

Grasp Anything: Combining Teacher-Augmented Policy Gradient Learning with Instance Segmentation to Grasp Arbitrary Objects

Malte Mosbach and Sven Behnke

Abstract—Interactive grasping from clutter, akin to human dexterity, is one of the longest-standing problems in robot learning. Challenges stem from the intricacies of visual perception, the demand for precise motor skills, and the complex interplay between the two. In this work, we present Teacher-Augmented Policy Gradient (TAPG), a novel two-stage learning framework that synergizes reinforcement learning and policy distillation. After training a teacher policy to master the motor control based on object pose information, TAPG facilitates guided, yet adaptive, learning of a sensorimotor policy, based on object segmentation. We zero-shot transfer from simulation to a real robot by using Segment Anything Model for promptable object segmentation. Our trained policies adeptly grasp a wide variety of objects from cluttered scenarios in simulation and the real world based on human-understandable prompts. Furthermore, we show robust zero-shot transfer to novel objects. Videos of our experiments are available at https://maltemosbach.github.io/grasp_anything.

I. INTRODUCTION

Grasping and retrieving a specific object is a fundamental, yet highly challenging sensorimotor skill, underpinning diverse manipulation tasks ranging from pick-and-place to tool-use. While seemingly trivial for humans, grasping has been remarkably difficult to automate. Consider the setup in Figure 1 where an anthropomorphic robot hand should perform grasping from a cluttered pile. The target item is user-selected among multiple objects of unknown geometry. Solving this task presents two key challenges: i) *identifying and discerning targets* based on user cues and ii) the motor capability required for *dexterous grasping and retrieval*.

The recent surge of vision foundation models (VFMs) has enabled semantic understanding and targeted segmentation of arbitrary scenes from human-understandable input. This can effectively address the identification of a target object, which can then be represented to a policy via its segmented point cloud. However, while VFMs have found widespread adoption in computer vision, their perceptual power and understanding are rarely used in reinforcement learning (RL).

To understand this apparent disconnect, consider the prevalent training procedure of RL methods. Using RL to learn sophisticated manipulation behaviors is notoriously sample-inefficient, requiring on the order of hundreds of millions of interactions—a problem that is compounded when learning from high-dimensional observations and in complex action spaces. Utilizing high-capacity vision models to produce

All authors are with the Autonomous Intelligent Systems group, Computer Science Institute VI – Intelligent Systems and Robotics – and the Center for Robotics and the Lamarr Institute for Machine Learning and Artificial Intelligence, University of Bonn, Germany; mosbach@ais.uni-bonn.de

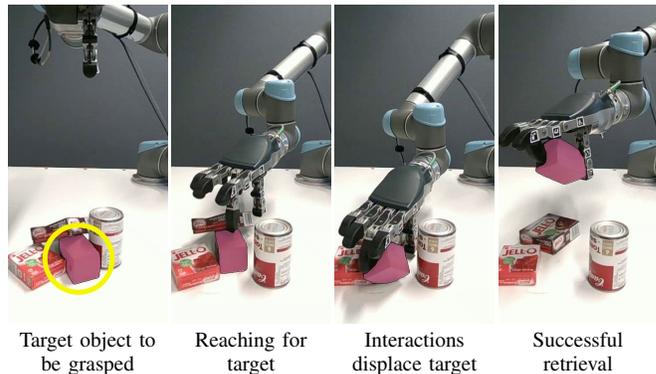


Fig. 1: Robotic setup demonstrating the retrieval of a specific item from a cluttered environment. The agent dynamically grasps the user-specified object, recovering from failures and changes in the scene resulting from object interactions.

observations pushes computational and sample complexities beyond currently feasible limits. To train visual policies more efficiently, policy distillation (PD) has recently been used to transfer knowledge from policies trained on low-dimensional states to high-dimensional visual observations [1], [2]. However, differences between the characteristics of these observation spaces and resulting optimal behaviors have not yet been considered. This neglect becomes particularly pronounced for tasks with substantial interplay between action and perception. For targeted grasping, self-induced occlusion is the primary example of this phenomenon. While the teacher learns to solve the task in an unrestricted manner, fully occluding an object during the grasping procedure may obstruct tracking and lead to missing information for the student policy. Hence, different behaviors become feasible depending on the policy’s observation-space – a disparity not considered in prior works.

