

Visual-Policy Learning Through Multi-Camera View to Single-Camera View Knowledge Distillation for Robot Manipulation Tasks

Cihan Acar^{1b}, Kuluhan Binici^{1b}, Alp Tekirdağ^{1b}, and Yan Wu^{1b}, *Senior Member, IEEE*

Abstract—The use of multi-camera views simultaneously has been shown to improve the generalization capabilities and performance of visual policies. However, using multiple cameras in real-world scenarios can be challenging. In this study, we present a novel approach to enhance the generalization performance of vision-based Reinforcement Learning (RL) algorithms for robotic manipulation tasks. Our proposed method involves utilizing a technique known as knowledge distillation, in which a “teacher” policy, pre-trained with multiple camera viewpoints, guides a “student” policy in learning from a single camera viewpoint. To enhance the student policy’s robustness against camera location perturbations, it is trained using data augmentation and extreme viewpoint changes. As a result, the student policy learns robust visual features that allow it to locate the object of interest accurately and consistently, regardless of the camera viewpoint. The efficacy and efficiency of the proposed method were evaluated in both simulation and real-world environments. The results demonstrate that the single-view visual student policy can successfully learn to grasp and lift a challenging object, which was not possible with a single-view policy alone. Furthermore, the student policy demonstrates zero-shot transfer capability, where it can successfully grasp and lift objects in real-world scenarios for unseen visual configurations.

Index Terms—Deep learning in grasping and manipulation, learning from experience, transfer learning.

I. INTRODUCTION

IN RECENT years, the field of deep reinforcement learning (DRL) has seen rapid progress in robot control and manipulation applications, with vision-based RL algorithms being

Manuscript received 14 July 2023; accepted 5 November 2023. Date of publication 23 November 2023; date of current version 4 December 2023. This letter was recommended for publication by Associate Editor M. Burke and Editor A. Faust upon evaluation of the reviewers’ comments. This work was supported in part by the Singapore Government’s Research under Grant A19E4a0101, in part by Innovation and Enterprise 2020 Plan (Advanced Manufacturing and Engineering domain) and administered by the Agency for Science, in part by Technology and Research, and in part by A*STAR Computational Resource Centre through the use of its high performance computing facilities. (*Corresponding author: Cihan Acar.*)

Cihan Acar and Yan Wu are with the Institute for Infocomm Research (I2R), A*STAR, Singapore 138632 (e-mail: acar_cihan@i2r.a-star.edu.sg; wuy@i2r.a-star.edu.sg).

Kuluhan Binici is with the School of Computing, National University of Singapore, Singapore 138632 (e-mail: kuluhan@comp.nus.edu.sg).

Alp Tekirdağ is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 138632 (e-mail: alp001@e.ntu.edu.sg).

[Video attachment: <https://youtu.be/CnDQK9ly5eg>].

This letter has supplementary downloadable material available at <https://doi.org/10.1109/LRA.2023.3336245>, provided by the authors.

Digital Object Identifier 10.1109/LRA.2023.3336245



Fig. 1. Illustration of a single-camera visual policy setup for robot manipulation tasks, with a front-facing third-person camera.

particularly noteworthy due to their ability to learn directly from high-dimensional visual inputs such as images, point clouds or videos [1], [2], [3]. However, one of the most significant challenges for these algorithms remains the generalization of learned skills to new and diverse situations. To address this issue, researchers have employed various data randomization [4] and augmentation methods, including changes to the background color, texture, and other parameters [5], [6], [7], [8]. While these techniques have shown effectiveness in learning representations of high-dimensional data, their generalization capabilities are limited to the features acquired from a fixed viewpoint perspective only.

To overcome this limitation, researchers have explored various techniques to enhance the generalization capabilities of vision-based learning algorithms [9], [10], [11]. One approach involves training the algorithm using multiple viewpoints to recognize objects from different perspectives. For example, to develop visuomotor manipulation skills, researchers often employ a third-person camera view that captures the robot and environment from the outside, as well as a first-person (egocentric) camera that is linked to the robot’s end-effector. The third-person camera offers a broad, fixed perspective of the environment and the robot, leading to a more comprehensive understanding. On the other hand, the egocentric camera provides a closer but limited observation since it cannot fully encompass the robot arm and the environment, potentially causing vital visual features to be obscured. However, by concurrently utilizing both egocentric and third-person cameras, these limitations can be overcome, resulting in enhanced generalization capabilities and improved

visual policy performance [12], [13]. In a simulation environment, it is possible to use multiple third-person and egocentric cameras together to obtain better results without having any limitations. In contrast, deploying multiple camera devices in real-world scenarios presents significant challenges, including the need for additional hardware components and design constraints. In addition, coordinating and synchronizing multiple cameras to capture frames simultaneously can be technically complex and time-consuming.

In this letter, we propose a method for learning a robust manipulation policy that can generalize to unseen visual configurations, using a single camera. Our approach leverages knowledge distillation to transfer privileged multiple-view information from a teacher policy to a student policy that observes the task using information only from a single camera. By learning from the teacher policy's rich multi-view information, the student policy can acquire robust features that can consistently locate the object of interest, regardless of the single-camera viewpoint. First, a teacher policy is trained with multiple views gathered from fixed viewpoints to learn more generalizable representations. Then, a student with a single-camera viewpoint is trained to imitate the action of the teacher policy. To enhance robustness against camera location, the camera locations and parameters are constantly changed, in addition to data augmentation techniques such as random cropping and color jiggling. Furthermore, we employ a straightforward curriculum training approach for continuous policy learning by gradually expanding the student's camera parameter range. Notably, camera randomization alone is insufficient on its own to ensure robust learning. Our experiments demonstrate that substantial camera randomization, without the guidance of knowledge distillation, fails to enable policies to acquire the essential knowledge and skills required for successful task performance. Consequently, the single-camera student can learn the same actions as the teacher from heavily augmented and perturbed camera positions. This enables the student policy to perform manipulation tasks using a single-camera observation and as a result, makes it more practical and efficient for real-world applications. To this end, the main contributions can be summarized as follows:

- A novel and robust visual-policy learning by transferring privileged multiple-view information of the teacher to the single-camera view student.
- The use of randomized camera views during training, allowing the student policy to learn to adapt to different camera angles and positions.
- Investigation on the effect of the number of third-person camera views for knowledge distillation from the multiple-view teacher to the single-view student.

