One-Shot Learning for Robotic Manipulators: Rapid Replication of Human Activities from a Single Demonstration

Jaime Duque-Domingo[1,*,†], Riccardo Caccavale[2,†], Alberto Finzi[2,†], Eduardo Zalama[1,3,†] and Jaime Gómez-García-Bermejo[1,3,†]

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

[2]*PRISMA Lab. Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione. Università degli Studi di Napoli "Federico II", Napoli, Italy*

[3]*Centro Tecnológico CARTIF, Boecillo, 47151 Valladolid, Spain*

Abstract

This paper presents a One-Shot Learning framework able to process a RGB-D video of a human task demonstration and to perform it on a robot manipulator. Learning from a single human demonstration is one of the most interesting challenges in robotics. The aim is to allow a robot to reproduce operator's activities after observing how they are performed just once. Although the work presented in this paper focuses on specific manipulation tasks, the proposed method can be extended to multi-stage operations carried out in different fields, both domestic and industrial. In the proposed approach, the demonstration is first segmented into primitives, which are then mapped into robot actions to be executed by a manipulator. This work also aims to ensure that the learning process is carried out rapidly. The paper provides an overview of the overall framework and illustrates the system at work in a use case.

Keywords

One-Shot Learning, robot learning, human demonstrations, activity segmentation

1. Introduction

This paper introduces a one-shot learning framework capable of processing RGB-D video of a human task demonstration involving known objects and replicating the task using a robotic manipulator. One-shot learning allows a robot to imitate tasks or activities after only a single observation. This problem is both relevant and challenging, as it offers the advantage of quickly acquiring new skills, while requiring effective use of prior knowledge to generalize from just one example. In robotics, one-shot learning represents a significant leap forward, in that it allows robots to quickly and efficiently learn tasks that would otherwise require extensive training, mirroring the adaptive learning process of humans.

In this work, we propose a novel framework that exploits real-time object detection and assumptions about manipulation actions to both segments human demonstrations and flexibly reproduce observed tasks. Specifically, in the proposed approach, a RGB-D recorded human demonstration is firstly segmented and then associated to action primitives, which are composed and adapted to be reproduced by a robotic manipulator acting on the same target. The framework employs Yolo-based 3D object segmentation, alongside human feature tracking (including key hand trajectories and gaze detection), to monitor human-object interactions, enabling the isolation, interpretation, and replication of action primitives. While our current proposal focuses on basic manipulation capabilities (e.g., grasp, drop,

11th Italian Workshop on Artificial Intelligence and Robotics (AIRO 2024)

*Corresponding author.

†These authors contributed equally.

jaime.duque@uva.es (J. Duque-Domingo)

0000-0001-6649-5550 (J. Duque-Domingo)

carry, etc.), the framework is intended to incrementally observe and reproduce multi-step operations across various domains, both domestic and industrial.

2. Overview of related work

In the field of collaborative robotics, significant research efforts aim to enhance our understanding of environmental interactions and object manipulation, aspiring to replicate human dexterity. Many studies highlight the integration of advanced perception tools, such as computer vision, which empower robots to interpret their surroundings with precision [1]. Additionally, there is an increasing focus on developing algorithms that mimic human proficiency in grasping and manipulating various objects, addressing challenges like handling diverse shapes [2, 3]. Techniques such as learning from demonstrations [4, 5], and reinforcement learning [6, 7] are also employed to promote more natural human-robot interactions during task learning. Collectively [8], these advancements enable robots to better understand their environment and skillfully manage objects, approaching human-like abilities.

One prominent robotic technique is reinforcement learning (RL), where robots autonomously develop control strategies through iterative experimentation. Lobbezoo *et al.* [9] combine traditional control methods with RL in both virtual and physical environments, advocating for RL in conventional industrial tasks such as reaching, grasping, and placing. In contrast to typical methods where robots identify and perform grasps, Kalashnikov *et al.* [10] propose a vision-based, closed-loop control system. In this approach, the robot continuously refines its grasp strategy based on new sensory data, optimising its success rates. To tackle the challenge of identifying optimal grasp locations, Mahler *et al.* [11] utilised a synthetic dataset containing a wide range of point clouds, grasps, and analytical metrics to train a predictive model for grasp success. Similarly, Guo *et al.* [12] developed a dataset featuring real-world manipulable objects, providing detailed pose information and affordance predictions. Another notable approach utilises a multi-stage grasp detection algorithm for Kinova robots in cluttered environments [13].

