2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Continuous Rapid Learning by Human Imitation using Audio Prompts and One-Shot Learning

Jaime Duque-Domingo[1], Miguel García-Gómez[1], Eduardo Zalama[1] and Jaime Gómez-García-Bermejo[1]

Abstract— **In the general field of collaborative robotics, one of the topics of greatest interest to the scientific community is the ability to learn to perform certain actions by imitating humans. If we think about humans, when someone teaches us how to perform a certain action, we often need to be shown just one time how to do it. Likewise, we believe that robotics should follow this line, using models that do not involve the capture of huge data sets or exhaustive training. Furthermore, while general models can typically be pretrained offline, the robot must quickly adapt to new knowledge without requiring an expensive retraining process. In this article we present a flexible neural learning architecture that allows a robot to learn how-to pick-up a given object just by watching how a human does it. Then, the robot will be able to pick up the current object, or other objects previously learned, anywhere in the work field, with a simple audible indication from the user. This is achieved based on continuous incremental learning techniques and generic segmentation networks integrated with Siamese network models according to the recently proposed CP-CVV method. Results are presented for the success rate in grasping a varied set of objects.**

I. INTRODUCTION

In the field of collaborative robotics, one of the topics of greatest interest to the scientific community is the ability to learn to perform certain actions by imitating humans from one or a few examples of how to do it. If we think about humans, when someone teaches us to do a certain task we usually need to be shown at least once how to do it. Within social robotics, tasks such as feeding a person using unfamiliar cutlery, learning to chop up food, or learning to pick up objects or products never seen by the robot are challenges of great scientific interest.

This paper presents a novel learning method that enables a collaborative robot to learn to pick up objects quickly from a single example of how a person picks it up. It will capture this learning by imitating a human. In continuous learning the robot will be able to learn to pick up objects during the operation phase. In other words, there is no need to train the models used prior to operation. Continuous learning is an ambitious challenge within Few-Shot Learning (FSL), where there are only a few examples in each category. One of the most restrictive cases of FSL is One-Shot Learning (OSL), where we only have one example of how the person picks up the object. Our object picking perspective is based on a data perspective [1], where we start from prior knowledge to augment the supervised experience. Gradually new classes appear and the model must learn them without falling into the so-called catastrophic forgetting of previous classes.

Additionally, our system is looking for fast learning, where the user has the feeling that the robot learns immediately. The system needs to be fast in integrating new classes, something that adds much more difficulty to the process. There are very few advances in the literature that integrate continuous learning and FSL. In the Few-Shot Continual Active Learning (FoCAL) model [2], they propose a uniform Gaussian mixture model and use pseudo-trial to mitigate catastrophic forgetting. However, the authors themselves point out several problems with their approach that can be studied, such as that a human assistant provides the robot with the correct object labels.

In our system, objects are registered from a name, an image of the object in isolation and an image of the object showing how a person grasps it. The object is registered by segmenting the object based on the grip coordinates of the person's fingers. Once a new object is registered, the person asks the robot by audio to pick up a particular object. Using an automatic speech recognition (ASR) method, we extract a text prompt. Specifically, the use of audio prompts improves learning through making it more natural and easier for humans, as they do not have to worry about launching specific commands to execute the robot's movement. This prompt is compared using semantic comparison techniques [3] to find out which object the person wants to pick up. We have used generic segmentation techniques [4] to capture unknown aspects of the scene integrated with FSL comparison techniques [5] to identify new elements. Specifically, we have used the Class Partitioning and Cross Validation Voting method (CP-CVV) [6] for segmented object matching. From the text prompt classification, we search for the closest object from the segmented ones. If the object is not found directly, the robot will evaluate different positions to locate it.

Generic segmentation models have the advantage that they have been previously trained with such a large number of images that they can operate under the narrow paradigm of zero-shot learning.These models are able to work even when they do not know a new object. To date, to the best of our knowledge, there is no previous study that makes use of these models with FSL based on CP-CVV, continuous and fast learning.