II. RELATED WORK

A. Vision-Based Reinforcement Learning

The overall goal of vision-based reinforcement learning is to enable agents to learn from visual inputs, such as images or videos, in order to perform tasks in complex environments. To extract high-level features and improve the sample efficiency of vision-based RL algorithms, CURL method is proposed [5]. This method employs contrastive loss to ensure consistency between an image and its augmented version. Many works also focus on data augmentation methods such as random shift, random crop, color jitter, applying Gaussian noise, and random convolutions

for learning robust representations from the raw input image [6], [7], [8]. Data augmentation and CURL are also successfully implemented for real-world robotic tasks [14]. One limitation of these methods is that they may not be adequate to handle more intricate forms of variation, such as variations in viewpoint or camera parameters.

B. Knowledge Distillation

The procedure of using the knowledge of a pre-trained neural network model to enhance the capabilities of another one is referred to as knowledge distillation (KD) [15]. Most commonly the pre-trained model, called the "teacher", is used to provide additional annotation for the data samples to be used along with the ground truth labels while training another model from scratch. This enriches the limited information conveyed by the ground truth labels. As a result, the model trained with such enhanced information called the "student", can achieve performance beyond its original capacity and even outperform the teacher.

In the context of reinforcement learning, KD has been employed for various tasks, such as policy distillation [16] which proposes distilling the knowledge of a large and accurate policy network, called the teacher, into a smaller and faster policy network, called the student. Additionally, it is worth noting that similar concepts have been extended to scenarios where a single student agent can leverage multiple expert teachers, each specializing in individual tasks, allowing for more diverse and versatile knowledge transfer and multi-task learning [17], [18].

C. Privileged Provision

Recently, KD is utilized in learning vision-based urban driving [19], [20], dexterous hand manipulation [21], in-hand object reorientation [22], stacking diverse objects [23] and quadrupedal locomotion over challenging environments [24], [25] where the teacher policy that has access to privileged information in a simulation environment is used to guide the student policy that has only access to sensory information available in the real world. The student policy is then deployed and successfully tested in a variety of challenging real-world environments. Another approach [26], [27] is to train the teacher in a simplified virtual environment with basic geometry and rendering, enabling faster learning. Then, a student is trained in a high-fidelity simulation featuring realistic sensors with the guidance of the teacher before being deployed in the real world. Similarly, KD is also applied for zero-shot generalization of visual policies [28]. At first, weak image augmentation is used to train an expert policy. Then, using imitation learning with strong augmentations, a student network learns to mimic the expert policy. Even though it has been demonstrated that strong augmentations can offer robust representation learning for a range of visual policy tasks in diverse simulation contexts, generalization capabilities are restricted to features acquired through a single fixed viewpoint.

III. PRELIMINARIES

The reinforcement learning environment is defined as a Markov Decision Process (MDP) tuple $\langle S, A, p, r, \gamma \rangle$: S is the continuous state space, A is continuous action space, $p(s_{t+1}|s_t, a_t)$ is the state transition model where probability density of the next state $s_{t+1} \in S$ given $s_t \in S$ and $a_t \in A$, r is the reward function, and $\gamma \in (0, 1]$ is the discount factor.

Episode length is denoted by T . The objective of an RL algorithm is to find an optimal policy $\pi : S \rightarrow A$ to maximize the expected return.

In order to determine the policy distribution $\pi(a_t|s_t)$, we employ an actor-critic model. As our algorithm of choice, we use the Soft Actor-Critic (SAC) [29], a state-of-the-art off-policy maximum entropy deep reinforcement learning algorithm due to its good performance in sparse reward setups. The optimal policy equation of SAC is shown below:

$$\pi^* = \arg \max_{\pi} \sum_t \mathbb{E}_{(s_t, a_t) \sim \rho_{\pi}} [r(s_t, a_t) + \alpha H(\pi(\cdot|s_t))] \quad (1)$$

where the temperature parameter, denoted as α , controls the stochasticity of the optimal policy by determining the relative importance of the entropy term compared to the reward.

IV. APPROACH

A. Problem Formulation

In visual policy learning, the agent perceives the environment through cameras, and therefore, the environment is only partially observable. This makes the problem formulation a Partial Observability Markov Decision Process (POMDP) instead of a standard MDP. As a result, the state vectors s_t in (1) are replaced with observations, denoted as o . These observations are obtained by encoding RGB images collected from the environment $[I^1, \dots, I^N]$ via a function f before being provided to the policy module:

$$o := f([I^1, \dots, I^N]) \quad \text{where } I \in \mathbb{R}^{C \times H \times W} \quad (2)$$

Typically, f is parameterized by a neural network (NN) ϕ and learned using gradient-based optimization techniques. Since neural networks generalize poorly when training data lacks diversity, increasing the number of unique views N by using more cameras can improve the performance. As a result, there is a natural trade-off between the hardware complexity of the robotics system and performance since utilizing multiple camera devices to process diverse views of the environment typically provides robust policies against domain variations at the expense of increased hardware demand and design complexity. To break such a trade-off, we propose training a privileged teacher policy $\pi_t(a|o_t)$ using multiple camera views from the simulation environment and later distilling the learned policy to train a non-privileged student policy $\pi_s(a|o_s)$ that works with only single-view observations. In this way, while the hardware requirement of deploying π_s is limited to a single camera, it can provide similar performance to π_t .