To address this limitation, we introduce a new guided student-teacher learning algorithm, Teacher-Augmented Policy Gradient (TAPG), that amalgamates ideas from RL and PD. TAPG centers on the idea that the majority of the dexterous behaviors required to solve a manipulation task can be learned from low-dimensional observations, such that only the adaptation of this knowledge to a new observation space is required from the student. As a result, only a small fraction of observations are generated in a high-dimensional, visual observation space. This methodology enables the learning of perception-aware control strategies without the need to learn the entire policy from visual inputs. We demonstrate this advantage by fine-tuning a vision-agnostic teacher with

an additional target visibility reward. This enables the real-world deployment of the derived policies using a pre-trained VFM promptable object segmentation model. To assess the robustness of the learned policies, we test them on unseen objects and in cluttered environments. These evaluations showcase strong zero-shot transfer and efficient recovery from failures. In summary, our main contributions are:

- We propose a novel two-stage learning framework, TAPG, that synergizes reinforcement learning and policy distillation to learn sensorimotor policies.
- We show how TAPG can be employed to integrate a pre-trained segmentation model with RL, resulting in a real-world executable, promptable grasping policy.
- We demonstrate the robustness of the learned behaviors in cluttered environments and on unseen objects.

II. RELATED WORK

A. Vision Foundation Models

Vision foundation models (VFMs), high-capacity deep learning models pre-trained on vast datasets, have become a cornerstone of computer vision (CV) research over the past decade. The early success of convolutional neural networks trained on visual datasets like ImageNet demonstrated the effectiveness of this paradigm [3]. Motivated by the transformative success of promptable large language models, there has been renewed interest to mirror this paradigm in CV. This has led to the emergence of powerful VFMs, such as the Segment-Anything-Model (SAM) [4], X-Decoder [5], and SegGPT [6]. SAM is an image segmentation model that can be conditioned with flexible prompts such as points or bounding boxes to segment arbitrary objects in images and has recently been extended to track segmentations in videos [7]. This provides a robust foundation for identifying and representing a target object to be grasped without limiting the scope of application to specific object sets.

B. Learning-based Robotic Grasping

The task of robotic grasping has been studied extensively in the literature, with particular challenges arising from unknown object geometries and cluttered environments. The advent of deep RL has brought forth systems that learn to grasp objects through trial and error. Pioneering work by Levine et al. [8] employs deep and reinforcement learning approaches to grasp objects from clutter, resulting in a continuous control policy. Building on this foundation, Kalashnikov et al. [9] used real-world and simulated data to train large-scale policies for grasping, outperforming prior methods. Zeng et al. [10] use deep Q-learning to learn synergies between pre-grasp manipulation (e.g. pushing) and grasping behaviors to retrieve specific objects from clutter. Another significant line of work focuses on combining RL with human demonstrations. Vecerik et al. [11] demonstrated the successful integration of human demonstrations and deep RL to accelerate training and enhance performance. Similarly, Rajeswaran et al. [12] proposed a method to incorporate demonstrations into policy gradient methods,

accelerating progress on manipulation tasks with sparse rewards. Behnke and Pavlichenko [13] designed a dense multi-component reward function for learning dexterous pre-grasp manipulation and functional grasping of novel instances of object categories with an anthropomorphic hand. Quillen et al. [14] introduced an approach that uses Q-learning with deep neural networks applied to grasping in clutter. This work emphasized real world scenarios and simulation-to-real transfer, highlighting the challenges and solutions related to cluttered environments.

C. Integrating Imitation with Reinforcement Learning

The integration of imitation learning (IL) objectives with RL has emerged as a powerful strategy. Two principal paradigms can be distinguished: imitating expert demonstrations and imitating another policy. Firstly, expert demonstrations can be used to streamline the exploration process in RL. Specifically, demonstrations can be employed to initialize a policy [12], or guide the learning process of on-policy [12] or off-policy [11], [15] methods. Rajeswaran et al. [12] propose a system that combines ideas from RL and behavior cloning (BC) [16]. Their work demonstrates how adding an objective that maximizes the probability of the expert actions can be used to foster the learning process of policy gradient methods on dexterous manipulation tasks. Mosbach and Behnke learn the use of tools from a single demonstration by employing non-rigid grasp-pose registration in a shape space [17].

Secondly, imitation can be used to transfer knowledge between policies, also known as policy distillation (PD) [18]. A specific formulation of this framework has recently proven useful in the context of robot learning [1], [2], wherein a teacher policy is trained from privileged, low-dimensional observations and then distilled into a visual student policy. The distillation from teacher to student is performed using DAgger [19]. Our proposed method amalgamates ideas from both RL and PD paradigms, facilitating knowledge transfer between observation spaces, while enabling unrestricted learning of the student policy.

III. METHOD

In sensory-motor robot learning, a key challenge is to devise a continuous control policy that can interpret sensory observations and act on them. Formally, we consider the problem of learning a policy, π_θ , which, given high-dimensional sensory observations \mathbf{s}_t , yields actions \mathbf{a}_t that control a robot. The objective is to adjust the parameter set θ to maximize the discounted sum of rewards

$$J(\pi_\theta) = \mathbb{E}_{\mathbf{s}_t, \mathbf{a}_t \sim \rho_\pi} [\gamma^t r_t], \quad (1)$$

where γ is a discount factor and ρ_π denotes the joint distribution of state-action pairs under the policy π . The rewards r_t encapsulate task-objectives like grasping an object.