II. OVERVIEW OF RELATED WORK

In the realm of collaborative robotics, extensive research aims to deepen our comprehension of environmental interactions and object handling, striving to replicate human-

[1]Institute of Advanced Production Technologies - Department of Systems Engineering and Automatics (ITAP-DISA), Prado de la Magdalena 3-5, 47011, Institute of Advanced Production Technologies - Department of Systems Engineering and Automatics (ITAP-DISA), School of Industrial Engineers, University of Valladolid, Spain jaime.duque@uva.es

like finesse. Pioneering studies underscore the adoption of sophisticated perception tools, like computer vision, empowering robots to grasp their surroundings with clarity [7]. Furthermore, there's a surge in crafting algorithms that mirror human adeptness in seizing and maneuvering diverse objects, tackling issues like accommodating varied shapes [8]. Techniques like demonstration learning [9] and reinforced learning [10] have also been harnessed to foster more intuitive robot-human interactions. Collectively [11], these strides propel robots towards discerning their environment and deftly handling objects, paralleling human capabilities.

One notable robotic strategy employs reinforcement learning (RL). Here, robots autonomously forge control strategies via iterative trials. Lobbezoo *et al.* [12] meld traditional and RL controls in both virtual and tangible settings, endorsing the RL technique for conventional industrial chores like reaching, grasping, and placing. Their ambition is to imbue robotic control with intelligence, enabling task completion sans meticulous environment, constraint, or action blueprint delineation. Diverging from prevalent methods where robots pinpoint and execute grasps, Kalashnikov *et al.* [13] advocate a vision-centric, closed-loop control paradigm. Herein, the robot continually refines its grasp tactics based on fresh data, optimizing success trajectories. Their model, termed QT-Opt, is grounded in self-guided vision reinforcement learning, harnessing vast real-world grasp trials to train a deep neural network.

Addressing the quandary of optimal grasp locations, Mahler *et al.* [14] harness a synthetic dataset encompassing myriad point clouds, grasps, and analytic metrics to train a predictive grasp success model. Their framework has since evolved, accommodating vacuum suction tools [15] and dual-arm robotic systems [16]. In a similar vein, [17] curated a dataset spotlighting real-world manipulable items, proffering detailed pose insights and affordance forecasts. Their annotation process, leveraging a standard camera and semi-automated techniques, yields pristine 3D annotations, bypassing crowd-sourcing. Another noteworthy approach employs a multi-stage grasp detection algorithm for Kinova robots in congested settings [18], eclipsing rival algorithms on the VMRD dataset [19].

Within the domain of robotic learning via demonstration, some endeavors emphasize gesture recognition via human skeletal data, employing neural networks and Markovian frameworks [20]. Concurrently, others delve into the nuances of human demonstrations across disparate contexts [21], championing imitation learning paradigms. A niche avenue explores robot eye-hand synchronicity [22], wherein robots glean task cues from human videos to guide real-time actions. Merging human demonstration insights with RL, San *et al.* [23] advocate for continuous robot-environment interactions to hone skills. Similarly, Kamali *et al.* [24] harness virtual reality, guiding robotic actions via hand gestures. Cabi *et al.* [25] curate diverse manipulation task policies through varied techniques, amalgamating human preferences for nuanced task rewards.

The main problems found in the reviewed previous methods are that they are not capable of learning without simulation or pre-training, and from a single example of how the person performs the action. In addition, audio/text-guided learning increases the naturalness of learning. Our research not only overcomes the inherent challenges of RL, but also transforms the paradigms of object handling without the need for extensive and intricate object representations, ensuring broad adaptability.