Our objective is to find the optimal student policy π_s^* that minimizes the knowledge distillation (KD) loss. The KD loss measures the discrepancy between the student policy and the teacher policy:

$$\pi_s^* := \underset{\pi_s}{\operatorname{argmin}} L_{KD}(\pi_s, \pi_t) \quad (3)$$

B. Implementation Details

The teacher operates with images collected from a single egocentric camera (I_e) and N stationary third-person view cameras ($[I_t^{\theta_1}, \dots, I_t^{\theta_N}]$) while the student uses the view of a single third-person view camera ($I_t^{\theta_k}$).

$$o_t := f_{\phi t}(I_e, [I_t^{\theta_1}, \dots, I_t^{\theta_N}])$$

$$o_s := f_{\phi s}(I_t^{\theta_k}) \quad \text{where } \theta_k \sim U(0, \theta_{\max}) \quad (4)$$

To make the student robust against environmental perturbations, we apply domain randomization and sample the camera angle θ_k randomly from a uniform distribution that ranges from 0 to θ_{\max} . Both the teacher and the student have the same network architecture comprised of a 12-layer convolutional encoder with a single-layer self-attention module and a SAC policy. We use PyTorch [30] implementation of SAC as [14].

Knowledge Distillation: We apply the principles of DAgger [31], which reduces distributional shifts in learned policies by incorporating data from an expert policy, to distill the knowledge of the teacher policy into the student policy. Different than DAgger, our expert is not a human, instead we use a pre-trained teacher policy. To train the student, we use two sources of information: the actions and encoded features provided by the teacher. To distill the first information into the student, the mean squared error (MSE) between the student π_s and teacher π_t policies is minimized:

$$L_{KD} := \|\pi_s(o_s) - \pi_t(o_t)\| \quad (5)$$

The student policy π_s is updated by minimizing this loss using the transition batch sampled from experiences replay buffer \mathcal{D} . The updated student policy is then used to collect more data, which is added to \mathcal{D} and used for the next iteration of the policy update. During the training of the teacher, only the loss of the critic is used to update the weights of the encoder network for learning the representation space. For training the student encoder, instead of using the gradients of critic to update the student encoder, we implement two methods. The first one is inspired by [32] where we align the pairwise similarity matrix of teacher observations C_t with that of the student observations C_s by minimizing the following loss:

$$L_{sim} := \sum_i \sum_j \|C_t^{ij} - C_s^{ij}\|_F^2 \quad (6)$$

Different to [32], We use the cosine distance to quantify the pairwise similarities between different observations:

$$C := \frac{\bar{C}}{\|o_{[i,:]} \| \|o_{[i,:]} \|_T^T}; \quad \bar{C} := oo^T; \quad o \in \mathbb{R}^{b \times (z)} \quad (7)$$

Here o is the flattened output tensor of the feature extractor f_{ϕ} . In the second one, the distillation objective is to minimize the Euclidean distance between the features of the teacher and student. Thus, the second method tries to minimize the sum of all the squared differences between the teacher encoder output and the predicted student output:

$$L_{MSE} := \|o_t - o_s\|_2 \quad (8)$$

Our results, to be presented in the next section, show that minimizing the Euclidean distance between features yields superior performance for the student policy when compared to the pairwise similarity method. This finding suggests that minimizing the Euclidean distance provides better distillation for learning high-level visual features.

Curriculum Learning: To make the student robust against environmental perturbations we apply domain randomization and permute the camera angle θ_k and camera parameters. We utilize a simple curriculum training that gradually increases the range of camera parameters of the student. Initially, when training begins, the camera parameter range closely aligns with

Algorithm 1: Multi-View to Single-View KD.

```

1: Input: multi-view teacher policy  $\pi_t$ , single-view
   student policy  $\pi_s$ , feature extractors  $f_{\phi_t}$  and  $f_{\phi_s}$ , a
   buffer containing fixed expert demonstrations  $\mathcal{B}_e$ .
2: #Train teacher policy  $\pi_t$  and  $f_{\phi_t}$ .
3: for  $t$  in  $1, \dots, T_{\text{train}}$  do
4:   Update  $\pi_t$  and  $f_{\phi_t}$  based on (1) using dataset  $\mathcal{B}_e$ .
5:   Execute policy  $\pi_t$  and obtain new transitions  $\mathcal{B}_t$ .
6:    $\mathcal{B}_e \leftarrow \mathcal{B}_e \cup \mathcal{B}_t$ .
7: end for
8: #Train student policy  $\pi_s$  and  $f_{\phi_s}$  via KD.
9: Use  $\pi_t$  to collect multi-camera view transitions  $\mathcal{B}$ .
10: for  $t$  in  $1, \dots, T_{\text{distill}}$  do
11:   Update  $\pi_s$  and  $f_{\phi_s}$  to minimize the loss in (5) and
      (8) using observations in  $\mathcal{B}$ .
12:   Execute policy  $\pi_s$  and obtain new transitions  $\mathcal{B}_s$ .
13:    $\mathcal{B} \leftarrow \mathcal{B} \cup \mathcal{B}_s$ .
14:   if Episode is done then
15:     Randomize camera viewpoint.
16:   end if
17: end for

```

that of the teacher. As training progresses, we incrementally widen the parameter range every 50 episodes (up to the maximum parameter range) and sample camera parameters randomly from this expanded distribution. Our intuition is that gradually increasing the camera parameters increases the difficulty level of the task. It is also possible to utilize more efficient automatic curriculum learning methods such as ALP-GMM [33], Automatic Curriculum Learning through Value Disagreement [34], where the complexity of the task is determined by the performance of student policy.

Data Augmentation: To improve the generalization ability of the policies, we employ two techniques for data augmentation during training: pixel shifting and color augmentation. These techniques help to expose the models to a more diverse range of visual inputs, which in turn can improve their ability to perform the task in a wider variety of scenarios. Specifically, we randomly shift the pixels of the images in two dimensions and apply color augmentation techniques such as changing the brightness, contrast, and saturation. By applying these techniques, we aim to make the policies more robust to changes in lighting, and other environmental factors that may be present in real-world scenarios.