In the area of robot learning from demonstrations, some studies focus on gesture recognition through human skeletal data, leveraging neural networks and Markov models [14]. Others examine human demonstrations across various contexts [15], promoting imitation learning frameworks [16] or kinesthetic teaching of structured tasks [17]. A specialised approach explores robot eye-hand coordination [18], where robots extract task-relevant information from human videos to guide real-time actions. By integrating human demonstration data with RL, San *et al.* [19] advocate for continuous robot-environment interaction to enhance skill acquisition. Similarly, Kamali *et al.* [20] utilise virtual reality to guide robotic actions through hand gestures. Cabi *et al.* [21] develop policies for diverse manipulation tasks using a variety of techniques, incorporating human preferences to refine task rewards.

Differently from these methods, in this paper, we address the challenge of rapid one-shot learning [22, 23, 24]. In particular, we are interested in learning structured robotic manipulation tasks, demonstrated through a single human demonstration captured by an RGB-D camera. In this respect, similarly to [22], the proposed approach focuses on quickly adapting the human demonstration to enable direct task replication, without the need for detailed or complex object models. This ensures broad adaptability while avoiding the extensive training typically required by reinforcement learning (RL) methods [25] or behavior cloning methods on a data-set of tasks [24, 26]. However, in contrast with [22], our approach introduces a novel method that leverages object and action segmentation from RGB-D video, allowing us to isolate and replicate manipulation primitives inferred from the demonstration.

3. Methodology

The proposed system is based on two main stages: *activity recording and processing*, *task reproduction.*

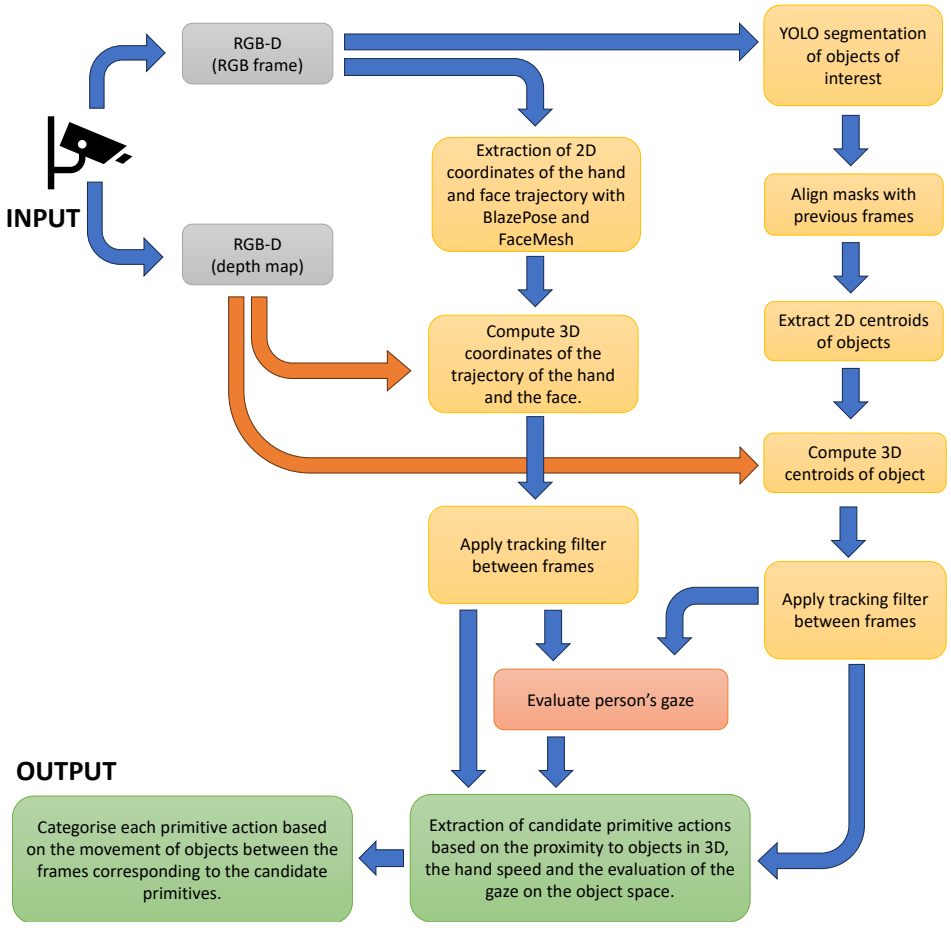

Figure 1: Scheme of the proposed system.