This direct formulation of sensorimotor learning intertwines two difficult problems: *deriving underlying state-representations* from high-dimensional data and *determining optimal behavior* in each state [20]. Fortunately, student-teacher learning has been shown to efficiently disentangle

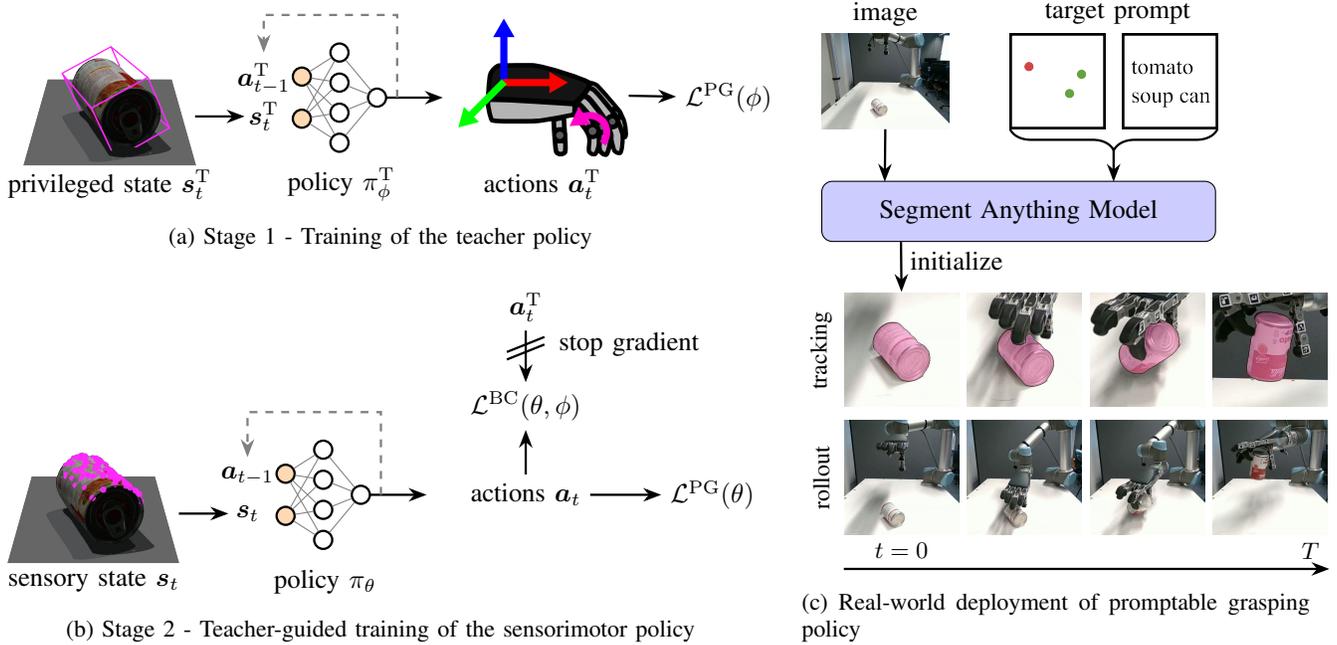


Fig. 2: Overview of the proposed two-stage learning framework. In the first stage (Fig. 2a), a teacher policy π_ϕ^T is trained to solve the grasping task from privileged state information s_t^T . Here, the goal is to frame the problem as a tractable RL task. In the second stage (Fig. 2b), the proficient teacher is used to guide the training of the sensorimotor policy π_θ . Replacing ground-truth segmentations with the output of a promptable segmentation model enables real world deployment (Fig. 2c).

both tasks [20], [1]. Initially, a teacher policy, π_ϕ^T , learns to solve the manipulation task given privileged access to the environment’s low-dimensional state s_t^T . The succinct nature of these inputs makes learning significantly more sample-effective than learning from images. In synergy, low-dimensional observations can be generated efficiently in massively parallelized simulators [21], providing abundant training data. The trained teacher policy offers ample supervision for the learning of the sensorimotor policy. We collect demonstrations (state-action tuples) from the teacher policy in a dataset, denoted $\mathcal{D}^T = \{(s_i, \mathbf{a}_i^T)\}$. Our method bridges the gap between imitation and reinforcement learning – a combination that allows for sensorimotor policies to adapt to the nuances of visual perception.