Setup Details: Multiple cameras can be employed to capture various viewpoints and enhance the generalization capability of visual-policy learning. Our study focuses on the utilization of two-camera and three-camera setups to train an expert visual policy. The two-camera setup comprises a first-person (egocentric) camera attached to the robot's end-effector and a third-person camera positioned in front of the robot to capture a bird's eye view of the table and robot. The three-camera setup consists of a first-person camera and two third-person cameras positioned in front and on the side of the robot, respectively. These cameras are positioned to offer orthogonal viewpoints. This arrangement minimizes information redundancy across cameras, maximizing the advantages of multi-view input.

During the training of the student policy, we employ only one of the third-person cameras to capture the visual information. More precisely, we use a single third-person camera placed

in front of the robot but the camera parameters and view are randomly changed. By constantly changing the camera view and parameters, the training data becomes more diverse, and the student policy can learn to adapt to different camera angles and positions. This allows the student policy to generalize better to new, unseen camera views, which is crucial for applications where the camera position and orientation may not be known or fixed.

In our context, the visual policy is responsible for generating control actions that enable the robot to perform a given task. The action space includes control over the 3-DoF end-effector position along the X, Y, and Z axes, 1-DoF rotation control on the yaw-axis, and 1-DoF gripper control. We train both the teacher and student policies using 84×84 RGB camera images as observations. However, for tasks requiring finer details, it is possible to utilize higher-resolution observations during student policy training. This approach offers the advantage of reducing computational costs and accelerating training times when compared to training students from scratch using higher-resolution image observations.

Reward Scheme: We train the teacher policies using a delayed reward scheme with a fixed time horizon, in which the agent gets rewarded only if the task is completed successfully. Achieving a task successfully results in a +100 reward at each subsequent time step, and accumulates until the end of the episode, which encourages the agent to complete the task as quickly as possible.

Task Definition: The tasks in our experiments require the robot to grasp either a cube or a mug and lift it to a predetermined height from its initial position. While cubes and spheres have a symmetrical shape that makes it easier for visual policies to learn grasping strategies, the lack of symmetry in the mug presents a greater challenge. The task is further complicated by visual localization, which requires the agent to infer object positions solely from high-dimensional RGB camera pixel data. To increase the diversity of training data, objects are randomly spawned at different positions on the table for each episode. This ensures that the agent learns to adapt to different object positions, enhancing its ability to perform the task in novel scenarios.

V. RESULTS

A. Simulation Experiments

In order to evaluate the effectiveness of our proposed approach, we conducted experiments using a 7°-of-freedom Franka Panda Emika manipulator equipped with a two-fingered parallel gripper in a PyBullet simulation environment [35], based on code from [36]. To train the teacher policies π_t , we utilized combinations of first-person, front, and side camera views, as depicted in Fig. 2(a). The two-camera view teacher policy comprised a front camera labeled as (1) and an egocentric camera attached to the robot's end-effector labeled as (2) in Fig. 2(a). Additionally, the three-camera view teacher policy included a side camera labeled as (3) in Fig. 2(a) in addition to these. Furthermore, we introduced a single-camera view teacher, which solely utilized a front camera. This single-camera view teacher provided a baseline for comparison and allowed us to evaluate the benefits of incorporating multiple camera views in the training process.

For each task, we used a 3D-mouse to collect 50 human demonstrations (total of 5 K transitions), which were added to the replay buffer to aid in training the multi-camera view expert

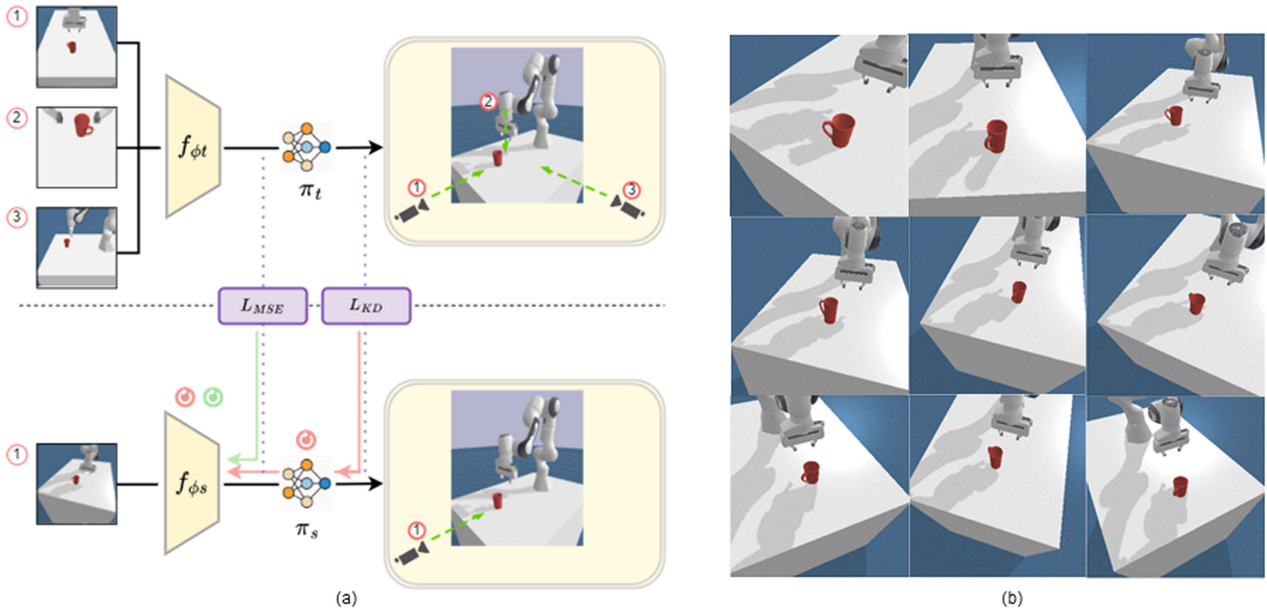


Fig. 2. Figure depicts the overview of our framework for training the teacher π_t and student π_s policies in part (a) and an example of randomized third-person camera views used to train the student policies in the visual-policy learning task in part (b). In part (a), we illustrate the flow of information and learning between the teacher and student policies. The teacher policy is trained using a delayed reward scheme and is used to guide the training of the student policy. In part (b), we show a collection of sampled third-person camera views used to train the student policy, which are randomized by camera position and parameters to introduce variability in the training data and improve the generalization capability of the student policy.