3.1. Activity recording and processing

In the first stage of the proposed pipeline we collect and process a RGB-D video capturing a human activity demonstration. The video is segmented to isolate primitive actions using several features, such as the proximity of the user's hands to relevant objects, the speed of hand movements, and the direction of the user's gaze. The segmentation process is outlined in Figure 1. Initially, RGB frames are analyzed to perform object segmentation and extract key points from the user's hands and face. Depth maps, combined with filters, are then used to derive 3D trajectories of the hand and face points, as well as the 3D centroids of detected objects. The orientation of the user's gaze is subsequently evaluated to identify potential target objects, which are then exploited to segment the human demonstration and isolate candidate primitive actions as interpretation of those segments. The primitive actions are finally classified based on the motion of the associated objects.

The action segmentation process introduced above relies on the proximity and velocity of the operator's hand relative to the detected objects in the scene, with additional reinforcement from the operator's gaze direction. More specifically, segmentation is based on three thresholds u_1, u_2, and u_3. These thresholds are used to determine the operator's intention to interact with objects in the 3D space through specific primitive manipulation actions. The first threshold, u_1, defines the maximum distance between the operator's hand and the centroid of a detected object for an interaction to be considered. If multiple objects fall within this distance, potential interactions are prioritized by proximity. The second threshold, u_2, specifies the maximum hand speed allowed for an interaction to be considered with a nearby object. The third threshold, u_3, sets the maximum allowable angle between the operator's gaze direction and a proximal object to consider a plausible intention to interact. The underlying assumption is that the user gaze should be directed towards the target of a manipulation action. For each primitive

action extracted by the segmentation process, the system tracks and records key positions of the hand trajectory with respect to the centroid of the objects participating in the action.

The overall pipeline described above is built on object segmentation, hand detection/tracking, and gaze direction monitoring. Additional details about these modules are provided below.

Object segmentation is performed using YOLOv8 [27] that exploits a deep convolutional neural network architecture to process images similar to the one of YOLO [28], enhanced with additional layers and a special branch to predict segmentation masks. The output of this module includes bounding boxes, object classes, class probabilities, and segmentation masks.

Hand points detection and tracking are based on the MediaPipe Hand Landmarker [29], which operates in real-time by first using a palm detection model to locate the hand and then predicting 21 key landmarks on the hand, including finger joints, tips, and the wrist. While this method efficiently tracks multiple hands and is optimized for gesture recognition, augmented reality (AR), and interactive applications, we found that the 3D coordinates returned by this model did not yield satisfactory results within our framework. Therefore, as we use a depth camera, we found it more reliable to directly leverage the depth information provided by the RGB-D camera.

Gaze orientation estimation is achieved using MediaPipe's FaceMesh [30], a real-time facial landmark tracking technology that detects 468 key points on the face using a standard camera. FaceMesh detects the face, maps 2D landmarks, and estimates 3D coordinates for each point. Though it provides high efficiency for facial feature tracking, FaceMesh does not directly return the center of the eyes. Thus, interpolation is used to calculate this point. To determine the direction of the gaze, a vector is computed between the center of the forehead and the eyes, giving a normal vector for the gaze direction. The distance from the RGB-D camera to each facial point is used to construct the 3D model of the face.

The basic primitives are fixed: grasping (TAKE), moving (MOVE OVER), waiting (WAIT) and releasing (RELEASE). The activities are decomposed into several primitives. Users can record the activities at different speeds, although there are thresholds for correct detection.

3.2. Task reproduction

Once the human demonstration has been segmented and processed, the interpreted manipulation task is to be reproduced by a robot manipulator operating in the same workspace. The robotic platform is assumed to be a manipulator equipped with a gripper.

For ease of demonstration, it is assumed that the robot and the human are positioned opposite each other in the workspace. Consequently, the trajectories and points collected during task segmentation must first be mirrored and then adapted for the robot's execution. Mirroring is achieved by applying a 180-degree transformation to the objects' axes and the hand interaction points relative to the camera and base marker, both of which are in the same plane as the table where the actions occur.