III. ANALYSIS OF THE SYSTEM

Our system allows the robot to learn to perform new unfamiliar activities by imitation in a fast way. In this paper we present the operation of the system for a grip case, although it is extensible to other more complex tasks that require continuous learning. The system uses audio prompts to guide the robot. On the one hand, if we want the robot to learn a new object, it will be enough to say "robot, learn screwdriver". On the other hand, if we want the robot to pick up an object, the audio prompt will be a message that semantically represents the action of picking up that object, for example: "please robot, pick up the screwdriver". The system has two distinct parts, the acquisition of new objects and the processing and localization of objects. For this second part we use the recent CP-CVV method which has been adapted here to a robotics problem.

A. Registration of new objects

For the registration of new objects, the user tells the robot through an audio prompt that he wants to learn a new object (see Figure 1). This prompt is converted into text by employing an automatic speech recognition (ASR) system based on a pre-trained OpenAI Whisper model [26]. In case the system identifies that the user wants to learn a new object, the robot then takes two consecutive images: one with the object in the scene and the other showing how a person grasps it. Using a model of obtaining the key points of the hand based on Mediapipe [27], we extract various grasping coordinates in 3D space. The coordinates are used to identify the new object to be recorded. Using a generic segmentation based on SAM [4], we extract the object. In the database we store the encoding of the object name, the segmented image of the object and its mask, and the relative hand landmarks.

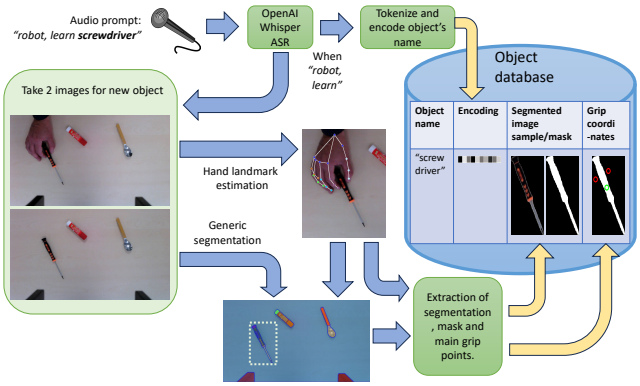


Fig. 1: Registration of new objects (e.g., screwdriver).

B. Locating and grasping objects

Figure 2 shows the general scheme of the system in operation. Objects in the database can be registered and added at different times. During the operation of the system, a person asks the robot to pick up some object by means of audio prompts such as "Could you pick up the toy?". This prompt is converted into text with OpenAI Whisper model [26]. Then, the text is analyzed to find out which is the object we are looking for. For this purpose, the text is encoded and by means of a semantic comparison [3] with the encoding of the different objects (e.g., "screwdriver", "toy" or "spoon") we obtain which is the object that the person is looking for. Then, we obtain an image with the robot's camera and launch a focused segmentation making use of a SAM [4].

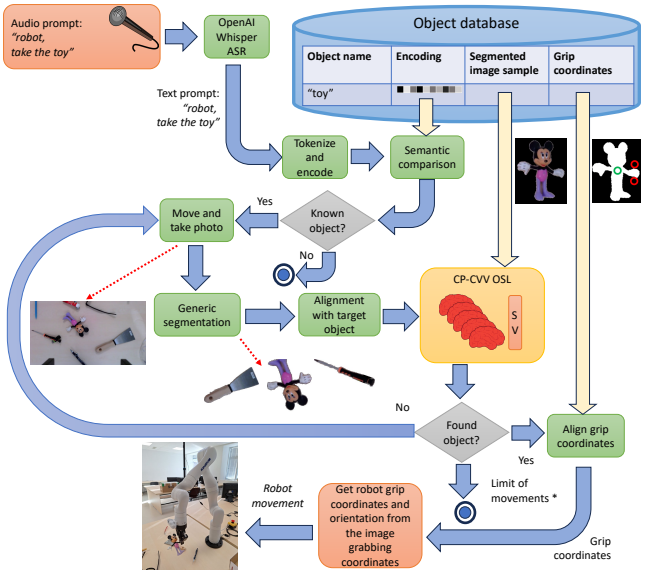


Fig. 2: Schematic of the system in operation.