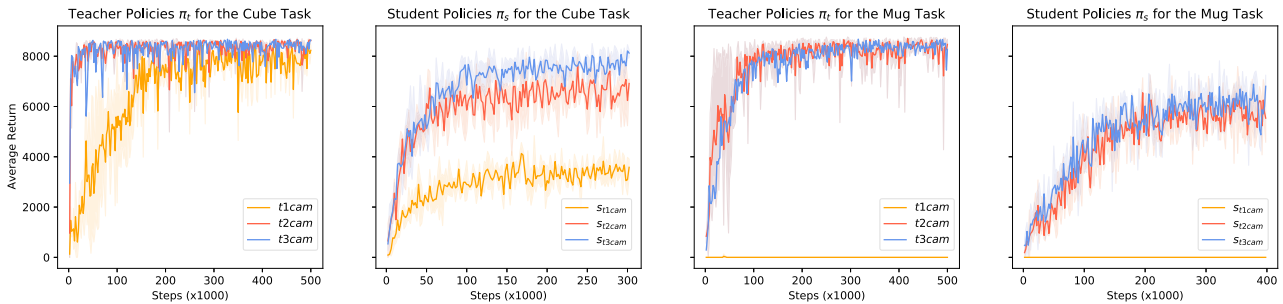


Fig. 3. Training results of teachers and distilled student policies for lifting cube and mug objects in three separate runs, each using different random seeds. The solid curves represent the mean returns, while the shaded regions indicate the range between the minimum and maximum returns. t_{1cam} , t_{2cam} , and t_{3cam} refer to the single-camera view, two-camera view, and three-camera view teacher policies, respectively. s_{t1cam} , s_{t2cam} , and s_{t3cam} refer to the student policies distilled from t_{1cam} , t_{2cam} , and t_{3cam} teacher policies, respectively.

teachers. It is important to note that these demonstrations were exclusively utilized for training the teacher policies and not for training the students. This method of using demonstrations has been shown to be effective in improving the learning process and accelerating the convergence of the RL algorithms [14]. During the training process, we keep only the most recent 50 K transitions in the replay buffer. The teacher policies were initially trained from fixed viewpoints. The student policies were subsequently distilled from these trained teacher policies. We introduced a significant level of randomization to the camera position and parameters, such as field of view (fov) and aspect ratio to increase the robustness of the student policies π_s . Py-Bullet facilitated this by employing a synthetic camera specified by two 4 by 4 matrices: the view matrix and the projection matrix, allowing for the rendering of images from arbitrary camera positions for each episode. We provide a few examples in Fig. 2(b) to showcase the different camera perspectives used during the training of the student policies in the simulation

environment. For all teachers and students trained, we used the same data argumentation methods described in Section IV-B. All simulation experiments are repeated 3 times, each utilizing a different random seed.

The training results of teacher policies π_t using different numbers of camera views and the corresponding results of the student policies distilled from them to lift the mug and cube objects are presented in Fig. 3. For the cube object, all teacher policies have successfully learned the task, but the single-camera view teacher policy π_{t1cam} had a lower performance than the multiple camera-view teacher policies π_{t2cam} and π_{t3cam} . More importantly, for the mug object, the single-camera view teacher policy π_{t1cam} was unable to effectively learn the task. These results demonstrate the importance of utilizing multiple camera views for learning complex tasks such as lifting asymmetric objects effectively.

Table I presents the results of evaluating the performance of teacher and student policies on lifting the cube and mug objects

TABLE I
AVERAGE REWARD AND SUCCESS RATES OF CUBE AND MUG LIFTING TASK ACROSS 100 RANDOM VIEW TRIALS FOR 3 DIFFERENT RUNS BLACKPER TASK EACH USING DIFFERENT RANDOM SEEDS

Evaluation	Methods	Cube		Mug	
		Average Reward (R)	Success Rate (%)	Average Reward (R)	Success Rate (%)
Fixed Angle	π_{t1cam}	7954.0 ± 189.5	96.5 ± 3.5	0	0.0
	π_{t2cam}	8320.6 ± 269.9	98.6 ± 2.3	8440 ± 111.4	99.3 ± 0.57
	π_{2cam} ([13] cross-view attn.)	8397.6 ± 142.5	99.3 ± 1.1	8574 ± 83.8	99.3 ± 1.1
	π_{t3cam}	8420.0 ± 64.5	99.3 ± 5.7	8543 ± 121.3	100 ± 0.0
	π_{st1cam} (ours-SCV)	4285.6 ± 321.3	93.0 ± 3.6	—	—
	π_{st2cam} (ours-MCV)	8015.6 ± 231.1	96.6 ± 0.5	7431 ± 244.4	89.6 ± 2.5
	π_{st3cam} (ours-MCV)	8388.6 ± 158.3	99.0 ± 1.7	7567 ± 112.8	91.6 ± 1.1
Random Angle	π_{t1cam}	112.0 ± 96.9	02.0 ± 2.0	0	0.0
	π_{t2cam}	1597.6 ± 352.5	23.3 ± 3.2	2216.3 ± 1409.1	30.6 ± 18.8
	π_{2cam} ([13] cross-view attn.)	2821.6 ± 2201.3	39.6 ± 29.1	2909.0 ± 672.2	37.3 ± 9.1
	π_{t3cam}	5419.0 ± 916.7	68.7 ± 12.2	2818.0 ± 171.6	41.3 ± 3.2
	π_{st1cam} (ours-SCV)	3673.0 ± 475.1	84.6 ± 2.3	—	—
	π_{st2cam} (ours-MCV)	7110.3 ± 219.4	88.7 ± 3.2	5450.6 ± 491.7	77.0 ± 6.0 π
	π_{st3cam} (ours-MCV)	7815.0 ± 149.2	96.0 ± 1.7	6334.3 ± 629.9	81.7 ± 8.3