To enable rapid task reproduction, the robotic system executes the sequence of detected manipulation actions step-by-step, operating on the target objects as demonstrated by the human. The robot first segments the scene using YOLOv8 to detect and locate task-relevant objects before deploying the demonstrated actions. Given the 3D locations of the objects, key points from the human hand trajectory - such as the wrist, index finger, and thumb - are mapped to reproduce the trajectory of the robot's end-effector and gripper movements relative to the target object. For instance, to replicate object grasping, the robot's end-effector follows the trajectory of the human hand to reach the pose necessary for approaching the object, followed by a grasp movement, where the gripper motion is adapted from the recorded motion of the human's index finger and thumb. If the task involves multiple steps or objects, the system continuously monitors the execution of each manipulation action and the status of the target object until the task is completed.

4. Experimentation

Experiments were conducted using a Kuka IIWA 7 robot, a Real Sense D415 camera, an i7 server with an RTX 3060 GPU and ROS 2 software. Human demonstrations were recorded and processed offline,

with a processing time of approximately 1 minute for a 15-second video. During the execution phase, the system receives the task to be performed and begins by segmenting the objects using YOLOv8. Once segmentation is completed, the system retreives each detected primitive action of the task and executes them by leveraging the key points and trajectories associated with the segmented actions.

Figure 2 illustrate the overall system in action, where a person demonstrates picking up a bottle of water and placing it on top of a small bowl. During this one-shot task demonstration, the system identifies two action primitives (Figure 2, second column) involving two target objects (the bottle and the bowl). The robot is then able to rapidly and flexibly reproduce the demonstrated task (Figure 2, second column), regardless of the objects' positions in the workspace, as it learns the relationships between the hand and the objects for each action segment. In this scenario, we observed reliable and precise task reproduction, with an error margin of $\pm$ 1cm during execution. Additional tasks such as grasping, carrying, placing, and pouring were also tested, yielding satisfactory results. However, challenges remain during task demonstration, particularly with depth camera precision and estimation errors when fingers are occluded, which need to be addressed to ensure more robust task detection. As for task replication, we currently assume a clear workspace where obstacles and potential collisions are neglected for simplicity. Future work will focus on developing methods that can handle more complex tasks and environments with obstacles in flexible and reliable manner.

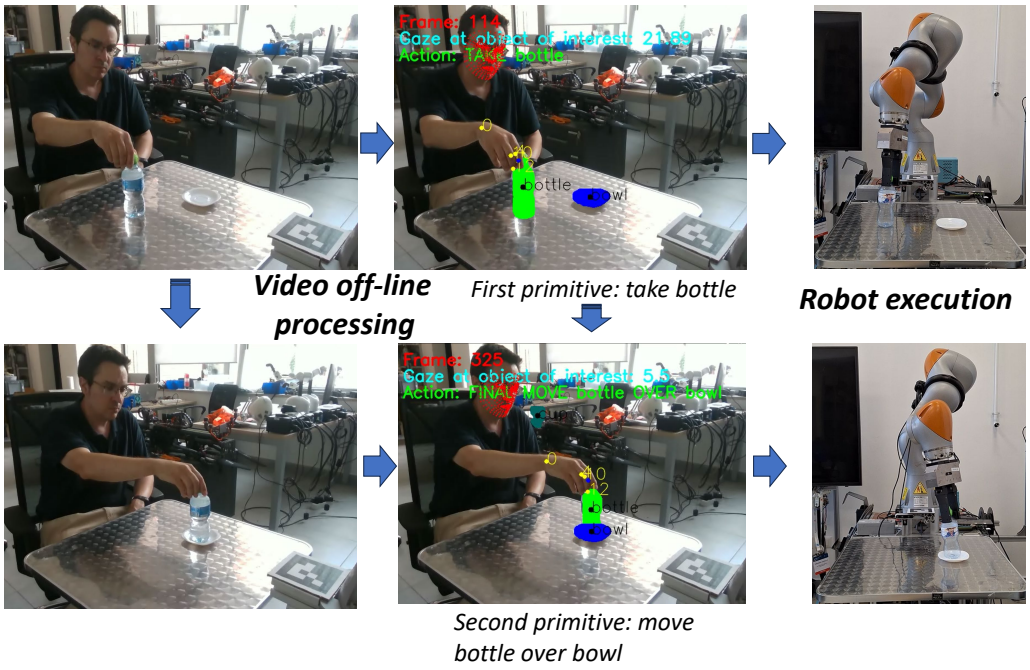

Figure 2: Action demonstration and task reproduction: The operator picks up a bottle and places it over a bowl. During the demonstration, primitive actions are segmented and categorized (left and center). During execution, the manipulator leverages the hand-object relationships captured during the demonstration to flexibly reproduce the sequence of actions (right).