Classification of segmented objects is carried out using CP-CVV [6]. In CP-CVV, a set of Siamese networks learn to distinguish whether two images belong to the same class. For this purpose, the networks are previously trained with data belonging to classes that have nothing in common with the classes that will later be used in operation. Our CP-CVV has been trained against Fss-1000 dataset [28], which includes 1,000 different classes of objects. To improve the classification of the objects, we carry out an alignment of the objects found in the image with the axis of inertia of the object searched in the database.

In CP-CVV, validation sets are formed for each k slot by distributing the n classes among k validation slots (refer to Figure 3). Specifically, for each of the k training instances of a model, the validation set is composed of approximately n/k classes. Prior to allocating validation slots, the order of the classes is randomized. This approach ensures that potentially related classes are not grouped together during a single training session.

CP-CVV model incorporates k Siamese neural networks through a soft/hard voting mechanism. Unlike the training

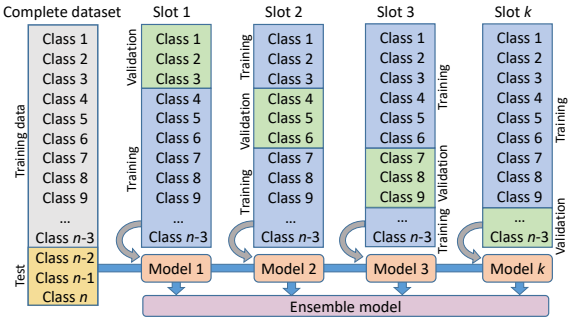


Fig. 3: Distribution of classes in k slots.

methodology of typical Siamese nets, it involves training of each of the k independent networks separately, utilizing distinct sets of training and validation classes. During inference, the model receives two images, signifying a positive pair if the images belong to the same class and negative otherwise. The pair of images is inputted into each of the k Siamese networks, all constructed with the same type of backbone, generating a feature vector. In our case we have performed experiments using a simple backbone based on ConvNeXt-small [29], which has allowed us to speed up training. While the weights within a Siamese network's backbone are similar, they differ across the k models. The output feature vector from each backbone undergoes element-wise multiplication. Subsequently, three dense layers are incorporated, including dropout and batch normalization. The initial dense layer employs a ReLU activation function, the second employs a sigmoid activation, and the final dense layer is responsible for classification using another sigmoid activation function. In contrast to conventional Siamese networks, which often use Euclidean distance for connecting backbone features, this model opts for multiple dense hidden layers. Each Siamese network's output approximates a value of 0 or 1, depending on whether the pair is classified as positive or negative. In the integration of multiple classifiers through hard-voting, positive and negative pairs are counted, and the final output is determined by the majority count. In the case of soft-voting, the output values are accumulated across different classifiers and divided by the total number of classifiers. If the result exceeds $\frac{1}{2}$, the pair is classified as positive; otherwise, it is deemed negative.

Let τ be the set that includes all the classes of the dataset, λ the set of classes used for training and β the set of classes for testing. Let T_i and V_i be the training and validation sets corresponding to the slot i and k the number of slots used in CP-CVV. These sets must verify Equations 1 to 4.

$$\tau = \lambda \cup \beta \tag{1}$$

$$\lambda = \bigcup_{i=1}^{k} V_i \tag{2}$$

$$\bigcap_{i=1}^{k} V_i = \emptyset \tag{3}$$

$$[T_i = \lambda - V_i] \forall i \in k \tag{4}$$

Each of the k models undergoes training with its respective training and validation slot. This data distribution strategy enhances the ensemble model's ability to generalize across various scenarios, mitigating the risk of validation overfitting. In the inference phase, voting is employed. For a given input sample, denoted as x, $p_i(x)$ represents the sigmoid output value generated by the Siamese network i. In the sigmoid output scenario, the output assumes the value 0 when the images belong to the same class and 1 otherwise. Equation (5) illustrates the soft voting mechanism we have used, achieved by aggregating the output values from all the classifiers. Each classifier i is associated with a weight w_i, set to $\frac{1}{k}$ in our case. If the cumulative result surpasses $\frac{1}{2}$, it signifies that the images belong to different categories, leading to a global output of 1. We have used soft instead of hard since the Fss-1000 training results were significantly better with that method.