In the fixed-angle setup, no perturbations were made to the front camera position. However, for the random angle settings, random front-camera views were introduced. π_{t1cam} , π_{t2cam} , and π_{t3cam} are the single, two, and three-camera view teacher policies, respectively. π_{st1cam} , π_{st2cam} , and π_{st3cam} denotes the student policies distilled from the corresponding π_{t1cam} , π_{t2cam} , and π_{t3cam} teacher policies, respectively. SCV: Single-Camera View, MCV: Multi-Camera View.

across 100 fixed and random front-camera views. All results are averaged across three different runs. The performances of seven different policies are compared, including single-camera teacher policy π_{t1cam} , two-camera multi-view teacher policy π_{t2cam} , three-camera multi-view teacher policy π_{t3cam} , and three student policies π_{st1cam} , π_{st2cam} , and π_{st3cam} , distilled from the corresponding teacher policies. Additionally, we established a baseline by comparing with π_{2cam} , a cross-view attention-based two-camera view visual policy introduced by Jangir et al. [13]. The results indicate that for fixed and random front angles settings the students distilled from multi-view teachers, π_{st2cam} and π_{st3cam} , outperform the students distilled from single-camera teachers, π_{st1cam} in both cube and mug lifting tasks. Furthermore, since the student policies are trained with extensive randomization of the camera positions and parameters, their performance surpasses that of the teacher policies and the cross-view attention policies in random camera view settings. However, it should be emphasized that without KD, applying only camera randomization resulted in a 0 success rate for the single-view agent in both tasks. Notably, the multi-camera view to single-camera view distillation approach enables the single-camera view student policy to learn and perform the task of lifting the mug object, despite the single front-view camera teacher was unable to accomplish this. This highlights the effectiveness of distillation in enhancing the capabilities of the single-camera view policy. In terms of teacher training time, multiple-view teacher policies require slightly more clock time due to increased data volume. The total training time for the students is the sum of distillation time and the time taken for training the teacher models. Consequently training π_{st3cam} and π_{st2cam} take 16% and 9.1% more time compared to π_{st1cam} , respectively.

Ablation Studies: In addition to our main experiments, we conducted ablation studies to investigate the effect of feature alignment between the teacher and student observations. As shown in Fig. 4, we compared two different alignment methods: minimizing the euclidean distance between the features of the teacher and student (8) and aligning the pairwise similarity matrix of teacher observations with that of the student observations (6). Our results indicate that the former method

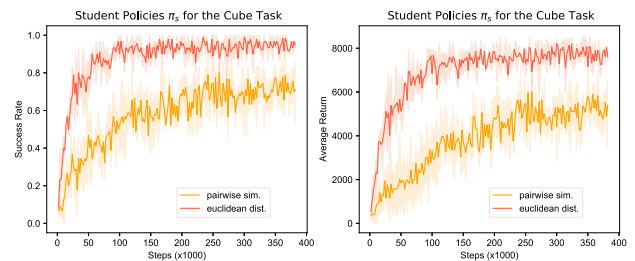


Fig. 4. Comparison of two feature alignment methods utilized for knowledge distillation in our framework for lifting the cube object.

provides better results, suggesting that direct feature alignment is more effective for knowledge distillation in our framework. This analysis provides insights into the importance of feature alignment in knowledge distillation and its impact on the performance of the student policy.

B. Real Robot Experiments

To validate the zero-shot policy transfer capability of the proposed framework, experiments were conducted in a real-world environment using the Panda robot with a Panda Hand gripper and an intel RealSense D435i RGBD camera, as shown in Fig. 1. Polymetis robot control framework [37] was used to control the robot and the gripper position. The visual policies trained for each of the aforementioned lifting tasks, i.e. mug and cube, in the simulation, were directly transferred to the real-world setting without any further training. Fig. 5 shows snapshots of the robot successfully executing these tasks with a random camera viewpoint.

In order to evaluate the performance of student policies, we tested them under random camera views for cube and mug tasks as shown in Figs. 6 and 7 respectively. As π_{st3cam} had demonstrated overall better performance than π_{st2cam} in simulation, we selected it for evaluation in real-world experiments. Tables II and III show the average success rates of cube and mug lifting tasks for different random views in these figures, respectively. The

