, O. Taylor, M. Liu, E. Romo, *et al.*, "Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching," *IJRR*, vol. 41, no. 7, pp. 690–705, 2022.

[15] 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.

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

[17] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *IEEE CVPR*, 2020, pp. 11 444–11 453.

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

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

[20] 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.

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

[22] 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 *The Conference on Robot Learning (CoRL)*, vol. abs/1806.10293, 2018.

[23] 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.

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

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

[26] H. Y. Li, M. Danielczuk, A. Balakrishna, V. Satish, and K. Goldberg, "Accelerating grasp exploration by leveraging learned priors," in *2020 IEEE 16th International Conference on Automation Science and Engineering (CASE)*. IEEE, 2020, pp. 110–117.

[27] 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 *2015 IEEE International Conference on Automation Science and Engineering (CASE)*. IEEE, 2015, pp. 572–579.

[28] Y. Shi, P. Schillinger, M. Gabriel, A. Kuss, Z. Feldman, H. Ziesche, and N. A. Vien, "Uncertainty-driven exploration strategies for online grasp learning," 2023.

[29] X. Zhu, D. Wang, O. Biza, G. Su, R. Walters, and R. Platt, "Sample efficient grasp learning using equivariant models," *Proceedings of Robotics: Science and Systems (RSS)*, 2022.

[30] Y. Wang, H. Chen, Q. Heng, W. Hou, M. Savvides, T. Shinozaki, B. Raj, Z. Wu, and J. Wang, "Freematch: Self-adaptive thresholding for semi-supervised learning," *ArXiv*, vol. abs/2205.07246, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:248811614

[31] H. Chen, R. Tao, Y. Fan, Y. Wang, J. Wang, B. Schiele, X. Xie, B. Raj, and M. Savvides, "Softmatch: Addressing the quantity-quality trade-off in semi-supervised learning," *arXiv preprint arXiv:2301.10921*, 2023.

[32] G. Li, X. Li, Y. Wang, Y. Wu, D. Liang, and S. Zhang, "Pseco: Pseudo labeling and consistency training for semi-supervised object detection," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*. Springer, 2022, pp. 457–472.

[33] ——, "Dtg-ssod: Dense teacher guidance for semi-supervised object detection," *arXiv preprint arXiv:2207.05536*, 2022.

[34] M. Xu, Z. Zhang, H. Hu, J. Wang, L. Wang, F. Wei, X. Bai, and Z. Liu, "End-to-end semi-supervised object detection with soft teacher," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3060–3069.

[35] S. S. Rambhatla, S. Suri, R. Chellappa, and A. Shrivastava, "Sparsely annotated object detection: A region-based semi-supervised approach," *arXiv preprint arXiv:2201.04620*, 2022.

[36] H. Wang, L. Liu, B. Zhang, J. Zhang, W. Zhang, Z. Gan, Y. Wang, C. Wang, and H. Wang, "Calibrated teacher for sparsely annotated object detection," *arXiv preprint arXiv:2303.07582*, 2023.

[37] J. Yoon, S. Hong, and M.-K. Choi, "Semi-supervised object detection with sparsely annotated dataset," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 719–723.

[38] T. Wang, T. Yang, J. Cao, and X. Zhang, "Co-mining: Self-supervised learning for sparsely annotated object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 4, 2021, pp. 2800–2808.

[39] E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, and K. McGuinness, "Pseudo-labeling and confirmation bias in deep semi-supervised learning," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[40] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical data augmentation with no separate search," *CoRR*, vol. abs/1909.13719, 2019. [Online]. Available: http://arxiv.org/abs/1909.13719

[41] K. Sohn, Z. Zhang, C. Li, H. Zhang, C. Lee, and T. Pfister, "A simple semi-supervised learning framework for object detection," *CoRR*, vol. abs/2005.04757, 2020.