$$S(x) = \begin{cases} 1 & \text{if } \left[\sum_{i=1}^{k} w_i \cdot p_i(x)\right] > \frac{1}{2} \\ 0 & otherwise \end{cases} \tag{5}$$

Siamese networks enable us to determine whether two images belong to the same class. However, in practical classification applications, the primary goal is often to accurately classify images into specific categories. For instance, in our robotics problem, there may be a need to introduce new objects and have the system automatically categorize images of incoming objects within these classes, with only one original image per class. In the CP-CVV framework, the outputs of multiple Siamese networks are mathematically comparable, indicating whether two images belong to the same category. Our focus shifts to analyzing the cumulative probability that an image belongs to each potential category. To achieve this, we conduct $k \cdot c$ inferences from the Siamese networks, resulting in a matrix with k rows and c columns. Each cell in the matrix denotes the probability that an image belongs to class c in the k slot. By accumulating the column values of a specific cell and dividing by k, we obtain the probability that an image belongs to that class across all slots.

Let P_c represent the cumulative result obtained by summing the sigmoid outputs of different Siamese networks for a test class c, where $c \in \beta$. Therefore, P_c signifies the value for a test image belonging to a particular category, assessed by selecting one random image per category. Let p_{ci} denote the sigmoid output of classifier i with an image from category c. For soft voting, P_c is calculated as per (6).

$$P_c(x) = \sum_{i=1}^{k} w_i \cdot p_{ci}(x) \tag{6}$$

To identify the most similar category, we select the one with the closest proximity to 0. This is accomplished through the use of the $arg_c min$ function, as depicted in Equation (7). This function yields the winning class.

$$C_{soft}(x) = arg_c min\,[P_c(x)] \tag{7}$$

Once our system locates the object, based on different thresholds of the CP-CVV classification, we align the hand gripper coordinates according to how the object was in the database and how it has been found. That allows us to obtain the orientation of the robotic gripper and move the robot, which has previously been calibrated. In case the robot does not find the object, a scanning work is carried out in different areas.

IV. EXPERIMENTS AND RESULTS

Before starting the operation of our system, we had to carry out a training of the Siamese networks used by the CP-CVV method to learn to distinguish new objects. To train them we have chosen a database of images with objects and segmentations and segmentations called Fss-1000 [28]. This dataset has 1,000 different types of objects along with their segmentations. Since our system works with the segmented object images, we have applied the masks to the objects to obtain their segmentations. Figure 4 shows two inputs during the training of the CP-CVV model, where we can see that the image injected into the Siamese is the segmented one. We have used a ConvNeXt-small backbone [29]. On the right we see the probabilities of each output of the Siamese networks and of the total ensemble model. For training we have divided the 1,000 classes into 100 test classes and 900 training classes. The 900 have been distributed using the above mentioned scheme. As we used $K = 5$ slots, each Siamese was trained with 720 training classes and 180 validation classes. The training was carried out with data augmentation with changes in perspective, rotation, translation and illumination. An Adam optimizer with a learning factor of 10^{-4} and Binary Cross Entropy loss has been used. The training plots are shown in Figure 5.

The classification results were evaluated against the test data. For each test image, random images were selected from its category and from other categories. The classification result of an object from the Fss-1000 database into its correct class can be seen in Table I. It shows after 5 runs the average results of an individual Siamese with ConvNeXt model and of the model using CP-CVV.