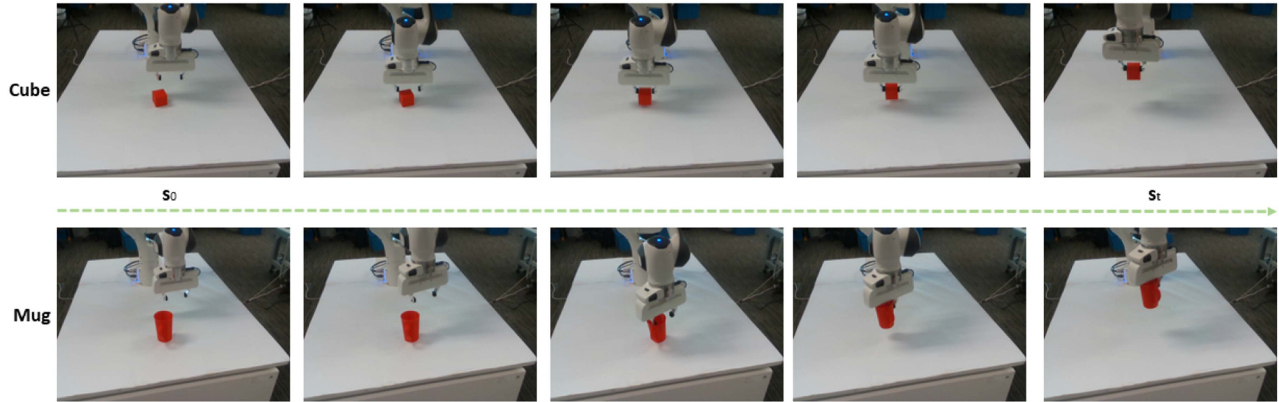


Fig. 5. Snapshots of the cube and mug lift task being performed by a robot in a real-world setup. The robot uses its single-view visual policy trained in simulation to successfully grasp and lift the objects. The robot’s end-effector is located at the starting position, and the cube and mug objects are placed randomly on the table.

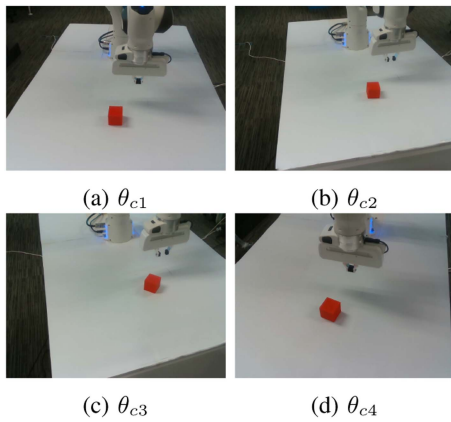


Fig. 6. Cube lifting task executed with different random camera views in real-world experiment. The results for each view are given in Table II.

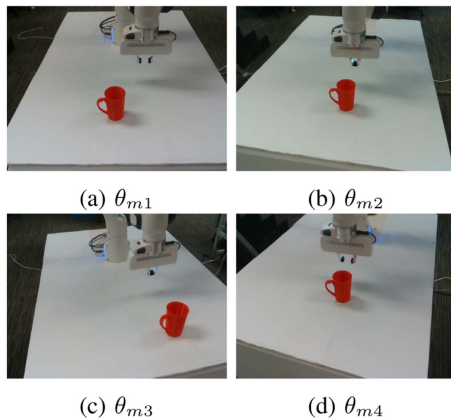


Fig. 7. Mug lifting task executed with different random camera views in real-world environment. The results for each view are given in Table III.

success rate is defined as the percentage of successful lifts out of the total number of trials.

The results show that the student policy π_{st3cam} outperforms the student policy trained with a single camera view π_{st1cam} in all camera viewpoints. As for the mug lifting task, there is no corresponding π_{st1cam} policy available as the single-camera view teacher was unable to learn the task in the simulation

TABLE II
ZERO-SHOT AVERAGE SUCCESS RATE OF CUBE LIFTING TASK FOR RANDOM CAMERA VIEWS IN FIG. 6 EVALUATED IN A REAL-WORLD SETUP THROUGH 20 TRIALS

Camera View	π_{st1cam}	π_{st3cam}
θ_{c1}	40%	80%
θ_{c2}	60%	80%
θ_{c3}	20%	60%
θ_{c4}	35%	50%

The student policies distilled from the three-camera view and one-camera view teacher policy are denoted as π_{st3cam} and π_{st1cam} , respectively.

TABLE III
ZERO-SHOT AVERAGE SUCCESS RATE OF MUG LIFTING TASK FOR RANDOM CAMERA VIEWS IN FIG. 7 EVALUATED IN A REAL-WORLD SETUP THROUGH 20 TRIALS

Camera View	π_{st1cam}	π_{st3cam}	$\pi_{st3cam+state}$
θ_{m1}	–	70%	80%
θ_{m2}	–	70%	80%
θ_{m3}	–	65%	75%
θ_{m4}	–	50%	65%

The student policy distilled from the three-camera view teacher policy is denoted as π_{st3cam} . The policy distilled from the three-camera view teacher using joint state information is denoted as $\pi_{st1cam+state}$.

environment. Notably, the average success rates were lower than those obtained in the simulation. This can be attributed to the domain gap between simulation and the real-world. The single-view teacher policy, on the other hand, was unable to lift the object in any of these random camera views, despite being trained in simulation with the data augmentation method as well. Moreover, we conducted an analysis to evaluate the impact of the robot state on the performance of the proposed approach. The joint states of the robot represented as $q \in \mathbb{R}^7$ for its seven degrees of freedom joint angles, were augmented with visual features to train both the teacher and student policies. The results presented in Table III demonstrate that incorporating robot joint angles leads to performance improvements for zero-shot policy transfer in a real-world setup. These findings highlight the need for further research to bridge the gap between simulation and reality and improve the transferability of policies learned in simulation to real-world scenarios. The video attachment showcases the successful lifting of objects from various random angles and positions, demonstrating the ability

of the student policy to handle camera location perturbations in real-time.

VI. CONCLUSION AND FUTURE WORK

In conclusion, we proposed a multi-camera view to single-camera view distillation approach for enhancing the performance of a single-view camera robot policy. The proposed method enables a single-view policy to learn from a pre-trained teacher policy, which has been trained with multiple camera viewpoints, ensuring robust feature learning. To make the student policy robust against unseen visual camera configurations, we utilized a combination of data augmentation techniques and significant variations in camera location and parameters. The experimental results demonstrated that our approach can enhance the performance of a single-view camera policy, and successfully lift the challenging object even when the single-camera view teacher policies fail. Hence, this approach holds promise for increasing the practicality of visual policy learning in scenarios where the use of multiple cameras may not be feasible or where camera calibration is unnecessary. Future work will focus on investigating the use of point-cloud and depth information to improve performance in zero-shot sim-to-real transfer scenarios.