In our experiments we have used a robot controller implementing obstacle-free movement of the end-effector toward the desired pose in the robot's operational space. The processing of the vision is the most computationally expensive part, both for the segmentation and for the execution itself. However, the processing of a new activity is carried out in a few minutes and its execution in the robot is carried out in a few seconds since only the processing of the first frame is required.

5. Conclusions

This work presents a system capable of learning multi-step tasks from human demonstrations and reproducing them on a robot manipulator. Operating under a one-shot learning paradigm, the system aims to enable rapid, flexible, and reliable reproduction of typical manipulation tasks across a dataset of known objects. Currently, the system is being tested on tasks such as picking, carrying, placing, and pouring, performed by a robot manipulator equipped with a gripper. Initial results are promising; however, several challenges remain, particularly in scaling and generalizing task interpretation and reproduction for more complex manipulation scenarios.

Acknowledgments

This research has received funding from projects ROSOGAR PID2021-123020OB-I00 funded by MCIN/AEI/10.13039/501100011033/FEDER, UE, EIAROB funded by Consejería de Familia of the Junta de Castilla y León - Next Generation EU, INVERSE (EU Horizon, grant 101136067), euROBIN (EU Horizon, grant 101070596), Melody (CUP E53D23017550001).

References

[1] M. Zhao, G. Zuo, S. Yu, D. Gong, Z. Wang, O. Sie, Position-aware pushing and grasping synergy with deep reinforcement learning in clutter, CAAI Transactions on Intelligence Technology (2023).

[2] K. Kleeberger, R. Bormann, W. Kraus, M. F. Huber, A survey on learning-based robotic grasping, Current Robotics Reports 1 (2020) 239–249.

[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] H. Ravichandar, A. S. Polydoros, S. Chernova, A. Billard, Recent advances in robot learning from demonstration, Annual review of control, robotics, and autonomous systems 3 (2020) 297–330.

[5] B. Fang, S. Jia, D. Guo, M. Xu, S. Wen, F. Sun, Survey of imitation learning for robotic manipulation, International Journal of Intelligent Robotics and Applications 3 (2019) 362–369.

[6] L. Brunke, M. Greeff, A. W. Hall, Z. Yuan, S. Zhou, J. Panerati, A. P. Schoellig, Safe learning in robotics: From learning-based control to safe reinforcement learning, Annual Review of Control, Robotics, and Autonomous Systems 5 (2022) 411–444.

[7] B. Singh, R. Kumar, V. P. Singh, Reinforcement learning in robotic applications: a comprehensive survey, Artificial Intelligence Review (2022) 1–46.

[8] Q. Zou, K. Xiong, Q. Fang, B. Jiang, Deep imitation reinforcement learning for self-driving by vision, CAAI Transactions on Intelligence Technology 6 (2021) 493–503. URL: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/cit2.12025. doi:https://doi.org/10.1049/cit2.12025. arXiv:https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/cit2.12025.

[9] A. Lobbezoo, H.-J. Kwon, Simulated and real robotic reach, grasp, and pick-and-place using combined reinforcement learning and traditional controls, Robotics 12 (2023). URL: https://www.mdpi.com/2218-6581/12/1/12. doi:10.3390/robotics12010012.

[10] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, et al., Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation, arXiv preprint arXiv:1806.10293 (2018).

[11] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, K. Goldberg, Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics, 2017. arXiv:1703.09312.

[12] A. Guo, B. Wen, J. Yuan, J. Tremblay, S. Tyree, J. Smith, S. Birchfield, Handal: A dataset of real-

world manipulable object categories with pose annotations, affordances, and reconstructions, 2023. arXiv:2308.01477.