TABLE I: Classification results of CP-CVV with FSS-1000

Model	Accuracy	Loss
Siamese with ConvNeXt small backbone	0.8604	2.8866
CP-CVV soft (K = 5 slots)	0.9230	2.1504

From the trained CP-CVV model, we performed the experiments on the robot. The experimentation has been conducted using a Kinova Gen 3 robot (see Figure 6). This robot integrates an RGB-D camera in the wrist that we used to take the images. The robot has been connected to a computer with RTX-3090 GPU for faster operation.

Semantic text comparison was performed by setting a threshold of 0.6 for text similarity based on the cosine

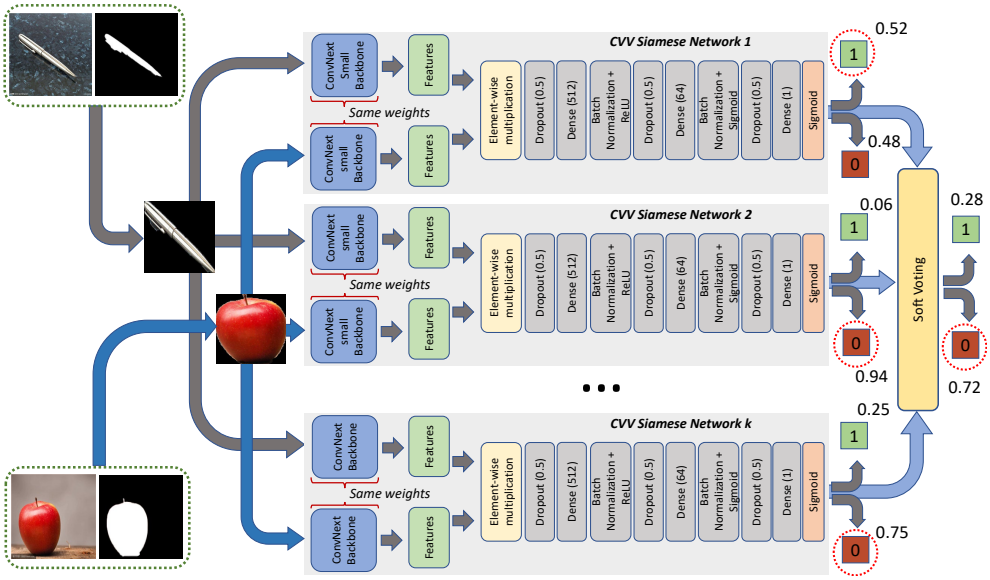
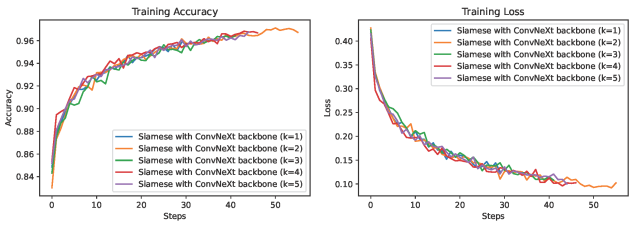


Fig. 4: CP-CVV model and FSS-1000 training image input.



(a) Accuracy (b) Loss

Fig. 5: Siamese nets training.

(a) Scene (b) Toy grip

Fig. 6: Experiments with Kinova Gen 3: Holding a toy by the hand. The robot has learned how a person grasps it. Unlike classical vision methods, the robot can learn complex grips by imitation. Objects may overlap.

similarity of the encoding vector produced. This way we managed to identify if the object exists in the database. In the CP-CVV method we also use a threshold of 0.9 as soft voting probability. This allows us to filter whether the object is identified by vision.

In the experiments with the Kinova we saw that the CP-CVV model confused some objects such as screws. To solve this problem, we saw that the results were significantly improved by aligning the input image of the Siamese with the inertial axis of the searched object, which as explained above was known from the semantic content of the text. We tested adding up to 15 different objects and obtained a 93% success rate in picking them up (after 100 pick-up tests and including some cases of overlapping objects). Learning each new object takes less than 2 seconds while locating an object takes 3 seconds. Gripping error is below $\pm 1cm$. This error is mainly caused by object segmentation variations.