A. Teacher-Augmented Policy Gradient (TAPG)

Policy gradient (PG) methods center on the idea of increasing the probabilities of actions that yield higher returns, and decreasing the probabilities of actions that lead to lower returns. Therefore, an estimate of the policy gradient is optimized with stochastic gradient ascent. Most commonly, a gradient estimator of the form

$$\hat{g} = \hat{\mathbb{E}}_t \left[\nabla_\theta \log \pi_\theta(\mathbf{a}_t | s_t) \hat{A}^{\pi_\theta}(s_t, \mathbf{a}_t) \right] \quad (2)$$

is employed, where $\hat{\mathbb{E}}_t[\cdot]$ represents the mean over a finite batch of samples and \hat{A} is an estimate of the advantage function [22]. The corresponding loss function is given by

$$\mathcal{L}^{\text{PG}}(\theta) = - \hat{\mathbb{E}}_t \left[\log \pi_\theta(\mathbf{a}_t | s_t) \hat{A}^{\pi_\theta}(s_t, \mathbf{a}_t) \right]. \quad (3)$$

Multiple methods have been proposed that take constrained updates on this objective to avoid policy collapse [22], [23].

Although PG approaches have achieved impressive results in robot learning from low-dimensional states, they generally require a vast number of samples. This problem is exacerbated in high-dimensional settings. Tackling complex sensorimotor manipulation tasks necessitates guidance, i.e., from demonstrations or an adept teacher policy. Cloning the behaviors from a teacher policy or dataset (BC) corresponds to solving the following maximum likelihood problem:

$$\max_{\theta} \hat{\mathbb{E}}_t \left[\log \pi_\theta(\mathbf{a}_t | s_t) \right]. \quad (4)$$

Dagger [19] improves the performance guarantees of BC by iteratively collecting new data from the current policy and adding it to the dataset. However, the teacher policy is assumed to be optimal. Given that student-teacher learning in RL is fundamentally concerned with transferring knowledge between different observation spaces, it is desirable for the student to adjust its behavior in response to novel information or challenges that arise from its expressive observations.

To formulate a guiding yet non-restrictive imitation objective, we make use of the fact that the teacher policy has already learned a value function, $V^{\pi_\phi^T}(s_t^T)$. Using this additional knowledge, we introduce a gating term that estimates when to trust the teacher:

$$\hat{A}^T = [V^{\pi_\phi^T}(s_t^T) - V^{\pi_\theta}(s_t)]_{>0}, \quad (5)$$

where $[\cdot]_{>0}$ evaluates to 1 if the argument is positive, and to 0 otherwise. Integrating this term into the BC objective

Algorithm 1 Teacher-Augmented Policy Gradient (TAPG)

Require: teacher policy π_ϕ^T , empty teacher dataset \mathcal{D}^T

- 1: **for** $k = 0$ to N **do**
 - 2: Collect trajectories \mathcal{D}_k by running the current policy $\pi_k = \pi(\theta_k)$.
 - 3: Compute policy advantage estimates \hat{A}^{π_θ} .
 - 4: Compute teacher advantage estimates \hat{A}_t^T .
 - 5: Update parameters θ_k by optimizing the combined loss $L^{PG}(\theta_k) + L^{BC}(\theta_k, \phi)$.
 - 6: **end for**
-

yields the following loss function:

$$\mathcal{L}^{BC}(\theta, \phi) = - \mathbb{E}_{\mathbf{s}_t, \mathbf{a}_t \sim \mathcal{D}^T} \left[\log \pi_\theta(\mathbf{a}_t | \mathbf{s}_t) \hat{A}_t^T \right], \quad (6)$$

which incentivizes the student policy to imitate the teacher policy only if the teacher is estimated to be better in a given state. Considering the relation

$$\begin{aligned} \mathcal{L}^{BC}(\theta, \phi) &= - \mathbb{E}_{\mathbf{s}, \mathbf{a}^T \sim \pi_\phi^T} \left[\log \pi_\theta(\mathbf{a}^T | \mathbf{s}) A^T \right] \\ &= H^\times(\pi_\phi^T(\mathbf{s}) || \pi_\theta(\mathbf{s})) A^T, \end{aligned} \quad (7)$$

it can be shown that for a sufficiently small imitation loss ($\mathcal{L}^{BC} \leq H(\pi_\phi^T(\mathbf{s})) \forall \mathbf{s} \in \mathcal{S}$), the student policy is guaranteed to be at least as good as the teacher, i.e., $\mathbb{E}_{\mathbf{s} \sim \mathcal{S}_0} [V^{\pi_\theta}(\mathbf{s})] \geq \mathbb{E}_{\mathbf{s} \sim \mathcal{S}_0} [V^{\pi_\phi^T}(\mathbf{s})]$. The full proof can be found in [18].

The resulting method, which we refer to as teacher-augmented policy gradient (TAPG), optimizes the combined objective $\mathcal{L}^{PG}(\theta) + \mathcal{L}^{BC}(\theta, \phi)$. The training procedure is outlined in Algorithm 1.