REFERENCES

- [1] S. Levine et al., "End-to-end training of deep visuomotor policies," *J. Mach. Learn. Res.*, vol. 17, pp. 1334–1373, 2016.
- [2] D. Kalashnikov et al., "Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation," 2018, *arXiv:1806.10293*.
- [3] S. Levine et al., "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection," *Int. J. Robot. Res.*, vol. 37, no. 4/5, pp. 421–436, 2018.
- [4] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2017, pp. 23–30.
- [5] D. Yarats, I. Kostrikov, and R. Fergus, "Image augmentation is all you need: Regularizing deep reinforcement learning from pixels," in *Proc. Int. Conf. Learn. Representations*, 2020.
- [6] M. Laskin, K. Lee, A. Stooke, L. Pinto, P. Abbeel, and A. Srinivas, "Reinforcement learning with augmented data," in *Proc. Adv. Neural Inf. Process. Syst.*, pp. 19884–19 895, 2020.
- [7] N. Hansen and X. Wang, "Generalization in reinforcement learning by soft data augmentation," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2021, pp. 13611–13617.
- [8] A. Srinivas, M. Laskin, and P. Abbeel, "Curl: Contrastive unsupervised representations for reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 5639–5650.
- [9] F. Sadeghi, A. Toshev, E. Jang, and S. Levine, "Sim2real view invariant visual servoing by recurrent control," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4691–4699.
- [10] B. C. Stadie, P. Abbeel, and I. Sutskever, "Third-person imitation learning," in *Proc. Int. Conf. Learn. Representations*, 2017.
- [11] J. Shang and M. S. Ryoo, "Self-supervised disentangled representation learning for third-person imitation learning," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2021, pp. 214–221.
- [12] K. Hsu, M. J. Kim, R. Rafailov, J. Wu, and C. Finn, "Vision-based manipulators need to also see from their hands," in *Proc. Int. Conf. Learn. Representations*, 2022.
- [13] R. Jangir, N. Hansen, S. Ghosal, M. Jain, and X. Wang, "Look closer: Bridging egocentric and third-person views with transformers for robotic manipulation," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 3046–3053, Apr. 2022.
- [14] A. Zhan, R. Zhao, L. Pinto, P. Abbeel, and M. Laskin, "A framework for efficient robotic manipulation," in *Deep RL Workshop NeurIPS*, 2021.
- [15] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *Comput. Sci.*, vol. 14, no. 7, pp. 38–39, 2015.
- [16] A. A. Rusu et al., "Policy distillation," in *Proc. Int. Conf. Learn. Representations*, 2016, pp. 1–13.
- [17] S. Schmitt et al., "Kickstarting deep reinforcement learning," 2018, *arXiv:1803.03835*.
- [18] T. Haarnoja et al., "Learning agile soccer skills for a bipedal robot with deep reinforcement learning," 2023, *arXiv:2304.13653*.
- [19] D. Chen, B. Zhou, V. Koltun, and P. Krährenbühl, "Learning by cheating," in *Proc. Conf. Robot Learn.*, 2020, pp. 66–75.
- [20] Z. Zhang, A. Liniger, D. Dai, F. Yu, and L. Van Gool, "End-to-end urban driving by imitating a reinforcement learning coach," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 15222–15232.
- [21] D. Jain, A. Li, S. Singhal, A. Rajeswaran, V. Kumar, and E. Todorov, "Learning deep visuomotor policies for dexterous hand manipulation," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2019, pp. 3636–3643.
- [22] T. Chen, J. Xu, and P. Agrawal, "A system for general in-hand object re-orientation," in *Proc. Conf. Robot Learn.*, 2022, pp. 297–307.
- [23] A. X. Lee et al., "Beyond pick-and-place: Tackling robotic stacking of diverse shapes," in *Proc. Conf. Robot Learn.*, 2021, pp. 1089–1131.
- [24] J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning quadrupedal locomotion over challenging terrain," *Sci. Robot.*, vol. 5, no. 47, 2020, Art. no. eabc5986.
- [25] T. Miki, J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning robust perceptive locomotion for quadrupedal robots in the wild," *Sci. Robot.*, vol. 7, no. 62, 2022, Art. no. eabk2822.
- [26] M. Sorokin, J. Tan, C. K. Liu, and S. Ha, "Learning to navigate side-walks in outdoor environments," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 3906–3913, Apr. 2022.
- [27] A. Agarwal, A. Kumar, J. Malik, and D. Pathak, "Legged locomotion in challenging terrains using egocentric vision," in *Proc. Conf. Robot Learn.*, 2023, pp. 403–415.
- [28] L. Fan et al., "Secant: Self-expert cloning for zero-shot generalization of visual policies," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 3088–3099.
- [29] T. Haarnoja et al., "Soft actor-critic algorithms and applications," 2018, *arXiv:1812.05905*.
- [30] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8026–8037.
- [31] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2011, pp. 627–635.
- [32] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1365–1374.
- [33] R. Portelas, C. Colas, K. Hofmann, and P.-Y. Oudeyer, "Teacher algorithms for curriculum learning of deep rl in continuously parameterized environments," in *Proc. Conf. Robot Learn.*, 2020, pp. 835–853.
- [34] Y. Zhang, P. Abbeel, and L. Pinto, "Automatic curriculum learning through value disagreement," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 7648–7659.
- [35] E. Coumans et al., "Bullet physics library," *Open Source: Bulletphysics. Org*, vol. 15, no. 49, p. 5, 2013.
- [36] J. Borja-Diaz, O. Mees, G. Kalweit, L. Hermann, J. Boedecker, and W. Burgard, "Affordance learning from play for sample-efficient policy learning," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2022, pp. 6372–6378.
- [37] Y. Lin, A. S. Wang, G. Sutanto, A. Rai, and F. Meier, "Polymetis," 2021. [Online]. Available: <https://facebookresearch.github.io/fairo/polymetis/>