[13] X. Dong, Y. Jiang, F. Zhao, J. Xia, A practical multi-stage grasp detection method for kinova robot in stacked environments, Micromachines 14 (2023). URL: https://www.mdpi.com/2072-666X/14/1/117. doi:10.3390/mi14010117.

[14] J. D. Domingo, J. Gómez-García-Bermejo, E. Zalama, Visual recognition of gymnastic exercise sequences. application to supervision and robot learning by demonstration, Robotics and Autonomous Systems 143 (2021) 103830.

[15] Z. Qian, M. You, H. Zhou, X. Xu, B. He, Robot learning from human demonstrations with inconsistent contexts, Robotics and Autonomous Systems 166 (2023) 104466. URL: https://www.sciencedirect.com/science/article/pii/S0921889023001057. doi:https://doi.org/10.1016/j.robot.2023.104466.

[16] R. Caccavale, M. Saveriano, G. A. Fontanelli, F. Ficuciello, D. Lee, A. Finzi, Imitation learning and attentional supervision of dual-arm structured tasks, in: Proc. of ICDL-EpiRob, 2017, pp. 66–71.

[17] R. Caccavale, M. Saveriano, A. Finzi, D. Lee, Kinesthetic teaching and attentional supervision of structured tasks in human-robot interaction, Auton. Robots 43 (2019) 1291–1307.

[18] J. Jin, L. Petrich, M. Dehghan, Z. Zhang, M. Jagersand, Robot eye-hand coordination learning by watching human demonstrations: a task function approximation approach, in: 2019 International Conference on Robotics and Automation (ICRA), 2019, pp. 6624–6630. doi:10.1109/ICRA.2019.8793649.

[19] X. Sun, J. Li, A. V. Kovalenko, W. Feng, Y. Ou, Integrating reinforcement learning and learning from demonstrations to learn nonprehensile manipulation, IEEE Transactions on Automation Science and Engineering 20 (2023) 1735–1744. doi:10.1109/TASE.2022.3185071.

[20] K. Kamali, I. A. Bonev, C. Desrosiers, Real-time motion planning for robotic teleoperation using dynamic-goal deep reinforcement learning, in: 2020 17th Conference on Computer and Robot Vision (CRV), 2020, pp. 182–189. doi:10.1109/CRV50864.2020.00032.

[21] S. Cabi, S. G. Colmenarejo, A. Novikov, K. Konyushkova, S. Reed, R. Jeong, K. Zolna, Y. Aytar, D. Budden, M. Vecerik, O. Sushkov, D. Barker, J. Scholz, M. Denil, N. de Freitas, Z. Wang, Scaling data-driven robotics with reward sketching and batch reinforcement learning, 2020. arXiv:1909.12200.

[22] Y. Wu, Y. Demiris, Towards one shot learning by imitation for humanoid robots, in: 2010 IEEE International Conference on Robotics and Automation (ICRA 2010), 2010, pp. 2889–2894.

[23] M. A. R. S. A. L. J. W. M Kopicki, R Detry, One-shot learning and generation of dexterous grasps for novel objects, The International Journal of Robotics Research 35 (2015).

[24] S. Dasari, A. Gupta, Transformers for one-shot visual imitation, in: J. Kober, F. Ramos, C. Tomlin (Eds.), Proceedings of the 2020 Conference on Robot Learning, volume 155 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 2071–2084.

[25] J. Fu, S. Levine, P. Abbeel, One-shot learning of manipulation skills with online dynamics adaptation and neural network priors, in: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2016, pp. 4019–4026.

[26] T. Z. P. A. S. L. Chelsea Finn, Tianhe Yu, One-shot visual imitation learning via meta-learning, in: In Conference on Robot Learning, 2017, pp. 357–368.

[27] G. Jocher, A. Chaurasia, J. Qiu, Ultralytics yolo v8, https://docs.ultralytics.com/models/yolov8/, 2023. Last access: 2024-07-15.

[28] J. Redmon, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.

[29] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, M. Grundmann, Mediapipe hands: On-device real-time hand tracking, arXiv preprint arXiv:2006.10214 (2020).

[30] Y. Kartynnik, A. Ablavatski, I. Grishchenko, M. Grundmann, Real-time facial surface geometry from monocular video on mobile gpus, arXiv preprint arXiv:1907.06724 (2019).