For the task at hand, we instantiate TAPG as outlined in Figure 2. In the first stage, we train the teacher policy to grasp objects represented by their oriented 3D bounding boxes. This representation is low-dimensional to allow for efficient learning while conveying enough information for specialized, geometry-aware behaviors to emerge [24]. We subsequently learn a student policy from segmented point clouds of the object to be grasped. Although operating in a guided learning setting, generating the amounts of data required to imitate a teacher policy or adapt its behaviors with RL is impractical with a VFM in the loop. Instead, we utilize the ground truth segmentations provided by the simulator. In the ideal case, the segmentation model will output identical masks in deployment. To nevertheless account for the fact that tracked segmentations are less robust to high degrees of occlusion, we introduce an auxiliary reward term that encourages behaviors benign to the tracking model by maximizing the visibility of the target object.

B. Perception-driven Reward Design

Avoiding difficulties in object tracking amounts to avoiding large occlusions to the target object. Therefore, in addition to task objectives, we optimize for behaviors that retain the visibility of the target object, as quantified by the visibility ratio r_v : defined as an object’s visible area

TABLE I: Rewards combine terms for **task completion**, **directed exploration**, and **safety**. The sensorimotor policy is additionally rewarded for ease of perception.

Term	Equation	Weight
Sparse task reward	$[\ \Delta \mathbf{p}_t\ _2]_{< \bar{p}}$	50.0
Dense task reward	$1/(\ \Delta \mathbf{p}_t\ _2 + \epsilon_p)$	1.0
Fingertips close to object	$\sum_{i=1}^F \ \mathbf{p}_t^{f_i} - \mathbf{p}_t^o\ _2^2$	0.1
Clearance reward	$1/(\ \Delta h\ + \epsilon_h)$	1.0
Large actions penalty	$\ \mathbf{a}_t\ _2^2$	-0.01
Contact penalty	$\ \mathbf{c}_t\ _2\ _{> \bar{c}}$	1.0
Target visibility reward	r_v	20.0

$\Delta \mathbf{p}_t$ is the distance of the object to the goal position, Δh is the clearance between the object and the table, \mathbf{c}_t is the vector of contact forces acting on the robot arm, and r_v is the visibility ratio of the target object.

over its total area in the image plane. To estimate r_v , we project points sampled on the object’s surface to the image plane of our camera. The visibility ratio is then estimated as the number of projected points that fall into the object’s segmentation mask over the total number of points.

C. Sim-To-Real Adaptation

A major impediment in sim-to-real deployment is the discrepancy between the simulated and real-world dynamics of the robot. While the robot arm reaches the desired targets precisely in both settings, the anthropomorphic hand presents intricate joint couplings due to its tendon-actuated control. To accurately emulate the hand’s dynamics, we perform a calibration process where the fingers are opened and closed at various velocities. Since the used Schunk SIH hand does not have joint encoders, we attach ArUco markers to the finger phalanges (the markers can be seen in Figure 1 and are only used during calibration), allowing for angle tracking via a camera. Subsequently, we fitted the simulated robot’s response curves through a least squares polynomial approximation to match the real-world data.

Robotic manipulation inherently necessitates contact with the environment. However, manipulation should be performed exclusively with the hand, while the arm should not collide with the environment. To encourage this behavior, we add a term to the reward function that penalizes the policy for applying excessive contact forces to the robot arm.

IV. TASK AND SYSTEM DESCRIPTION

We study the problem of learning to grasp objects with an anthropomorphic robot hand in both the isolated single-object case and from more cluttered environments. Figure 1 shows the robot setup, which we replicate in Isaac Gym [21].

A. Task Formulation

1) *Observation and action space:* The teacher policy takes as input proprioceptive information from the robot, the ground truth state of the object, and its own previous action. For the student policy, the ground truth state of the object is replaced by a segmented point cloud of the target

TABLE II: Success rates (%) of the teacher policies.

Object set	Number of objects		
	1	3	5
train	95.1 ± 0.5	85.3 ± 0.7	79.0 ± 1.1
test	89.7 ± 0.7	82.2 ± 0.6	75.9 ± 1.4

object. Actions $\mathbf{a}_t \in \mathbb{R}^{11}$ specify desired changes to the end-effector’s 6D pose and the joint angles of the fingers. The control frequency of the system is 7.5 Hz and an episode terminates after 75 steps.

2) *Reward function and success criterion*: The reward function combines incentives for task completion with auxiliary objectives to guide exploration and maintain safety. Exploration is facilitated by rewarding the policy for bringing the fingertips close to the target object and for lifting the object off the table. To learn safe behaviors, we penalize large actions and contact forces acting on the robot arm. The full reward function is detailed in Table I.