Although it difficult to quantify and compare metrics of different methods under exact conditions, such as success rate and time, most of the methods in the state of the art offer results around 70-90% of correct reproduction of the activity (according to their experiments). For example, in [5] the authors show results between 76.7% and 91.7% in successful grasps, but with a FSL approach. Our method obtained 93% success rate for our selected activity type with a OSL approach.

V. CONCLUSIONS

We have presented in this paper a fast continuous learning system initially focused on grasping objects. The system uses audio prompts to operate and learns by imitation to grasp objects just as a person does. We use semantic comparison models for object selection based on text, and generic segmentation models and the CP-CVV method for object classification and selection from images. Our model was pretrained with a segmentation dataset of images totally different from those used in operation. The application of generic segmentation with the CP-CVV method avoids the costly work of having to create specific datasets and carry out lengthy training.

The experimentation has been carried out on a Kinova Gen 3 robot, obtaining an accuracy close to 93% of object grasping without catastrophic forgetting. The robot performs a search for the object in different positions and the learning of new objects is very fast, taking only a few seconds. Unlike classical vision methods, the robot can learn complex grips by imitation. In addition, objects may overlap. Our future line of research seeks to integrate this technique in the learning of more complex tasks, such as feeding a person.

Future work will consist of integrating SAM and CP-CVV into a single model that selects the winning mask from a single entry of an object type. However, this approach requires a large dataset to generalize to new objects. Moreover, intensive experimentation with other kinds of objects could be addressed, provided that an adequate robot grip is attached to the robot. Results are expected to be satisfactory, as long as rigid objects are used.

ACKNOWLEDGMENT

This research has received funding from projects ROSOGAR PID2021-123020OB-I00 funded by MCIN/AEI/10.13039/501100011033/FEDER, UE, and EIAROB funded by Consejería de Familia of the Junta de Castilla y León - Next Generation EU.

REFERENCES

[1] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020.

[2] Ali Ayub and Carter Fendley. Few-shot continual active learning by a robot. *Advances in Neural Information Processing Systems*, 35:30612–30624, 2022.

[3] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[4] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[5] Pengyuan Wang, Fabian Manhardt, Luca Minciullo, Lorenzo Garattoni, Sven Meier, Nassir Navab, and Benjamin Busam. Demograsp: Few-shot learning for robotic grasping with human demonstration. In *2021 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 5733–5740. IEEE, 2021.

[6] Jaime Duque-Domingo, Roberto Medina Aparicio, and Luis Miguel González Rodrigo. One shot learning with class partitioning and cross validation voting (cp-cvv). *Pattern Recognition*, 143:109797, 2023.

[7] Min Zhao, Guoyu Zuo, Shuangyue Yu, Daoxiong Gong, Zihao Wang, and Ouattara Sie. Position-aware pushing and grasping synergy with deep reinforcement learning in clutter. *CAAI Transactions on Intelligence Technology*, 2023.

[8] Rhys Newbury, Morris Gu, Lachlan Chumbley, Arsalan Mousavian, Clemens Eppner, Jürgen Leitner, Jeannette Bohg, Antonio Morales, Tamim Asfour, Danica Kragic, et al. Deep learning approaches to grasp synthesis: A review. *IEEE Transactions on Robotics*, 2023.

[9] Harish Ravichandar, Athanasios S Polydoros, Sonia Chernova, and Aude Billard. Recent advances in robot learning from demonstration. *Annual review of control, robotics, and autonomous systems*, 3:297–330, 2020.

[10] Lukas Brunke, Melissa Greeff, Adam W Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P Schoellig. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 5:411–444, 2022.

[11] Qijie Zou, Kang Xiong, Qiang Fang, and Bohan Jiang. Deep imitation reinforcement learning for self-driving by vision. *CAAI Transactions on Intelligence Technology*, 6(4):493–503, 2021.