We consider an episode as successful if the target object is lifted to within 5 cm of its target position central above the table. The initial poses of the objects are randomized by dropping them sequentially until all objects are in the workspace of the robot.

3) *Evaluated objects*: We utilize the YCB object and model set [25], which contains diverse every-day items, to train our policies. Some YCB objects are too large to be grasped with our Schunk SIH hand. Therefore, we select a subset of graspable objects with at least two bounding-box dimensions measuring under 10 cm. This results in a total of 60 objects, which we randomly divide into 48 training objects and 12 test objects.

B. System Description

1) *System setup*: Our setup employs a Universal Robots UR5 robotic arm paired with a Schunk SIH five-finger hand. The hand features 11 degrees of freedom, of which 5 are tendon-actuated. For visual perception, we use a RealSense D455 RGB-D camera mounted statically above the workspace.

2) *Policy architecture*: The teacher policy consists of an MLP with three hidden layers, comprising 1024, 512, and 512 units, respectively. Each layer uses ELU activations [26]. The student policy architecture splits the processing of vectorized and point cloud-based observations to obey the permutation invariant nature of this observation type. We employ a PointNet-like encoder [27] to process the point cloud and concatenate the resulting embedding with the remaining observations. The result is fed into an MLP with the same architecture as the teacher policy.

V. RESULTS

Our experimental design is set up to evaluate the efficacy of our approach in dexterous robotic grasping. Specifically, we target the following research questions: (1) Is the teacher policy able to master dexterous grasping using privileged

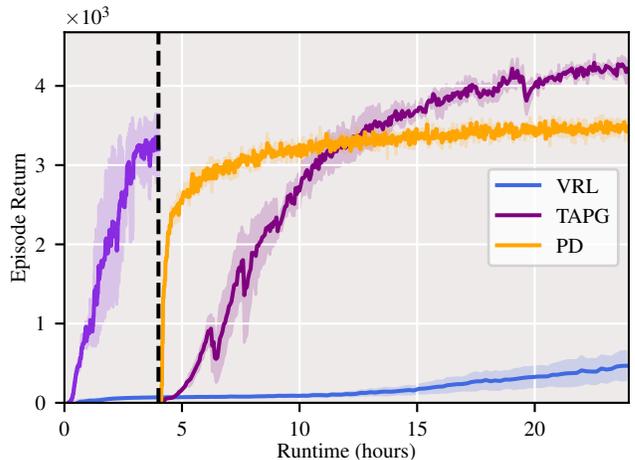


Fig. 3: Episode return during training illustrating learning speed and final performance of different methods for obtaining a sensorimotor grasping policy. Vertical line indicates transition from privileged teacher (in light purple) to student policy. We limit the maximum training time to 24 hours and report the mean and standard deviation of three runs.

state information? (2) Can TAPG distill the teacher’s expertise into a sensorimotor student policy and demonstrate a tangible advantage over the baselines? (3) Are the resultant policies capable of real-world deployment?

The rationale of our proposed method naturally suggests a comparison with two methods for acquiring a sensorimotor policy. The first is to apply RL directly from visual observations, which we denote as VRL. The second, labeled PD, leverages policy distillation. However, unlike our method, it solely relies on cloning the teacher’s behavior, without any further fine-tuning of the sensorimotor behaviors.

A. Training the Teacher Policy

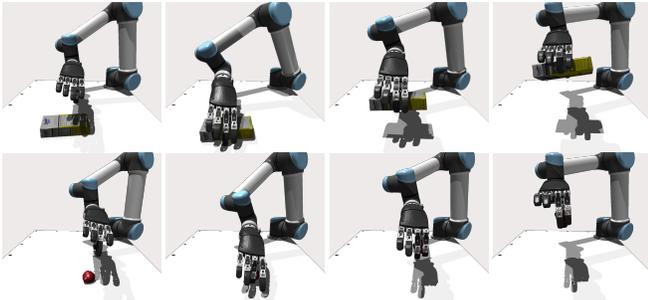
We train our teacher policies on YCB objects selected uniformly from the training set. As reported in Table II, the resultant policies achieve a success rate of over 95% for the single-object case. Moreover, our evaluations of zero-shot transfer to the test set underline the robustness of the learned behaviors. We further probe their robustness by placing 3 or 5 items in the workspace to induce unseen interactions, resulting in success rates of 85.3% and 79.0%, respectively, when using training objects and 82.2% and 75.9% when using test objects. A qualitative investigation of failure modes showed that the majority of unsuccessful trials result from objects that are very flat, making it difficult to pick them up.

Overall, the privileged teacher policies adeptly solve all task configurations, answering our first research question in the affirmative. We can confirm that model-free RL is able to synthesize proficient control policies for the investigated problem given access to privileged simulator state information and ample environment interactions.

B. Training the Sensorimotor Policy

In this section, we analyze different methods for obtaining a sensorimotor policy. Specifically, we contrast the

Policy distillation (PD)



Teacher-Augmented Policy Gradient (TAPG, ours)

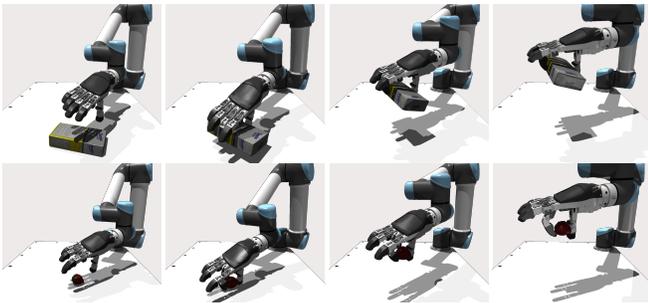


Fig. 4: Grasping behaviors learned by vanilla PD versus TAPG as viewed from the camera used for tracking. We see that TAPG assumes the learned grasping skills and successfully adapts them to the visual observation space. For larger objects, such as the sugar box (top rows), this amounts to retracting the fingers not actively involved in holding the object and grasping at an angle. Here, both methods learn behaviors that retain the visibility of the majority of the object. In contrast, for smaller objects such as the strawberry (bottom rows), modifying the behaviors becomes crucial. While the PD policy severely occludes the object, TAPG maintains object visibility throughout the grasping process.

performance of two established frameworks (VRL and PD) with our proposed method TAPG. The progression of the policy performance is plotted in Figure 3. RL from vision (VRL), while making some progress, cannot solve the studied task in any acceptable time. Both PD and TAPG learn capable sensorimotor policies. The fully supervised learning paradigm of PD results in fast and stable convergence of the student policy. Although TAPG requires longer to recover the teacher’s performance, it is subsequently able to surpass it. As such, TAPG is the only paradigm able to sufficiently adapt to the intricacies and requirements of visual perception.

To elucidate the improvement of TAPG over PD, we depict the grasping behaviors learned by both paradigms in Figure 4. TAPG is able to improve the visibility of the target object while maintaining high grasping performance. This effect is most pronounced for small objects that fit inside the hand. In summary, our findings show that TAPG offers substantial benefits over the baselines for the studied task, thereby affirmatively answering our second question.

C. Real-world Deployment

The proposed pipeline transfers well to the real world, without any adaptation (zero-shot). After manually placing

TABLE III: Quantitative zero-shot real-world evaluation.

Method	Number of objects			
	1	3	5	ALL
Visual RL (VRL)	0/60	0/10	0/10	0/80
Policy distillation (PD)	18/60	2/10	0/10	20/80
TAPG (ours)	35/60	5/10	4/10	44/80

Reported are successful trials / number of trials.

TABLE IV: Object-wise zero-shot real-world evaluation.

Object	PD	TAPG (ours)
 021_bleach_cleanser	4/5	5/5
 004_sugar_box	5/5	5/5
 061_foam_brick	0/5	5/5
 065_cups-a	0/5	4/5

Reported are successful trials / number of trials.

one or multiple objects in the robot’s workspace, we deem a trial successful if the robot retrieves and holds the target object above the table. We perform five runs for each test object and ten runs for cluttered configurations with three and five objects on the table. The results are reported in Table III. While VRL is not successful, PD succeeds in 25% of the trials. Our method TAPG clearly outperforms PD by succeeding in 55% of the trials.

Next, to verify the benefit of the adapted behaviors, we report an object-wise performance assessment in Table IV. For large objects that are never fully occluded (bleach cleanser and sugar box), maintaining the behaviors of the teacher policy proves to be sufficient. In contrast, the strength of TAPG can really be seen on the smaller objects (foam brick and smallest cup) that can fit inside the hand. Occluding them fully during the grasping process causes the vision pipeline to lose track of the object, resulting in failed executions.

VI. DISCUSSION AND CONCLUSION

Our experimental evaluation has shown that the proposed pipeline is able to learn grasping diverse objects in an interactive, human-like manner. The agent acquires interesting emergent behaviors, such as pre-grasp manipulation and recovery from failures and perturbations. Further, we have demonstrated how TAPG can adjust learned strategies to a new observation-space, a capability neglected in prior works. These adjustments resulted in significant improvements to the task’s success by obeying the restrictions of the adopted segmentation and tracking model. Finally, we have shown that our method zero-shot transfers well to the real robot.

An interesting direction for future research is learning to grasp from a deep, cluttered container, where complete occlusions may be unavoidable.