[12] Andrew Lobbezoo and Hyock-Ju Kwon. Simulated and real robotic reach, grasp, and pick-and-place using combined reinforcement learning and traditional controls. *Robotics*, 12(1), 2023.

[13] Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation. *arXiv preprint arXiv:1806.10293*, 2018.

[14] Jeffrey Mahler, Jacky Liang, Sherdil Niyaz, Michael Laskey, Richard Doan, Xinyu Liu, Juan Aparicio Ojea, and Ken Goldberg. Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics, 2017.

[15] Jeffrey Mahler, Matthew Matl, Xinyu Liu, Albert Li, David Gealy, and Ken Goldberg. Dex-net 3.0: Computing robust vacuum suction grasp targets in point clouds using a new analytic model and deep learning. In *2018 IEEE International Conference on robotics and automation (ICRA)*, pages 5620–5627. IEEE, 2018.

[16] Jeffrey Mahler, Matthew Matl, Vishal Satish, Michael Danielczuk, Bill DeRose, Stephen McKinley, and Ken Goldberg. Learning ambidextrous robot grasping policies. *Science Robotics*, 4(26):eaau4984, 2019.

[17] Andrew Guo, Bowen Wen, Jianhe Yuan, Jonathan Tremblay, Stephen Tyree, Jeffrey Smith, and Stan Birchfield. Handal: A dataset of real-world manipulable object categories with pose annotations, affordances, and reconstructions, 2023.

[18] Xuefeng Dong, Yang Jiang, Fengyu Zhao, and Jingtao Xia. A practical multi-stage grasp detection method for kinova robot in stacked environments. *Micromachines*, 14(1), 2023.

[19] Hanbo Zhang, Xuguang Lan, Xinwen Zhou, Zhiqiang Tian, Yang Zhang, and Nanning Zheng. Visual manipulation relationship network for autonomous robotics. In *2018 IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids)*, pages 118–125, 2018.

[20] Jaime Duque Domingo, Jaime Gómez-García-Bermejo, and Eduardo Zalama. Visual recognition of gymnastic exercise sequences. application to supervision and robot learning by demonstration. *Robotics and Autonomous Systems*, 143:103830, 2021.

[21] Zhifeng Qian, Mingyu You, Hongjun Zhou, Xuanhui Xu, and Bin He. Robot learning from human demonstrations with inconsistent contexts. *Robotics and Autonomous Systems*, 166:104466, 2023.

[22] Jun Jin, Laura Petrich, Masood Dehghan, Zichen Zhang, and Martin Jagersand. Robot eye-hand coordination learning by watching human demonstrations: a task function approximation approach. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 6624–6630, 2019.

[23] Xilong Sun, Jiqing Li, Anna Vladimirovna Kovalenko, Wei Feng, and Yongsheng Ou. Integrating reinforcement learning and learning from demonstrations to learn nonprehensile manipulation. *IEEE Transactions on Automation Science and Engineering*, 20(3):1735–1744, 2023.

[24] Kaveh Kamali, Ilian A. Bonev, and Christian Desrosiers. Real-time motion planning for robotic teleoperation using dynamic-goal deep reinforcement learning. In *2020 17th Conference on Computer and Robot Vision (CRV)*, pages 182–189, 2020.

[25] Serkan Cabi, Sergio Gómez Colmenarejo, Alexander Novikov, Ksenia Konyushkova, Scott Reed, Rae Jeong, Konrad Zolna, Yusuf Aytar, David Budden, Mel Vecerik, Oleg Sushkov, David Barker, Jonathan Scholz, Misha Denil, Nando de Freitas, and Ziyu Wang. Scaling data-driven robotics with reward sketching and batch reinforcement learning, 2020.

[26] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.

[27] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214*, 2020.

[28] Xiang Li, Tianhan Wei, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. Fss-1000: A 1000-class dataset for few-shot segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2869–2878, 2020.

[29] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *arXiv preprint arXiv:2201.03545*, 2022.