ACKNOWLEDGEMENT

This work has been funded by the German Ministry of Education and Research (BMBF), grant no. 01IS21080, project “Learn2Grasp: Learning Human-like Interactive Grasping based on Visual and Haptic Feedback”.

REFERENCES

- [1] T. Chen, J. Xu, and P. Agrawal, "A system for general in-hand object re-orientation," in *Conference on Robot Learning (CoRL)*. PMLR, 2022, pp. 297–307.
- [2] T. Chen, M. Tippur, S. Wu, V. Kumar, E. H. Adelson, and P. Agrawal, "Visual dexterity: In-hand reorientation of novel and complex object shapes," *Science Robotics*, vol. 8, no. 84, 2023.
- [3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 25, 2012.
- [4] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.
- [5] X. Zou, Z.-Y. Dou, J. Yang, Z. Gan, L. Li, C. Li, X. Dai, H. Behl, J. Wang, L. Yuan *et al.*, "Generalized decoding for pixel, image, and language," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE/CVF, 2023, pp. 15 116–15 127.
- [6] X. Wang, X. Zhang, Y. Cao, W. Wang, C. Shen, and T. Huang, "SegGPT: Segmenting everything in context," *arXiv preprint arXiv:2304.03284*, 2023.
- [7] Y. Cheng, L. Li, Y. Xu, X. Li, Z. Yang, W. Wang, and Y. Yang, "Segment and track anything," *arXiv preprint arXiv:2305.06558*, 2023.
- [8] 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 (IJRR)*, vol. 37, no. 4-5, pp. 421–436, 2018.
- [9] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, and S. Levine, "Scalable deep reinforcement learning for vision-based robotic manipulation," in *2nd Annual Conference on Robot Learning (CoRL)*, ser. Proceedings of Machine Learning Research, vol. 87. PMLR, 2018, pp. 651–673.
- [10] 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," in *International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 4238–4245.
- [11] M. Vecerik, T. Hester, J. Scholz, F. Wang, O. Pietquin, B. Piot, N. Heess, T. Rothörl, T. Lampe, and M. Riedmiller, "Leveraging demonstrations for deep reinforcement learning on robotics problems with sparse rewards," *arXiv preprint arXiv:1707.08817*, 2017.
- [12] A. Rajeswaran, V. Kumar, A. Gupta, G. Vezzani, J. Schulman, E. Todorov, and S. Levine, "Learning complex dexterous manipulation with deep reinforcement learning and demonstrations," in *Robotics: Science and Systems (RSS)*, 2018.
- [13] D. Pavlichenko and S. Behnke, "Deep reinforcement learning of dexterous pre-grasp manipulation for human-like functional categorical grasping," in *19th IEEE International Conference on Automation Science and Engineering (CASE)*, 2023.
- [14] D. Quillen, E. Jang, O. Nachum, C. Finn, J. Ibarz, and S. Levine, "Deep reinforcement learning for vision-based robotic grasping: A simulated comparative evaluation of off-policy methods," in *International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 6284–6291.
- [15] A. Nair, B. McGrew, M. Andrychowicz, W. Zaremba, and P. Abbeel, "Overcoming exploration in reinforcement learning with demonstrations," in *International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 6292–6299.
- [16] D. A. Pomerleau, "ALVINN: An autonomous land vehicle in a neural network," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 1, 1988.
- [17] M. Mosbach and S. Behnke, "Learning generalizable tool use with non-rigid grasp-pose registration," in *19th IEEE International Conference on Automation Science and Engineering (CASE)*, 2023.
- [18] W. M. Czarnecki, R. Pascanu, S. Osindero, S. Jayakumar, G. Swirszcz, and M. Jaderberg, "Distilling policy distillation," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR, 2019, pp. 1331–1340.
- [19] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*. JMLR Workshop and Conference Proceedings, 2011, pp. 627–635.
- [20] D. Chen, B. Zhou, V. Koltun, and P. Krähenhühl, "Learning by cheating," in *Conference on Robot Learning (CoRL)*. PMLR, 2020, pp. 66–75.
- [21] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa *et al.*, "Isaac Gym: High performance GPU-based physics simulation for robot learning," *arXiv preprint arXiv:2108.10470*, 2021.
- [22] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.
- [23] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *International Conference on Machine Learning (ICML)*. PMLR, 2015, pp. 1889–1897.
- [24] M. Mosbach and S. Behnke, "Efficient representations of object geometry for reinforcement learning of interactive grasping policies," in *International Conference on Robotic Computing (IRC)*, 2022.
- [25] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar, "The YCB object and model set: Towards common benchmarks for manipulation research," in *International Conference on Advanced Robotics (ICAR)*. IEEE, 2015, pp. 510–517.
- [26] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," *arXiv preprint arXiv:1511.07289*, 2015.
- [27] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE/CVF, 2017, pp. 652–660.